How your Cultural Dataset is Connected to the Rest Linked Open Data?

Michalis Mountantonakis[1] and Yannis Tzitzikas[1,2]

[1]Institute of Computer Science, FORTH-ICS, Greece, and
[2]Computer Science Department, University of Crete, Greece
{mountant,tzitzik}@ics.forth.gr

Abstract. More and more publishers tend to create and upload their data as digital open data, and this is also the case for the Cultural Heritage (CH) domain. For facilitating their Data Interchange, Integration, Preservation and Management, publishers tend to create their data as Linked Open Data (LOD) and connect them with existing LOD datasets that belong to the popular LOD Cloud, which contains over 1,300 datasets (including more than 150 datasets of CH domain). Due to the high amount of available LOD datasets, it is not trivial to find all the datasets having commonalities (e.g., common entities) with a given dataset at real time. However, it can be of primary importance for several tasks to connect these datasets, for being able to answer more queries and in a more complete manner (e.g., for better understanding our history), for enriching the information of a given entity (e.g., for a book, a historical person, an event), for estimating the veracity of data, etc. For this reason, we present a research prototype, called `ConnectionChecker`, which receives as input a LOD Dataset, computes and shows the connections to hundreds of LOD Cloud datasets through `LODsyndesis` knowledge graph, and offers several measurements, visualizations and metadata for the given dataset. We describe how one can exploit `ConnectionChecker` for their own dataset, and we provide use cases for the CH domain, by using two real linked CH datasets: a) a dataset from the National Library of Netherlands, and b) a dataset for World War I from the Universities of Aalto and Helsinki.

Keywords: Linked Data, Digital Heritage, Cultural Datasets, Connectivity Analytics, Data Integration, Data Enrichment, Verification

1 Introduction

There is a high proliferation of publishers that decide to provide their data as digital open data, since such data can be a valuable asset for scientists and users, and this is also the case for Cultural Heritage (CH) domain [5,9,11]. However, given the high volume of data (e.g., see an example in CH domain [1]), it a strong requirement that the data are Findable, Accessible, Interoperable, and Reuseable (FAIR) [24], for easing their interchange, integration, preservation and management. Therefore, an emerging challenge is to link and integrate these data at large scale, for aiding users to find all the data about an entity, to discover relationships, answer queries and better understanding the past in general, and to estimate data veracity. This need is quite important for CH domain comparing to other domains, since cultural data cover various disciplines, i.e., digital libraries (such as Europeana [9]), archaeological data [1], museums and galleries [14], cultural databases for identifying images [22], visual collections for 3D reconstruction [13], and others.

One way to achieve this is by publishing the data in a structured way by using Linked Open Data (LOD) techniques [3], and a typical Data Publishing and Integration scenario is introduced in the upper side of Figure 1. In particular, different providers produce data in many formats, e.g., CSV files, relational databases, and these data are usually transformed by using a specific model such as international standards like CIDOC-CRM [6] (see an example in [10]), for creating and publishing a central knowledge base as LOD. Existing approaches, such as FAST CAT [8] and Synthesis [7], can be exploited for performing the above process for cultural datasets. A further important step is to create links with existing LOD datasets, i.e., for enabling its publishing and connectivity to the popular LOD Cloud[1], which contains over 1,300 LOD Datasets (including over 150 datasets of CH domain), and for enabling the production of more advanced data access services.

However, due to the high and increasing number (and volume) of available LOD datasets and given the distributed nature of LOD, it is quite challenging to find all the datasets having commonalities with a given one. The major problems are that (i) it is inefficient and very time-consuming to discover and analyze every other LOD dataset (they can be even thousands), and (ii) publishers tend to use different URIs, names, schemas, languages and techniques for creating their data [19].

For tackling these difficulties, one can create cross-dataset relationships between entities and schemas, e.g., by using `owl:sameAs` relationships, however, it is not trivial since the `owl:sameAs` relationships a) model an equivalence relation and their transitive and symmetric closure has to be computed, and b) this presupposes knowledge of all datasets. For assisting this task at large scale, we have created `LODsyndesis` [16,17] knowledge graph, which has pre-computed the equivalence relationships among hundreds of LOD datasets (including 94 CH LOD datasets) and provides fast access to all the available information about an entity, through global scale entity-centric indexes and offers connectivity measurements for each underlying dataset. However, `LODsyndesis` contains a set of pre-collected manually fetched datasets, thereby it is not feasible to have access to such services for a new dataset before its actual publishing, although they can be an important asset for a dataset owner.

