Why we need **Ontology-specific Data Portals**: A **Case Study** for **CIDOC-CRM**

59th CIDOC CRM & 52nd FRBR/LRMoo SIG Meeting

**Presented in SWODCH Workshop 2023*

Michalis Mountantonakis, Ioannis Theocharakis and Yannis Tzitzikas

FORTH
INSTITUTE OF COMPUTER SCIENCE

FORTH-ICS
Information Systems Laboratory

University of Crete
Computer Science Department

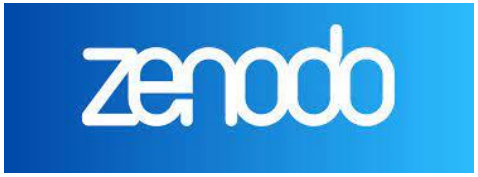
Outline

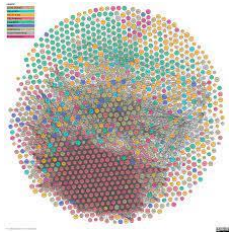
- Context - Motivation (4 m)
- Collecting CIDOC-CRM Datasets and Computing Statistics (4 m)
- The CIDOC-CRM Web Portal (4 m)
- Experimental Evaluation (3 m)
- Concluding Remarks (1 m)

Context – Motivation

Context – **Where to Publish**

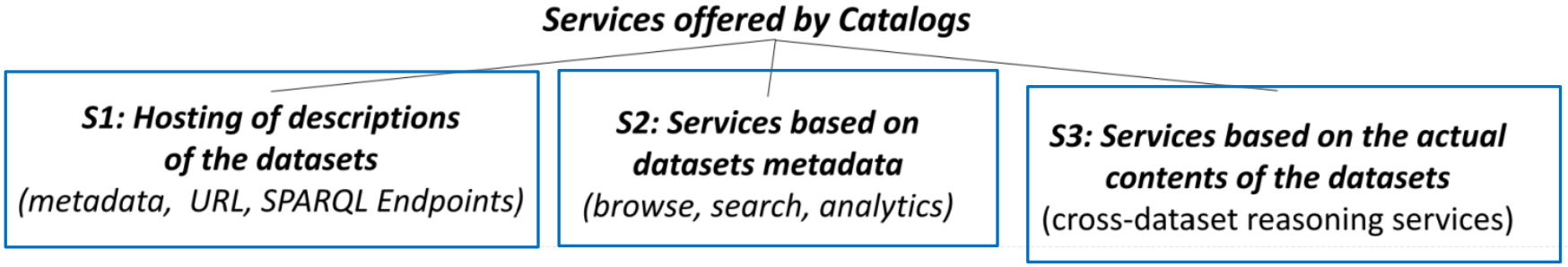
- **There are numerous ways for publishing data on the web**
  - Web Pages
  - GitHub
  - Zenodo
  - As Linked Open Data
  - Online dataset catalogs





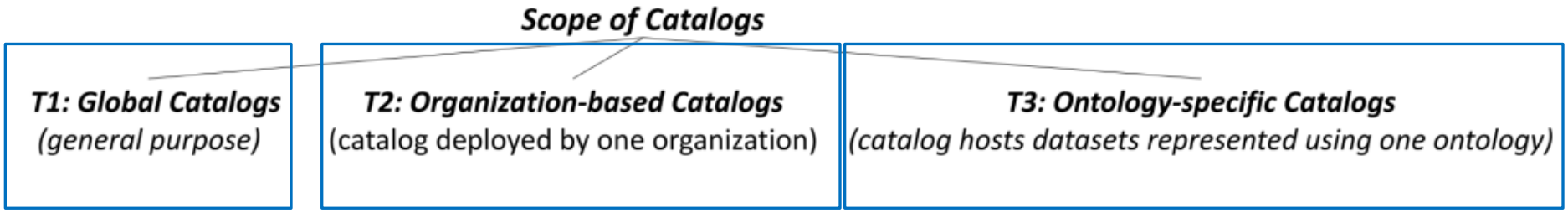
- ***The Key Notion:*** *Dataset Catalogs can offer more services comparing to the alternative ways*

Context – **Dataset Catalogs (1/2)**



- **Services of Catalogs**
  - **Hosting of datasets descriptions**
    - Metadata about these datasets (their URL, SPARQL endpoints and availability)
  - **Services based on dataset's metadata**
    - Browsing, searching and offering analytics
  - **Services based on the actual contents (triples) of the datasets**
    - Cross-dataset reasoning services, e.g., for finding all the datasets of a URI.

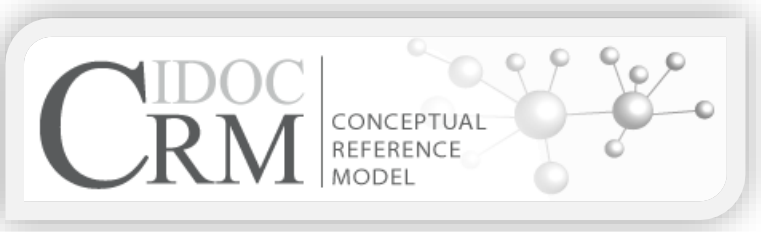
Context – **Dataset Catalogs (2/2)**



- **Scope of Catalogs**
  - **Global Catalogs**
    - general purpose catalogs
  - **Organization-based Catalogs**
    - a CKAN instance deployed by one organization, a Zenodo channel, etc.
  - **Ontology-specific Catalogs**
    - hosted datasets are represented using a single ontology (and its specializations)

Context – **Where we focus**

- We focus on
  - **Ontology-specific Catalogs** since **we restrict the datasets of the catalog** only on them using the **ISO 21127 Standard CIDOC Conceptual Reference Model (CIDOC-CRM)**
  - **Services** based on **Metadata** (mainly)

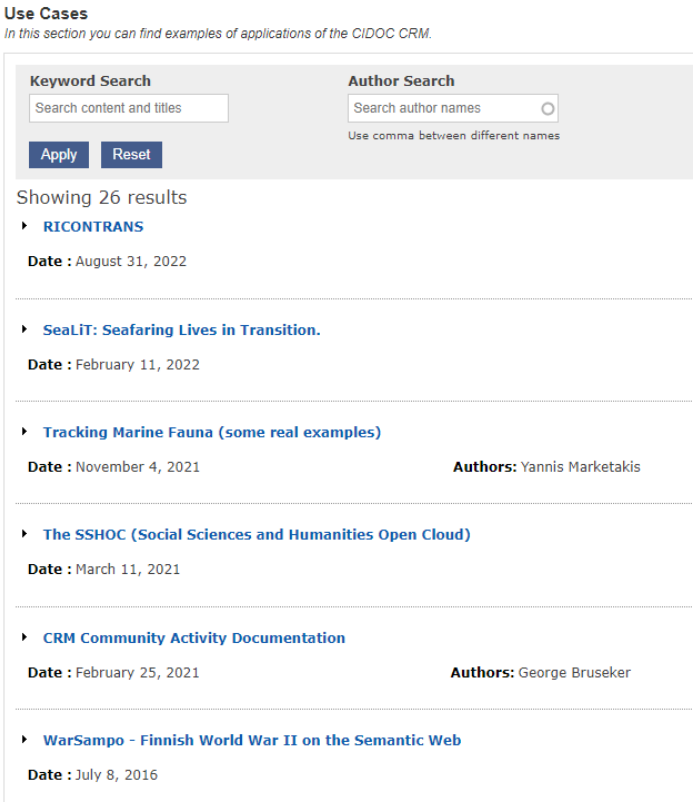


- Objectives of **Ontology-specific Catalogs:**
  - It is important for the **community** of the **focused ontology** to **publish their datasets in that catalog**
  - it is more **sustainable** to **achieve completeness for one ontology**, **than being complete for all the available ones**

