IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 January 2023, accepted 21 February 2023, date of publication 27 February 2023, date of current version 6 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250105

RESEARCH ARTICLE

Observing LOD: Its Knowledge Domains and the Varying Behavior of Ontologies Across Them

LUIGI ASPRINO AND VALENTINA PRESUTTI
Department of Modern Languages, Literatures, and Cultures, University of Bologna, 40126 Bologna, Italy

Corresponding authors: Luigi Asprino (luigi.asprino@unibo.it) and Valentina Presutti (valentina.presutti@unibo.it)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Agreement 101004746, and in part by the Gruppo per l'Armonizzazione della Rete della Ricerca (GARR) Consortium (GARR provides high-performance connectivity and computational resources to the Italian research and education community).

ABSTRACT Linked Open Data (LOD) is the largest, collaborative, distributed, and publicly-accessible Knowledge Graph (KG) uniformly encoded in the Resource Description Framework (RDF) and formally represented according to the semantics of the Web Ontology Language (OWL). LOD provides researchers with a unique opportunity to study knowledge engineering as an empirical science: to observe existing modelling practices and possibly understanding how to improve knowledge engineering methodologies and knowledge representation formalisms. Following this perspective, several studies have analysed LOD to identify (mis-)use of OWL constructs or other modelling phenomena e.g. class or property usage, their alignment, the average depth of taxonomies. A question that remains open is whether there is a relation between observed modelling practices and knowledge domains (natural science, linguistics, etc.): do certain practices or phenomena change as the knowledge domain varies? Answering this question requires an assessment of the domains covered by LOD as well as a classification of its datasets. Existing approaches to classify LOD datasets provide partial and unaligned views, posing additional challenges. In this paper, we introduce a classification of knowledge domains, and a method for classifying LOD datasets and ontologies based on it. We classify a large portion of LOD and investigate whether a set of observed phenomena have a domain-specific character.

INDEX TERMS Intelligent web services and semantic web, knowledge engineering methodologies, knowledge and data engineering tools and techniques.

I. INTRODUCTION

Linked Open Data (LOD) is based on the Resource Description Framework (RDF), therefore its formal interpretation is assumed to comply with the semantics of the Web Ontology Language (OWL).[1] Nevertheless, OWL constructs are often used with a different semantics than the prescribed one. For example, `owl:sameAs` is used with several intended meanings (e.g. similar to) as observed in [1], [2], and [3]. From a formal-logic perspective, these are errors. But, as observed by [1], the *misuse* of `owl:sameAs` is so diffuse and systematic that it may suggest requirements from ontology and knowledge graph designers that are overlooked by knowledge representation languages (e.g. additional constructs). Following a similar approach, relevant effort has been spent to observe ontology design practices. For example, in [4], resources such as DBpedia [5] and YAGO (Yet Another Great Ontology) [6] are analysed to identify common modelling patterns that sometimes show anti-patterns or errors. The results of these studies is crucial to inform knowledge engineering methodologies and to improve the quality of common practices and guidelines. This line of research is called *empirical semantics*: the study of meaning and conceptual modelling using observational tools as opposed to only using the prescriptive approach of formal semantics based on logic and theoretical computer science.[2]

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

[1] Or with RDF Schema

[2] Inspired by https://wouterbeek.github.io/presentation/Empirical-Semantics# and https://www.slideshare.net/Frank.van.Harmelen/empirical-semantics

A. EMPIRICAL KNOWLEDGE ENGINEERING

With LOD, the largest *knowledge graph* ever [7] encoded with a logic-based language (OWL), there is an unprecedented opportunity to empirically studying knowledge engineering. However, only few studies analyse LOD at large. One of them is [8] that analyses the emerging, global LOD ontology resulting from the network of interlinked ontologies published in LOD. The focus of [8] is on class and property hierarchies to assess their usage in instance data and their level of reuse (alignments). The results show that the number of explicit alignments between classes/properties is very low and that there is a significant number of classes/properties with empty extension. It is also observed that property hierarchies are mainly flat, while class hierarchies have varying depth degree, although most of them are flat too. This paper is inspired by the open question whether these are generalised phenomena. For example, why are there so many classes and properties with empty extension? We could speculate that ontology designers tend to overgeneralise their ontologies including concepts that are never directly instantiated. Another hypothesis can be that the phenomenon (empty extension) characterises some community of practice representing specific knowledge domains. For example, many ontologies in the biomedical knowledge domain have empty extension and a possible explanation could be that they are - so far - mostly used as terminological resources to support natural language processing tasks or to annotate, index/retrieve, and exchange documents e.g. clinical records of patients [9]. This leads to reformulate the question as: have these observed phenomena a domain-specific character? If this is the case, it might suggest that certain practices have developed and are established in certain communities instead of others, motivating additional research to understand why.

B. KNOWLEDGE DOMAINS COVERED BY LINKED OPEN DATA

Addressing this question requires to cope with challenging preliminary steps: to identify the knowledge domains covered by LOD and to classify LOD datasets[3] according to them. Only by knowing to which knowledge domains a class/property belongs to, will it be possible to associate an observed phenomenon with them. There is valuable research to address LOD dataset classification [10], [11], [12], [13], [14], [15], [16], [17]. Most of these methods are experimented on small scale data and use limited classification systems (e.g. keywords associated with the Linked Open Vocabularies (LOV) [18], LOD cloud labels [19]. They address single-label classification with the exception of [13], which shows low performance (cf. Section II). We aim at performing multi-label classification of datasets at LOD-scale and we adopt a machine learning based approach. Therefore, we need both a classification system and a dataset for training and testing classification algorithms. We build on top of the resources and lessons learned from existing work and develop a classification system of knowledge domains represented by 59 concepts, with a top-level of (additional) 6 concepts. This classification system is available online as a Simple Knowledge Organisation System (SKOS).[4] We reduce the dataset classification task to a text classification task by creating a ''Virtual Text Document'' (VTD) for each LOD dataset. To create a reference training/testing corpus we extend the Topic Profiling Benchmark (TPB) [13] (which includes 198 datasets) to maximise its represented knowledge domains, with datasets retrieved from LOD. The resulting corpus contains 1002 datasets and is manually annotated with three knowledge domains per dataset.[5] We experiment with six multi-label classification algorithms, all of them show very high precision. The best performance (in terms of F1 score[6]) is obtained with eXtreme Gradient Boosting (XGBoost), a tree-boosting approach for learning classification models. If we compare the F1 scores, our method outperforms the existing approaches to multi-label classification of LOD datasets.

C. OBSERVING LOD THROUGH KNOWLEDGE DOMAINS

We use XGBoost trained on our corpus to classify all datasets included in LOD Laundromat [20]. This allows us to observe whether and how the *behaviour* of LOD ontologies vary as the knowledge domain varies. A set of metrics are defined to assess the knowledge domains addressed by LOD and to observe, across domains, how entity reuse is performed, how empty extensions are distributed, and how the depth of taxonomies varies. It is also noticed that there are significant correlations between pairs of metrics: some of them are obvious, others show interesting behaviour that may be worth further investigation.

The results of this research may impact on multiple lines of research. The empirical observations presented in this paper give an overview on how standard languages are used in practice, which may inform standardisation effort as well as the development of tools to facilitate adoption and application of good practices e.g. alignment. The classification method and its associated classification system can be used to enrich metatada collected by central repositories (e.g. LOV [18], LOD Cloud [19]). This, in turn, may improve indexing, summarising and searching ontologies and knowledge graphs (KG) on the web.

Contribution: The main contribution of this paper can be summarised as follows:

- a LOD-scale analysis of knowledge domains coverage;
- a classification system of knowledge domains[4];

[3] In this paper the terms LOD dataset and knowledge graph are used as synonyms. A knowledge graph contains RDF triples expressing instance data, called Assertion Box (ABox), as well as OWL axioms expressing ontology statements, called Terminological Box (TBox). Sometimes we use RDF dataset (ABox) or ontology (TBox).

[4] https://w3id.org/eke/kds

[5] https://dx.doi.org/10.21227/7h2p-7b38

[6] F1 is a measure of the accuracy of the classification. It is defined as the harmonic mean of precision and recall.

- a curated corpus of annotated datasets (with knowledge domains)[5];
- a novel method for multi-label classification of knowledge graphs;
- a set of metrics to perform observations on the changing behaviour of LOD knowledge graphs as the knowledge domain changes.

The reminder of the paper is organised as follows. Section II discusses related work. Section III gives an overview on the conceptual framework introduced in [8] and illustrates its extension, proposed in this paper. The knowledge domain classification system is presented in Section IV. Section V describes a novel method for classifying LOD datasets according to their Knowledge Domains. Section VI presents the manually annotated corpus of LOD datasets. A LOD-scale analysis of LOD datasets based on their knowledge domain is presented in Section VIII. Section IX concludes the paper and discusses future research.

II. RELATED WORK

Relevant work related to our research include the classification of LOD datasets according to their knowledge domain (often referred to as topical classification) and the empirical analysis of LOD.

A. TOPICAL CLASSIFICATION OF LOD DATASETS

Topical classification is the task of associating a LOD dataset (or an ontology) with one or multiple knowledge domains (or topics). Domain annotation has been used for enhancing semantic web archives and search engines [10], [12], for assessing provenance and quality of data [21], [22], and for improving performance in learning tasks [23]. We identify two types of topical classification approaches: machine learning-based and alignment-based.

1) MACHINE LEARNING-BASED APPROACHES

Machine learning-based approaches involve training and testing a classifier by applying a supervised learning algorithm. They differ in the classification system they use - a taxonomy of knowledge domains (i.e. target labels for the classifier) - the feature engineering method, the learning algorithm and the training dataset.

Patel et al. [10] propose to treat ontologies as plain text and to train a single-label classifier (with Naïve Bayes, Probabilistic Indexing or K-Nearest Neighbours) over the DMOZ dataset.[7] They show good results but the experiments use a limited number (five) of knowledge domains, making the classification too general for many applications e.g. indexing/searching. It is unaddressed how the method would perform with a larger classification system. Meusel et al. [11] suggest feeding a single-label classifier (trained using learning algorithms such as K-Nearest Neighbours, J48 or Naïve Bayes) with a set of features extracted from the dataset under study, such as the URI (Universal Resource Identifier) of the vocabularies used by the dataset, the URIs of the classes and properties occurring in it, the labels and the local names of the entities of the dataset, the top-level web domain of the dataset (e.g. com, gov), and its incoming/outgoing link degree. The method is evaluated on a corpus containing 1014 datasets crawled from LOD and annotated with LOD cloud labels [19] (cf. Section IV). The experiments show a maximum accuracy of ~80% in the single-label classification. Our method performs multi-label classification with high performance on a finer-grained classification system of knowledge domains (8 vs. 65 knowledge domains). Pister and Atemezing [12] present a method for classifying LOV ontologies according to the 43 LOV keywords [18]. LOV is a project collecting LOD ontologies along with a set metadata, including a set of keywords, provided by the ontology designers (cf. Section IV). The approach consists of the following steps: *(i)* building a Virtual Text Document (VTD) for each ontology from labels and comments associated with their entities; *(ii)* computing the Term Frequency-Inverse Document Frequency (TF-IDF) of the terms in the VTDs; *(iii)* lowering the dimensionality of the corpus using the Truncated Single Value Decomposition; *(iv)* training (and testing) a multi-label classifier by experimenting with four classification algorithms in a One-vs-rest strategy (Support Vector Machine, Multi-Layer Perceptron, K-Nearest Neighbours and Random Forest). The performance of the method is low with F1=0.36. In our method, we use a similar notion of VTD but create it with a different approach. We also introduce a classification system that encompasses the main existing ones, including LOV, and show high performance on large scale data. Spahiu et al. [13] introduce the Topic Profiling Benchmark (TPB) for the task and experiment with both single- and multi-label classification. The benchmark consists of 198 datasets crawled from LOD and annotated with the LOD cloud classification system [19]. They use the same set of features used in in [11] and show a maximum F1 = 0.43 in the multi-label classification experiments. We reuse TPB (cf. Section VI) to create a larger (1002 datasets) training and testing corpus (Section VII). Nogales et al. [14] propose a deep learning architecture for assigning a (single) label to LOV ontologies. The authors build a corpus of Wikipedia articles annotated with Wikipedia Categories. The corpus (its word embedding representation) is used to train and test a classifier with the Random Multimodel Deep Learning approach. Although the authors demonstrates high accuracy on the Wikipedia corpus (~93%), the accuracy drops significantly in the classification of LOV ontologies (~44%).

