heritage

Article

LODsyndesis: Global Scale Knowledge Services

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

[1] Institute of Computer Science, FORTH-ICS, 70013 Heraklion, Greece
[2] Department of Computer Science, University of Crete, 70013 Heraklion, Greece
[*] Correspondence: mountant@ics.forth.gr (M.M.); tzitzik@ics.forth.gr (Y.T.)

Received: 27 October 2018; Accepted: 14 November 2018; Published: 17 November 2018

Abstract: In this paper, we present LODsyndesis, a suite of services over the datasets of the entire Linked Open Data Cloud, which offers fast, content-based dataset discovery and object co-reference. Emphasis is given on supporting scalable cross-dataset reasoning for finding all information about any entity and its provenance. Other tasks that can be benefited from these services are those related to the quality and veracity of data since the collection of all information about an entity, and the cross-dataset inference that is feasible, allows spotting the contradictions that exist, and also provides information for data cleaning or for estimating and suggesting which data are probably correct or more accurate. In addition, we will show how these services can assist the enrichment of existing datasets with more features for obtaining better predictions in machine learning tasks. Finally, we report measurements that reveal the sparsity of the current datasets, as regards their connectivity, which in turn justifies the need for advancing the current methods for data integration. Measurements focusing on the cultural domain are also included, specifically measurements over datasets using CIDOC CRM (Conceptual Reference Model), and connectivity measurements of British Museum data. The services of LODsyndesis are based on special indexes and algorithms and allow the indexing of 2 billion triples in around 80 min using a cluster of 96 computers.

Keywords: Semantic web; connectivity analytics; british museum; cultural RDF datasets; CIDOC CRM

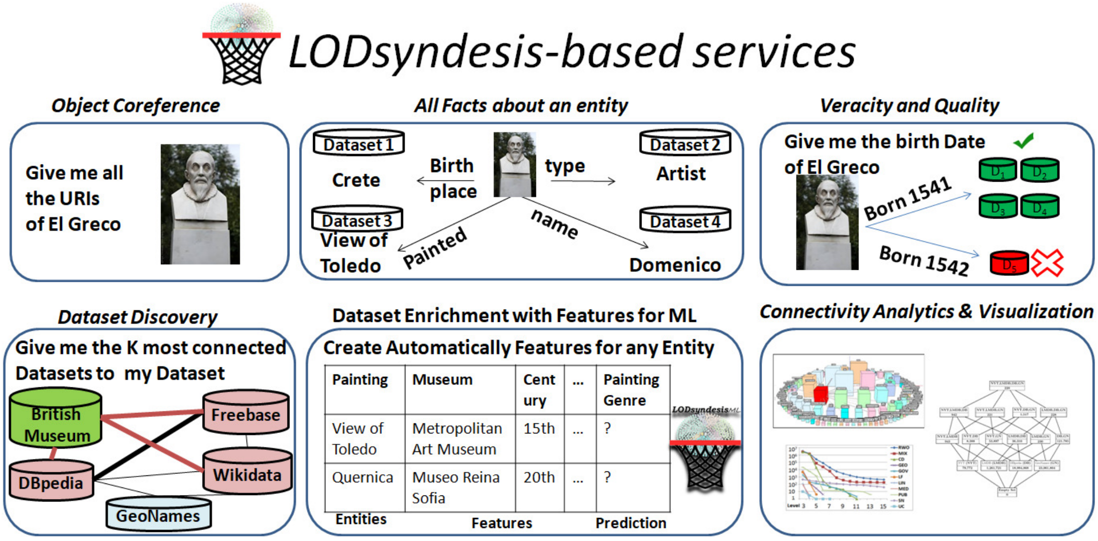
1. Introduction

In recent years, a large volume of open data has been published and this number keeps increasing. However, it is necessary such open data to be Findable, Accessible, Interoperable and Reusable (FAIR; see more information for the FAIR principles in [1]), and for this reason there is an attempt for using standards and good practices, to achieve these targets. Moreover, one major objective is to link and integrate these data, to enable fast access to all the available information about an entity (by also preserving their provenance), and to estimate the veracity and correctness of these data. One way to achieve linking and integration is to publish such data in a structured way, by using Linked Data, and thousands of datasets from various domains that use linked data techniques have already been published, i.e., approximately 10,000 datasets according to [2]. However, the semantic integration of data at a large scale is not a straightforward task, since there are various difficulties that should be tackled to achieve such a target. The main difficulties follow: (i) publishers tend to use different models and formats for the representation of their data; (ii) different URIs (Uniform Resource Identifiers) or languages are used for describing the same entities; (iii) publishers describe their data by using different concepts, e.g., CIDOC CRM (Conceptual Reference Model) [3] represents the birth date of a person as an event, while DBpedia [4] uses a single triple for the same fact; (iv) data from different sources can be inconsistent or conflicting; (v) a lot of complementary information occur in different sources; and (vi) many datasets are updated very frequently.

For instance, suppose that one desires to describe a specific real fact, say "Heraklion is the birth place of El Greco". Even for this simple example, two or more datasets can use different URIs to describe the entities of that fact, i.e., "Heraklion" and "El Greco", and the schema element "birth place". By using RDF (Resource Description Framework) and Linked Data, the difficulties of different schemas and URIs for the same concepts and entities can be partially tackled by creating equivalence relationships between entities and schemas of different datasets. In particular, it can be achieved through the exploitation of some predefined equivalence relationships (or properties), such as owl:sameAs, owl:equivalentProperty and owl:equivalentClass. However, the aforementioned relations are symmetric and transitive, therefore it is mandatory to compute their transitive and symmetric closure, in order to collect all the available data for an entity, without missing entities and facts that occur in two or more datasets. Moreover, this presupposes knowledge from all the available datasets, otherwise we would fail to find all the URIs that are used (from different datasets) for representing an entity. As a consequence, in order to find all URIs and facts about an entity, say El Greco, we have to index and enrich numerous datasets, through cross-dataset inference.