Motivation – **CIDOC-CRM Datasets (1/3)**

- **Problem**: It is not trivial to create such a catalog for CIDOC-CRM.
  - **Numerous available ways to publish a dataset and in different places**
  - **Quite challenging** even to **discover all the datasets using a popular model.**

- In 2022, we **decided to write a survey about CIDOC-CRM and Machine Learning [1]** and we observed that **was too difficult to discover the available CIDOC-CRM datasets!**
  - In the official website we found a page with 26 use cases, **but for many of them the datasets were not reachable**!



Motivation – **CIDOC-CRM Datasets (2/3)**

- Then we **further searched for CIDOC-CRM datasets** and we **found an online excel file** in the **webpage of CIDOC-CRM** containing information **for 17 CIDOC-CRM datasets.**
  - Again **some of them were not reachable**!!

A	B	C	D	E	F	G	H
Name	End Point Address	Reusability for Federated Search (Y/N)	API (Yes/No)	Link to API	Maintainer	Maintainer Contact	Maintained y/n?
ADS	http://data.archaeologydataservice.ac.uk/query/	N			Archaeological Data Service	https://archaeologydataservice.ac.uk	
~~BM~~	https://collection.britishmuseum.org/resource/sparql (down)	Y			British Museum	https://www.britishmuseum.org/about_us/news_and_press/press_releases/2011/semantic_web_endpoint.aspx	
Beni Culturali	http://dati.beniculturali.it/sparql	Y	No		MIBAC	http://dati.beniculturali.it/il-progetto/	n
Foundation Zeri	http://data.fondazionezeri.unibo.it/sparql	Y	No		Fondazione Federico Zeri University of Bologna	http://www.fondazionezeri.unibo.it/it/fototeca/fototeca-zeri/zeri-lode	y

Motivation – **CIDOC-CRM Datasets (3/3)**

- As a first step **we created a GitHub page** for the available CIDOC-CRM datasets (**At that time we had found 18 datasets**) that provide **either an online data dump or a running SPARQL Endpoint.**
  - However, **it was again not enough**, since
    - **Such a list was not so practical – no statistics and visualizations**
    - We were sure that **more CIDOC-CRM datasets were available.**

ID	Dataset	Link	Domain	Triples	SPARQL Endpoint/ API	Data Dump
1	Archaeology Data Service	http://data.archaeologydataservice.ac.uk	Heritage Data of United Kingdom	1,559,912	✓	
2	Auckland Museum	https://api.aucklandmuseum.com/	Auckland Museum, New Zealand	>10,000,000	✓	
3	Beni Culturali	https://dati.cultura.gov.it/linked-open-data/	Cultural Institutions in Italy	755,702,389	✓	✓
4	Corago LOD	https://zenodo.org/record/3377586	Italian Opera,1600 to 1900	22,399,698		✓

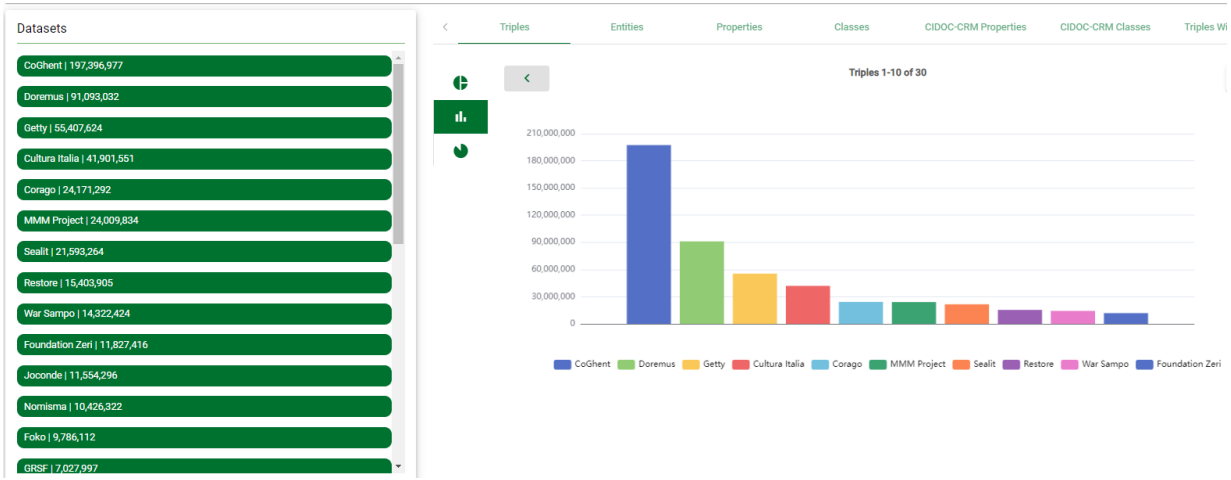
Objective – **CIDOC-CRM Portal**

- **Our target is to create an ontology-specific portal** (or catalog), by focusing on **CIDOC-CRM** that will
  - Enable the **discoverability**, **reusability** and **preservation** of all the **CIDOC-CRM datasets**
  - Offer an **interactive way** to **browse statistics** and **visualizations** for the **CIDOC-CRM datasets**, by computing **ontology-based descriptions**

- The objective is the portal to be important for several use cases including ***data discovery, data integration, data publishing and ontology evaluation***

Contribution – **CIDOC-CRM Portal**

- We first collect **30 CIDOC-CRM datasets** and we **compute ontology-based descriptions** by using the **VoID vocabulary [2]**
- We **present an online web portal** (https://demos.isl.ics.forth.gr/CIDOC-CRM_Portal) that offers
  - **Browsing** and **analytics of all the collected CIDOC-CRM datasets** through **statistics** and **visualizations**