2) ALIGNMENT-BASED APPROACHES

Alignment-based methods exploit the linking structure of LOD, in particular the existing alignments between a dataset-to-classify and the knowledge domains in the classification system.

OntClassifire [15] selects a sample of datasets as representative for each domain. The classification is based on a score indicating the similarity between the dataset-to-classify

[7] https://dmoz-odp.org/

and the samples. The similarity score is based on linguistic, structural and axiomatic features extracted from the dataset. The approach was evaluated over a corpus containing 34 ontologies belonging to 6 domains. The approach shows good performance in terms of precision and recall. Although promising, the experiment is conducted on a small scale and the classification system is very limited. It is unclear how the sample datasets representing the domain are selected. Lalithsena et al. [16] propose to classify datasets on the basis of their (existing or discovered) alignments to a background knowledge graph (KG) e.g., Freebase. The assumption is that individuals of the dataset-to-classify are aligned with the individuals of the background KG, and the individuals of the background KG belong to classes associated with a knowledge domain. The authors reported a good classification performance (F1=0.72), but the experiments involved a limited number of datasets (30), thus demanding further investigation to demonstrate the generalisability of the approach. Similarly, the method introduced by [17] assigns a ranked list of topics to a dataset by linking its individuals to DBPedia categories. It consists of three steps: *(i)* Extracting a sample of instances from the dataset; *(ii)* Concatenating literal values associated with each entity (essentially, building a VTD for each entity); *(iii)* Identifying DBpedia entities mentioned in the text using an Entity Linker (i.e. DBpedia Spotlight); *(iv)* Retrieving and ranking the categories associated with the mentioned DBpedia entities. The authors evaluated the method over a collection of 129 datasets and demonstrated a maximum Normalised Discounted Cumulative Gain (NDCG) of 0.4. The NDCG measures the relevance (or gain) of a document topic on its position in the result list. Even if the NDCG is not directly comparable to F1 and accuracy (which are the metrics used in the other studies), it gives us a rough indication of the precision of the method.

The advantage of alignment-based methods is their unsupervised nature. Their main disadvantage is the potential limited coverage and errors that may accompany the alignments, which affects the performance.

B. EMPIRICAL ANALYSES OF LOD

Large-scale analyses of LOD have been performed since the early years of the Semantic Web. We provide an overview of the approaches most relevant to this paper.

1) OBSERVING PATTERNS AND MODELLING STYLE IN LOD

This class of works targets: *(i)* methods and tools for observing modelling practices; *(ii)* insights on common modelling practices in LOD. This paper extends a framework presented in [8] for observing the linking structure of classes and properties in LOD, the depth of their hierarchies, and their usage in instance data.

A technique for extracting common *conceptual components* (CC) from a corpus of ontologies is proposed in [24]. CCs are general concepts such as membership, participation, authorship, etc. that many (if not all) ontologies have in common although they implement them in different ways. The authors show that CCs can be used for indexing ontologies to support knowledge engineering tasks such as ontology understanding, ontology reuse, and ontology alignment. Another notion of pattern is extracted by ABSTAT-HD [25] which summarises large knowledge graphs as sets of Abstract Knowledge Patterns (AKPs) of the form ⟨subjectType, predicate, objectType⟩. In [4], the authors introduce an approach to identify anti-patterns and errors in large KGs such as DBpedia and YAGO. The method developed in this paper may be used in combination with these approaches e.g. to observe domain-specific CCs, AKPs and anti-patterns.

2) ASSESSING (MIS)USE OF STANDARD KNOWLEDGE REPRESENTATION LANGUAGES

A strand of research focuses on assessing the use and misuse (i.e. the improper or incorrect use of classes and properties with respect to their intended semantics) of standard representation languages, such as RDF and OWL. Several studies [1], [2], [3], [26], [27] analyse the use of `owl:sameAs` in practice. Mallea et al. [28] show that blank nodes, although discouraged by guidelines, are prevalent on the Semantic Web. Paulheim and Gangemi [29] analyse the coherence of large LOD datasets, such as DBpedia, by leveraging foundational ontologies. Observations on the presence of foundational distinctions in LOD are reported in [30].

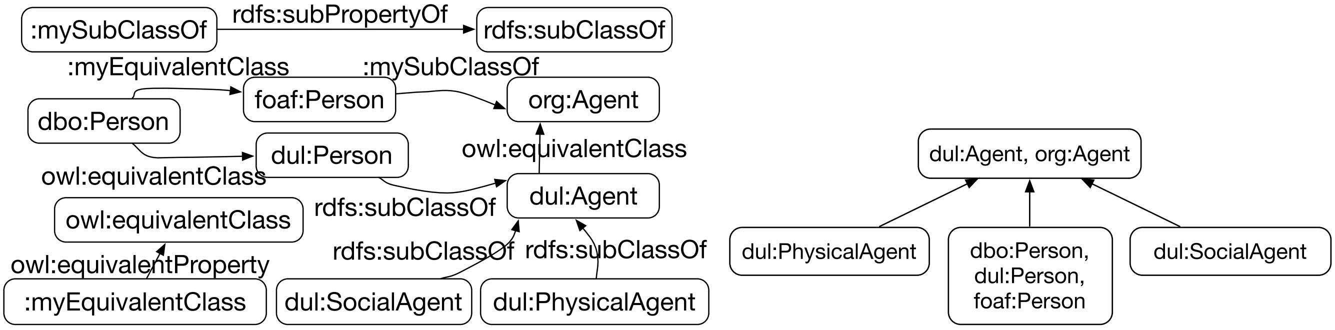
We share a common goal with these studies: to answer how knowledge representation is used in practice, in the Semantic Web. All existing work overlook the potential dependency or association of phenomena with the knowledge domains addressed by the analysed datasets. An exception is the analysis performed by Schmachtenberg et al. [21]. Considering a small number of knowledge domains (LOD cloud labels), the authors contextualise their analysis on how linking, vocabulary usage and metadata provision are performed in LOD. They only use domains provided by the dataset owner and require manual annotation of datasets without explicit domain.

We introduce a general methodology and tool support for performing this type of analyses automatically and by covering a larger spectrum of knowledge domains.

III. EQUIVALENCE SET GRAPHS AND METRICS

To conduct large-scale semantic analyses on LOD, it is necessary to calculate the deductive closure of very large hierarchical structures. To this end, [8] introduces the formal notion of *Equivalence Set Graph* (ESG). Figure 1 shows an example of ESG (Figure 1b) extracted from an RDF Knowledge Graph (Figure 1a).

ESGs enable: *(i)* reducing the dimension of large hierarchical structures while keeping the desired semantic information to be analysed; *(ii)* implementing efficient algorithms to perform large-scale semantic analysis on LOD; *(iii)* defining a set of metrics that synthesise the semantic dimensions that we



(a) RDF Knowledge Graph

(b) Equivalence Set Graph

FIGURE 1. An example of RDF knowledge graph and its corresponding equivalence set graph.

aim to observe; *(iv)* observing the usage of classes, properties and individuals in LOD.

Formally, an Equivalence Set Graph (ESG) is a tuple

$$\langle \mathcal{V}, \mathcal{E}, p_{eq}, p_{sub}, p_e, p_s \rangle$$

For example, the ESG in Figure 1b is computed from the RDF KG in Figure 1a by analysing equivalence and specialisation relations between its classes. The nodes $\mathcal{V}$ of the ESG are equivalence sets of entities e.g., {`dbo:Person`, `dul:Person`, `foaf:Person`}. The directed edges $\mathcal{E}$ of the ESG are specialisation relations (e.g. {`dbo:Person`, `dul:Person`, `foaf:Person`} *specialises* {`dul:Agent`, `foaf:Agent`}). Notice that the edges of an ESG can be designed to also capture other meanings e.g. disjointness.

p_{eq} is an equivalence relation (`owl:equivalentClass`) that must hold between entities that belong to a same ESG node. p_{sub} is a partial order relation that determines the specialization relation between the equivalence sets (e.g. `rdfs:subClassOf`).

It is possible to define equivalences and specialisations of p_{eq} and p_{sub} by using p_e and p_s, respectively. This allows to build ESGs based on more than one relation. All properties that are equivalent to/specialisation of p_{eq} and p_{sub} will be retrieved and used for building the ESG. For example, it is possible to define

$$p_e = \texttt{owl:equivalentProperty}$$
$$p_s = \texttt{rdfs:subPropertyOf}$$

and then assert that

$$\texttt{:myEquivalentClass}\ p_e\ \texttt{owl:equivalentClass}$$
$$\texttt{:mySubClassOf}\ p_s\ \texttt{rdfs:subClassOf}$$

This will cause all triples using `:myEquivalentClass` to contribute to the build the equivalent sets of the ESG (along with `owl:equivalentClass`) and all triples using `:mySubClassOf` to contribute to the specialisation relations between the nodes of the ESG (alogn with `rdfs:subClassOf`) - cf. Figure 1.

In [8], two ESGs are computed from LOD-a-lot [31] (LOD-a-lot integrates the 650K datasets crawled by LOD Laundromat [20] into a single self-indexed RDF-HDT [32] file which can be queried with a limited memory footprint): one with nodes containing classes and one with nodes containing properties. We reuse these ESGs and compute over them a set of measures to observe the semantic structure of LOD hireachies, the level of their alignments, and the level of usage/population of their classes. Observations reported in Section VIII enrich the set of measures computed over these ESGs and explore the relation between equivalence sets and knowledge domains.

In this paper, we extend the ESG framework with the knowledge domains addressed by entities and equivalence sets of an ESG (cf. Section V). We also introduce the following metrics that can be computed by querying the ESG and the knowledge domain annotations:

Percentage of observed entities belonging to a domain (%OE): the number of entities (indicated as OE) of a domain divided by the total number of entities in the ESG (expressed as decimal).

Percentage of equivalence sets belonging to a domain (%ES): the number of equivalence sets (indicated as ES) of a domain divided by the total number of equivalence in the ESG (expressed as decimal).

Ratio (R) between the number of equivalence sets and the number of entities belonging to a certain domain (i.e. $\frac{ES}{OE}$): it indicates to what extent equivalence is used among the observed entities of a certain domain. If equivalence is rarely used, R approaches 1.0.

Percentage of non-instatiated entities of a domain (%Ud): the ratio between non-istantiated entities of a domain d and the number of entities belonging to d (expressed as decimal). This indicates the tendency of a certain domain (as observed in the considerd sample) to have non-instatiated entities.

Percentage of non-instatiated entities (%Ut): the number of non-istantiated entities of a domain divided by the total number of non-instatiatied entities in the ESG (expressed as decimal). This metric measures the impact of the non-istantiated entities of a domain on the whole graph.

Density (D). The graph density is the ratio between the number of edges connecting nodes (in a certain domain) and the total number of edges in the ESG. The density indicates how much the specialisation relation is used in a domain.