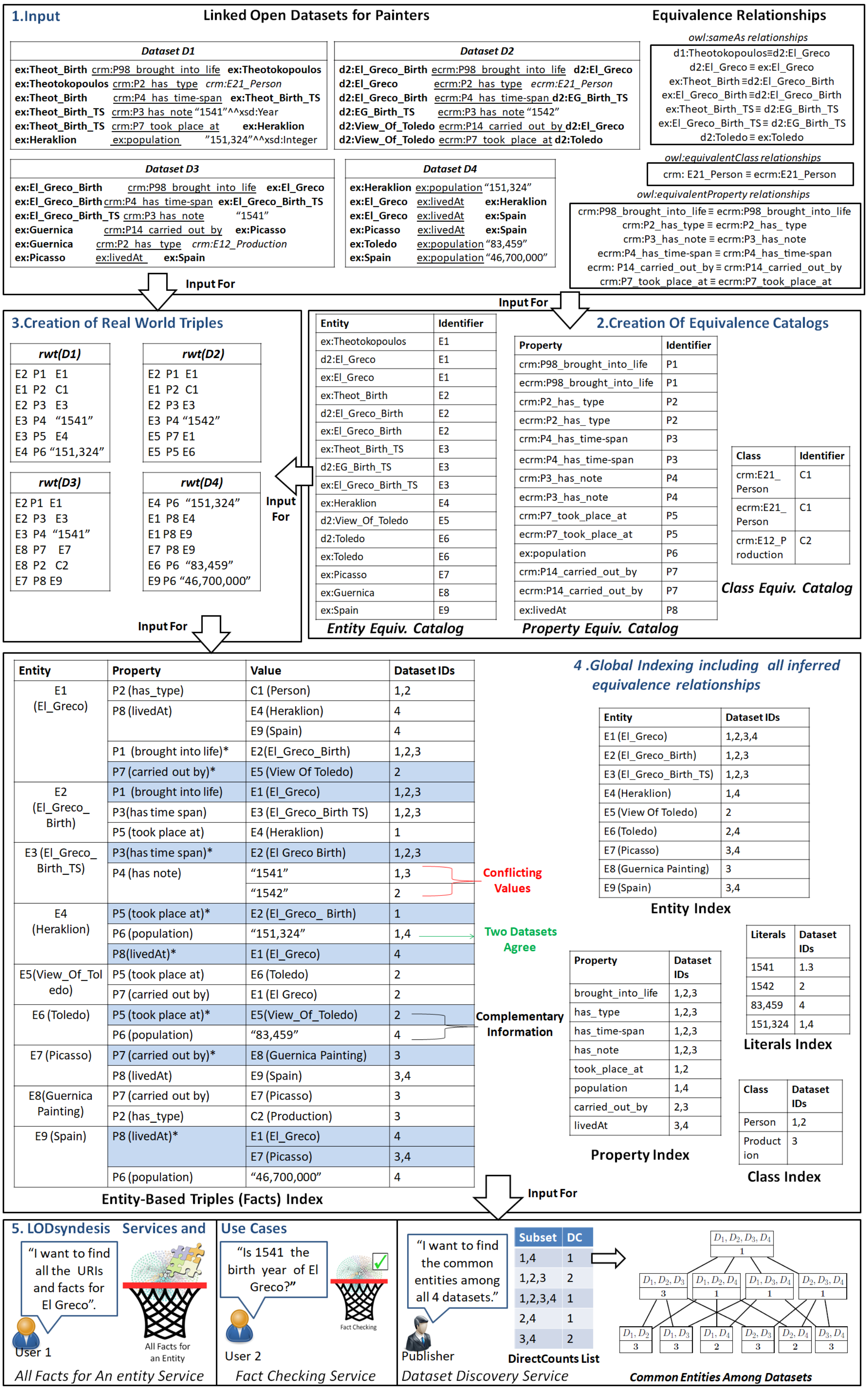
For this reason, i.e., assisting the process of semantic integration of data at large scale, we have designed and developed novel indexes, methods and tools [5–7]. The current suite of services and tools that have been developed are known as "LODsyndesis" (http://www.ics.forth.gr/isl/LODsyndesis). The major characteristic of LODsyndesis is that it indexes the whole content of hundreds of datasets in the Linked Open Data cloud, by taking into consideration the closure of equivalence relationships, and to the best of our knowledge LODsyndesis is the "largest knowledge graph of Linked Data that includes all inferred equivalence relationships". All these semantics-aware indexes are exploited, to perform fast connectivity analytics and to offer advanced connectivity services that are of primary importance for several real world tasks. An overview of the available LODsyndesis-based services for these tasks can be seen in Figure 1. With respect to the work that have been made [5–7], in this paper:

- We describe in brief the process of constructing semantic indexes and performing connectivity measurements for any subset of datasets.
- We introduce specific use cases and services, we mention how they can also be important in cultural domain by showing specific examples, whereas we show ways to exploit them (e.g., programmatically through a REST API or through an HTML page).
- We report connectivity analytics for hundreds of LOD Cloud datasets, by focusing on publications (and cultural heritage) domain, and we show measurements for datasets that use CIDOC CRM model, such as British Museum.


LODsyndesis-based services
Object Coreference
Give me all the URIs of El Greco
All Facts about an entity
Dataset 1
Crete
Birth place
type
Dataset 2
Artist
Dataset 3
View of Toledo
Painted
name
Dataset 4
Domenico
Veracity and Quality
Give me the birth Date of El Greco
Born 1541
Born 1542
D1
D2
D3
D4
D5
Dataset Discovery
Give me the K most connected Datasets to my Dataset
British Museum
Freebase
DBpedia
Wikidata
GeoNames
Dataset Enrichment with Features for ML
Create Automatically Features for any Entity
Entities
Features
Prediction
LODsyndesisML
Connectivity Analytics & Visualization

Painting	Museum	Century	...	Painting Genre
View of Toledo	Metropolitan Art Museum	15th	...	?
Quernica	Museo Reina Sofia	20th	...	?

Figure 1. The services provided by LODsyndesis with examples from cultural domain.


1.Input
Linked Open Datasets for Painters
Equivalence Relationships
Input For
3.Creation of Real World Triples
2.Creation Of Equivalence Catalogs
Input For
Input For
4 .Global Indexing including all inferred equivalence relationships
Conflicting Values
Two Datasets Agree
Complementary Information
Input For
5. LODsyndesis Services and Use Cases
"I want to find all the URIs and facts for El Greco".
User 1
All Facts for an Entity
All Facts for An entity Service
"Is 1541 the birth year of El Greco?"
Fact Checking
User 2
Fact Checking Service
"I want to find the common entities among all 4 datasets."
Publisher
Dataset Discovery Service
DirectCounts List
Common Entities Among Datasets
D_1, D_2, D_3, D_4 1
D_1, D_2, D_3 3
D_1, D_2, D_4 1
D_1, D_3, D_4 1
D_2, D_3, D_4 1
D_1, D_2 3
D_1, D_3 3
D_1, D_4 2
D_2, D_3 3
D_2, D_4 2
D_3, D_4 3

Dataset D1

ex:Theot_Birth	crm:P98_brought_into_life	**ex:Theotokopoulos**
ex:Theotokopoulos	crm:P2_has_type	*crm:E21_Person*
ex:Theot_Birth	crm:P4_has_time-span	**ex:Theot_Birth_TS**
ex:Theot_Birth_TS	crm:P3_has_note	"1541"^^xsd:Year
ex:Theot_Birth_TS	crm:P7_took_place_at	**ex:Heraklion**
ex:Heraklion	ex:population	"151,324"^^xsd:Integer

Dataset D2

d2:El_Greco_Birth	ecrm:P98_brought_into_life	**d2:El_Greco**
d2:El_Greco	ecrm:P2_has_type	*ecrm:E21_Person*
d2:El_Greco_Birth	ecrm:P4_has_time-span	**d2:EG_Birth_TS**
d2:EG_Birth_TS	ecrm:P3_has_note	"1542"
d2:View_Of_Toledo	ecrm:P14_carried_out_by	**d2:El_Greco**
d2:View_Of_Toledo	ecrm:P7_took_place_at	**d2:Toledo**

Dataset D3

ex:El_Greco_Birth	crm:P98_brought_into_life	**ex:El_Greco**
ex:El_Greco_Birth	crm:P4_has_time-span	**ex:El_Greco_Birth_TS**
ex:El_Greco_Birth_TS	crm:P3_has_note	"1541"
ex:Guernica	crm:P14_carried_out_by	**ex:Picasso**
ex:Guernica	crm:P2_has_type	*crm:E12_Production*
ex:Picasso	ex:livedAt	**ex:Spain**