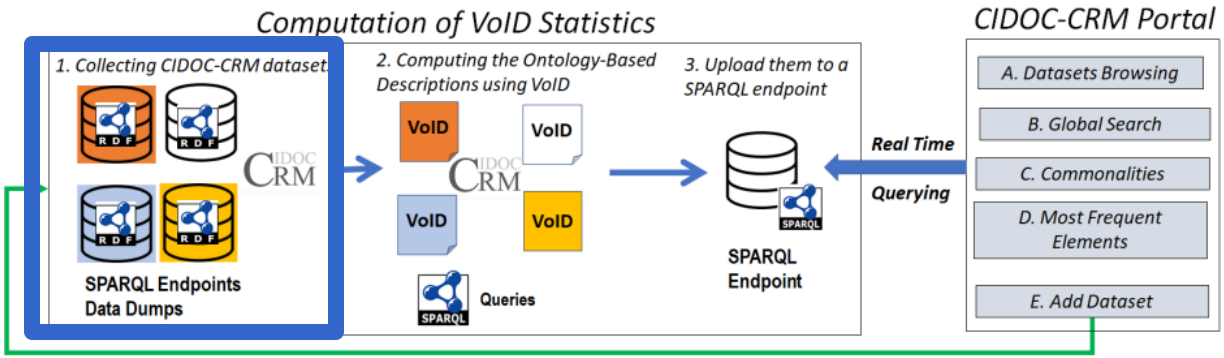
- We offer an **analysis** for the **30 collected CIDOC-CRM datasets**

Collecting CIDOC-CRM Datasets and Computing Statistics

Important Note – **CIDOC-CRM Version**

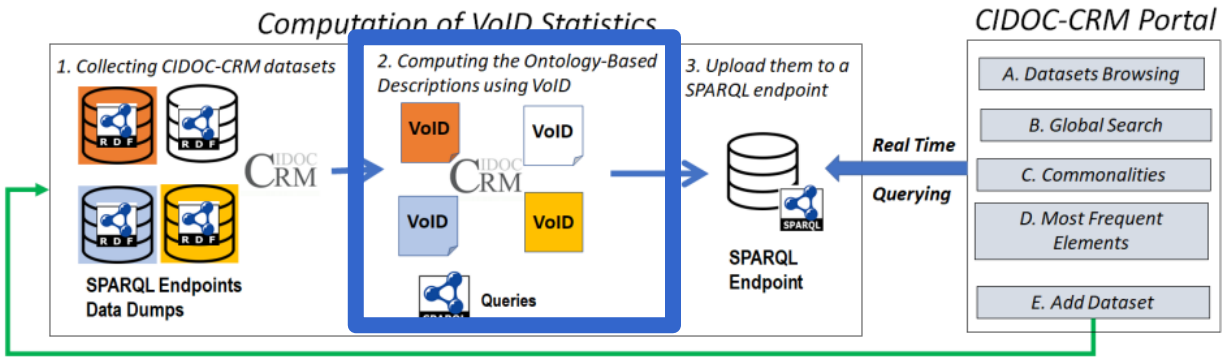
- When we refer to **CIDOC-CRM properties and classes**, we mean
  - all the properties and classes of the RDF file of **CIDOC-CRM version 7.1.2** which contains
    - **309 CIDOC-CRM properties** (including inverse properties)
    - **76 CIDOC-CRM classes**
      - https://cidoc-crm.org/rdfs/7.1.2/CIDOC_CRM_v7.1.2.rdfs
  - We do not refer in properties and classes that extend the above CIDOC-CRM properties and classes.

Step A. Collecting CIDOC-CRM datasets.



- **We tried to collect all the available CIDOC-CRM datasets by using**
  - the list of **18 datasets provided in the GitHub page**
  - By further **searching in google scholar** and **catalogs** like Zenodo and search engines:
    - with the keywords "**CIDOC-CRM dataset/endpoint/data dump**"
- **We collected 30 real RDF datasets (having in total 560 million RDF triples) that have been modelled by using CIDOC-CRM**
  - 21 of them offer a public SPARQL endpoint
  - 9 of them only an RDF data dump

Step B. Computing the Ontology-Based Descriptions using VoID



- For the **computation of the VoID statistics we send SPARQL queries** to the SPARQL endpoint of each dataset
- For the datasets **that do not offer a SPARQL Endpoint**, we **downloaded the data dumps** and we **uploaded them to our SPARQL endpoint** for performing the computations.
  - **(-) Time consuming** in some cases, due to
    - **the large size of some datasets**
    - **syntax errors in some RDF files**
- For each dataset, we **produced a single file** containing all the **statistics by using the VoID vocabulary**.

Step B. Computing the Ontology-Based Descriptions using VoID – **An example file for the dataset Open Archaeo**

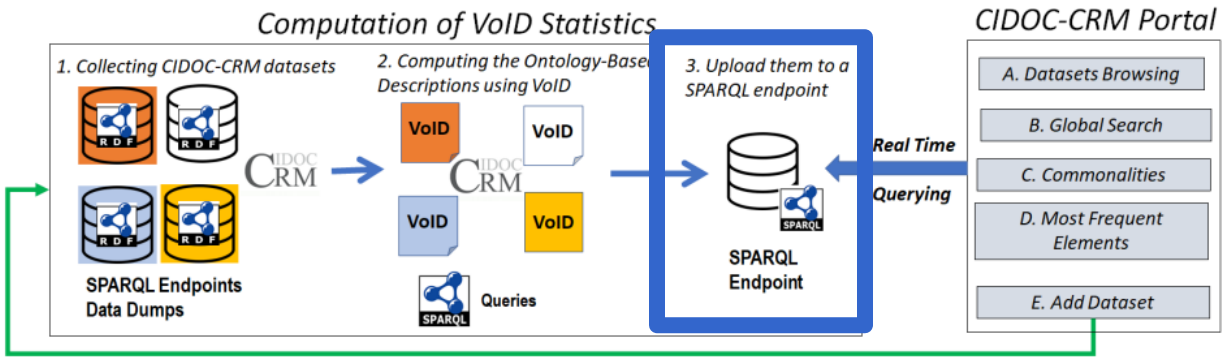
```
@prefix void: <http://rdfs.org/ns/void#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix void-crm: <http://www.ics.forth.gr/isl/void-crm/>.
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/>.

<http://openarchaeo.huma-num.fr/explorateur/home> rdf:type void:Dataset;
        dcterms:title "Open Archaeo";
        dcterms:description "A semantic mediator for archaeological datasets";
        void:triples "1424168";
        void:entities "266454";
        void:properties "61";
        void:classes "23";

        void:propertyPartition [
                void:property crm:P4_has_time-span;
                void:triples "24344";
        ];

        void:classPartition [
                void:class crm:E53_Place;
                void:triples "4368";
                ]; ...

        void-crm:propertiesCIDOC "30";
        void-crm:classesCIDOC "14";
        void-crm:triplesWithCIDOCproperty "652201";
        void-crm:triplesWithCIDOCpropertyPercentage "45.80%";
        void-crm:triplesWithCIDOCinstance "1195837";
        void-crm:triplesWithCIDOCinstancePercentage "83.97%";
```

General VoID Statistics

VoID Statistics For properties and classes

Dedicated CIDOC-CRM statistics

Step C. Upload the Ontology-based Descriptions to a SPARQL Endpoint



- The produced files of all the datasets are uploaded in an **online SPARQL Endpoint**
  - 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
  - the **visualization** of the already **computed statistics**
  - the **computation** of even **more statistics through more SPARQL queries**
  - the **easy addition** of any **CIDOC-CRM dataset**

The CIDOC-CRM Web Portal

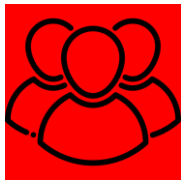
The Desired Users of the Portal

Simple Users

Simple (RDF) users **familiar** with **Semantic Web technologies**

CIDOC-CRM Dataset owners

Users that **have published at least one dataset** by using the **CIDOC-CRM model**.