Domain Assortativity (A) indicates the likelihood of a node (i.e. an equivalence set) to be connected to others of the same domain. It ranges between 1 (assorative graph, i.e. nodes of the same domain tend to be connected together) and -1 (disassortative graph). In assortative graphs, there is high likelihood that two nodes of the same domain are connected, while in disassortative ones the likelihood is low. When A is close to zero, the graph is neither assortative nor disassortative.

Height of Nodes (H). The height $h(v)$ of a node v is defined as the length of the shortest path from a leaf to v. The maximum height of an ESG is defined as $H = \mathrm{argmax}_{v \in V}\, h(v)$.

IV. A REFERENCE CLASSIFICATION SYSTEMS OF KNOWLEDGE DOMAINS

A Knowledge Domain (KD) is a topic identifying the subject area covered by a knowledge graph. KDs can be hierarchically related by means of specialisation/generalisation relations, i.e. narrower than/broader than. Several hierarchies of KDs have been proposed as classification systems for books, articles, web pages, lexical senses, or datasets. They differ in their expressivity, coverage, structure, design principles and purpose. A classification system should have a large coverage of KDs, however it is a challenge to find a good balance for their dimension: a too large classification system makes the classification task hard for both humans and machines, while a too small one makes the classification less effective in concrete applications such as searching, indexing, etc. Instead of defining yet a new hierarchy from scratch, we align and integrate popular KD classification systems: Wikipedia Categories,[8] Dewey Decimal Classification,[9] Library of Congress Classification,[10] WordNet domains [33], BabelNet domains [34], LOD cloud classification [19], LOV keywords [18].

A. WIKIPEDIA CATEGORIES

Wikipedia is a web-based encyclopedia that anyone can edit. Each Wikipedia entry is associated with at least one category. Wikipedia categories are often interpreted as KDs (e.g. [14]) since they form one of the richest KD hierarchies available today containing over 2M concepts. Wikipedia categories are hierarchically related, meaning that each category is linked to broader and narrower categories, forming a conceptual network often used in knowledge engineering for classifying entities. This network can be formally represented as a directed and non-acyclic graph. However, this graph is noisy: it has many cycles thus hindering the possibility of applying a full-fledged taxonomical reasoning and it contains concepts used only for administrative purposes (e.g. Articles created by bots) or very narrow (e.g. 21st-century Roman Catholics) which are unsuitable for our purpose. Therefore, similarly to previous approaches (e.g. [14], [34]), we consider only a limited set of manually curated categories that classify a subset of Wikipedia "featured articles"[11] (the articles serving as guide for Wikipedia editors). These categories change over time (at the time of [34] they were 34, at the time of writing this article they are 30).

B. DEWEY DECIMAL CLASSIFICATION

The Dewey Decimal Classification (DDC) is one of the most used library classification systems for organising the contents of a library, it divides all KDs into 10 groups. The DDC assigns 100 numbers to each group: 000-099 computer science, information and general works; 100-199 philosophy and psychology; 200-299 religion; 300-399 social sciences; 400-499 language; 500-599 science; 600-699 technology; 700-199 arts and recreation; 800-899 literature; 900-999 history and geography. These groups are further divided into more specific subgroups (e.g. science is divided into: 510-519 Mathematics; 520-529 Astronomy; 530-539 Physics etc.) and subgroups are further divided (e.g. Mathematics is divided into: 511 General principles; 512 Algebra etc.). The DDC is also the basis of the Universal Decimal Classification (UDC) which further specialises DDC numbers.

C. LIBRARY OF CONGRESS

Similarly to DCC, the Library of Congress Classification (LCC) is a system of library classification developed by the Library of Congress in the United States aimed at organising the content of a library. The LCC defines 21 classes knowledge domains: (A) General Works; (B) Philosophy, Psychology, Religion; (C) Auxiliary Sciences of History; (D) World History; (E) History of America; (F) Local History of the Americas; (G) Geography, Anthropology, Recreation; (H) Social Sciences; (J) Political Sciences; (K) Law; (L) Education; (M) Music; (N) Fine Arts; (P) Language and Literature; (Q) Science; (R) Medicine; (S) Agriculture; (T) Technology; (U) Military Science; (V) Naval Science; (Z) Bibliography, Library Science.

D. WordNet DOMAINS AND BabelDomains

The WordNet Domains Hierarchy [33] is a collection of ∼200 hierarchically organised KDs exhibiting specific terminology and lexical coherence. These KDs are used for annotating WordNet synsets (each synset is associated with at least one KD). Similarly, Camacho-Collado and Navigli [34] propose a method for automatically annotating BabelNet's synsets, with the categories of the Wikipedia's featured articles page. Differently from WordNet Domains, in BabelDomain the association synset-domain is weighted with a score ranging from 0 (unrelated) to 1 (highly related).

[8] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories

[9] https://www.gutenberg.org/files/12513/12513-h/12513-h.htm

[10] https://www.loc.gov/catdir/cpso/lcco/

[11] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

E. LOD CLOUD CLASSIFICATION

The LOD Cloud website [19] collects metadata describing LOD datasets published on the web voluntarily submitted from their maintainers, and publishes a graphical representation of them: the LOD cloud. These datasets are categorised by the contributors according to a classification system consisting of nine classes: 1) Cross Domain, 2) Geography, 3) Government, 4) Life Sciences, 5) Linguistics, 6) Media, 7) Publications, 8) Social Networking, 9) User Generated.

F. LOV KEYWORDS

Linked Open Vocabularies (LOV) [18] is a project collecting vocabularies (ontologies) used for publishing Linked Data. It provides an indexing/searching service based on metadata provided by the vocabulary maintainers. Among the metadata fields, the contributors are required to provide a set of keywords describing the submitted vocabulary. Therefore, the set of keywords may expand over time. In this paper, we refer to a LOV snapshot downloaded on 2022-01-09 (available online at[12]) which contains 43 keywords. In most cases the keywords indicate the KDs addressed by the vocabulary (e.g. Biology, Music, Geography). Some keywords may indicate the project in which the ontology has been defined (e.g. SPAR, SSDesk) or the type of the vocabulary (e.g. RDF, W3C's Recommendation).

G. DESIGN PROCESS

Our goal is to define a classification system that encompasses all the KDs defined by the existing, just described, ones. Such a classification system consists in a collection of concepts `D`, and a set of hierarchical relations among them `H`. We approach the design of `D` and `H` as an alignment problem. We perform the following process. *(i)* We add to `D` all the KDs belonging to the top level of each classification system. As for the WordNet classification we also include the second level of KDs since the first level is very general compared to the other classification systems. *(ii)* We discard catch-all concepts (e.g. Cross-domain from LOD Cloud, General&Upper from LOV, Factotum from WordNet, General Works from LCC). This strategy is also adopted for building Topic Profiling Benchmark (TPB) [13]. *(iii)* We manually align the concepts in `D`. Similar concepts are merged (e.g. Geographic from LOD Cloud expresses the same concept denoted by Geography from LOV). Hierarchical relations are specified in `H` to link broader/narrower concepts. *(iv)* We homogenise the labels of the concepts and define a description for all of them. *(v)* We remove from `D` narrow concepts (e.g. History of America and History of Americas from Library of Congress). The resulting classification system is summarised in Table 3 and is available online as a SKOS vocabulary[4]. It consists of 59 concepts generalised by 6 (additional) concepts. It is used, in the remainder of the paper, as the reference classification system to automatically classify LOD datasets.

V. AUTOMATIC CLASSIFICATION OF LOD DATASETS BASED ON KNOWLEDGE DOMAINS

We approach the problem of LOD dataset classification as a multi-label classification task using a machine learning method. Multi-label classification is the problem of predicting a set of categories to which an entity belongs, given a set of examples for each category, called a training set. The training set is processed by an algorithm to learn a predictive model based on the observation of a number of features (features can be categorical, ordinal, integer-valued or real-valued).

Most of existing approaches require the extraction of several features such as the URIs of the classes, properties and vocabularies occurring in the dataset, the name of the host publishing the dataset etc. - cf. Section II). Our method uses a different approach and relies on the extraction of a textual representation for each dataset, called *Virtual Text Document* (VTD). A VTD is used to classify its corresponding dataset, thus reducing the dataset classification task to a text classification problem. We experiment with six multi-label classification algorithms: Random Forest, K-nearest neighbours, Extra Trees, XGBoost, Ada Boost and Multi-layer Perceptron. Details and results of the experiments are provided in Section VII. This approach (VTD) improves the generality of the predictive model, since tokens are more likely to be found across datasets than URIs. It also reduces the dependency of the features from the modelling style of the creator of the dataset. For example, in the Linnean Taxonomy pattern,[13] the ranks of the taxon are classes (e.g. `lt:Species`, `lt:Genus`, `lt:Family` etc.), while the uniprot ontology[14] models the rank as a property (i.e. `up:rank`) connecting a `up:Taxon` (e.g. Engraulis encrasicolus[15]) to a `up:Rank` (e.g. `up:Species`). If we would train a classifiers only with URIs of classes and properties, we would not consider `up:Species` as a feature of the uniprot ontology since in uniprot it is an individual of the class `up:Rank`. This would reduce the similarity with the Linnean Taxonomy pattern.

EXTRACTING VTDs FROM DATASETS

To extract the VTD from a dataset we parse the dataset twice. The first parsing collects all the entities (i.e. URIs and Blank Nodes) occurring in the dataset with their labels:

(i) the process iterates over the triples to find a matching with the following pattern: `?entity ?labelPredicate ?label` where `?labelPredicate` is `rdfs:label` or one of its equivalents or sub-properties (a query for retrieving equivalent or sub-properties is needed in case those properties are not provided as input of the process - for example by retrieving them from a pre-computed Equivalence Set Graph [8]); *(ii)* if multiple labels are associated with the same entity these are concatenated; *(iii)* if a dataset does not provide a label for an entity, the ID of the entity (the rightmost part

[12] https://w3id.org/eke/KDA/InputRDF/lov.nq

[13] http://ontologydesignpatterns.org/wiki/Submissions:LinnaeanTaxonomy

[14] https://www.uniprot.org/

[15] http://purl.uniprot.org/taxonomy/184585

of its URI or the id of a Blank Node) is indexed instead. The second parsing builds, for each triple, a sentence of the form "`subject predicate object`" where: *(i)* `subject` and `predicate` are the labels of the subject and predicate of the triple, respectively; *(ii)* `object` is either the label of the object of the triple (if the object is a URI or a blank node) or a string (in case the object is a literal). The concatenation of these sentences forms the VTD for a dataset. For example, consider the sample dataset below:

```
@prefix ex:  <http://example.org/>.
@prefix rdfs: <http://www.w3.org/2000/
               01/rdf-schema#>.
@prefix rdf: <http://www.w3.org/1999/02/
              22-rdf-syntax-ns#>.
@prefix foaf: <http://xmlns.com/foaf/
               0.1/>.

ex:monitor rdfs:label ``monitor''.
ex:monitor rdfs:comment ``A Computer
     monitor is an output device that
     displays information in pictorial
     or text form''.
ex:label  rdfs:subPropertyOf rdfs:label.
ex:monitor ex:label ``screen''.
ex:monitor rdf:type ex:Device
```

The corresponding VTD is the following:

```
monitor screen label monitor
monitor comment A Computer monitor is an
   output device that displays
   information in pictorial or text form
label subPropertyOf label
monitor screen label screen
monitor screen type Device
```

EXTRACTING FEATURE VECTORS FROM VTDs

The process for extracting feature vectors from the VTDs consists of three activities: (i) Pre-processing; (ii) Vectorisation; (iii) TF-IDF computation.