Dataset D4

ex:Heraklion	ex:population	"151,324"
ex:El_Greco	ex:livedAt	**ex:Heraklion**
ex:El_Greco	ex:livedAt	**ex:Spain**
ex:Picasso	ex:livedAt	**ex:Spain**
ex:Toledo	ex:population	"83,459"
ex:Spain	ex:population	"46,700,000"

owl:sameAs relationships

d1:Theotokopoulos≡d2:El_Greco
d2:El_Greco ≡ ex:El_Greco
ex:Theot_Birth ≡d2:El_Greco_Birth
ex:El_Greco_Birth ≡d2:El_Greco_Birth
ex:Theot_Birth_TS ≡ d2:EG_Birth_TS
ex:El_Greco_Birth_TS ≡ d2:EG_Birth_TS
d2:Toledo ≡ ex:Toledo

owl:equivalentClass relationships

crm: E21_Person ≡ ecrm:E21_Person

owl:equivalentProperty relationships

crm:P98_brought_into_life ≡ ecrm:P98_brought_into_life
crm:P2_has_type ≡ ecrm:P2_has_ type
crm:P3_has_note ≡ ecrm:P3_has_note
ecrm:P4_has_time-span ≡ crm:P4_has_time-span
ecrm: P14_carried_out_by ≡ crm:P14_carried_out_by
crm:P7_took_place_at ≡ ecrm:P7_took_place_at

rwt(D1)

E2	P1	E1
E1	P2	C1
E2	P3	E3
E3	P4	"1541"
E3	P5	E4
E4	P6	"151,324"

rwt(D2)

E2	P1	E1
E1	P2	C1
E2	P3	E3
E3	P4	"1542"
E5	P7	E1
E5	P5	E6

rwt(D3)

E2	P1	E1
E2	P3	E3
E3	P4	"1541"
E8	P7	E7
E8	P2	C2
E7	P8	E9

rwt(D4)

E4	P6	"151,324"
E1	P8	E4
E1	P8	E9
E7	P8	E9
E6	P6	"83,459"
E9	P6	"46,700,000"

Entity	Identifier
ex:Theotokopoulos	E1
d2:El_Greco	E1
ex:El_Greco	E1
ex:Theot_Birth	E2
d2:El_Greco_Birth	E2
ex:El_Greco_Birth	E2
ex:Theot_Birth_TS	E3
d2:EG_Birth_TS	E3
ex:El_Greco_Birth_TS	E3
ex:Heraklion	E4
d2:View_Of_Toledo	E5
d2:Toledo	E6
ex:Toledo	E6
ex:Picasso	E7
ex:Guernica	E8
ex:Spain	E9

Entity Equiv. Catalog

Property	Identifier
crm:P98_brought_into_life	P1
ecrm:P98_brought_into_life	P1
crm:P2_has_ type	P2
ecrm:P2_has_ type	P2
crm:P4_has_time-span	P3
ecrm:P4_has_time-span	P3
crm:P3_has_note	P4
ecrm:P3_has_note	P4
crm:P7_took_place_at	P5
ecrm:P7_took_place_at	P5
ex:population	P6
crm:P14_carried_out_by	P7
ecrm:P14_carried_out_by	P7
ex:livedAt	P8

Property Equiv. Catalog

Class	Identifier
crm:E21_ Person	C1
ecrm:E21_ Person	C1
crm:E12_P roduction	C2

Class Equiv. Catalog

Entity	Property	Value	Dataset IDs
E1 (El_Greco)	P2 (has_type)	C1 (Person)	1,2
	P8 (livedAt)	E4 (Heraklion)	4
		E9 (Spain)	4
	P1 (brought into life)*	E2(El_Greco_Birth)	1,2,3
	P7 (carried out by)*	E5 (View Of Toledo)	2
E2 (El_Greco_ Birth)	P1 (brought into life)	E1 (El_Greco)	1,2,3
	P3(has time span)	E3 (El_Greco_Birth TS)	1,2,3
	P5 (took place at)	E4 (Heraklion)	1
E3 (El_Greco_ Birth_TS)	P3(has time span)*	E2 (El Greco Birth)	1,2,3
	P4 (has note)	"1541"	1,3
		"1542"	2
E4 (Heraklion)	P5 (took place at)*	E2 (El_Greco_ Birth)	1
	P6 (population)	"151,324"	1,4
	P8(livedAt)*	E1 (El_Greco)	4
E5(View_Of_Tol edo)	P5 (took place at)	E6 (Toledo)	2
	P7 (carried out by)	E1 (El Greco)	2
E6 (Toledo)	P5 (took place at)*	E5(View_Of_Toledo)	2
	P6 (population)	"83,459"	4
E7 (Picasso)	P7 (carried out by)*	E8 (Guernica Painting)	3
	P8 (livedAt)	E9 (Spain)	3,4
E8(Guernica Painting)	P7 (carried out by)	E7 (Picasso)	3
	P2 (has_type)	C2 (Production)	3
E9 (Spain)	P8 (livedAt)*	E1 (El_Greco)	4
		E7 (Picasso)	3,4
	P6 (population)	"46,700,000"	4

Entity-Based Triples (Facts) Index

Entity	Dataset IDs
E1 (El_Greco)	1,2,3,4
E2 (El_Greco_Birth)	1,2,3
E3 (El_Greco_Birth_TS)	1,2,3
E4 (Heraklion)	1,4
E5 (View Of Toledo)	2
E6 (Toledo)	2,4
E7 (Picasso)	3,4
E8 (Guernica Painting)	3
E9 (Spain)	3,4

Entity Index

Literals	Dataset IDs
1541	1.3
1542	2
83,459	4
151,324	1,4

Literals Index

Property	Dataset IDs
brought_into_life	1,2,3
has_type	1,2,3
has_time-span	1,2,3
has_note	1,2,3
took_place_at	1,2
population	1,4
carried_out_by	2,3
livedAt	3,4

Property Index

Class	Dataset IDs
Person	1,2
Product ion	3

Class Index

Subset	DC
1,4	1
1,2,3	2
1,2,3,4	1
2,4	1
3,4	2

Figure 2. Running Example-The process of Global Indexing.

The rest of this paper is organized as follows. Section 2 introduces the background and discusses related work, Section 3 shows in brief the process of semantics-aware indexes construction and how to perform connectivity analytics. Section 4 introduces real use cases where the aforementioned indexes and connectivity analytics can be exploited, whereas Section 5 reports connectivity analytics for hundreds of LOD (Linken Open Data) Cloud datasets, especially from publications domain and for the dataset of British Museum. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Context and Related Work

Here, in Section 2.1 we introduce some information about RDF and Linked Data, while in Section 2.2, we mention related approaches and services.

2.1. RDF and Linked Data

Resource Description Framework (RDF) [8,9] is a graph-based data model. It uses Uniform Resource Identifiers (URIs), or anonymous resources (blank nodes) to denote resources, and constants (Literals), while triples are used to relate a resource with other resources or constants. A triple is a statement of the form subject-predicate-object (s,p,o), and it is any element of $T = (U \cup Bn) \times (U) \times (U \cup Bn \cup L)$, where U, Bn and L denote the sets of URIs, blank nodes and literals, respectively, whereas an RDF graph (or dataset) is any finite subset of T. For instance, the triple (d1:El_Greco, d1:birthPlace, d1:Heraklion), contains three URIs, where the first one (i.e., d1:El_Greco) is the subject, the second one (i.e., d1:birthPlace) is the predicate (or property) and the last one (i.e., d1:Heraklion) is the object. Moreover, the set of URIs can be distinguished in three different subsets, (i) entities (e.g., El Greco), (ii) properties (e.g., birthPlace) and (iii) classes (e.g., Human, Painter, Museum). An entity can occur as a subject or object in a triple, a property occurs always as a predicate, while a class can be found in the object of a triple and corresponds to the type/category where an entity belongs to, e.g., the triple (El_Greco, rdf:type, Painter) can be used to denote that the type of El_Greco is Painter. By using Linked Data, the linking of datasets can be achieved by the existence of common URIs or Literals, or by defining equivalence relationships among entities, e.g., by using owl:sameAs relationships, or among schema elements (properties and classes), e.g., by using owl:equivalentProperty and owl:equivalentClass relationship, respectively.