The key motivation (see the lower side of Fig. 1), is to connect the new dataset (e.g., a central knowledge base) to existing LOD datasets (through `LODsyndesis`) before its actual publishing, for ensuring its connectivity, for fixing possible connectivity errors, and for enriching its contents by discovering related datasets. For

[1] `https://lod-cloud.net`

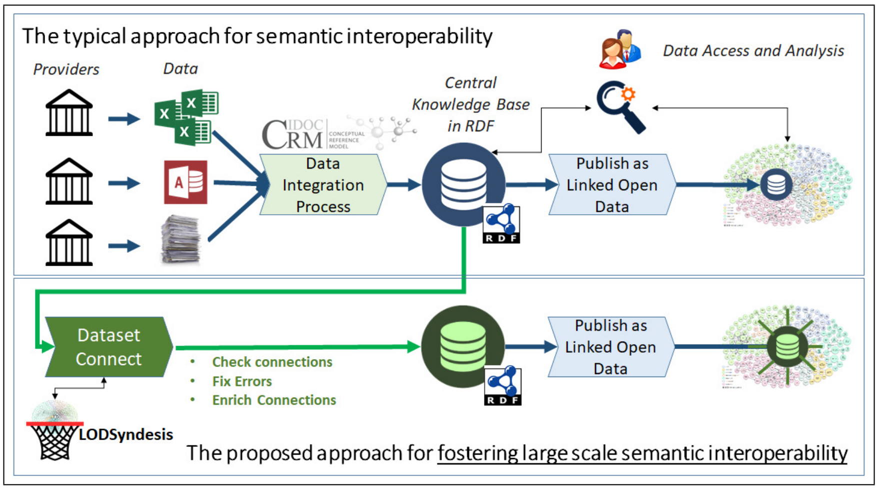

Fig. 1. The typical versus the proposed approach for fostering semantic interoperability

achieving this target, we introduce `ConnectionChecker` application, which extends `LODsyndesis` for providing such services directly at no cost for the dataset owner at real time. `ConnectionChecker` exploits the results of the transitive and symmetric closure of `owl:sameAs` relationships of `LODsyndesis`, for connecting the new dataset to `LODsyndesis` and for offering several measurements, visualizations and metadata for the input dataset, which can be exploited for evaluating its connectivity.

As a running example, suppose that in Fig. 2 we create a dataset containing data about "Jerusalem Old City Heritage" (JER), since we desire to perform an analysis about heritage sites, e.g., an analysis for "Holy Sepulchre" (e.g., see [1]). For making "JER" dataset more discoverable and reusable, we create links with one other CH dataset, say VIAF[2]. Our target is to connect "JER" dataset to `LODsyndesis`, for finding for each entity all its equivalent URIs (and their provenance), for enriching their information, and for discovering the top-K connected datasets to "JER" dataset.

Concerning our contribution, `ConnectionChecker` is accessible online[3], where one can easily check the connectivity of a dataset even in a few seconds for thousands of RDF triples. Moreover, we evaluate `ConnectionChecker` by using two real LOD datasets from CH domain, i.e., a) a dataset from the National Library of Netherlands[4], and b) a dataset for World War I [14] from the Universities of Aalto and Helsinki, Finland.

The rest of this paper is organized as follows. In §2, we give more details about Linked Data and `LODsyndesis` knowledge graph, and we present related approaches. In §3, we describe the `ConnectionChecker` application, and in §4

[2] https://viaf.org
[3] https://demos.isl.ics.forth.gr/ConnectionChecker/
[4] http://data.bibliotheken.nl

we evaluate ConnectionChecker through use cases from CH domain. Finally, §5 concludes the paper and discusses directions for future work.

2 Background & Related Work

Here, in §2.1 we provide background information about Linked Open Data and RDF, in §2.2 we describe LODsyndesis knowledge graph, whereas in §2.3 we present related approaches.