The *pre-processing* phase reduces the randomness and inflectional forms of the text and it involves the following activities: *(i) Camel case removal* (e.g. "subPropertyOf" → "sub Property Of"); *(ii) Tokenisation* (by evaluating the regular expression `[a-zA-Z][a-zA-Z]+`) and *conversion to lowercase* (e.g. "label sub Property Of label" → "label", "sub", "property", "of", "label"); *(iii) Lemmatisation* with simplelemma[16] (e.g. "displays" → "display" or "is" → "be"); *(iv)Stopword removal* with python stop-words package[17] and by also stripping a list of tokens very common in RDF datsets and OWL ontologies (such as: label, comment, ontology, class, property etc.) which contribution is irrelevant to the KD of the dataset.

The *vectorisation* phase transforms a VTD into a integer-valued vector whose values are associated with tokens of the document. A dictionary containing all the tokens mentioned in the VTDs is built, and a vector of the size of the dictionary is created for each document (i.e. each position in the vector corresponds to a token of the dictionary, uniquely). To compute the values of the vectors, we experiment with two kinds of vectorisation: *(i)* Binary in which the value 1 is assigned to the positions associated with tokens contained in the VTD, 0 otherwise. *(ii)* Count in which the value represents the number of times the token is mentioned in the VTD.

Finally, for each token of the VTD we compute its TF-IDF score. The TF-IDF is a real number indicating how relevant a token is to a document with respect to a collection of documents. The TF-IDF score is the product of the token frequency within a document (TF) with the frequency inverse of the token within the whole corpus (IDF).

Binary and TF-IDF vectors are used for training and testing the predictive model, while the Count vectors serve for computing TF-IDF scores only.

A simplified vectorisation of a VTD is the following.

$$\begin{array}{c}\vdots\\ \text{device}\\ \text{display}\\ \text{precipitation}\\ \vdots\end{array}\begin{bmatrix}\vdots\\1\\1\\0\\\vdots\end{bmatrix}\quad\begin{array}{c}\vdots\\ \text{device}\\ \text{display}\\ \text{precipitation}\\ \vdots\end{array}\begin{bmatrix}\vdots\\2\\1\\0\\\vdots\end{bmatrix}\quad\begin{array}{c}\vdots\\ \text{device}\\ \text{display}\\ \text{precipitation}\\ \vdots\end{array}\begin{bmatrix}\vdots\\0.3\\0.1\\0\\\vdots\end{bmatrix}$$

BINARY COUNT TF-IDF

A. PROPAGATION OF KNOWLEDGE DOMAINS

The classification of the datasets according to their KDs allows us to observe the varying behaviour of LOD, at dataset level. To observe the behaviour at entity level (classes and properties), LOD entities need to be classified according to the KDs. To this end, we propagate KDs assigned to a dataset to its entities (classes and properties occurring in the dataset). It is worth noticing that an entity, formally identified by a URI, may occur multiple times in multiple datasets. Therefore, an entity is classified by a set of all KDs assigned to all the datasets in which the entity occurs in.

To observe the varying behaviour of knowledge domains through equivalence and specialisation relations, we propagate domain classification to equivalence sets. Specifically, KDs assigned to an entity are propagated to the equivalence set which includes the entity.

VI. A REFERENCE CORPUS FOR LOD CLASSIFICATION

To experiment with our classification approach (cf. Section V), we create a reference corpus of LOD datasets manually classified according to the classification system presented in Section IV. The resulting corpus includes datasets from three sources: *(i)* ontologies indexed by LOV [18], *(ii)* datasets crawled by LOD Laundromat [20],

[16] https://pypi.org/project/simplemma/
[17] https://pypi.org/project/stop-words/

(iii) the Topic Profiling Benchmark (TPB) [13]. The corpus is available online [5].

A. SELECTING A SAMPLE OF LOD DATASETS

A representative sample of LOD datasets satisfies the following requirements: *(i)* it must contain a similar amount of RDF graphs (ABox) and ontologies (TBox); *(ii)* it must provide examples for (ideally all) the KDs of the reference classification. To the best of our knowledge, the Topic Profiling Benchmark (TPB) [13] is the only collection that *partially* meets these requirements. TPB consists of 198 datasets crawled from the LOD Cloud and only covers the KDs from the LOD cloud classification. Therefore, to cover as many KDs as possible, we add datasets and ontologies crawled from the Linked Open Data. We also consider the possibility of filling this gap with text articles, e.g. from Wikipedia. Nevertheless, as reported by Nogales et al. [14], a predictive model learned from a corpus of text articles is likely to show low performance when applied on VTDs (from ∼93% of accuracy on text articles to ∼44% on VTDs).

Therefore, we collect datasets from two additonal sources: *(i)* LOV [18], which contains 773 vocabularies used for publishing Linked Data; *(ii)* LOD Laundromat [20], which contains ∼650K RDF datasets crawled from LOD.

We add all ontologies from LOV having a textual description and associated with at least one keyword. Keywords and descriptions are the minimal information needed for understanding the KD of an ontology. As a result, 751 ontologies are added to the reference corpus.

To have an equal number of RDF graphs and ontologies, we select 751 RDF graphs from LOD Laundromat. Differently from LOV which provides descriptive metadata in natural language (e.g. keywords and descriptions), LOD Laundromat stores only numerical indicators of the quality of the collected datasets. Although these indicators are valuable for assessing the syntactic quality of a dataset, it is useless in identifying its KD. As a result, we face two challenges: *(i)* A natural language description is to be automatically generated from datasets in order to avoid the annotators to go through the triples of the dataset during their job; *(ii)* A representative sample of datasets is to be selected from the 650K datasets crawled by LOD Laundromat.

To generate a dataset's description, we resort to the notion of VTD: we extract VTDs from the LOD Laundromat's datasets (as described in Section V), we compute the TF-IDF of the terms within the VTDs and we associate with each dataset a description consisting of the list of the 50 most significative terms (those with the highest TF-IDF score).

While defining a strategy to select a representative sample of LOD datasets, we observe that many datasets crawled by LOD Laundromat come from few sources (e.g. ∼139K from the Open Data of the World Bank,[18] ∼10K from DBpedia, ∼18K from Eurostat[19] etc.). A random selection would cause a strong unbalance towards most represented sources. Therefore, we cluster the datasets according to their source and we go through all cluster randomly selecting one dataset at each iteration until we reach the desired amount: 751. For clustering the datasets we use the KMeans algorithm. The optimal number of clusters, i.e. 128, was computed with the Elbow method.

B. CLASSIFYING DATASETS WITH KNOWLEDGE DOMAINS

The datasets included in TPB are classified according to the LOD cloud labels [13]. We exploit the links between LOD cloud labels and the KDs in our classification system.

The datasets collected from LOV and LOD Laundromat are manually annotated. The annotation involves 9 junior researchers in knowledge engineering who were trained on the reference KD classification system. Each dataset is classified by three researchers, independently. Each annotator is asked to indicate up to three KDs. No classification was indicated in case of general-purpose, cross-domain datasets (e.g. the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)[20]).

We compute the inter-rater agreement using the Krippendorff's alpha coefficient (α_k). Values of α_k range from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement and negative values indicate inverse agreement. Krippendorff suggests that it is customary to require α_k greater than 0.8 and consider 0.667 as the lowest conceivable limit. We measure both the α_k considering all the annotations of all datasets together, which indicates the overall agreement of the annotators, and the α_k for each dataset, which expresses the inter-rater agreement on each dataset, individually. As for the LOV datasets, the measured α_k was 0.64 (slightly under the threshold of 0.667) thus indicating that the agreement is nearly sufficient. To select a reliable collection of classified datasets for experimenting with classification methods, we exclude the datasets classified with low inter-rater agreement (i.e. $\alpha_k < 0.667$) and those (54) identified by all annotators as cross-domain. The overall inter-rater agreement associated with the remaining 315 classified datasets increased to 0.90. The classification sets provided by different contributors on a dataset may differ (e.g. the Video Game Ontology, abbreviated VGO,[21] was classified as `Recreation` by two researchers and `Recreation` and `Sport` by the third). In such cases the union of all the annotation sets (e.g. `Recreation` and `Sport` for the VGO) is kept for the training data. From this sample, we observe that the most common domains in LOV are `Computer Science` (60/82 - 60 are the datasets classified as `Computer Science` only, 82 in combination with other domains), `Geography` (16/44), `Linguistics` (16/19), `Bibliography` (12/17), `Law` (10/27).

As for LOD Laundromat, the value of α_k is 0.85. We exclude 67 datasets annotated as cross-domain by all

[18] https://data.worldbank.org/
[19] https://ec.europa.eu/eurostat
[20] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl
[21] http://purl.org/net/VideoGameOntology

the contributors and those (199) having α_k lower than 0.667. The α_k increased to 0.98, showing a nearly perfect agreement between the contributors on classifying the selected datasets (489). Most common domains in the observed sample from LOD Laundromat are `Meteorology` (388/401), `Geography` (14/19), `Sociology` (18/22), `Economy` (8/17).

OVERVIEW OF THE REFERENCE CORPUS OF DATASETS

The reference corpus consists of 1002 datasets: 198 from TPC, 315 from LOV, and 489 from LOD Laundromat. We add general KDs to the manual classifications by using the hierarchical relations from the classification system. For example, if a dataset is classified by `Meteorology` we also include its super-domains `Earth` and `Pure Science`. We compute the number of datasets belonging to each KD. Such distribution is summarised in Table 3 (column T). We observe that:

- 54 out of 65 KDs have at least one representative dataset. 11 KDs are not represented: `Dance`, `Plastic Arts`, `Astrology`, `Astronomy`, `Paleontology`, `Physics`, `Religion`, `Fashion`, `Naval Science`, `Royalty and nobility`, `Physics`, `Astronomy`;
- The most represented KD is `Pure Science` which comprises 523 datasets, mainly belonging to `Earth` (494), `Meteorology` (405) and `Geography` (89). `Social Science` (315), `Applied Science` (128), and `Bibliography` (104) are also quite frequent in the collection;
- 8 domains (`Theatre`, `Archaeology`, `Oceanography`, `Philosophy`, `Psychology`, `Sexuality`, `Literature`, `Drawing`) are only addressed by one dataset.

Although there are techniques for learning predictive models from unbalanced training sets with classes having few examples (e.g. ML-SMOTE [35]), further research is needed to handle unrepresented domains. As for this study, we experiment with KDs that have at least one representative in the reference corpus and leave to future work the investigation of strategies to cope with this issue. The classified datasets are available online[5].

VII. EVALUATING THE AUTOMATIC CLASSIFICATION OF LOD

We evaluate our classification method (cf. Section V) against the reference corpus presented in Section VI. The datasets from the corpus are vectorised according to the procedure described in Section V. We test both the Binary and the TF-IDF vectorisations. In both cases the corpus was pre-processed as follows.

A. PREPROCESSING

The vectors have 1,492,822 components; therefore, the reference corpus is represented by a matrix of 1,002 rows and 1,492,822 columns. This matrix is very sparse (the sparsity scored 0.99[22]) and unsuitable for classification algorithms. A countermeasure to this issue is to perform a linear dimensionality reduction by means of truncated singular value decomposition (SVD). After this process, the number of columns reduces to 100 (which is the recommended dimension for a corpus of 1K documents [36]) while preserving the latent semantic of the matrix.

To validate the stability of the machine learning model, we perform a 10-fold cross-validation. To this end, we extract 10 folds from the corpus using a stratified sampling technique tailored for the multi-label classification problem [37] (the implementation of the method is provided by the iterative stratification library[23]). Each fold consists of a set of training examples (90% of the original collection) and a set of test examples (10%).