CIDOC-CRM experts

CIDOC-CRM Special Interest Group (SIG): An active community where **many organizations and researchers participate**

The Corresponding Use Cases

Simple Users

Q1. I want all the **CIDOC-CRM datasets** and **statistics** about them.

Q2. I want all the **CIDOC-CRM datasets describing places** for training a **Machine Learning Model**

Dataset Discovery And Selection

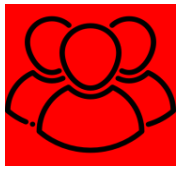
CIDOC-CRM Dataset owners

Q3. **Where and how to publish** my (CIDOC-CRM) **dataset** for **improving** its **discoverability**?

Data Publishing

Q4. I want to **integrate** my (CIDOC-CRM) **dataset** with a **dataset** sharing the **most common CIDOC-CRM properties.**

Data Integration



CIDOC-CRM experts

Q5. I want to see how the **CIDOC-CRM properties and classes are used** (e.g., for detecting possible errors).

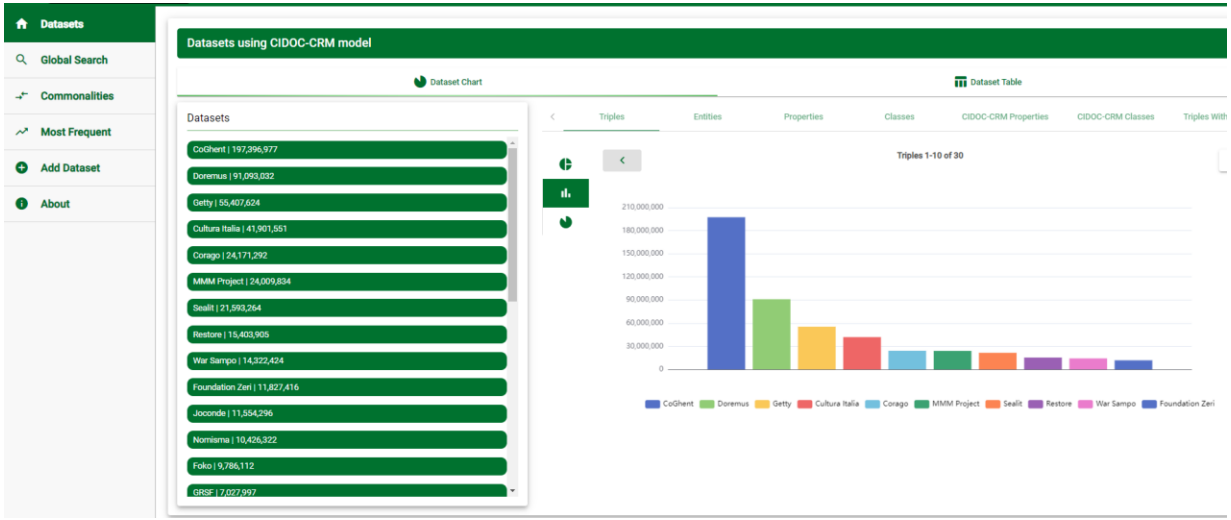
Q6. I want **the most frequently used CIDOC-CRM properties and classes**. Is there a **power-law distribution**?

Ontology Evaluation

The Modes of the CIDOC-CRM Portal

The portal offers **five interactive modes:**

- ***A. Datasets Browsing***
  - For Dataset Discovery and Selection, Ontology Evaluation
- ***B. Global Search***
  - For Dataset Discovery and Selection
- ***C. Commonalities***
  - For Data Integration
- ***D. Most Frequent Elements***
  - For Ontology Evaluation
- ***E. Add Dataset***
  - For Data Publishing



Mode A. Datasets Browsing (1/3)

- The user can **browse statistics** and **visualizations** for all the **datasets**.
  - Ranking lists (using HTML tables) and visualizations through charts.

Simple Users

Q1. I want **all the CIDOC-CRM datasets** and **statistics about them.**

Datasets

- CoGhent | 197,396,977
- Doremus | 91,093,032
- Getty | 55,407,624
- Cultura Italia | 41,901,551
- Corago | 24,171,292
- MMM Project | 24,009,834
- Sealit | 21,593,264

All the Datasets

Title	Triples	Entities	Properties	Classes	CIDOC-CRM Properties	CIDOC-CRM Classes	Triples With CIDOC-CRM Property	Triples With CIDOC-CRM Property Percentage	Triples With CIDOC-CRM Instance	Triples With CIDOC-CRM Instance Percentage
CoGhent	197,396,977	61,437,653	412	96	32	11	43,082,959	21.83%	94,036,071	47.64%
Doremus	91,093,032	18,531,445	507	151	44	15	12,721,751	13.97%	18,789,384	20.63%
Getty	55,407,624	9,773,325	58	25	42	24	35,174,151	63.48%	54,855,980	99.00%
Cultura It	41,901,551	9,951,056	181	177	33	29	21,927,494	52.33%	41,896,526	99.99%
Corago	24,171,292	5,102,527	115	58	26	21	7,655,451	31.67%	16,147,649	66.81%

Mode A. Datasets Browsing (2/3)

- The user can **browse statistics** and **visualizations** for **each single dataset**.
  - Ranking lists (using HTML tables) and visualizations through charts.

Simple Users

Q1. I want **a specific CIDOC-CRM dataset** and **statistics about it.**

Browse a Single Dataset

Seafaring Lives in Transition

Dataset Title: Sealit | URL: https://zenodo.org/record/6460841

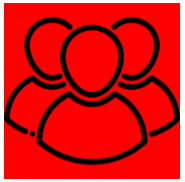
Basic Statistics

Title: Sealit
Triples: 21,593,264
Properties: 140
Triples With CIDOC-CRM property: 10,662,888 (49.38%)
CIDOC-CRM Properties: 71

Description: SeaLiT Knowledge Graphs - Maritime History Data in RDF
Entities: 4,466,105
Classes: 81
Triples With CIDOC-CRM Instance: 18,220,946 (84.38%)
CIDOC-CRM Classes: 38

Mode A. Datasets Browsing (3/3)

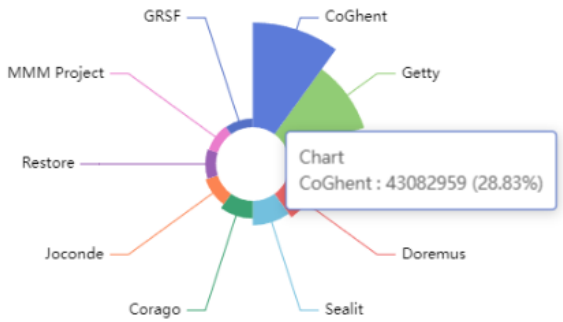
- The user can browse **specialized statistics for CIDOC-CRM.**
  - Ranking lists (using HTML tables) and visualizations through charts.