2.1 Background: Linked Data and RDF

"Linked Data refers to a method of publishing structured data, so that it can be interlinked and become more useful through semantic queries, founded on HTTP, RDF and URIs" [3]. The major principles of Linked Data, are the following: "(1) use URIs as names for things, (2) use HTTP URIs so that people can look up those names, (3) when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL), and (4) include links to other URIs, so that they can discover more things."

Concerning RDF, it is a knowledge base that can be represented as a graph. It identifies resources with URIs (Uniform Resource Identifiers), e.g., the URI of "Holy Sepulchre" in German National Library (DNB) is `http://d-nb.info/gnd/4073018-9`. RDF describes resources with triples, where each triple is a statement of the following form: subject-predicate-object. A subject describes an entity, a predicate corresponds to a property of that entity, and an object to the value of that property for the entity occurring as subject, e.g., the upper left side of Fig. 2 shows an example with 4 triples. One triple is the following: "Holy Sepulchre, founder, Constantine the Great", where "Holy Sepulchre" is the subject, "founder" the predicate and "Constantine the Great" the object.

In the running example, the prefix of each URI (or node), i.e., the text before ":", indicates the provenance of each URI, e.g., "dbp" means that the provenance is DBpedia knowledge base [12]. Finally, owl:sameAs relationships are used for denoting that two URIs (or nodes) refer to the same real world entity, e.g., in the upper left side of Fig. 2, we connected the URI of "Holy Sepulchre" of "JER" dataset (i.e., "jer:Holy_Sepulchre") with the equivalent VIAF URI ("viaf:Holy_Sepulchre") through an owl:sameAs property.

2.2 Background: LODsyndesis Knowledge Graph

The current version of LODsyndesis[5] [16, 17] contains over 400 million entities from 400 LOD datasets and two billion facts. It has pre-computed the transitive and symmetric closure of 44 millions owl:sameAs relationships among all the underlying datasets, for storing for each entity in global entity-centric indexes all its URIs, its triples and their provenance. The upper right side of Fig. 2

[5] https://demos.isl.ics.forth.gr/lodsyndesis

depicts the graph representation of `LODsyndesis` for "Holy Sepulchre" entity, e.g., `LODsyndesis` has computed all its equivalent URIs in all the datasets (see the nodes inside the big node) and has replaced all the URIs of "Holy Sepulchre" with a unique node (i.e., see the big node). It has also collected all its triples and their provenance (see the labels under each node in Fig. 2), e.g., the fact "Holy Sepulchre, style, Romanesque" occurs in DNB and DBpedia datasets.

By exploiting this knowledge graph the following services are offered [16, 17, 19]: (i) a service for finding all the available URIs, the provenance and all the facts about an entity (e.g., "Find all the URIs of Holy Sepulchre"), (ii) a fact checking service for estimating the veracity of a fact (e.g., "Is 335 AC the consecrated year of "Holy Sepulchre?"), (iii) Dataset Discovery services, for ensuring the connectivity of each dataset and for discovering the top-K relevant datasets to a given one, e.g., "Which are the top-K datasets having common entities with the National Library of France?". Moreover, it provides (iv) measurements among any subset of datasets, e.g., for finding the number of common entities among three datasets, e.g., "How many entities share VIAF, German and British Library?". Finally, it offers (v) several other services, e.g., machine-learning based services.

Limitation. The key limitation of `LODsyndesis` is that one is not feasible to add a new dataset before its actual publishing, e.g., for ensuring its connectivity, and for this reason we introduce `ConnectionChecker` application.

2.3 Related Work

There are several approaches focusing on data management for CH domain. [2] proposes an approach for making historical research data reusable according to the FAIR principles, through a collaborative ontology management environment. [5] introduces ArCo, a knowledge graph of Italian Cultural Heritage, which consists of several ontologies that model the CH domain. Furthermore, [25] presents an approach for enhancing knowledge management for Heritage Building Information Modeling (H-BIM) in CH domain through LOD techniques, whereas [11] presents a knowledge graph for Finland in the Second World War by using an infrastructure containing shared ontologies. [15] presents ARIADNE infrastructure for registering and connecting archaeological data and offers several data access services for the integrated resources. [4] introduces a framework for enriching the contents of CH datasets by using knowledge bases such as Wikidata [23], whereas [21] presents tools and methodologies for enriching and publishing CH data through LOD techniques. Finally, [9] describes a workflow for aiding linked CH data analysis and integration and a case study for the Europeana network.