As illustrated in Table 3, the number of examples varies a lot across domains. This would make the classifier tend towards the most represented classes. Therefore, we re-sample the training datasets using ML-SMOTE [35], a technique for producing synthetic examples for the less represented classes.

We standardise the training and testing datasets using: *(i)* the sklearn's Standard Scaler, which removes the mean value of each feature and scale it by dividing by their standard deviation (as a result each feature has 0 mean and unit variance); *(ii)* the sklearn's Normalizer which scales each sample to have unit norm.

Finally, we train and test, on each of the 10 fold (i.e. 90% of examples for training, 10% for testing), six multi-label classification algorithms: Random Forest, K-nearest neighbours, Extra Trees, XGBoost, Ada Boost and Multi-layer Perceptron. We refer to the scikit-learn[24] and XGBoost[25] libraries for the implementation of the algorithms. All the algorithms are initialised with default hyperparameters. The code and the instructions for reproducing the classification experiments are available online.[26]

B. RESULTS

Table 1 reports the results of the classification experiments. The results are expressed in terms of the mean and the standard deviation (indicated as σ) over the 10 folds of the micro-averaged precision, recall, and F1. We remark that the Multi-Layer Perceptron model failed to converge, hence, the reported precision, recall and F1 are unreliable. If we look at the precision only, all the classification algorithms perform well. In fact, precision ranges from 0.79 and 0.90. The recall falls significantly in most cases. This suggests that the trained models return a limited number of labels but most of them are correct. Considering the purpose of our study, high precision is desirable.

[22] The sparsity is calculated as $1 - \frac{n}{c}$ where n is the number of non-zero elements in the matrix and c is the total number of elements in the matrix.

[23] https://github.com/trent-b/iterative-stratification

[24] https://scikit-learn.org/

[25] https://xgboost.readthedocs.io/en/stable/

[26] https://github.com/empirical-knowledge-engineering/kda

TABLE 1. Results of the classification.

		Precision		Recall		F1	
	Classifier	Mean	σ	Mean	σ	Mean	σ
BINARY	Random Forest	0.896	0.027	0.582	0.021	0.705	0.021
	K-nearest neighbours	0.819	0.031	0.637	0.018	0.716	0.020
	Extra Trees	0.898	0.037	0.586	0.022	0.709	0.022
	XGBoost	0.871	0.028	0.631	0.022	**0.732**	0.019
	Ada Boost	0.787	0.029	**0.643**	0.025	0.707	0.024
	Multi-layer Perceptron[27]	0.680	0.022	**0.677**	0.030	0.678	0.018
TF-IDF	Random Forest	**0.903**	0.035	0.550	0.016	0.683	0.016
	K-nearest neighbours	0.833	0.039	0.637	0.033	0.721	0.031
	Extra Trees	0.878	0.032	0.554	0.016	0.679	0.016
	XGBoost	0.845	0.038	0.618	0.031	0.714	0.029
	Ada Boost	0.793	0.029	0.628	0.017	0.700	0.017
	Multi-layer Perceptron[27]	0.683	0.030	**0.673**	0.026	0.678	0.025

We also observe that both vectorisation strategies (Binary and TF-IDF) have comparable performance, but the Binary performs slightly better (a similar behaviour has been observed by [13]). Since Binary is also simpler than TF-IDF, it can be considered as a default vectorisation strategy for this classification task.

XGBoost shows the highest F1 measure, thus indicating that it performs well in terms of both precision and recall. Moreover, the metrics calculated across the different folds are also the most stable (since the standard deviation is one of the lowest for all the metrics). This designates XGBoost as the preferable classification algorithm for annotating LOD datasets.

VIII. OBSERVATIONS AND DISCUSSION

This section reports the observations that we perform on Linked Open Data relying on the metrics summarised in Table 3, presented in Section III. Note that, the table reports twice the domains (i.e. `Medicine`, `Biology`, `Engineering`, `Architecture`) that inherits from multiple domains and percentage values are expressed as decimal. Coherently with the general research question introduced in Section I, whether the observed phenomena reported in [8] have a domain-specific character, we want to address the following: (i) what KDs are covered by LOD; (ii) if and how ontology reuse varies as the KDs vary; (iii) if and how the empty-extension phenomenon is KD-dependent; (iv) if the depth degree of class/property hierarchies varies as the KDs vary.

A. KNOWLEDGE DOMAINS ADDRESSED BY LOD

The predictive model trained with XGBoost is used for classifying all datasets included in LOD Laundromat [20], according to the reference classification system of KDs defined in Section IV. The results are summarised in Table 3. L is the number of datasets per domain and $\%L$ indicates, the percentage of LOD-Laundromat's datasets classified by a given KD (for completeness the table reports the value T: the number of datasets classified by that domain that are present in the training set). Moreover, Figure 2 depicts the distribution of KDs in LOD Laundromat as a treemap. Each rectangle in the Figure is proportional to the (logarithm of the) number of datasets in LOD Laundromat classified by that domain. Note that the logarithmic scale improves the readability of the low represented domains, but flattens the differences between domains (e.g. `Pure Science` and `Social Science` rectangles have nearly the same area, but they count $\sim$ 604K and $\sim$ 35K datasets respectively).

1) KNOWLEDGE DOMAINS OF DATASETS

We observe that LOD Laundromat covers 35 KDs. Considering an individual dataset as our observation unit, 91% of them (604K out of ~658k) are classified by the `Pure Science` KD. These include 594K `Earth` datasets, which, in turn, are mostly from `Meteorology` (582K). Other sub-domains of `Pure Science` that show good coverage are `Geography` (10K), `Chemistry` (8.9K) and `Geology` (7.8K). `Social Science` also shows a wide coverage with 35K datasets mainly from `Sociology` and `Economy`. `Applied Science` and `Life Science` are covered by ~10K datasets each.

There are 8.3K unclassified datasets. A manual inspection (performed by one of the authors) reveals that ~5.6k of them are *Cross-Domain* - mainly coming from DBPedia and the Billion Triple Challenge (BTC) [38]). The remaining ones pertain `Social Science` (1,925), e.g. datasets from Eurostats - and `Scientific Research & Academy` (455), e.g. Semantic Web Dog food[28]). The presence of ~1% of cross-domain datasets in LOD confirms previous observations in [11]. A possible explanation of a large number of meteorological datasets is that it is a common publication strategy of this KD to release datasets daily (e.g. [29] and[30] report on the air temperature observed on 2008-9-2 and 2009-9-17 by the KNOESIS Laboratory[31]). This raises a question on how this aspect can be taken into account to compute KD coverage of a distributed knowledge graph such as LOD in a more precise way.

Considering that `Chemistry`, `Geography` and `Social Science`, `Medicine`, and `Social Networking` KDs have a long tradition in LOD, it is reasonable to see them among the most covered in LOD Laundromat. There are nineteen KDs that are under-represented in LOD Laundromat (cf. Section VI): eight that are only addressed by one dataset (each) and eleven that are not (yet) addressed.

2) KNOWLEDGE DOMAINS OF CLASSES AND PROPERTIES

Although LOD datasets are *de facto* coherent thematic units constituting LOD, to assess KD coverage of LOD it is important to analyse KDs at the level of knowledge graph entities (classes and properties), as observational units. As explained in Section V, an entity inherits (is classified by) the KDs

[27] Multi-Layer Perceptron failed to converge, hence, the reported precision, recall and F1 are not reliable.

[28] https://old.datahub.io/dataset/semantic-web-dog-food

[29] https://w3id.org/eke/KDA/InputRDF/Laundromat/85/852b4b986c5f50306c8a52f28377f68e/data.nq.gz

[30] https://w3id.org/eke/KDA/InputRDF/Laundromat/79/793bbbfdedd8aa69457beedbef98242f/data.nq.gz

[31] https://engineering-computer-science.wright.edu/lab/knoesis

TABLE 2. Pearson's correlation coefficient domain-by-domain between pairs of metrics.

Metric 1	Metric 2	Correlation	Interpretation
%OE	%OE_p	0.931	Very High Positive Correlation
%OE	%ES_p	0.930	Very High Positive Correlation
%OE	%OE_c	0.932	Very High Positive Correlation
%OE	%ES_c	0.883	High Positive Correlation
%OE	D_p	0.742	High Positive Correlation
%OE	D_c	0.864	High Positive Correlation
%OE_p	%ES_p	1	Very High Positive Correlation
%OE_p	%OE_c	0.767	High Positive Correlation
%ES_p	%OE_c	0.765	High Positive Correlation
%OE_c	%ES_c	0.952	Very High Positive Correlation
%OE_c	Ut_c	0.772	High Positive Correlation
%OE_c	D_c	0.952	Very High Positive Correlation
%ES_c	Ut_p	0.728	High Positive Correlation
%ES_c	Ut_c	0.912	Very High Positive Correlation
%ES_c	D_p	0.709	High Positive Correlation
%ES_c	A_p	0.705	High Positive Correlation
%ES_c	D_c	0.958	Very High Positive Correlation
Ut_p	Ut_c	0.701	High Positive Correlation
Ut_p	D_p	0.957	Very High Positive Correlation
Ut_p	A_p	0.885	High Positive Correlation
Ut_p	H_c	0.794	High Positive Correlation
Ut_c	D_c	0.807	High Positive Correlation
D_p	A_p	0.788	High Positive Correlation
D_p	H_c	0.914	Very High Positive Correlation

from the datasets that contain it, independently of whether it is defined or reused in them. Based on this information, the metrics presented in Section III are computed. They are summarised here for the sake of readability: *%OE* (% of entities in a KD), *%ES* (% of equivalent sets per KD), *R* (equivalences among entities in a same KD), *%Ud* (% non-intantiated entities in a KD), *%Ut* (% of non-instantiated entities in a KD on the total of non-instantiated entities), *D* (usage of specialisation relation in a KD), *A* (connections between nodes of the same KD) and *H* (height of a node in the ESG). The results are reported in Table 3.

Using an entity-driven perspective, we observe that LOD Laundromat covers `Pure Science` and `Social Science` in a similar way. If compared with the dataset perspective, it can be noticed that `Pure Science` datasets are 0.919 of the total number of datasets in LOD Laundromat, while the number of `Pure Science` entities is 0.214 of the total number of LOD-Laundromat entities. The `Social Science` domain is instead the largest one, with 0.262 of the total number of entities. This indicates that the datasets classified by `Pure Science` are generally smaller but numerous. The dimension and numerousness of datasets may be KD-dependent aspects. An interesting question is whether they depend on dataset design strategies or other practice followed by that domain's community, e.g. the frequency of update publications.

B. ONTOLOGY REUSE IN LOD

Following the definition by [39], we distinguish between direct and indirect reuse of ontology entities. Direct reuse is performed when an existing ontology is used to encode a RDF graph. For example, instead of defining a class `my:Organisation` to type individuals in a RDF graph, the class `dbo:Organisation` from the DBpedia ontology[32] is reused. An ontology entity is indirectly reused when an ontology links (aligns) to it by means of equivalence or specialisation relations. For example, when a class `ex:MyOrganisation` is defined in an ontology and then linked as equivalent to (i.e. `owl:equivalentClass`) `foaf:Organisation`.