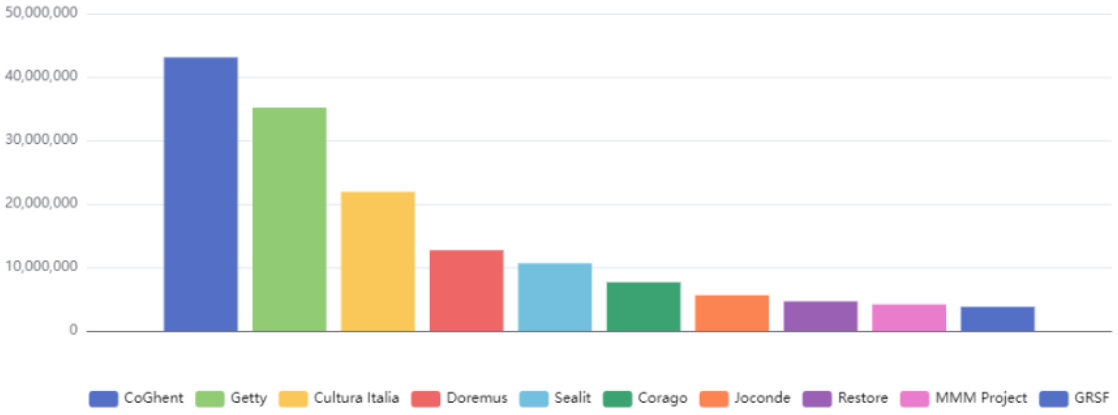


CIDOC-CRM experts

Q5. I want to see **how the CIDOC-CRM properties and classes are used** (e.g., for detecting possible errors)

Number of Triples including CIDOC-CRM properties per Dataset



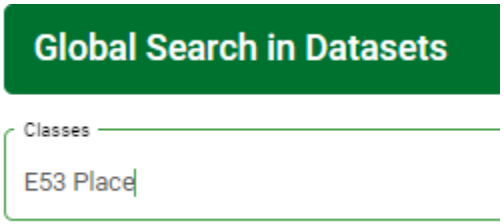


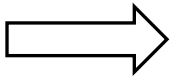
Mode B. Global Search

- **The user can search for any property/class:**
  - The **portal returns all the datasets containing the desired property/class** and the **number of triples**.
  - We provide **autocomplete services** and a **drop-down list** including all the **CIDOC-CRM properties and classes**

Simple Users

Q2. I want **all the CIDOC-CRM datasets** describing **places** for training a **Machine Learning Model**





Title	Triples ↓
Cultura Italia	91,558
Sealit	51,229
Nomisma	50,314
Doremus	30,547
ADS	19,144
Foundation Zeri	12,821
LINCS	8,171
Getty	6,354
MMM Project	5,077
Open Archaeo	4,368

23 datasets contain Places

Mode C. Commonalities

- The user can discover **all the common properties and classes** between **any pair of datasets**!

Q4. I want **to integrate my (CIDOC-CRM) dataset (SeaLiT)** with a dataset **sharing the most common CIDOC-CRM properties.**

SeaLiT Dataset Owner

Seafaring Lives in Transition

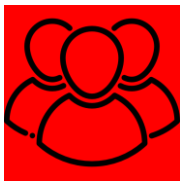
LINCS XTriples

SeaLiT *has*
44 common CIDOC-CRM properties
with the ***LINCS Dataset***

Property
http://www.cidoc-crm.org/cidoc-crm/P01i_is_domain_of
http://www.cidoc-crm.org/cidoc-crm/P02_has_range
http://www.cidoc-crm.org/cidoc-crm/P02i_is_range_of
http://www.cidoc-crm.org/cidoc-crm/P1_is_identified_by
http://www.cidoc-crm.org/cidoc-crm/P100i_died_in
http://www.cidoc-crm.org/cidoc-crm/P106_is_composed_of
http://www.cidoc-crm.org/cidoc-crm/P106i_forms_part_of
http://www.cidoc-crm.org/cidoc-crm/P107_has_current_or_former_member
http://www.cidoc-crm.org/cidoc-crm/P107i_is_current_or_former_member_of
http://www.cidoc-crm.org/cidoc-crm/P108i_was_produced_by
http://www.cidoc-crm.org/cidoc-crm/P11_had_participant
http://www.cidoc-crm.org/cidoc-crm/P11i_participated_in

Mode D. Most Frequent Elements

- The user can find the most **frequent properties and classes** according to
  - the **number of datasets**
  - the **number of triples**



CIDOC-CRM experts

Q6. I want **the most frequently used CIDOC-CRM properties and classes**. Is there a power-law distribution?

Class	Datasets
http://www.cidoc-crm.org/cidoc-crm/E53_Place	23
http://www.cidoc-crm.org/cidoc-crm/E52_Time-Span	20
http://www.cidoc-crm.org/cidoc-crm/E21_Person	19
http://www.cidoc-crm.org/cidoc-crm/E74_Group	17
http://www.cidoc-crm.org/cidoc-crm/E42_Identifier	17

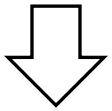
Most Frequent CIDOC-CRM Classes

Mode E. Add Dataset

- The user **can fill and submit a form** including **some very basic details** of the dataset, for **requesting to be published in the portal**

CIDOC-CRM Dataset owners

Q3. **Where and how to publish my (CIDOC-CRM) dataset** for improving its discoverability?



Experimental Evaluation

General Statistics

- We provide statistics and measurements for the **30 collected CIDOC-CRM datasets**
  - 30% of triples contain a **CIDOC-CRM property**
  - 53.5% of triples contain a **CIDOC-CRM instance**
  - Each dataset contains on average ~37 CIDOC-CRM properties and ~19 CIDOC-CRM classes