2.2. Related Work

Here, we first introduce Linked Data approaches that focus on cultural domain, but also generic approaches at scale that contain datasets from multiple domains. Europeana [10] is a digital platform for cultural heritage and contains data from 3,000 institutes (e.g., museums, digital libraries). It uses a common model, called Europeana Data Model (EDM) [11], for mapping the different sources (that contain different schemas and/or formats) to a global common standard. LODLaundromat [12] is a set of services, which crawls and indexes over 650,000 documents from several domains (including cultural ones). One can exploit the offered services in order to find all the documents of a given URI or namespace. LOD-a-LOT [13] is a service that collects all the documents of LODLaundromat in a single file, i.e., it enables query answering at large scale. LODStats [2] collects metadata for approximately 10,000 RDF datasets and offers statistics and metadata for them, while LOV (Linked Open Vocabularies) [14] has collected hundreds of ontologies (including CIDOC CRM), and offers a keyword search for finding the most relevant schema elements for a given keyword, e.g., by typing Birth, it can return as relevant class the "E67_Birth" of CIDOC CRM. Moreover, datahub.io (http://datahub.io) is a portal that contains thousands of datasets from several domains. An organization can publish and share their datasets through that portal, for making them findable, accessible and reusable. Comparing to the above services, the presented suite of services, i.e., LODsyndesis, computes the transitive and symmetric closure of equivalence relationships, while it offers connectivity measurements among any possible combination of datasets.

3. The Process for Performing Semantic Indexing and Connectivity Analytics

Here, in Section 3.1, we describe in brief the process of semantic indexing at large scale, while in Section 3.2, we mention briefly the methods that we follow for performing connectivity analytics.

3.1. Semantic Indexing Process

The process of global indexing comprises of five steps, which can be seen in the running example of Figure 2 and are described below.

Step 1. Input. In the first step, we collect several datasets and equivalence catalogs (see the upper left side of Figure 2). Our running example contains four datasets (each one having six triples), which include mainly information for two Painters, and several equivalence relationships. Each dataset contains entities (they are represented in bold), properties (they are underlined), classes (they are in italics) and literals (they are written by using quotes). As we can see, these datasets either use different schemas or different versions of CIDOC CRM schemas (e.g., CIDOC CRM or Erlangen CRM) and different URIs for the same entities (e.g., d1:Theotokopoulos, d2:El_Greco, ex:El_Greco for the painter El Greco). Moreover, we can observe the equivalence relationships for entities (i.e., owl:sameAs relationships) and schema elements (i.e., owl:equivalentProperty and owl:equivalentClass relationships).

Step 2. Creation of Equivalence Catalogs. The second step includes the computation of transitive and symmetric closure of schema and instance equivalence relationships, where catalogs containing for each URI an identifier are produced (see the upper right side of Figure 2). As we can see, all the URIs that refer to the same entity are assigned the same identifier, e.g., in the Entity Equivalence Catalog that all the URIs of El Greco are assigned the identifier E1. The same holds for the properties and the classes, e.g., in the Property Equivalence Catalog, all the properties for crm:P98_brought_into_life are getting the identifier P1, and in the Class Equivalence Catalog the two URIs for class E21_Person are getting the identifier C1.

Step 3. Creation of Real World Triples. In the third step, we use the initial datasets and the produced equivalence catalogs for creating a set of "real world" triples (see the third step in the left side of Figure 2), i.e., we replace each URI (entity, property or class) with its corresponding identifier (i.e., we replace each URI referring El_Greco with E1). Moreover, we perform a simple conversion to Literals (e.g., in our example, we removed the data type "xsd:Year" of "1541"xsd:Year in dataset D_1).

Step 4. Global Indexing including all inferred equivalence relationships. In the fourth step, we use the aforementioned set of real world triples, to create semantically enriched inverted indexes for different sets of elements (e.g., triples, entities), and we store the dataset IDs (i.e., a posting list) where they occur (see the indexes in the middle part of Figure 2). In particular, we create an Entity-Triples Index, where we store together all the facts for a specific entity (e.g., for El Greco), and since some entities occur as an object in a triple, we store such triples twice in that index (they are represented in blue in Figure 2). For instance, we store the triple (El_Greco,livedAt,Spain) twice, one time in the entry of El_Greco and one time in the entry of Spain. Moreover, we also store together all the values of a property for a given entity (e.g., the birth date of El Greco), for enabling the comparison of the values of each property. As an example, in Figure 2, we can see that we can easily compare the values for the birth date of "El Greco", i.e., two datasets state that the birth date of that person was "1541", while one dataset mentions the year "1542" as the birth date of that person. Moreover, we can easily check which information are common in two or more datasets (e.g., two datasets agree about the population of the city of Heraklion) and to find complementary information for the same entity, e.g., in Figure 2, one dataset contains information about the population of Toledo and another one for a painting which was produced in that city. In addition to that index, we create also some smaller indexes (see the indexes in the right side of Figure 2), for storing the datasets where a specific element occurs, i.e., an entity (see Entity Index), a schema element (see Property and Class Index) and a constant (see Literals Index).

Step 5. LODsyndesis Services and Use Cases. After the creation of these indexes, we can exploit them for several purposes, as it can be seen in Figure 1 and in the bottom part of Figure 2.

The methods and the construction algorithms of the above indexes can be found in [5–7], whereas all the offered services, which depend on the aforementioned indexes are described in Section 4.

3.2. Performing Connectivity Analytics

The constructed semantically enriched indexes are given as input for performing connectivity analytics. We have performed measurements about the commonalities (i.e., intersection) among any combination of datasets, for several measurement types, which are the following: number of common entities, common properties, common classes, common literals and common facts (or triples) [5–7]. An example can be seen in the lower right part of Figure 2, where we can observe the number of common entities among any combination of the four datasets of our running example. For instance, in our running example, all the datasets have one entity in common (i.e., El Greco), while the datasets D_3 and D_4 share three entities: El Greco, Picasso and Spain. The process that we follow comprises of two different steps, which are described in brief below.

Specifically, we first scan an index, e.g., Entity Index, and we measure the frequency of a subset of datasets in the posting lists of each index, and we create a table, called directCounts list, for storing that information. For instance, in our running example (in the lower right part of Figure 2), we can see a possible query "Give me the most common entities among all the datasets". For answering such a query, we traverse the Entity Index, and we measure the frequency of each subset in that index, e.g., "3,4" exists two times in the posting lists of that index. Afterwards, we use the produced list as an input to an incremental algorithm, which is based on lattices and set theory properties, for computing the number of common elements among any subset of datasets. In particular, for finding the common elements of a subset of datasets B, we should find which supersets of B occur in directCounts list, and then we just sum their scores (which are stored in directCounts list). As an example, for finding the common entities between datasets D_3 and D_4, the supersets of D_3, D_4 that can be found in that list are the following: ("3,4" ,"1,2,3,4"). If we take the sum of their score in the directCounts list, i.e., the score of "3,4" is 2, and the score of "1,2,3,4" is 1, we will find that these two datasets contain three common entities. For performing the measurements fast, for any combination of datasets, we have proposed two different incremental lattice-based algorithms, which can compute the commonalities between millions of subsets in less than 3 s. All the technical details about the aforementioned algorithms can be found in [5–7].