Novelty of `ConnectionChecker`. Comparing to the above approaches, to the best of our knowledge `ConnectionChecker` is the first research prototype that assists a data publisher (e.g., from CH domain) to evaluate the connectivity and to discover new connections for their dataset before its actual publishing. In this way, the data publishers can enrich (or verify) the contents of their dataset before publishing it to the LOD cloud.

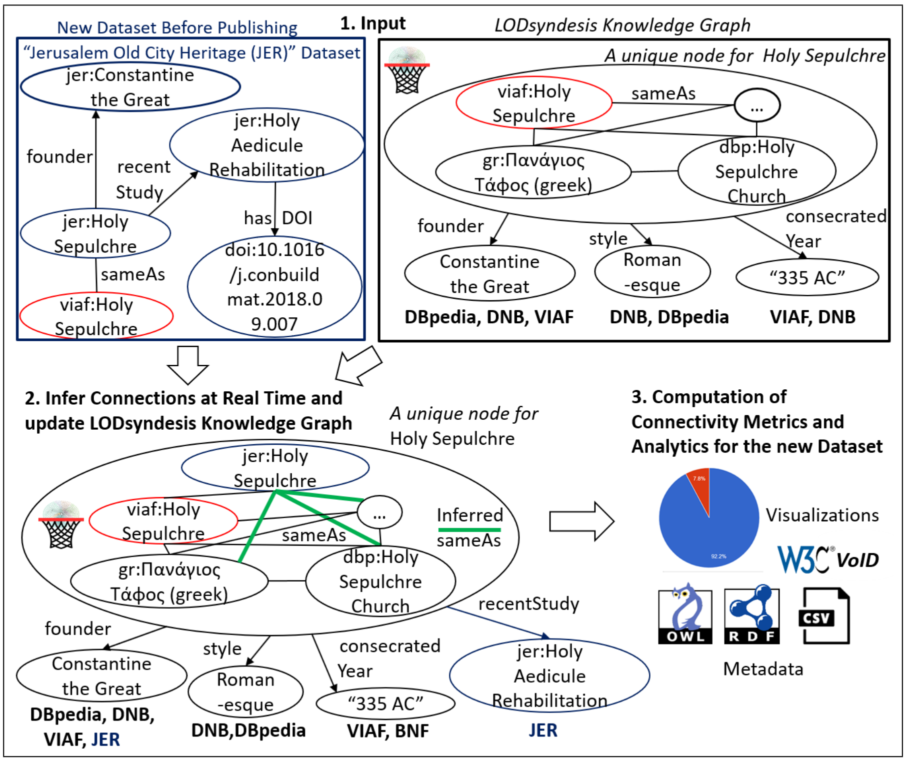

Fig. 2. Running example: Connecting the dataset "Jerusalem Old City Heritage" (JER) to the rest LOD Cloud before its actual publishing through `ConnectionChecker`

3 The Steps of `ConnectionChecker`

The process of `ConnectionChecker` consists of there different steps. In particular, in §3.1 we describe how to receive the input from a dataset publisher (i.e., Step 1), in §3.2 we analyze how to infer new connections by using `LODsyndesis` (i.e., Step 2), and in §3.3 we describe how we compute the connectivity measurements and what analytics and visualizations are offered (i.e., Step 3). Finally, in §3.4 we provide details about the current status of `ConnectionChecker`.

3.1 Step 1. Input from a Dataset Publisher

The dataset owner fills a form about their dataset (see the left side of Fig. 3). Specifically, the user/publisher should provide a link containing the RDF triples in N-Triples format[6], the name of the dataset, its URL and its domain. Finally, for having a very fast overview for the connectivity of a given dataset, one can optionally select to perform the measurements for a smaller part of their dataset (e.g., for the first 10,000 triples).