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

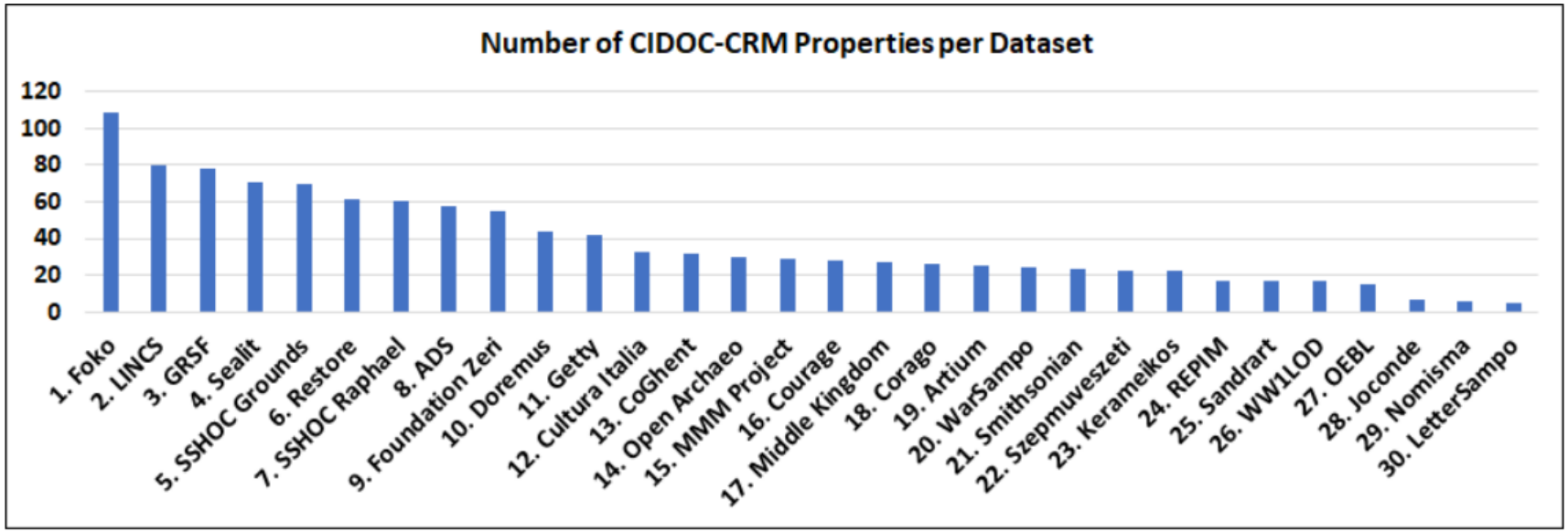
Statistics about the CIDOC-CRM Datasets

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

Average Values for the CIDOC-CRM Datasets

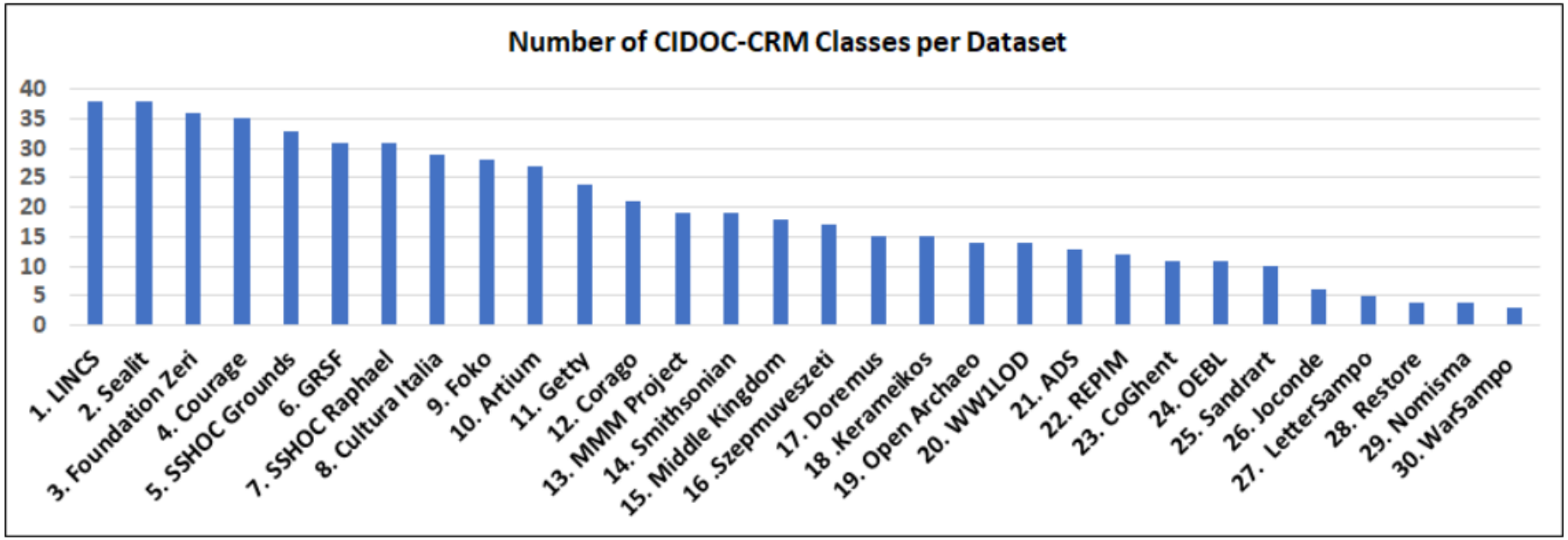
Number of CIDOC-CRM Properties per Dataset

- From the 30 datasets
  - The **top-1** uses over **100 CIDOC-CRM properties**
  - **7 datasets** use **≥ 60 CIDOC-CRM properties**
  - **Only 3 datasets** use **≤ 10 CIDOC-CRM properties.**



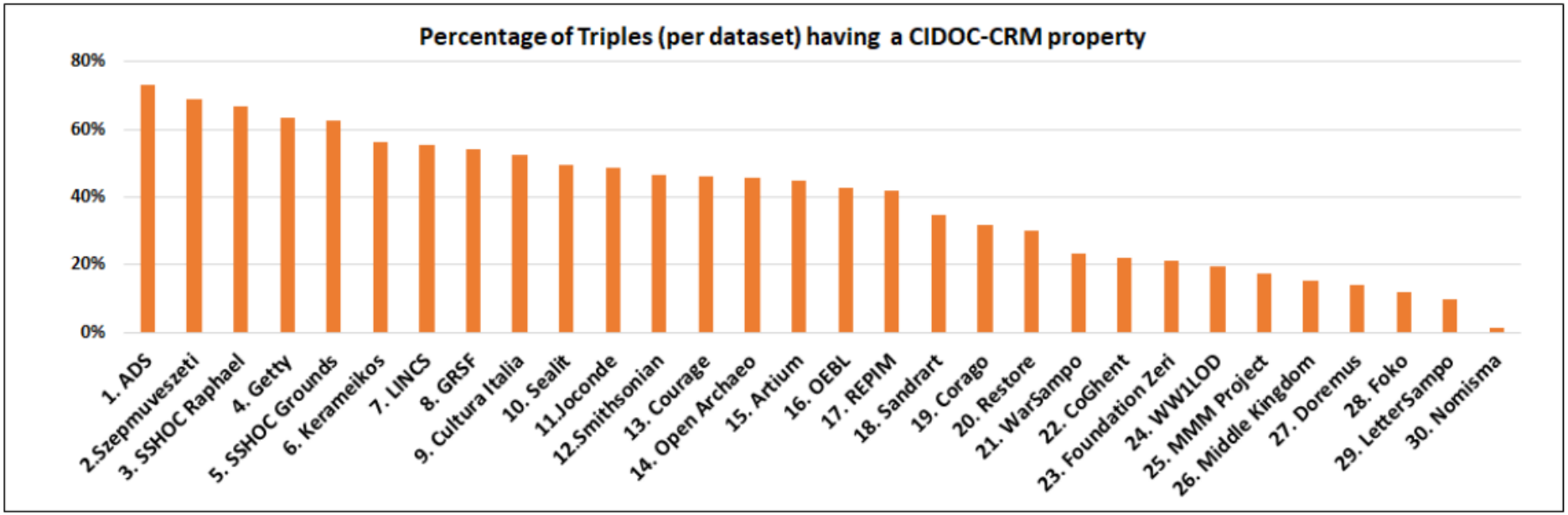
Number of CIDOC-CRM Classes per Dataset

- From the 30 datasets
  - The **Top-1** uses **38 CIDOC-CRM Classes**
  - **12 datasets** use **≥ 20 CIDOC-CRM classes**
  - **25 datasets** use **≥ 10 CIDOC-CRM classes**



