Doc2KG: Transforming Document Repositories to Knowledge Graphs

N. Stylianou[a]*, D. Vlachava[b], I. Konstantinidis[b], N. Bassilades[a] and V. Peristeras[b c]

[a] *School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[b] *School for Science and Technology, International Hellenic University, Thessaloniki, Greece*

[c] *European Commission, DG Informatics*

Corresponding author at: School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece.
Email addresses: nstylia@csd.auth.gr (Nikolaos Stylianou), d.vlachava@ihu.edu.gr (Danai Vlachava), i.konstantinidis@ihu.edu.gr (Ioannis Konstantinidis), nbassili@csd.auth.gr (Nick Bassiliades), v.peristeras@ihu.edu.gr (Vassilios Peristeras)

Document Management Systems (DMS) are used for decades to store large amounts of information in textual form. Their technology paradigm is based on storing vast quantities of textual information enriched with metadata to support searchability. However, this exhibits limitations as it treats textual information as black box and is based exclusively on user-created metadata, a process that suffers from quality and completeness shortcomings. The use of knowledge graphs in DMS can substantially improve searchability, providing the ability to link data and enabling semantic searching. Recent approaches focus on either creating knowledge graphs from document collections or updating existing ones. In this paper, we introduce Document-to-Knowledge-Graph (Doc2KG), an intelligent framework that handles both creation and real-time updating of a knowledge graph, while also exploiting domain-specific ontology standards. We use DIAVGEIA (clarity), an award winning Greek open government portal, as our case-study and discuss new capabilities for the portal by implementing Doc2KG.

Keywords: Semantic Web, Linked Data, Open Data, Machine Learning, Natural Language Processing, eGovernment, Government Portals

Introduction

A huge amount of new data is created and stored every minute by users in order to be retrievable and discoverable. In modern organisations, both in the private and public sector, textual information in electronic documents is stored in big volumes in Document Management Systems (DMS). DMS were first introduced in enterprise environ-

ments both in the private and the public sector over 30 years ago to receive, track, manage and store documents. Over time, along with the dramatic increase in the pace of data creation and increasing storage needs, these systems saw little improvement concerning information retrieval functionalities. This resulted in difficulties in locating, identify, retrieve information in collections that often expand to millions of documents. That is due to the fact that these systems cannot "look into" the textual information they store, but rather treat it as a black-box described by user-provided metadata. Inevitably, this human-created metadata often suffers from low quality.

With the rise of open Government and open data rhetorics and practices, some of these public sector DMS publish their content as open data to the Web to improve transparency and accessibility. Benefits of open data besides increased transparency also include democratic control, improved or new public products and services, improved government services, innovation and new knowledge creation from combined data sources and the possibility to identify patterns in large data volumes, among others (Pereira et al., 2017). For these reasons, in the last decade, open government policies have started to gain ground in an increasing number of countries globally, while several projects based on open data are executed all over the world (Mohamed et al., 2020; Zuiderwijk & Janssen, 2014).

This is the case of the Greek portal DIAVGEIA[1] (in English: Clarity) in which all public sector administrative decisions are published, as mandated by law, forming a huge and fast-growing collection of more than 43 million documents. Essentially, DIAVGEIA is providing access to the governmental DMS, which stores all the documents. The huge volume of this textual information, combined with the lack of high quality and standardised metadata, poses several problems and processing challenges, justifying the use of the term "big data" to describe such a corpus of information.

Open (big) data must be available in a convenient and modifiable form, in order to be easy to exploit, i.e., to increase data interoperability, be able to combine different datasets together. Towards improving information and knowledge extraction, Semantic Web technologies like RDF and OWL were developed and standardized in the form of (meta-)data graphs consisting of elementary vertice-edge-vertice triples (subject, predicate, object) (Zaveri et al., 2016). Tim Berners-Lee (2010), the inventor of the Web and linked data initiator, suggested a 5-star deployment scheme for open data quality that constitutes the status quo in Semantic Web best practices (Hasnain & Rebholz-Schuhmann, 2018). This scheme proposes publishing machine-readable structured data and using open standards from W3C that are also linked to other linked open data. These linked data principles can also provide the basis for complying data to other recommendations employed by the research community like the Findable Accessible Interoperable Reusable (FAIR) principles indicating that data resources should support discovery and reusability by different stakeholders (Garijo & Poveda-Villalón, 2020).