1) DIRECT REUSE

As *L* approaches *%OE* (e.g. in the `Pure Science` domain), classes and properties are more likely to be directly reused by different datasets. This might indicate that the community associated with the KDs showing this behaviour prefer direct reuse (over indirect) of classes and properties, although further investigation is needed to support this conclusion (e.g. by inspecting specific usage of classes and properties). In light of this observation, we further investigate the distribution of entities in LOD Laundromat datasets. For each entity, we compute the number of datasets in which it occurs. The results are available online at.[33] We inspect the 100 most frequent entities. The entities showing the largest (direct) reuse are from the meteorological domain (after excluding the entities from RDF and OWL standard languages). For example, predicates and classes defined in the Knoesis's Sensor Observation ontology [40] occur in half of LOD Laundromat datasets. This observation is inline with the dataset classification results, since these entities belong to the `Meteorology` KD. We also notice the occurrence of entities from standard or very popular ontologies, such as

[32] https://dbpedia.org/ontology/

[33] https://github.com/empirical-knowledge-engineering/kda/blob/main/EntityReuse.md

TABLE 3. Summary of the statistics discussed throughout the paper.

	Datasets			All		Classes								Properties							
Knowledge Domain	T	L	%L	OE	%OE	%OE	%ES	R	%Ud	%Ut	D	A	H	%OE	%ES	R	%Ud	%Ut	D	A	H
Pure Science	523	604801	0.919	1290064	0.214	0.202	0.136	0.558	0.132	0.053	0.192	0.741	5	0.567	0.568	1.0	0.001	0.005	0.009	0.220	4
├- Biology	12	9	∼0	46410	0.008	0.009	0.011	0.998	0.997	0.018	0.014	0.962	5	0.002	0.002	0.985	0.069	0.002	0.002	0.223	3
├- Mathematics	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Physics, Astronomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
│ ├- Physics	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
│ └- Astronomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Earth	494	594729	0.904	640018	0.106	0.036	0.013	0.302	0.082	0.006	0.008	0.080	6	0.449	0.450	0.999	0.001	0.003	0.006	0.187	4
│ ├- Meteorology	405	582858	0.886	83700	0.014	0.034	0.011	0.261	0.062	0.004	0.005	0.048	6	0.032	0.032	0.993	0.007	0.002	0.005	0.197	3
│ ├- Geography	89	10684	0.016	227891	0.038	0.026	0.003	0.081	0.029	0.002	0.001	0.014	5	0.166	0.167	0.999	∼0	∼0	0.003	0.150	2
│ ├- Oceanography	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
│ ├- Paleontology	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
│ └- Geology	11	7878	0.012	3135	0.001	0.003	∼0	0.114	0.023	∼0	∼0	0.004	2	0.001	0.001	0.949	0.008	∼0	∼0	0.097	2
└- Chemistry	8	8905	0.014	1049	∼0	∼0	∼0	0.334	0.012	∼0	∼0	0.013	4	∼0	∼0	0.768	0.057	∼0	0.001	0.087	3
Social Science	315	35141	0.053	1579563	0.262	0.167	0.144	0.717	0.474	0.134	0.112	0.410	33	0.760	0.762	1.0	0.019	0.150	0.184	0.669	4
├- Anthropology	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
│ └- Royalty and nobility	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Sociology	181	22497	0.034	541991	0.090	0.088	0.060	0.569	0.337	0.049	0.042	0.196	33	0.227	0.227	0.999	0.035	0.080	0.108	0.455	4
│ └- Social Networking	153	9112	0.014	27069	0.004	0.001	0.001	0.724	0.308	∼0	∼0	0.004	2	0.018	0.018	0.995	0.001	∼0	0.003	0.188	2
├- Economy	23	3754	0.006	29212	0.005	0.006	0.004	0.495	0.185	0.002	0.002	0.035	4	0.012	0.011	0.983	0.056	0.007	0.007	0.176	3
├- Political Science	50	185	∼0	96630	0.016	∼0	∼0	0.965	0.424	∼0	∼0	0.001	1	0.073	0.074	1.0	∼0	∼0	∼0	0.031	1
├- Mediology	39	111	∼0	10104	0.002	0.026	0.002	0.073	0.031	0.002	∼0	∼0	2	0.001	0.001	0.993	0.003	∼0	∼0	0.070	1
├- Pedagogy	4	30	∼0	8	∼0	∼0	∼0	0.667	-	-	∼0	0.143	0	∼0	∼0	0.778	-	-	∼0	0.014	0
├- Industry	20	30	∼0	18072	0.003	0.003	0.001	0.372	0.456	0.001	0.001	0.025	4	0.010	0.009	0.980	0.068	0.007	0.007	0.174	3
├- Law	29	12	∼0	4255	0.001	∼0	∼0	0.647	0.088	∼0	∼0	0.017	1	0.003	0.003	0.996	0.001	∼0	∼0	0.061	1
├- Sexuality	1	1	∼0	27	∼0	∼0	∼0	0.667	0.067	∼0	∼0	0.312	0	∼0	∼0	0.750	0.062	∼0	∼0	0.039	1
├- Telecommunication	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Fashion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Military	2	1	∼0	26	∼0	∼0	∼0	0.455	0.045	∼0	∼0	0.017	1	∼0	∼0	0.727	0.091	∼0	∼0	0.022	0
│ └- Naval Science	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Transport, Travel	11	1	∼0	147	∼0	∼0	∼0	0.757	0.081	∼0	∼0	0.086	3	∼0	∼0	0.737	0.152	∼0	∼0	0.047	2
├- Transport	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Tourism	4	1	∼0	147	∼0	∼0	∼0	0.757	0.081	∼0	∼0	0.086	3	∼0	∼0	0.737	0.152	∼0	∼0	0.047	2
Applied Science	128	10546	0.016	104168	0.017	0.041	0.023	0.460	0.038	0.003	0.003	0.120	4	0.010	0.007	0.691	0.701	0.069	0.041	0.423	3
├- Medicine	20	10028	0.015	91143	0.015	0.017	0.020	0.973	0.038	0.001	0.001	0.055	3	0.007	0.004	0.571	0.819	0.055	0.026	0.615	3
├- Alimentation	9	2	∼0	9	∼0	∼0	∼0	0.667	-	-	∼0	0.143	0	∼0	∼0	0.800	-	-	∼0	0.015	0
├- Architecture	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Engineering	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Computer Science	86	188	∼0	5016	0.001	0.001	0.001	0.779	0.692	0.001	0.001	0.065	4	0.001	0.001	0.916	0.262	0.003	0.004	0.198	3
└- Agriculture	3	1	∼0	24	∼0	∼0	∼0	0.667	0.067	∼0	∼0	0.312	0	∼0	∼0	0.724	0.069	∼0	∼0	0.039	1
Life Science	63	10140	0.015	757216	0.125	0.153	0.177	0.960	0.833	0.249	0.159	0.619	7	0.031	0.028	0.909	0.263	0.082	0.066	0.599	3
├- Medicine	20	10028	0.015	91143	0.015	0.017	0.020	0.973	0.038	0.001	0.001	0.055	3	0.007	0.004	0.571	0.819	0.055	0.026	0.615	3
└- Biology	12	9	∼0	46410	0.008	0.009	0.011	0.998	0.997	0.018	0.014	0.962	5	0.002	0.002	0.985	0.069	0.002	0.002	0.223	3
Art, archi., and arche.	30	270	∼0	9267	0.002	0.001	0.001	0.748	0.263	∼0	∼0	0.003	5	0.004	0.004	0.977	0.005	∼0	0.002	0.102	3
├- Architecture	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Archeology	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Art & Culture	25	151	∼0	5222	0.001	0.001	∼0	0.639	0.182	∼0	∼0	0.002	3	0.003	0.003	0.975	0.005	∼0	0.001	0.074	2
└- Drawing	1	2	∼0	12	∼0	∼0	∼0	0.889	-	-	∼0	0.035	0	∼0	∼0	0.467	-	-	∼0	0.013	0
└- Music	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Plastic Arts	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Photography	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Theatre	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Dance	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Philosophy, Psychology	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Philosophy	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Psychology	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sport, Recreation	9	119	∼0	19786	0.003	0.025	0.002	0.054	0.020	0.001	∼0	0.001	2	0.010	0.010	0.990	0.002	∼0	0.001	0.098	1
└- Recreation	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Sport	9	119	∼0	19786	0.003	0.025	0.002	0.054	0.020	0.001	∼0	0.001	2	0.010	0.010	0.990	0.002	∼0	0.001	0.098	1
Engineering, Technology	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
├- Engineering	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
└- Technology	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Astrology	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bibliography	104	393	0.001	116574	0.019	0.028	0.003	0.103	0.026	0.001	0.001	0.030	2	0.079	0.079	0.998	0.003	0.002	0.006	0.273	2
History	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Linguistics	20	2	∼0	25384	0.004	0.004	0.004	0.822	0.522	0.004	0.003	0.070	4	0.007	0.006	0.959	0.074	0.005	0.008	0.239	3
Literature	1	1	∼0	22	∼0	∼0	∼0	0.733	-	-	∼0	0.067	1	∼0	∼0	0.538	-	-	∼0	0.024	1
Religion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Scientific Research	18	3	∼0	258	∼0	∼0	∼0	0.135	0.344	∼0	∼0	0.012	2	∼0	∼0	0.737	0.505	0.001	0.001	0.085	2
Time	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Dublin Core,[34] Friend Of A Friend (FOAF)[35] and the W3C's Time Ontology.[36] This observation also gives insights on the most used ontologies across domains.

2) INDIRECT REUSE

Indirect reuse of entities through equivalence relations is measured by R (see Table 3): the closer *R* is to 0 the higher the reuse through equivalence relations, the closer *R* is to 1 the lower the reuse through equivalence relations.

We observe that *R* varies as the KDs vary as far as reuse of classes is concerned. For properties, low reuse by equivalence looks a KD-independent practice, with a couple of exceptions.

Looking at class reuse, Life Science ($R = 0.960$) shows the highest values for *R*. The Pure Science KD ($R = 0.558$) shows a very variable behaviour for R when looking at its subdomains: the subdomain showing the lowest reuse is Biology ($R = 0.998$), which is also a subdomain of Life Science. The other subdomains have significantly higher level of reuse ($R <= 0.334$). Low reuse is also observed in Linguistics ($R = 0.822$) and Art, Architecture and Archeology

[34] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[35] http://xmlns.com/foaf/0.1/