4. LODsyndesis Services and Use Cases

In this section, we introduce specific use cases and services that are offered by the webpage of LODsyndesis (http://www.ics.forth.gr/isl/LODsyndesis) and from the REST [15] API of LODsyndesis. First, in Section 4.1, we show how to find the URI of one or more keywords (e.g., the URI of Pablo Picasso), while in the remaining sections, we show five different use cases. In particular, in Sections 4.2 and 4.3, we show services and use cases (UC) that can be exploited to find all the available information or to check facts about an entity, while in Section 4.4, we show how the connectivity analytics can be exploited for the creation of advanced dataset discovery and selection services. In Section 4.5, we show how to exploit LODsyndesis, in order to create features for machine learning datasets, while in Section 4.6 we introduce a global namespace service.

4.1. How to Find the URI of an Entity

For most of the services that are described in Sections 4.2–4.6, the input is a URI, e.g., http://dbpedia.org/resource/Pablo_Picasso. For this reason, we offer a keyword to entity service, which can be used in order to find the URI for one or more keywords (e.g., Pablo Picasso).

How to use it: in the services that are offered through an HTML page, users can type one or more keywords and automatically, the webpage shows to the users a list of URIs, containing that keywords. For instance, by typing "Pablo Picasso", it will automatically show to the users the URI http://dbpedia.org/resource/Pablo_Picasso. Moreover, one can use our REST API (see the first service in Table 1) to find the corresponding URIs for a specific keyword, e.g., one can send the following GET request: LODsyndesis/rest-api/factChecking?keywordEntity=Pablo_Picasso, to find the URI of Pablo Picasso. The REST API provides the output in CSV [16], JSON [17] or XML [18] format.

Table 1. LODsyndesis REST API - GET Requests.

ID	Service URL	Description	Parameters	Response Types
1	LODsyndesis/rest-api/ keywordEntity	Finds all the URIs, containing one or more keywords.	**keyword**: Put one or more keywords.	text/csv, application/json, application/xml
2	LODsyndesis/rest-api/ objectCoreference	Finds all the equivalent entities of a given URI or the datasets where it occurs.	**uri**: Put any URI (Entity or Schema Element). **provenance**: It is an optional parameter. Put true for showing the datasets where the selected entity occurs.	application/n-triples, application/json, application/xml
3	LODsyndesis/rest-api/ allFacts	Finds all the facts (and their provenance) for a given URI (or an equivalent one).	**uri**: Put a URI that represents an entity.	application/n-quads, application/json, application/xml
4	LODsyndesis/rest-api/ factChecking	Checks a specific fact for a given entity.	**uri**: Put a URI that represents a single entity. **fact**: Put a fact, separate words by using space. **threshold:** Ratio of how many words of the fact should exist in the triple (optional).	application/n-triples, application/json, application/xml
5	LODsyndesis/rest-api/ datasetDiscovery	Finds the most connected datasets to a given one for several measurement types.	**dataset**: Put a URI of an RDF Dataset. **connections_number**: It is optional. It can be any integer greater than zero, i.e., for showing the top-k connected datasets. **subset_size**: It can be any of the following: [pairs, triads, quads] (e.g., select pairs for finding the most connected pairs of datasets). **measurement_type**: It can be any of the following: [Entities, Literals, Properties, Triples, Classes, SubjectObject].	application/n-triples, application/json, application/xml
6	LODsyndesis/rest-api/ namespaceLookup	Finds all the datasets where a namespace occurs.	**namespace**: Put any namespace.	text/csv, application/json, application/xml

4.2. UC1. Object Coreference and All Facts for an Entity Service

It is important to find all the available URIs and information for a given entity, e.g., suppose a scenario where a user would like to find all the available information about El Greco or which museums contain paintings of El Greco (e.g., see in Figure 2 that there exists three different URIs for El Greco). For this reason, we offer an object co-reference service for finding all the equivalent URIs for a given URI, and all its triples by showing also their provenance. In particular, we offer these services for 412 millions of URIs and over 2 billions of triples from 400 datasets. By using such a service, one can browse or export all the available information for an entity and possibly use these data for various purposes, e.g., creating an application for an entity (or a set of entities), finding complementary information for a set of entities, etc. Except for entities, one can also find all the datasets where specific schema elements occur, e.g., a CIDOC CRM property or class, and all the equivalent URIs for those schema elements.

How to use it: we offer an HTML page where a user can type a URI and select whether they desire to find equivalent URIs, all the triples of that URI or/and datasets where that entity occurs. Moreover, one can use our REST API (see the second and the third service in Table 1), to exploit these services programmatically. For instance, all the equivalent URIs for Pablo Picasso can be found by sending to LODsyndesis the following GET request: LODsyndesis/rest-api/objectCoreference?uri=http://dbpedia.org/resource/Pablo_Picasso, while for finding all the facts (i.e., all the available data) for this entity, one can send the following GET request to LODsyndesis: LODsyndesis/rest-api/allFacts?uri=http://dbpedia.org/resource/Pablo_Picasso. The results of these services can be seen either as an HTML page, or in N-Triples [19] (for object coreference), N-Quads [20] (for all the facts of an entity), JSON or XML format through our REST API. This service needs on average less than one second for deriving the equivalent URIs of an entity (e.g., 0.1 s to retrieve in JSON format the equivalent URIs of Pablo Picasso), and less than 10 seconds to retrieve all the facts for an entity (e.g., 3.5 s for collecting all the triples for Pablo Picasso in JSON format).

4.3. UC2. Fact Checking Service

By collecting all the available information for an entity, one can easily search whether a specific fact is verified from one or more datasets for a given entity (i.e., to verify the correctness and veracity of information). For example,"Had El Greco lived in Venice?", or check all the values for a specific fact, e.g., "I want to find the birth date of El Greco". For the first type of questions, we can see which datasets verify that fact, while for the second type of questions, two or more datasets can provide conflicting answers, therefore, we can compare them for deciding which is the correct one. Moreover, one can submit comparative questions, e.g., "Which is the relationship between El Greco and Jack Levine?" and the possible answer would be that the painter Jack Levine was influenced by El Greco. For answering all these questions, we offer a fact checking service, that contains over 2 billions of facts.

How to use it: we offer an HTML page where a user can type a URI and a set of words representing a fact. Moreover, one can use our REST API (see the fourth service in Table 1) to check for a fact programmatically. For instance, in order to find which is the birth date of Pablo Picasso, one can send the following GET request: LODsyndesis/rest-api/factChecking?uri=http://dbpedia.org/resource/Pablo_Picasso&fact=birth date. The output of this service can be seen as tables in HTML page, while the REST API offers the output in N-Quads, JSON or XML format. This service needs on average less than 10 s to check a fact, e.g., to find the birth date of Pablo Picasso we needed 3.5 s.

4.4. UC3. Dataset Discovery and Selection Services

The proposed connectivity measurements can be directly used for dataset discovery and selection services. In particular, it is crucial to collect information for the same real world entities from many datasets, for enabling the comparison of their values, in order to verify that information and produce a more accurate dataset. Moreover, it is also important to explore even more information for the entities of a given dataset. For instance, queries like "find the K datasets that are most connected to British Museum" can be important, if we want to select the datasets that are worth to be integrated with British Museum (since they contain information for the same entities). As a result, one can produce a dataset with high "pluralism factor", i.e., a dataset where the number of datasets that offer information about each entity is high, e.g., suppose a use case where one wants to "find the K datasets that maximize the pluralism factor of the entities of British Museum". For this reason, we offer a Dataset Discovery and Selection Service, where one can find all the connections for a specific dataset (we provide measurements for 400 real world datasets) for assessing its connectivity for several measurement types (entities, literals, schema elements, etc.).