[6] `https://www.w3.org/TR/n-triples/`

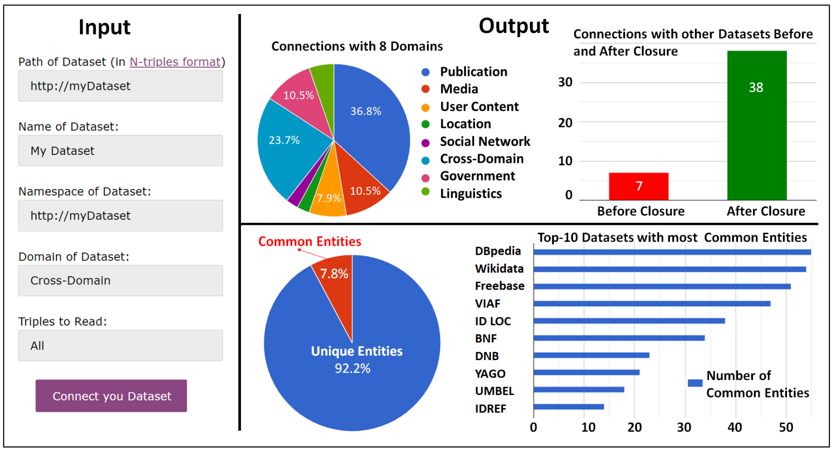

Fig. 3. Input and output of `ConnectionChecker`

3.2 Step 2. Inferring new connections by updating `LODsyndesis`.

`ConnectionChecker` retrieves the dataset, and merges the `LODsyndesis` knowledge graph with the input dataset. In particular, it detects the equivalence relationships of the new dataset and computes the transitive and symmetric closure of equivalence relationships between the input dataset and the already existing datasets in `LODsyndesis`, for discovering inferred equivalence relationships. For instance, in the upper left side of Fig. 2, we can see that the URI "jer:Holy_Sepulchre" of "JER" dataset is connected with a `owl:sameAs` relationship with the corresponding URI of VIAF dataset, i.e., "viaf:Holy_Sepulchre".

In `LODsyndesis`, i.e., see the upper right side of Fig. 2, the latter VIAF URI, is connected with several URIs from other datasets that contain information about "Holy_Sepulchre" (e.g., see the connections with DBpedia and with a dataset in Greek language). By merging these relationships (see the lower left side of Fig. 2), we inferred new `owl:sameAs` relationships for the URI of "Holy_Sepulchre" in "JER" dataset (see the green thick connections), and we discovered commonalities with other datasets (except for VIAF). For example, we inferred that "JER" is connected with DBpedia dataset, since the URI "jer:Holy_Sepulchre" refers to the same entity as the URI "dbp:Holy Sepulchre Church".

By finding such connections, it is feasible to enrich and to verify the contents of a dataset. For example, in the lower left side of Fig. 2, a) we can enrich the contents of "JER" dataset about "Holy Sepulchre" from other datasets, e.g., the information about the style and consecrated year of "Holy Sepulchre" were not included in "JER" dataset, and b) we can verify the fact "Constantine the Great was the founder of Holy Sepulchre" from three other datasets.

3.2.1 Detecting Possible Errors.

`ConnectionChecker` can detect possible `owl:sameAs` errors by checking if an entity of the new dataset is connected with two or more real entities. For example, suppose that in Fig. 4 we have created the

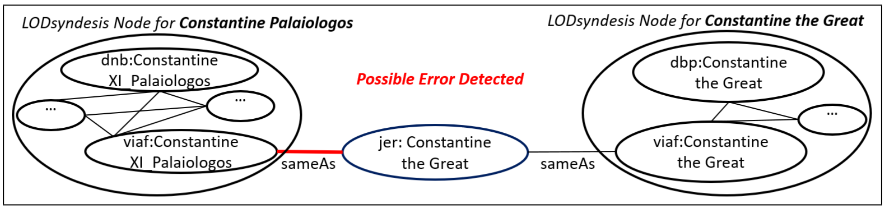

Fig. 4. Detect possible connectivity errors through `ConnectionChecker`

following `owl:sameAs` relationships (e.g., through an instance matching tool) for Constantine the Great: "jer:Constantine the Great `owl:sameAs` viaf:Constantine the Great" and "jer:Constantine the Great `owl:sameAs` viaf:Constantine XI Palaiologos". Since `LODsyndesis` has computed that "viaf:Constantine the Great" and "viaf:Constantine XI Palaiologos" refer to different entities, the user is informed that at least one of these `owl:sameAs` relationships is probably incorrect.