[36] https://www.w3.org/TR/owl-time/

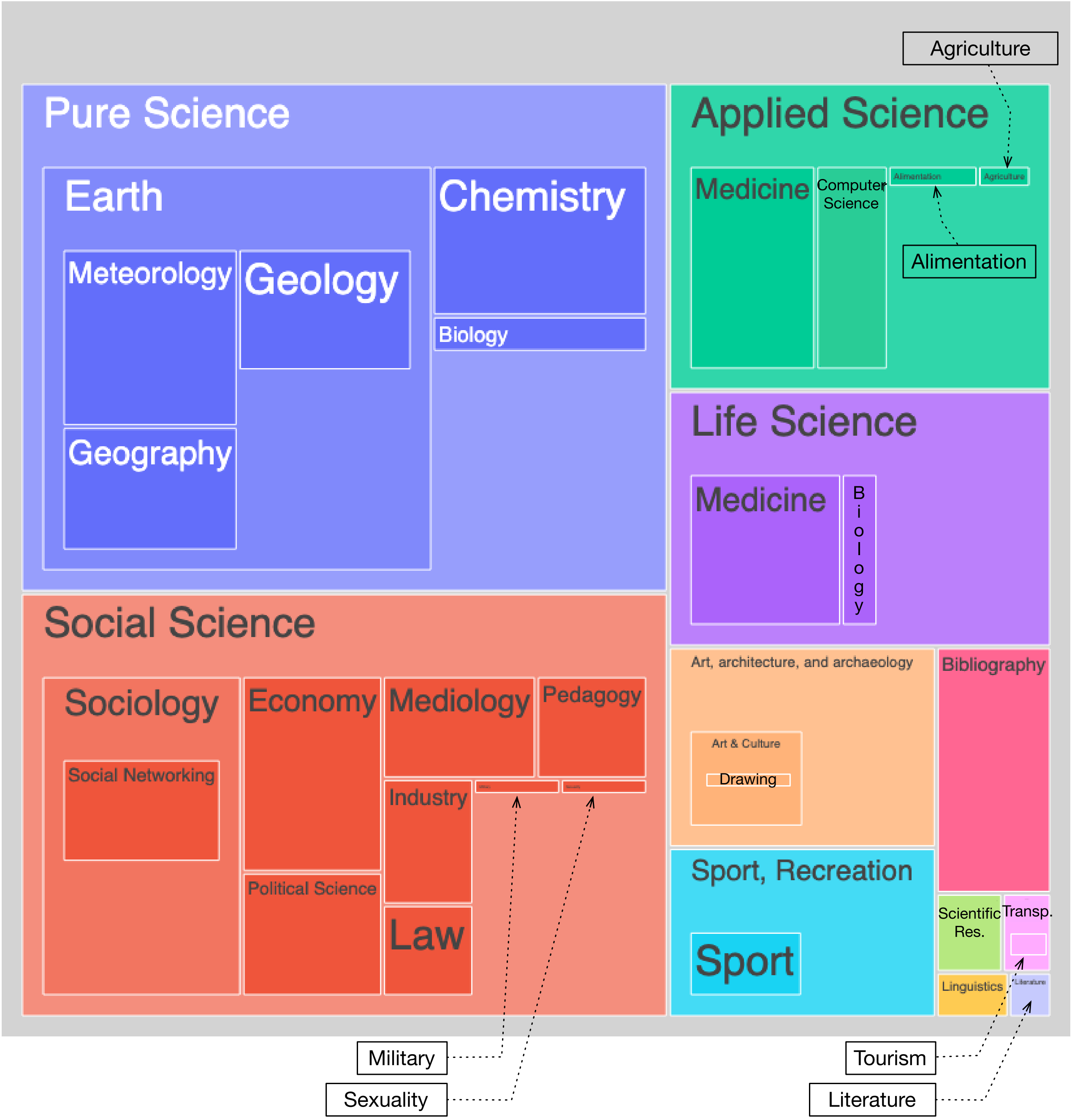


FIGURE 2. The distribution of KDs in LOD Laundromat.

($R = 0.748$). The Social Science KD also show a rather low level of reuse ($R = 0.717$), its subdomains have a very variable behaviour: from Political Science ($R = 0.965$) to Mediology ($R = 0.073$), which is among the ones that shows the highest reuse of classes. In summary, we can recognise a cluster of KDs showing a high reuse of ontologies through class equivalence ($R <= 0.14$): Geography, Geology, Mediology, Political Science, Sport, Recreation, Scientific Research, and Bibliography). And a cluster of KDs showing low reuse of ontologies through class equivalence ($R >= 0.9$): Biology, Medicine, and Political Science. This observation gives a good motivation to specialised ontology alignment effort as these KDs may highly benefit from it.

Looking at properties, it is confirmed that equivalence is rarely used (as noted in [8]), with notable exceptions: the Chemistry and Medicine KDs. These two KDs are pioneers in publishing their data as LOD. This deviation in the usage of the equivalence between properties could be partly explained with practices used for guaranteeing backward compatibility of URIs.[37] For instance, this behaviour is observed (by manual inspection) in the healthcare vocabulary.[38] It is noticed that `Chemistry` shows high indirect reuse for both classes and properties, while `Medicine` is among the KDs with lower indirect reuse, for classes.

3) USAGE OF SPECIALISATION RELATION

The specialisation relation is more used between classes than between properties (D significantly differs for the majority of KDs). Exceptions to this behaviour are observed for the `Medicine` and `Bibliography` domains, in which inheritance is more frequent between properties than classes.

We find that entities tend to attach to others of different KDs (A scores are usually less than 0.2). A different behaviour is observed in the `Biology` KD in which classes tend to attached within the same KD, exclusively ($A = 0.962$).

The common modelling practice in defining huge hierarchies of taxa (e.g. Linnean Taxonomy) may contribute to increment the assortativity in this KD. Nevertheless, it is also observed that `Biology` datasets are often published in central repositories (e.g. BioPortal[39]) and we can speculate that this practice fosters the findability of entities, thus providing LOD practitioners with better support for linking entities in their same KD.

C. EMPTY EXTENSIONS

We observe that among the ∼3M classes with empty extensions there are 987K blank nodes. For them, the lack of individuals is easily justified. After removing these entities, the majority of classes with empty extension ($Ut = 0.249$) are from `Life Science`. We remind that this domain is the third most covered in LOD Laundromat ($\%OE = 0.125$), with ∼757k entities and ∼10k datasets. It is also among the ones showing the lowest ontology reuse. Let us consider its subdomains `Biology` and `Medicine`. Although we cannot provide a comprehensive explanation of this phenomenon, we can inspect some of the most popular datasets to formulate hypotheses. As for `Biology`, most classes have empty extension ($Ud = 0.997$), while properties are instantiated. This is compatible, for example, with the design practice of the Gene ontology[40] where all *terms* are defined as classes under a top level of three classes (`Molecular Funcion`, `Cellular Component`, and `Biological Process`). These classes are used, in practice, as subjects and objects of triples applying *de facto* OWL punning. One question is whether the modelling practice of the Gene ontology represents a general modelling pattern from this KD and why. The other subdomain `Medicine` shows a different behaviour: classes are instantiated while most of the properties have empty extension ($Ud = 0.819$).

The rest of non-inistantiated entities mainly belong to the `Social Science` KD. A question that raises is whether the two KDs (`Life Science` and `Social Science`) use LOD for similar types of applications and whether this impacts on their modelling practice. And similarly, whether the other KDs showing a different behaviour are characterised by different application usage of LOD.

D. DEPTH DEGREE OF HIERARCHIES

Finally, we observe that, except the `Sociology` KD, which hierarchies reaches up to the 33 levels, hierarchies have maximum 7 levels (H is maximum 7). We manually inspect the deepest hierarchy in the `Sociology` domain to find that its entities occur in a dataset from the BTC.[41] This dataset collects information about people (mainly represented with FOAF). It also contains a hierarchy of 33 levels of classes defined in the FlyBase ontology[42] (FlyBase is a database of Drosophila genes, hence, the ontology belongs to the `Biology` domain). The FlyBase's classes included in the BTC's dataset have little contextual information (just the URIs of the entities and the specialisation relations), therefore, the classifier fails in associating them to the correct domain (i.e. `Biology`). With the exception of this hierarchy, the depth of the class taxonomies from the `Sociology` KD is in line with the other KDs (i.e. 6). This result confirms that LOD hierarchies are mostly flat as observed in [8], which may suggest a generalised design practice. A question is whether this observation can suggest a reference practice for the design of class and property taxonomies.

E. METRICS CORRELATION

To understand whether certain phenomena are related by potential causation, we compute the correlation between pairs of metrics, and analyse them domain-by-domain. We calculate the Pearson's correlation (ρ) coefficient over pairs of metrics computed for each domain and we assess the significance (p) of the result. Table 2 reports the pairs of metrics having high/very high correlation ($\rho > 0.7$) with significance $p < 0.001$, i.e. extremely significant (note that the notation "Metric_c" ("Metric_p") indicates that the metric has been computed considering only classes (properties)).

Most of the reported correlations are predictable. For example, the percentage of classes ($\%OE\_c$) and properties ($\%OE\_c$) increase as the total number of entities ($\%OE$) increases. It is also expected that the higher the number of entities in a domain ($\%OE$) the higher the (indirect) reuse of classes and properties: through equivalence ($\%ES\_p$ and $\%ES\_c$) and specialisation relations (D_c and D_p).

Interestingly, we report that the density of specialisations increases as the percentage of equivalence sets classes increases (since the high positive correlation between D_p

[37] To maintain semantic coherence, entities that change URIs across different releases of the same dataset are aligned by equivalence axioms.

[38] https://purl.org/healthcarevocab/v1

[39] https://bioportal.bioontology.org/

[40] http://geneontology.org/

[41] http://km.aifb.kit.edu/projects/btc-2010/btc-2010-chunk-243.gz

[42] http://purl.obolibrary.org/obo/fbsp.owl

and ES_c). Further investigation is required to interpret this phenomenon. For example, we can assess whether the properties of the aligned classes (i.e. the properties having as domain/range the aligned classes) are connected via specialisation relations. If this is the case, a possible explanation can be that the semantics of classes is more often compatible among ontologies developed for different use cases, than the semantics of properties, which may tend to be more specific. Another interesting phenomenon is the high positive correlation between Ut_p (i.e. percentage of properties with empty extension) and Ut_c (i.e. percentage of classes with empty extension). This may suggest that the ''empty extension'' phenomenon concerns entire modules of ontologies rather than being scattered and distributed among independent entities.

IX. CONCLUSION

In this paper, we investigate the relation between modelling practices (e.g. linking, ontology design and ontology population) with knowledge domains. Our goal is to assess whether the observed phenomena in LOD have a domain-specific character. To this end, we introduce a knowledge domain classification system and a novel method for multi-label topical classification of LOD datasets. As additional contribution, we manually curate a corpus of LOD datasets classified according to the classification system that can be used to reproduce our study as well as for further research on similar tasks. We also introduce a set of metrics (Section III) to perform observations on knowledge domain-specific modelling practices. The developed framework allowed us to report on the changing behaviour of LOD as the knowledge domains change, using LOD Laundromat as empirical basis.

We are working on additional metrics that can be computed on ESGs, and on extending the framework to analyse other kinds of relations (e.g. disjointness). We plan to develop quality indicators aiming at automatically assessing the design quality of datasets on the basis of the proposed metrics.

ACKNOWLEDGMENT

The authors would like to thank Luana Bulla, Stefano De Giorgis, Maria Chiara Frangipane, Anna Sofia Lillipolis, Delfina Sol Martinez Pandiani, Andrea Poltronieri, Margherita Porena, and Bruno Sartini for their help in annotating the collection of LOD datasets; and Aldo Gangemi for his enthusiastic support and precious hints.

REFERENCES

[1] W. Beek, S. Schlobach, and F. Van Harmelen, ''A contextualised semantics for owl: SameAs,'' in *Proc. Eur. Semantic Web Conf.*, H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, Eds. Cham, Switzerland: Springer, 2016, pp. 405–419, doi: 10.1007/978-3-319-34129-3_25.

[2] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson, ''When owl: SameAS isn't the same: An analysis of identity in linked data,'' in *Proc. Int. Semantic Web Conf.* (Lecture Notes in Computer Science), vol. 6496, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds. Cham, Switzerland: Springer, 2010, pp. 305–320, doi: 10.1007/978-3-642-17746-0.

[3] J. Raad, W. Beek, F. Van Harmelen, J. Wielemaker, N. Pernelle, and F. Sais, ''Constructing and cleaning identity graphs in the LOD cloud,'' *Data Intell.*, vol. 2, no. 3, pp. 323–352, 2020, doi: 10.1162/dint_a_00057.

[4] T. De Groot, J. Raad, and S. Schlobach, ''Analysing large inconsistent knowledge graphs using anti-patterns,'' in *Proc. 18th Int. Conf. (ESWC)*, R. Verborgh, K. Hose, H. Paulheim, P. Champin, M. Maleshkova, O. Corcho, P. Ristoski, and M. Alam, Eds. 2021, pp. 40–56, doi: 10.1007/978-3-030-77385-4_3.

[5] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, and C. Bizer, ''DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia,'' *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015, doi: 10.3233/SW-140134.