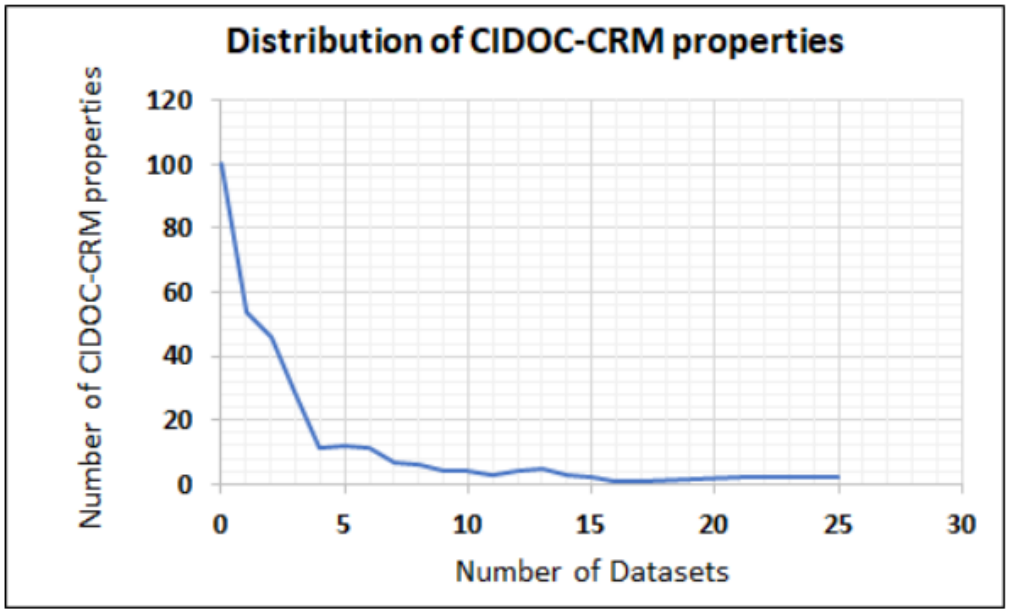
Percentage of triples having a CIDOC-CRM property

- From the 30 datasets
  - **5 datasets** use CIDOC-CRM **properties** in **at least 60% of their triples.**
  - **20 datasets** in at least **30% of their triples**.



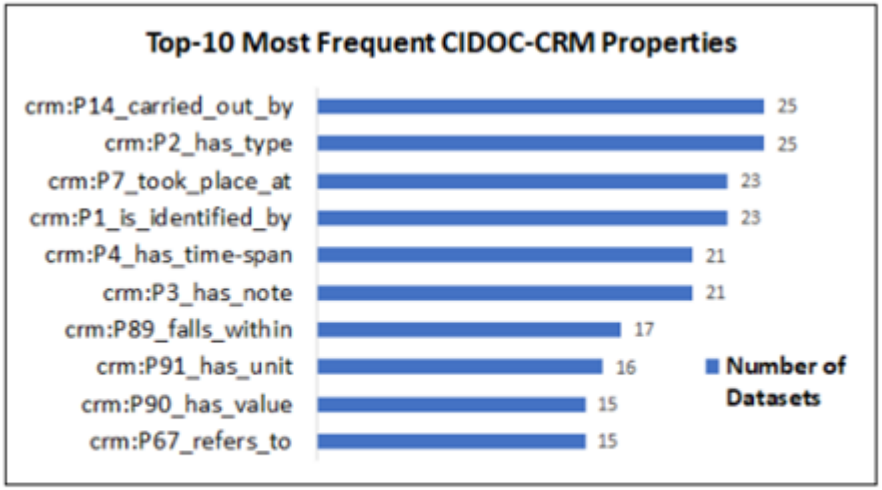
CIDOC-CRM Properties - Distribution measurements

- We can see a **power-law distribution** for the **309 available CIDOC-CRM properties (of CIDOC-CRM version 7.1.2)**
  - **100 properties** are used **only by a single or two datasets**.
  - There are also 100 properties (out of 309) that are not used by the collected datasets
  - Only **6 CIDOC-CRM properties** are used **from ≥ 20 datasets**

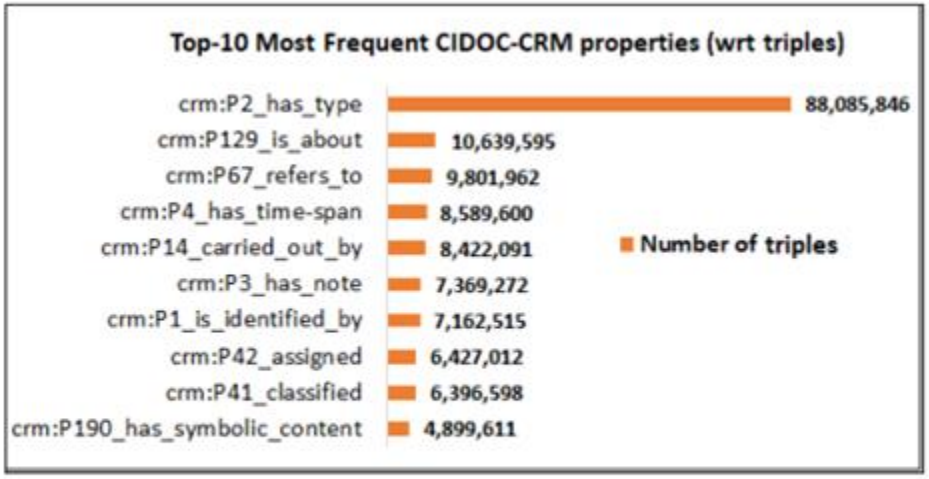


CIDOC-CRM Properties - Frequency

- We show the **most popular CIDOC-CRM properties** according to the number of a) datasets and b) triples
  - **The most popular properties** are "**crm:P14_carried_out_by**" and "**crm:P2_has_type**" that appear in **25 datasets**
  - The property **"crm:P2_has_type"** is the **top concerning the number of triples**, appearing in **88M triples.**