How to use it: we offer an HTML page, where a user can select a dataset from a list for finding its most connected datasets for different measurement types. Moreover, one can use the offered REST API (see the fifth service in Table 1) for retrieving the most connected datasets to a given one, programmatically. For instance, for discovering the five most connected triads (according to the number of common entities [5–7]) of datasets that contain the dataset of British Museum, one should send the following GET request: LODsyndesis/rest-api/datasetDiscovery?dataset=http://collection.britishmuseum.org/&connections_number=5&subset_size=triads&measurement_type=Entities. One can see the output of this service, either as a table in HTML format, or in N-Triples, CSV, JSON and XML format by using the aforementioned REST API. This service needs on average less than 5 s for returning the most connected datasets for a given one, e.g., for retrieving the most connected triads containing British Museum, we needed 1.5 s. Finally, we have published the results of the connectivity measurements in datahub in CSV and N-Triples format (http://old.datahub.io/dataset/connectivity-of-lod-datasets).

4.5. UC4. Dataset Enrichment for Machine Learning Based Tasks

First, we mention a possible use case for the cultural domain. Suppose that one wants to classify a set of paintings according to their genre [21], e.g., Impressionist, Renaissance and others, by using a machine learning algorithm. However, there are either few or even no available information for these entities, therefore, one should search on the web for those paintings to create more features. Such a process can be time-consuming, while the discovered data often should be transformed before being used in a Machine-Learning task. For this reason, we have created a tool, called LODsyndesisML [22], which can be used for the enrichment of Machine-Learning datasets.

The first step of LODsyndesisML is to discover datasets and URIs containing information for a set of entities (e.g., for paintings) by exploiting LODsyndesis. Afterwards, it sends SPARQL queries to the selected datasets for discovering and showing to the user a large number of possible features (belonging in nine different categories) [22], that can be created for that entities and finally it produces automatically, by sending SPARQL queries, a dataset that contains the features that have been selected by the user. Moreover, it is worth mentioning that LODsyndesisML can create features even for direct and indirect related entities of any path, e.g., to classify the genre of a painting, it could be also important to create features for the painter (of that painting). The produced dataset can be directly used as input for any machine learning algorithm. We have tested the aforementioned tool to classify how popular are a set of books and movies according to the number of their Facebook likes [23] and the accuracy of predictions was improved in both cases [22].

4.6. UC5. Global Namespace Service

In many cases, one would like to find fast all the datasets that contain a specific namespace (or prefix). A namespace is the first part of the URI, and it usually indicates the provenance of a URI, e.g., for the URI http://dbpedia.org/resource/Pablo_Picasso, the namespace is http://dbpedia.org. In such a way, one can find the datasets which contain terms from a specific ontology (e.g., http://www.cidoc-crm.org). For this reason, we offer such a service containing approximately 1 million namespaces. A user can type a namespace, and it returns back all the datasets that contain it, and the number of distinct URIs, where the given prefix occurs.

How to use it: we offer an HTML page, where a user can type a namespace for finding all the datasets where it occurs. Alternatively, through our REST API (see the last service in Table 1) one can send such a GET request to LODsyndesis: LODsyndesis/rest-api/namespaceLookup?namespace=http://www.cidoc-crm.org. The output can be an HTML page, or N-Triples, XML and JSON (through the REST API). It needs less than 1 s for deriving all the datasets for a namespace (for the previous example, 0.1 s was needed).

5. Connectivity Analytics over Hundreds of Linked Datasets (Focus on Datasets of Publications and Cultural Domain)

Here, we report measurements about several linked datasets, mainly from publications domain, since the datasets of cultural domain also belong in that category. In particular, we have indexed 400 real RDF datasets containing 2 billion of triples and we have computed the transitive and symmetric closure of 45 million of equivalence relationships [5–7]. This set of 45 million of equivalence relationships has been collected from those 400 datasets. They have been created either by domain experts or by the owners of datasets, e.g., manually or by using entity matching techniques and tools [24]. Therefore, we do not create equivalence relationships, we just use the predefined ones and we compute their transitive and symmetric closure [5–7]. For performing all these computations and tasks, we needed 81.5 min by using a cluster of machines in okeanos cloud service [25], which consists of 96 virtual machines, each one having 1 GB memory and a single core. From the 400 datasets, 94 of them belong to publications domain (which consists of 667 million of triples). First, in Section 5.1, we show how connected are the datasets from publications domain, which vocabularies (e.g., CIDOC CRM) are used from datasets of that domain, and how connected specific datasets are. Moreover, in Section 5.2, we introduce measurements for British Museum dataset [26] (which uses CIDOC CRM), while in Section 5.3, we introduce conclusions about the connectivity in the LOD Cloud.

5.1. Connectivity Analytics for Publications Domain

From the 94 datasets belonging to publications domain that we have indexed, five of them use CIDOC CRM for describing their entities (however, we have observed that different versions of CRM are used in some cases). In particular, the five datasets are: British Museum [26], ARTIUM [27], Sandrart [28], Szépművészeti Múzeum [29] and Data Archives Hub [30]. Concerning vocabulary usage, we found there are some vocabularies that are used in most datasets, such as RDF, RDFS and FOAF (Friend of a Friend), while most vocabularies (72% of vocabularies) are used only in one dataset. Finally, only 14% of vocabularies (including CIDOC CRM) are used from five or more sources. In Table 2, we can see how connected the publication domain is, while we compare its connectivity with the average connectivity of the LOD Cloud datasets. As we can see, only 14.3% of pairs of datasets belonging in Publications domain share common entities, while 88.7% contain common literals. Comparing to the average connectivity of LOD datasets, we can observe that this domain is more connected, almost for all the difference cases (except for the case of common triples), and especially in the case of schema elements. In particular, 44.8% of pairs of datasets of that domain share properties, and 11.1% of pairs share classes, while the corresponding average percentages for all the pairs of datasets (of any domain) is 24.4% and 5.4%, respectively. Concerning triads of datasets, again the datasets of that domain are more connected, however, the percentage in some cases is very low, e.g., only 2% of triads of datasets share common entities.

Table 2. Connectivity of Datasets belonging in Publications Domain and comparison with the average connectivity in LOD (Linked Open Data Cloud.

Category	% of Connected Pairs in Publications Domain (Average Connectivity of All the Datasets)	% of Connected Triads in Publications Domain (Average Connectivity of All the Datasets)
Entities	14.3% (11%)	2% (1.24%)
Literals	88.7% (78%)	68.2% (46.44%)
Properties	44.8% (24.45%)	16.87% (5.38%)
Classes	11.1% (5.42%)	2.1% (0.5%)
Triples	8.2% (10%)	0.5% (1%)

Concerning the commonalities among different datasets, it is worth noting that Library of Congress [31] and VIAF [32] share more than 9 millions of entities. Moreover, in Table 3, we show which triads of datasets of publications domain contain the most common entities. The most connected triad (with 1.3 millions of common entities) contains DNB [33], Library of Congress and VIAF, while the last two datasets and British Library [34] share a million of entities. Moreover, the above datasets share common entities with BNF [35] dataset (see the positions 3–5 in Table 3).

Table 3. The five most connected triads of Datasets of Publications Domain.