[6] F. M. Suchanek, G. Kasneci, and G. Weikum, ''YAGO: A core of semantic knowledge,'' in *Proc. 16th Int. Conf. World Wide Web*, C. Williamson, M. E. Zurko, P. Patel-Schneider, and P. Shenoy, Eds. 2007, pp. 697–706, doi: 10.1145/1242572.1242667.

[7] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, ''Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371),'' *Dagstuhl Rep.*, vol. 8, no. 9, pp. 29–111, 2018, doi: 10.4230/DagRep.8.9.29.

[8] L. Asprino, W. Beek, P. Ciancarini, F. van Harmelen, and V. Presutti, ''Observing LOD using equivalent set graphs: It is mostly flat and sparsely linked,'' in *Proc. 18th Int. Semantic Web Conf.*, C. Ghidini, O. Hartig, M. Maleshkova, V. Svatek, I. F. Cruz, A. Hogan, J. Song, M. Lefrancois, and F. Gandon, Eds. Cham, Switzerland: Springer, 2019, pp. 57–74, doi: 10.1007/978-3-030-30793-6_4.

[9] O. Bodenreider, ''Biomedical ontologies in action: Role in knowledge management, data integration and decision support,'' *Yearbook Med. Informat.*, vol. 17, no. 1, pp. 67–79, 2008.

[10] C. Patel, K. Supekar, Y. Lee, and E. K. Park, ''OntoKhoj: A semantic web portal for ontology searching, ranking and classification,'' in *Proc. 5th ACM Int. Workshop Web Inf. Data Manag.*, R. H. L. Chiang, A. H. F. Laender, and E. Lim, Eds. Nov. 2003, pp. 58–61, doi: 10.1145/956699.956712.

[11] R. Meusel, B. Spahiu, C. Bizer, and H. Paulheim, ''Towards automatic topical classification of LOD datasets,'' in *Proc. 24th Int. World Wide Web Conf.*, vol. 1409, C. Bizer, S. Auer, T. Berners-Lee, and T. Heath, Eds. 2015, pp. 1–6. [Online]. Available: http://ceur-ws.org/Vol-1409/paper-03.pdf

[12] A. Pister and G. A. Atemezing, ''Towards automatic domain classification of LOV vocabularies,'' in *Proc. Joint 6th Int. Workshop Dataset Proflling Search 1st Workshop Semantic Explainability Co-Located 18th Int. Semantic Web Conf.*, E. Demidova, S. Dietze, J. G. Breslin, S. Gottschalk, P. Cimiano, B. Ell, A. Lawrynowicz, L. Moss, and A. N. Ngomo, Eds. 2019, pp. 18–28.

[13] B. Spahiu, A. Maurino, and R. Meusel, ''Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned,'' *Semantic Web*, vol. 10, no. 2, pp. 329–348, Jan. 2019, doi: 10.3233/SW-180323.

[14] A. Nogales, M.-A. Sicilia, and A. J. Garcia-Tejedor, ''A domain categorisation of vocabularies based on a deep learning classifier,'' *J. Inf. Sci.*, 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1177/01655515211018170, doi: 10.1177/01655515211018170.

[15] M. Fahad, N. Moalla, A. Bouras, M. A. Qadir, and M. Farukh, ''Towards classification of web ontologies for the emerging semantic web,'' *J. Universal Comput. Sci.*, vol. 17, no. 7, pp. 1021–1042, 2011, doi: 10.3217/jucs-017-07-1021.

[16] S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain, ''Automatic domain identification for linked open data,'' in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, Nov. 2013, pp. 205–212, doi: 10.1109/WI-IAT.2013.206.

[17] B. Fetahu, S. Dietze, B. P. Nunes, M. A. Casanova, D. Taibi, and W. Nejdl, ''A scalable approach for efficiently generating structured dataset topic profiles,'' in *Proc. 11th Int. Conf.*, 2014, pp. 519–534, doi: 10.1007/978-3-319-07443-6_35.

[18] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, ''Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web,'' *Semantic Web*, vol. 8, no. 3, pp. 437–452, Dec. 2016, doi: 10.3233/SW-160213.

[19] *Lod Cloud.* Accessed: Sep. 6, 2022. [Online]. Available: https://lod-cloud.net/

[20] W. Beek, L. Rietveld, S. Schlobach, and F. Van Harmelen, ''LOD laundromat: Why the semantic web needs centralization (even if we don't like it),'' *IEEE Internet Comput.*, vol. 20, no. 2, pp. 78–81, Mar. 2016, doi: 10.1109/MIC.2016.43.

[21] M. Schmachtenberg, C. Bizer, and H. Paulheim, "Adoption of the linked data best practices in different topical domains," in *Proc. 13th Int. Semantic Web Conf.*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, Eds. 2014, pp. 245–260, doi: 10.1007/978-3-319-11964-9_16.

[22] M. Wylot, P. Cudre-Mauroux, M. Hauswirth, and P. Groth, "Storing, tracking, and querying provenance in linked data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1751–1764, Aug. 2017, doi: 10.1109/TKDE.2017.2690299.

[23] Q. Dai, X.-M. Wu, J. Xiao, X. Shen, and D. Wang, "Graph transfer learning via adversarial domain adaptation with graph convolution," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 19, 2022, doi: 10.1109/TKDE.2022.3144250.

[24] L. Asprino, V. A. Carriero, and V. Presutti, "Extraction of common conceptual components from multiple ontologies," in *Proc. 11th Knowl. Capture Conf.*, A. L. Gentile and R. Goncalves, Eds. Dec. 2021, pp. 185–192, doi: 10.1145/3460210.3493542.

[25] R. A. A. Principe, A. Maurino, M. Palmonari, M. Ciavotta, and B. Spahiu, "ABSTAT-HD: A scalable tool for profiling very large knowledge graphs," *VLDB J.*, vol. 31, pp. 1–26, Sep. 2021, doi: 10.1007/s00778-021-00704-2.

[26] L. Ding, J. Shinavier, Z. Shangguan, and D. McGuinness, "SameAS networks and beyond: Analyzing deployment status and implications of owl: SameAS in linked data," in *Proc. 9th Int. Semantic Web Conf.* (Lecture Notes in Computer Science), vol. 6496, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds. Cham, Switzerland: Springer, 2010, pp. 145–160, doi: 10.1007/978-3-642-17746-0_10.

[27] W. Beek, J. Raad, J. Wielemaker, and F. van Harmelen, "SameAS.cc: The closure of 500M owl: SameAS statements," in *Proc. 15th Int. Conf.*, 2018, pp. 65–80, doi: 10.1007/978-3-319-93417-4_5.

[28] A. Mallea, M. Arenas, A. Hogan, and A. Polleres, "On blank nodes," in *Proc. 10th Int. Semantic Web Conf.* (Lecture Notes in Computer Science), vol. 7031, L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, Eds. Cham, Switzerland: Springer, 2011, pp. 421–437, doi: 10.1007/978-3-642-25073-6_27.

[29] H. Paulheim and A. Gangemi, "Serving DBpedia with DOLCE—More than just adding a cherry on top," in *Proc. 14th Int. Semantic Web Conf.*, M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, and S. Staab, Eds. Cham, Switzerland: Springer, 2015, pp. 180–196, doi: 10.1007/978-3-319-25007-6_11.

[30] L. Asprino, V. Basile, P. Ciancarini, and V. Presutti, "Empirical analysis of foundational distinctions in linked open data," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3962–3969, doi: 10.24963/ijcai.2018/551.

[31] J. D. Fernandez, W. Beek, M. A. Martinez-Prieto, and M. Arias, "Lod-a-lot: A queryable dump of the LOD cloud," in *Proc. 16th Int. Semantic Web Conf.*, C. d'Amato, M. Fernandez, V. A. M. Tamma, F. Lecue, P. Cudre-Mauroux, J. F. Sequeda, C. Lange, and J. Heflin, Eds. Cham, Switzerland: Springer, 2017, pp. 75–83, doi: 10.1007/978-3-319-68204-4_7.

[32] J. D. Fernández, M. A. Martínez-Prieto, C. Gutiérrez, A. Polleres, and M. Arias, "Binary RDF representation for publication and exchange (HDT)," *J. Web Semantics*, vol. 19, pp. 22–41, Mar. 2013, doi: 10.1016/j.websem.2013.01.002.

[33] B. Magnini and G. Cavaglia, "Integrating subject field codes into wordnet," in *Proc. 2nd Int. Conf. Lang. Resour. Eval.*, 2000, pp. 1–6.

[34] J. Camacho-Collados and R. Navigli, "BabelDomains: Large-scale domain labeling of lexical resources," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 223–228.

[35] F. Charte, A. J. Rivera, M. J. D. Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowl.-Based Syst.*, vol. 89, pp. 385–397, Nov. 2015, doi: 10.1016/j.knosys.2015.07.019.

[36] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing applications," in *Proc. 17th ACM Conf. Inf. Knowl. Manag.*, J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K. Choi, and A. Chowdhury, Eds. Oct. 2008, pp. 153–162, doi: 10.1145/1458082.1458105.

[37] K. Sechidis, G. Tsoumakas, and I. P. Vlahavas, "On the stratification of multi-label data," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. 2011, pp. 145–158, doi: 10.1007/978-3-642-23808-6_10.

[38] J. Herrera, A. Hogan, and T. Kafer, "BTC-2019: The 2019 billion triple challenge dataset," in *Proc. 18th Int. Semantic Web Conf.*, C. Ghidini, O. Hartig, M. Maleshkova, V. Svatek, I. F. Cruz, A. Hogan, J. Song, M. Lefrancois, and F. Gandon, Eds. 2019, pp. 163–180, doi: 10.1007/978-3-030-30796-7_11.

[39] V. Presutti, G. Lodi, A. Nuzzolese, A. Gangemi, S. Peroni, and L. Asprino, "The role of ontology design patterns in linked data projects," in *Proc. 35th Int. Conf.*, I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, and M. Saeki, Eds. Cham, Switzerland: Springer, 2016, pp. 113–121, doi: 10.1007/978-3-319-46397-1_9.

[40] C. A. Henson, J. K. Pschorr, A. P. Sheth, and K. Thirunarayan, "SemSOS: Semantic sensor observation service," in *Proc. Int. Symp. Collaborative Technol. Syst.*, 2009, pp. 44–53, doi: 10.1109/CTS.2009.5067461.

LUIGI ASPRINO received the Ph.D. degree in computer science and engineering from the University of Bologna, in 2019. He is currently an Assistant Professor with the University of Bologna, where he teaches artificial intelligence and basic informatics (bachelor's degree in humanities). He has coauthored multiple scientific papers for international journals and conferences, including IJCAI, AAAI, and ISWC. His research interests include artificial intelligence, knowledge engineering, and empirical semantics.

VALENTINA PRESUTTI received the Ph.D. degree in computer science from the University of Bologna, in 2006. She is currently an Associate Professor with the University of Bologna. She is also an Associate Researcher with the Institute of Cognitive Science and Technologies of CNR and a Coordinator with STLab. She coordinates the EU H2020 Project Polifonia (2021–2024). She was responsible for several national and EU projects (e.g. MARIO, IKS, and ArCo). During her Post-doctoral degree, she created ontologydesignpatterns.org and the workshop series WOP, and reference resources for semantic web researchers. She has published more than 150 articles in international peer-reviewed venues. Her research interests include AI, semantic web, knowledge extraction and engineering, empirical semantics, social robotics, and music AI. She is a Editorial Board Member of *Data Intelligence* (MIT Press), *JASIST* (Wiley), *Intelligenza Artificiale* (IOS Press), and Semantic Web Studies (IOS Press). She is the Co-Director of International Semantic Web Research Summer School (ISWS). She is the Editor-in-Chief of the *Journal of Web Semantics* (Elsevier).

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI CARE Agreement