3.3 Step 3. Computation of Connectivity Measurements and Production of Analytics and Visualizations

`ConnectionChecker` exploits the updated `LODsyndesis` knowledge graph for computing content-based connectivity measurements (which are described in [18, 20]), and for producing several connectivity analytics and visualizations, see the lower right side of Fig. 2. In brief, for the input dataset the following connectivity analytics and dataset discovery measurements are computed and visualized (examples of visualizations are shown in the right side of Fig. 3):

- The number of common entities, i.e., indicates how many entities of the new dataset can be found in at least one other LOD dataset, and the average number of datasets containing each common entity.
- The number of unique entities, i.e., how many entities exist only in the new (input) dataset, and the corresponding percentage.
- The number of inferred `owl:sameAs` relationships (and the corresponding increase percentage), and the number of possible `owl:sameAs` errors.
- The number of connections before and after connecting to `LODsyndesis` (and the corresponding increase percentage), i.e., for estimating the gain of transitive and symmetric closure of `owl:sameAs` relationships.
- The dataset ranking according to the number of connections in descending order, i.e., if the ranking of the dataset is 1, it means that it is the most connected dataset (it has more connections than any other LOD dataset).
- The top-K datasets having the most common entities with the entities of the new dataset (i.e., the most relevant datasets to the given dataset).
- The number of connections of the new dataset with each domain (e.g., how many connections exist with datasets from CH domain).
- The top K triads and quads of datasets with the most common entities that contain the input dataset, based on lattice-based measurements [16, 20].

How to exploit the results. The user can either browse the above results through visualizations or export these results, all the inferred equivalence

Table 1. Measurements for *PNLN* and *WW1LOD* datasets.

Measurement	**Results for *PNLN***	**Results for *WW1LOD***
# of unique entities (% wrt all entities of dataset)	582,234 (79.6%)	10,270 (92.6%)
# of common entities (% wrt all entities of dataset)	149,222 (20.4%)	825 (7.4%)
avg # of datasets containing each common entity	4.9	7.9
# of `owl:sameAs` relationships	535,407	547
# of inferred `owl:sameAs` relationships	615,813	2,172
Increase % of `owl:sameAs` relationships	115%	397%
# of possible errors in `owl:sameAs` relationships	0	0
# of connections before `owl:sameAs` closure	3	5
# of connections after `owl:sameAs` closure	38	29
Increase % of connections	1166%	480%
# of connections with CH datasets	15	12
Dataset ranking in connections (all datasets)	184 (out of 401)	212 (out of 401)
Dataset ranking in connections (CH datasets)	38 (out of 95)	43 (out of 95)

`owl:sameAs` relationships of each entity, the provenance of each entity, and metadata according to specific standards, i.e., in VoID[7] and CSV[8] formats.

3.4 The current status of `ConnectionChecker`

`ConnectionChecker` is an online web application[9]. It uses standard technologies, e.g., JavaScript and JAVA Servlets, and Google Charts[10] for creating the visualizations. A tutorial video for `ConnectionChecker` is accessible online[11].

4 Evaluation: Use Cases in Cultural Heritage Domain

We present use cases for two CH datasets for estimating their connectivity and for evaluating the gain of connecting them to `LODsyndesis` through `ConnectionChecker`. In particular, we use a) a subset of the dataset "Persons of National Library of Netherlands (*PNLN*)"[12], which contains the following data about persons (e.g., writers, historians): 3,000,000 triples (facts), 731,456 entities, and 535,407 `owl:sameAs` relationships. The initial dataset contains links to three LOD datasets: *Wikidata* [23], *VIAF* and ISNI[13]. Moreover, we use b) the dataset "World War I as LOD (*WW1LOD*)" [14] from Aalto and Helsinki Universities, Finland, which contains data about World War I, i.e., 47,616 triples, 11,095 entities and 547 `owl:sameAs` mappings. The initial dataset contains links to five LOD datasets.