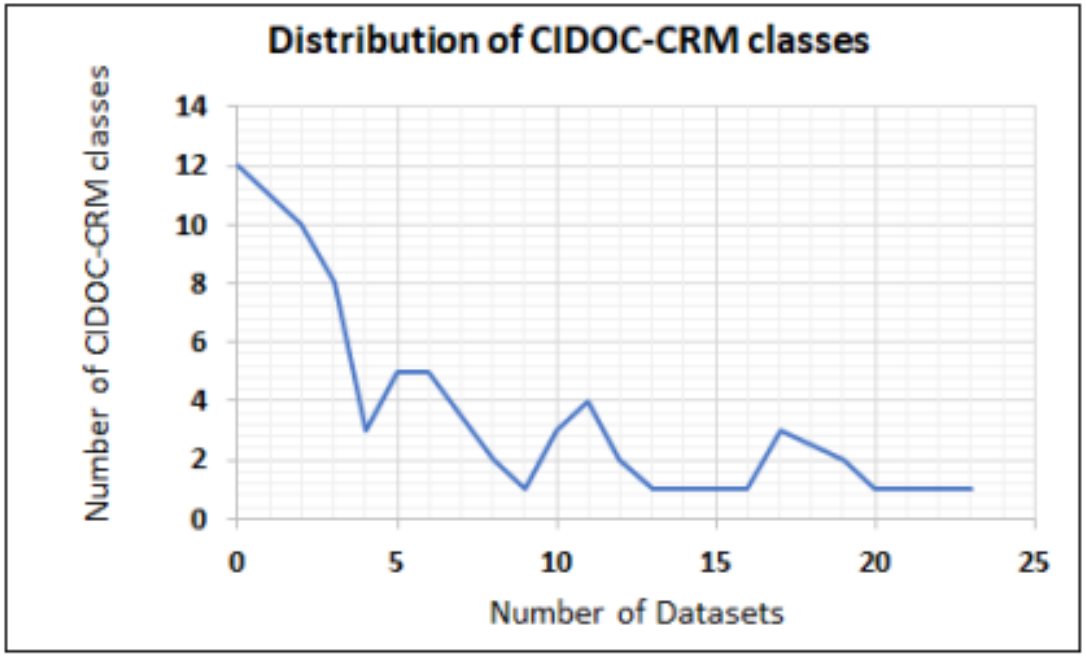
Top-10 most frequent CIDOC-CRM properties wrt the number of datasets



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

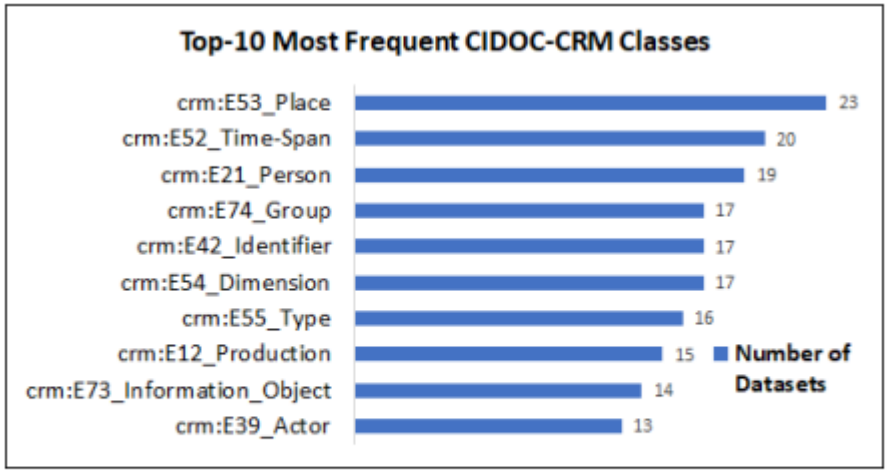
CIDOC-CRM Classes - Distribution measurements

- Most of classes are also used by a low number of datasets. Indeed, **from the 76 CIDOC-CRM classes of CIDOC-CRM version 7.1.2**
  - **21 classes** are used **only by a single or two datasets.**
  - There are **12 classes (out of 76)** that are **not used.**
  - **Only 19 CIDOC-CRM classes** are used from **≥ 10 datasets**
  - **Only 2 CIDOC-CRM classes** are used from **≥ 20 datasets**

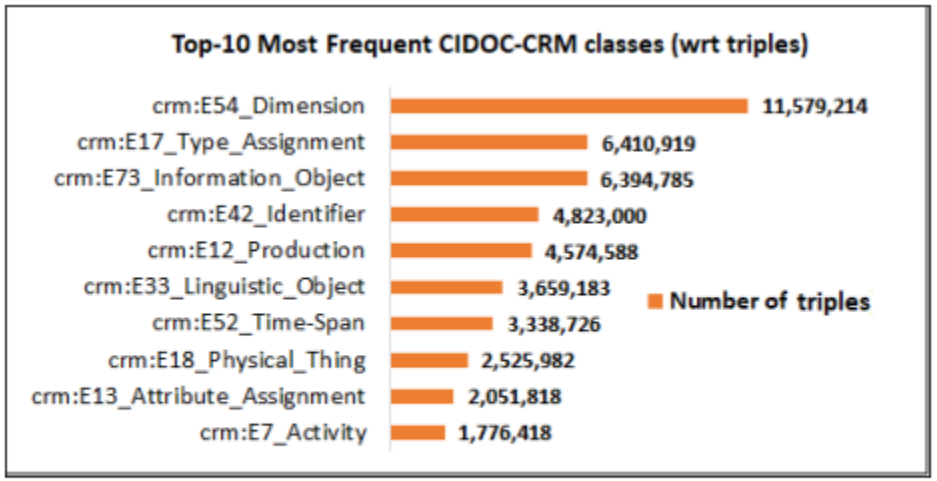


CIDOC-CRM Classes - Frequency

- We show the most popular CIDOC- CRM classes **according to the number of datasets and triples:**
  - The most popular classes are "**crm:E53_Place**" and "**crm:E52_Time_Span**" that appear in >**20 datasets**
  - The class "**crm:E54_Dimension**" is the **top concerning the number of triples**, appearing in **11M triples.**



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



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

Webpage and More Details

Much **more statistics, experiments and visualizations** can be found in the online page of the web portal:

- **Webpage of the CIDOC-CRM Portal**
  - https://demos.isl.ics.forth.gr/CIDOC-CRM_Portal/
  - There is **no need to download any software**

- **Github page** with the code and SPARQL Queries
  - **https://github.com/mountanton/CIDOC-CRM_Portal**

- **Tutorial Video** in YouTube
  - **https://youtu.be/ar8JEty94_w**

Conclusion and Future Work

Concluding Remarks

- 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 dataset.**

- We described **use cases**, **details about how the statistics are computed**, and the **modes** of the **portal**.

- We offered **measurements** about **30 real CIDOC-CRM datasets that** 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.

Future Work

- We plan to
  - **exploit triple/path patterns** since they can be used for natural Question Answering tasks through LLMs (ongoing work)
    - crm:E39_Actor→crm:P92i_was_brought_into_existence_by→ crm:E63_Beginning_of_Existence →crm:P4_has_time span→crm:E52_Time-Span
  - **compute/visualize more complex statistics**
  - offer mechanisms for **monitoring** the **changes** in **datasets** and **recomputing the statistics.**

Thank You!

References

[1] Y. Tzitzikas, M. Mountantonakis, P. Fafalios, Y. Marketakis, CIDOC-CRM and machine learning: a survey and future research, Heritage 5 (2022) 1612–1636

[2] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, Describing linked datasets with the VoID vocabulary (2011)

[3] https://datasetsearch.research.google.com/

[4] https://datahub.io/

[5] https://lod-cloud.net/

[6] 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

[7] 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).

[8] 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.

[9] 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

[10] 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.