Position	Subset of Datasets	# of Common Entities
1	{DNB, Library of Congress ,VIAF}	1,333,836
2	{British Library, Library of Congress ,VIAF}	1,040,862
3	{BNF, Library of Congress,VIAF}	1,007,312
4	{DNB, BNF, VIAF}	592,367
5	{DNB, BNF, Library of Congress}	516,323

In Figure 3, we can see the 15 most popular datasets from the publications (and cultural) domain, i.e., the datasets that contain common entities with a lot of datasets. In particular, Library of Congress contain common entities with 188 datasets (43 of them belong to publications domain), while DNB, BNF, VIAF and British Library have over 150 connections. Moreover, eight other datasets, i.e., IDREF [36], ZBW [37], EUScreen [38], Worldcat [39], LOD GESIS [40], COLINDA [41], mcu.es [42] and Data Archives Hub are connected with at least 100 datasets. Finally, the last two datasets that are shown in Figure 3, i.e., British Museum and SUDOC [43], have over 80 connections. Regarding datasets that are using CIDOC CRM and are highly connected, e.g., British museum has 95 connections, "Sandrart" is connected with 71 datasets, while ARTIUM has 51 connections. Concerning the popular datasets from any domain, it is worth mentioning that the 15 most connected datasets include seven datasets from cross-domain (especially some very popular ones, such as Freebase [44], DBpedia [4], Wikidata [45] and YAGO [46]), seven datasets from publications domain and one from geographical domain, which means that most publishers tend to connect their data with sources from cross or publications domain.

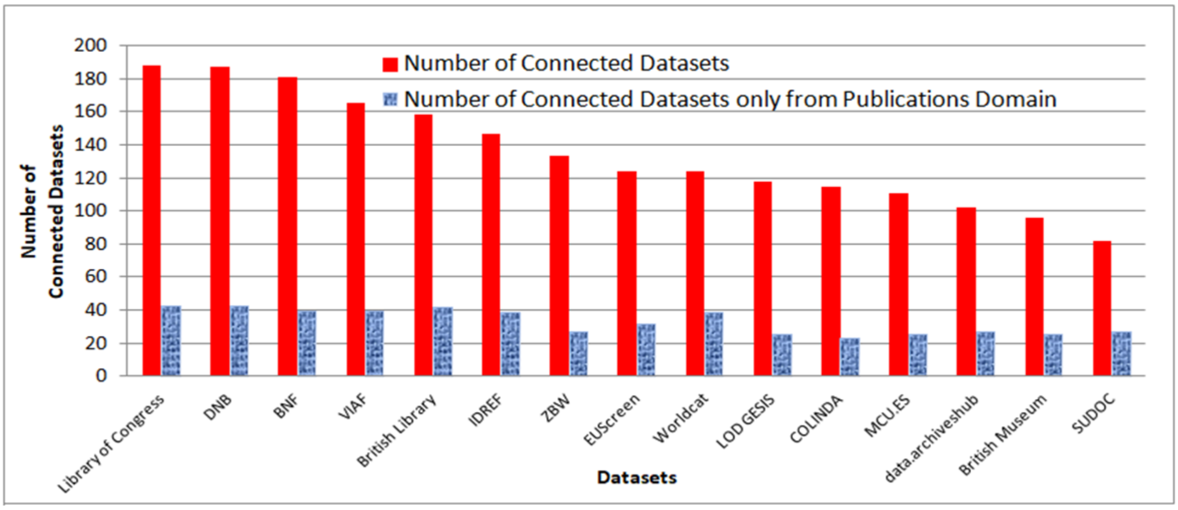


Figure 3. The 15 most popular datasets from publications (and cultural) domain and the number of their connections.

5.2. Connectivity Analytics for British Museum

In Table 4, we can see the number of connected datasets for British Museum concerning entities, literals, classes, properties and triples. This dataset has common entities with 95 datasets (26 of them from the publications domain), however, British museum shares over 1000 entities with only 13 of them, whereas it contains 21,119 common entities with Wikidata (which is a cross-domain dataset). Concerning other measurements, it shares literals with a almost all the datasets, i.e., 393 out of 400, and especially with YAGO. In particular, it shares 2,317,234 literals with that source. Concerning schema elements, it mainly shares properties and classes with datasets containing information about museums, i.e., Szépművészeti Múzeum and datos.artium.org, since they both use CIDOC CRM model for describing their data. In particular, it shares 14 properties with Szépművészeti Múzeum dataset, and 14 classes with ARTIUM dataset. Moreover, it shares thousands of triples with eight datasets, while it contains 15,305 common triples with Library of Congress dataset.

Table 4. Number of Connections of British Museum.

Measurement Type	Number of Connected Datasets	Datasets with at Least 10 Commonalities	Datasets with at Least 100 Commonalities	Datasets with at Least 1000 Commonalities	Most Connected Dataset to British Museum for Each Measurement Type
Entities	95	44	27	13	Wikidata
Literals	393	359	263	132	YAGO
Classes	4	2	0	0	Szepmuveszeti Muzeum
Properties	143	2	0	0	Datos ARTIUM
Triples	48	26	14	8	Library of Congress

In Table 5, we can see the top five triads of datasets that share common entities and contain British Museum dataset. As we can see, the triad of datasets, consisting of British Museum, VIAF and Wikidata, shares 18,865 entities. Moreover, a lot of entities, i.e., 15,317, are common in a triad that contains only datasets from publications domain, i.e., British Museum, Library of Congress and VIAF, while the dataset of British Museum has also a lot of common entities with DBpedia dataset.

Table 5. The five most connected triads of Datasets including British Museum that contain common entities.

Position	Subset of Datasets	# of Common Entities
1	{British Museum, VIAF, Wikidata}	18,865
2	{British Museum, DBpedia, Wikidata}	17,172
3	{British Museum, DBpedia, Yago}	16,039
4	{British Museum, Yago, Wikidata}	16,036
5	{British Museum, Library of Congress, VIAF}	15,317

5.3. Conclusions about Connectivity of the LOD Cloud

We observed that publications domain (which includes datasets from cultural domain) is more connected comparing to the average connectivity in LOD Cloud. In particular, datasets even from other domains contain common entities with these datasets, while 13 datasets from publications domain have common entities with more than 100 datasets. However, only a small percentage of pairs of datasets of that domain (i.e., 14.3% of pairs of datasets) share common entities, which shows the sparsity of LOD cloud. We have also observed that publishers tend to use different vocabularies to describe their data, which makes it difficult to integrate their information. In particular, most ontologies (i.e., 72%) are used only in one dataset, while some popular standard ontologies (such as RDF and FOAF) are used in almost all the datasets. Regarding datasets that use CIDOC CRM (e.g., British museum, Sandrart, etc.), they are highly connected with other datasets. For example, British Museum dataset contains common entities with 95 datasets, while it is highly connected with datasets from cross and publications domain. Moreover, British museum dataset shares schema elements with other datasets that use CIDOC CRM model for describing their data.

Concerning the challenges for publications and cultural heritage domain, there is a need to create more connections among the datasets of that domain, since only a small percentage of datasets share commonalities. Moreover, we observed that most datasets of that domain use different ontologies for representing their data, thereby, it should be good, most of these datasets, to use some standard ontologies (like CIDOC CRM). Generally, two processes, which are required to achieve data integration and data enrichment, are schema matching and entity matching [47], and both of them require a big effort. However, by using the same standard ontology (or ontologies) and by increasing the connections among datasets of that domain, the aforementioned effort of those two processes can be decreased, thereby, the tasks of data integration and data enrichment can be highly benefited.