[7] `https://www.w3.org/TR/void/`
[8] `https://en.wikipedia.org/wiki/Comma-separated_values`
[9] `https://demos.isl.ics.forth.gr/ConnectionChecker/`
[10] `https://developers.google.com/chart`
[11] `https://youtu.be/vwKu5nVnjoM`
[12] `http://data.bibliotheken.nl/doc/dataset/persons`
[13] `http://isni.org/`

Table 2. Top-5 Datasets having common entities with *PNLN*

Rank	Dataset having common entities with *PNLN*	Domain	Number of Common Entities
1	VIAF	Cultural Heritage	148,756
2	Library of Congress	Cultural Heritage	131,607
3	Germany National Library (DNB)	Cultural Heritage	67,174
4	France National Library (BNF)	Cultural Heritage	59,998
5	British National Library	Cultural Heritage	42,861

Table 3. Top-5 Connected Triads of Datasets including *PNLN*

Rank	Triad of Datasets	# of Common Entities
1	*PNLN*,*VIAF*,*Library of Congress*	131,578
2	*PNLN*,*VIAF*,*DNB*	67,173
3	*PNLN*,*VIAF*,*BNF*	59,914
4	*PNLN*,*DNB*,*Library of Congress*	55,458
5	*PNLN*,*Library of Congress*,*BNF*	54,586

For the measurements, we used a single computer with 8 cores, 8 GB main memory and 60 GB disk space. For *PNLN* dataset, `ConnectionChecker` needed 45 minutes for computing the results (due to its high number of `owl:sameAs` relationships), whereas for *WW1LOD* it needed only 30 seconds (since it is much smaller in size). The results of the measurements are accessible in a catalog[14]. Some indicative results are presented in Tables 1-3 and are analyzed below.

4.1 Results for *PNLN* dataset

In the second column of Table 1, we can see that *PNLN* shares over 149,000 entities with at least one other LOD dataset, i.e., 20.4% of all entities of *PNLN*, whereas each of these entities can be found on average in 4.9 datasets. Moreover, the gain of connecting *PNLN* to `LODsyndesis` is obvious, i.e., we managed to infer 615,813 new `owl:sameAs` relationships (i.e., 115% increase), which resulted in 35 new "inferred" connections with other datasets for *PNLN* (i.e., 1166% increase). We identified that *PNLN* shares entities with 15 other CH datasets.

Table 2 shows the 5 datasets (and their domain) having the most common entities with *PNLN*. We can see that it shares thousands of entities with other CH datasets, indicatively 67,174 entities with *DNB*, although there were no connections between these two datasets in the initial *PNLN* dataset. Finally, Table 3 shows measurements among triads of datasets that include *PNLN*, e.g., we can see that *PNLN*, *VIAF* and *Library of Congress* share over 131,000 entities.

4.2 Results for *WW1LOD* dataset

In the third column of Table 1, we can see that *WW1LOD* dataset shares over 800 entities with at least one other dataset. Moreover, we inferred 2,172 new

[14] `http://islcatalog.ics.forth.gr/el/dataset/connectionchecker`

`owl:sameAs` mappings and we discovered 24 new connections for *WW1LOD*, whereas we found that it is connected with 12 CH datasets. Finally, the most connected dataset to *WW1LOD* is *Wikidata* with 792 commons entities, whereas the most connected dataset from CH domain is *VIAF* with 369 common entities.

5 Concluding Remarks

We presented a research prototype called `ConnectionChecker`, which exploits `LODsyndesis` for connecting any dataset, e.g., from Cultural Heritage (CH) domain, to the rest Linked Open Data (LOD). In particular, `ConnectionChecker` can be exploited for ensuring the connectivity of a LOD dataset before its actual publishing, i.e., for discovering the most relevant datasets to this dataset, for better understanding the past, for enriching the information of its entities, and for estimating the veracity of its data.

Regarding evaluation, we introduced use cases from CH domain, e.g., for a dataset derived from the National Library of Netherlands. Indicatively, with `ConnectionChecker` we found that this dataset shares common entities with 38 other datasets, although the initial dataset included links to only 3 other datasets (i.e., 1166% increase). Moreover, we found that it shares even thousands of entities with several CH datasets (such as the National Library of France), although the initial dataset did not contain links to these datasets.

As a future work, we plan (a) to provide more connectivity measurements for an input dataset, e.g., for its schema elements and for its triples, (b) to describe all the technical details of `ConnectionChecker`, and (c) to evaluate its efficiency.

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

References

1. E. Alexakis, E. Kapassa, M. Touloupou, D. Kyriazis, A. Georgopoulos, and A. Moropoulou. Innovative methodology for personalized 3d representation and big data management in cultural heritage. In *International Conference on TMM-CH*, pages 69–77. Springer, 2018.
2. F. Beretta. A challenge for historical research: making data FAIR using a collaborative ontology management environment (ontome). *Semantic Web*, 1(12):279–294, 2021.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global, 2011.
4. G. Candela, P. Escobar, R. C. Carrasco, and M. Marco-Such. A linked open data framework to enhance the discoverability and impact of culture heritage. *Journal of Information Science*, 45(6):756–766, 2019.
5. V. A. Carriero, A. Gangemi, M. L. Mancinelli, A. G. Nuzzolese, V. Presutti, and C. Veninata. Pattern-based design applied to cultural heritage knowledge graphs. *Semantic Web*, 12(2):313–357, 2021.

6. M. Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75, 2003.
7. P. Fafalios, K. Konsolaki, L. Charami, K. Petrakis, M. Paterakis, D. Angelakis, Y. Tzitzikas, C. Bekiari, and M. Doerr. Towards semantic interoperability inhistorical research: Documenting research data and knowledge withsynthesis. In *International Semantic Web Conference*, pages 682–698. Springer, 2021.
8. P. Fafalios, K. Petrakis, G. Samaritakis, K. Doerr, A. Kritsotaki, Y. Tzitzikas, and M. Doerr. FAST CAT: Collaborative data entry and curation for semantic interoperability in digital humanities. *J. Comput. Cult. Herit.*, 14(4), July 2021.
9. N. Freire, R. Voorburg, R. Cornelissen, S. de Valk, E. Meijers, and A. Isaac. Aggregation of linked data in the cultural heritage domain: A case study in the europeana network. *Information*, 10(8):252, 2019.
10. I. Karvasonis. Study of the heterogeneity in cultural databases and transformation of examples from CIMI to the CIDOC CRM. Technical report, Citeseer, 2001.
11. M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, and E. Hyvönen. Warsampo knowledge graph: Finland in the second world war as linked open data. *Semantic Web*, (Preprint):1–14, 2021.
12. 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(2):167–195, 2015.
13. K. Makantasis, A. Doulamis, N. Doulamis, and M. Ioannides. In the wild image retrieval and clustering for 3d cultural heritage landmarks reconstruction. *Multimedia Tools and Applications*, 75(7):3593–3629, 2016.
14. E. Mäkelä, J. Törnroos, T. Lindquist, and E. Hyvönen. WW1LOD: an application of CIDOC-CRM to World War 1 linked data. *IJDL*, 18(4):333–343, 2017.
15. C. Meghini, R. Scopigno, J. Richards, H. Wright, G. Geser, S. Cuy, J. Fihn, B. Fanini, H. Hollander, F. Niccolucci, et al. ARIADNE: A research infrastructure for archaeology. *ACM JOCCH*, 10(3):1–27, 2017.
16. M. Mountantonakis. *Services for Connecting and Integrating Big Numbers of Linked Datasets*. IOS PRESS, 2021.
17. M. Mountantonakis and Y. Tzitzikas. LODsyndesis: global scale knowledge services. *Heritage*, 1(2):335–348, 2018.
18. M. Mountantonakis and Y. Tzitzikas. Scalable methods for measuring the connectivity and quality of large numbers of linked datasets. *JDIQ*, 9(3):1–49, 2018.
19. M. Mountantonakis and Y. Tzitzikas. Large-scale semantic integration of linked data: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40, 2019.
20. M. Mountantonakis and Y. Tzitzikas. Content-based union and complement metrics for dataset search over RDF knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, 12(2):1–31, 2020.
21. N. Simou, A. Chortaras, G. Stamou, and S. Kollias. Enriching and publishing cultural heritage as linked open data. In *Mixed Reality and Gami cation for Cultural Heritage*, pages 201–223. Springer, 2017.
22. E. Valle, M. Cord, and S. Philipp-Foliguet. CBIR in cultural databases for identification of images: A local-descriptors approach. 2008.
23. D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.
24. M. D. Wilkinson, M. Dumontier, et al. The FAIR guiding principles for scientific data management and stewardship. *Scienti c data*, 3(1):1–9, 2016.
25. A. E. Ziri, P. Bonsma, I. Bonsma, E. Iadanza, F. Maietti, M. Medici, F. Ferrari, and P. M. Lerones. Cultural heritage sites holistic documentation through semantic web technologies. In *International Conference on TMM-CH*, pages 347–358, 2018.