6. Conclusions

In this paper, we presented LODsyndesis, a suite of services over the datasets of the entire Linked Open Data Cloud. We described in brief how we construct semantics-aware indexes that take into account the cross-dataset reasoning, and how we perform connectivity analytics. Moreover, we introduced several services that are offered through LODsyndesis, such as services for object coreference, dataset discovery and selection, fact checking, dataset enrichment for machine learning-based tasks and others, whereas we described how one can exploit them (through an HTML page or by using the provided REST API). Moreover, we showed measurements from the current LOD Cloud that reveals the sparsity of the current datasets, as regards their connectivity, which in turn justifies the need for advancing the current methods for data integration. We focused on measurements over datasets from the publications and cultural domain, while we showed connectivity analytics about the British Museum dataset. In particular, British Museum has several connections with datasets from cross-domain and publications domain, while the most "popular" dataset (having the most connections) of publications domain is the dataset of Library of Congress. As future work, we plan to improve and extend this suite of services, while we plan to propose more services for improving the veracity of data and for offering more advanced dataset discovery and selection services.

Author Contributions: Conceptualization, M.M. and Y.T.; Data Curation, M.M.; Funding acquisition, Y.T.; Investigation, M.M.; Methodology, M.M. and Y.T.; Project administration, Y.T.; Software, M.M.; Supervision, Y.T.; Writing-original draft, M.M. and Y.T.

Funding: The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (GA. No. 31443).

Conflicts of Interest: The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**. [CrossRef] [PubMed]
2. Ermilov, I.; Lehmann, J.; Martin, M.; Auer, S. LODStats: The data web census dataset. In Proceedings of the International Semantic Web Conference, Kobe, Japan, 17–21 October 2016; Springer: Berlin, Germany, 2016; pp. 38–46.
3. Doerr, M. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag.* **2003**, *24*, 75.
4. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195.
5. Mountantonakis, M.; Tzitzikas, Y. On Measuring the Lattice of Commonalities Among Several Linked Datasets. *Proc. VLDB Endow.* **2016**, *9*, 1101–1112. [CrossRef]
6. Mountantonakis, M.; Tzitzikas, Y. Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets. *J. Data Inf. Q. (JDIQ)* **2018**, *9*. [CrossRef]
7. Mountantonakis, M.; Tzitzikas, Y. High Performance Methods for Linked Open Data Connectivity Analytics. *Information* **2018**, *9*, 134. [CrossRef]
8. Antoniou, G.; Van Harmelen, F. *A Semantic Web Primer*; MIT Press: Cambridge, MA, USA, 2004.
9. W3C RDF Specification. RDF 1.1 Concepts and Abstract Syntax. Available online: http://www.w3.org/TR/rdf11-concepts/ (accessed on 12 November 2018).
10. Antoniou, G.; Van Harmelen, F. Europeana linked open data–data.europeana.eu. *Semant. Web.* **2013**, *4*, 291–297.
11. Doerr, M.; Gradmann, S.; Hennicke, S.; Isaac, A.; Meghini, C.; Van de Sompel, H. The Europeana Data Model (EDM). In Proceedings of the World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg, Sweden, 10–15 August 2010; pp. 10–15.
12. Rietveld, L.; Beek, W.; Schlobach, S. LOD lab: Experiments at LOD scale. In Proceedings of the International Semantic Web Conference, Bethlehem, PA, USA, 11–15 October 2015; Springer: Berlin, Germany, 2015; pp. 339–355.
13. Fernández, J.D.; Beek, W.; Martínez-Prieto, M.A.; Arias, M. LOD-a-lot. In Proceedings of the International Semantic Web Conference, Vienna, Austria, 21–25 October 2017; pp. 75–83.
14. Vandenbussche, P.Y.; Atemezing, G.A.; Poveda-Villalón, M.; Vatant, B. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semant. Web* **2017**, *8*, 437–452. [CrossRef]
15. Richardson, L.; Ruby, S. *RESTful Web Services*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2008.
16. Common Format and MIME Type for Comma-Separated Values (CSV) Files. Available online: http://tools.ietf.org/html/rfc4180 (accessed on 12 November 2018).
17. The JavaScript Object Notation (JSON) Data Interchange Format. Available online: http://buildbot.tools.ietf.org/html/rfc7158 (accessed on 12 November 2018).
18. Extensible Markup Language (XML). Available online: http://www.w3.org/XML/ (accessed on 12 November 2018).
19. RDF 1.1 N-Triples. Available online: http://www.w3.org/TR/n-triples/ (accessed on 12 November 2018).
20. RDF 1.1 N-Quads. Available online: http://www.w3.org/TR/n-quads/ (accessed on 12 November 2018).

21. Siddiquie, B.; Vitaladevuni, S.; Davis, L. Combining multiple kernels for efficient image classification. In Proceedings of the Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA, 7–8 December 2009; pp. 1–8.
22. Mountantonakis, M.; Tzitzikas, Y. How Linked Data can Aid Machine Learning-Based Tasks. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, Thessaloniki, Greece, 18–21 September 2017; Springer: Berlin, Germany, 2017; pp. 155–168.
23. Ristoski, P.; de Vries, G.K.D.; Paulheim, H. A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International Semantic Web Conference*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 186–194.
24. Nentwig, M.; Hartung, M.; Ngonga Ngomo, A.; Rahm, E. A survey of current link discovery *Semant. Web* **2017**, *8*, 419–436.
25. Okeanos Cloud Computing Service. Available online: http://okeanos.grnet.gr (accessed on 12 November 2018).
26. British Museum Collection. Available online: http://collection.britishmuseum.org/ (accessed on 12 November 2018).
27. Datos Artium. Available online: http://biblioteca.artium.org (accessed on 12 November 2018).
28. Sandrart. Available online: http://ta.sandrart.net/en/ (accessed on 12 November 2018).
29. Szépművészeti Múzeum. Available online: http://www.szepmuveszeti.hu/ (accessed on 12 November 2018).
30. Data Archives Hub. Available online: http://data.archiveshub.ac.uk/ (accessed on 12 November 2018).
31. Library of Congress Linked Data Service. Available online: http://id.loc.gov/ (accessed on 12 November 2018).
32. The Virtual International Authority File. Available online: http://viaf.org (accessed on 12 November 2018).
33. Deutschen National Bibliothek. Available online: http://www.dnb.de (accessed on 12 November 2018).
34. The British Library. Available online: http://bl.uk (accessed on 12 November 2018).
35. Bibliothèque Nationale de France. Available online: http://www.bnf.fr (accessed on 12 November 2018).
36. IdRef-Identifiants et référentiels. Available online: http://www.idref.fr (accessed on 12 November 2018).
37. German National Library of Economics. Available online: www.zbw.eu/en/ (accessed on 12 November 2018).
38. EUscreen. Available online: http://www.euscreen.eu/ (accessed on 12 November 2018).
39. WorldCat.org: The World's Largest Library Catalog. Available online: http://www.worldcat.org/ (accessed on 12 November 2018).
40. LOD Gesis. Available online: http://lod.gesis.org (accessed on 12 November 2018).
41. Conference Linked Data. Available online: http://colinda.org (accessed on 12 November 2018).
42. Lista de Encabezamientos de Materia para las Bibliotecas Públicas en SKOS. Available online: http://id.sgcb.mcu.es (accessed on 12 November 2018).
43. SUDOC Catalogue. Available online: http://punktokomo.abes.fr/2011/07/04/le-sudoc-sur-le-web-de-donnees/ (accessed on 12 November 2018).
44. Freebase. Available online: http://developers.google.com/freebase/ (accessed on 12 November 2018).
45. Wikidata. Available online: http://www.wikidata.org (accessed on 12 November 2018).
46. Yago. Available online: http://yago-knowledge.org (accessed on 12 November 2018).
47. Kruse, S.; Papotti, P.; Naumann, F. Estimating Data Integration and Cleaning Effort. In Proceedings of the International Conference on Extending Database Technology, Brussels, Belgium, 23–27 March 2015; pp. 61–72.