Apart from the area of open data, linked data availability and integration as described above provide important benefits for enterprise data, including "not open" data, as well. RDF graphs with existing domain standards can improve data quality and facilitate data migration and cooperation between different enterprises but also within the same enterprise while addressing interoperability, access, legal and encryption issues (Hu & Svensson, 2010). Interestingly, while information in large governmental DMS is particularly valuable and qualifies to be published as open data, it can also be used for internal administrative processes to improve internal efficiencies. Thus, data from DMS,

---

[1] diavgeia.gov.gr/en/

like in DIAVGEIA, is highly valuable both for open and internal purposes and use cases.

There is a dizzying amount of low-quality data already in existence, especially in large governmental DMS. The manual upgrade of such huge amounts of data to 4, 5 stars would be an enormous error-prone and multi-year project leading to high costs. Therefore, there is a clear need for this transformation to be supported with appropriate tools to enable the automated extraction of structured information from big textual data. The use of modern Machine Learning (ML) and Natural Language Processing (NLP) combined with linked data and semantic web techniques can facilitate this process. In the last decade, staggering advancements have taken place in these fields. Despite of this, little work has been done towards the combination of open and linked data with intelligent functionalities, services and features. This is a promising combination that could transform large and static document collections, like large governmental DMS, into dynamic and user-friendly knowledge graphs. According to Hogan et al. (2021), a knowledge graph is "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities" (p. 2).

In this paper, we propose Doc2KG (Document-to-Knowledge-Graph), an intelligent framework that transforms a low-quality, from an information retrieval and reusability point of view, DMS to a knowledge graph. The proposed framework further supports the perpetual expansion of the knowledge graph, as it is enriched with additional documents when these become available. In summary, our contributions are as follows:

- A complete framework that handles DMS to knowledge graph transition and further enrichment. Previous work is focused only either on publishing linked data or on NLP approaches for information extraction.
- A novel methodology to creating knowledge from a document repository, using domain-specific ontology standards to handle the "cold-start" problem of entity linking and ensure the quality of the resulting Knowledge Graph.
- A modern Information Extraction pipeline, leveraging pre-trained Language Models (LM) and state-of-the-art Information Extraction (IE) models. We eliminate the need for huge amounts of training data and hard to maintain resources by exploiting Transfer Learning approaches and require less overhead to extract all the necessary information.
- Increased accessibility, retrievability and searchability of the available data while maintaining a reduced maintenance cost from constantly updating and indexing a DMS.

In the remainder of the paper, we first discuss related literature for knowledge graph construction and enrichment and our application domain. We then introduce our Doc2KG framework, analysing its components and detailing its architecture. For our case study we use DIAVGEIA, to demonstrate the benefits of our framework in a real-world portal. Conclusively, we discuss the limitations of our framework and indicate directions for future research.

Related Work

There has been limited research on end-to-end knowledge graph construction frameworks from a document repository. Most of the studies focus on specific subtasks, including entity recognition, entity disambiguation, entity linking, relation extraction, and

linking and publishing linked data. Yu et al. (2020) proposed a framework for domain knowledge graph creation by leveraging structured information from Wikipedia. However, in many document repositories with domain-specific knowledge, the entities are not related to information from Wikipedia due to insufficient data in low-resource languages. Many research studies combined existing Named Entity Recognition (NER) models with rule-based approaches for relation extraction and linking (Kertkeidkachorn & Ichise, 2018), and statistical approaches on specific domains of literature (Buscaldi et al., 2019). Yet, these methods did not consider restrictions in order to validated the triples in the knowledge graph creation process.

Martinez-Rodriguez et al. (2018) integrated open information extraction methods for triple extraction, enhanced with semantic role labeling and in combination with aggregated results from ensemble models. However, the proposed frameworks did not deal with entity linking to existing domain standards for semantic interoperability. Other methods focusing on entity linking considered the existence of a knowledge graph without dealing with the cold start problem by constructing a knowledge graph from the beginning, where there is no available information on an instance level (Ganea & Hofmann, 2017; Raiman & Raiman, 2018). Towards this regard, Rossanez et al. (2020) proposed an approach for knowledge graph construction from biomedical scientific articles. The framework combined various NLP techniques to identify all possible triples in the document and mapped them to Unified Modeling Language (UML) concepts that are subsequently linked to the domain ontology. However, the approach was highly dependent on the biomedical domain.

At the same time, other methods focused on how to publish a DMS as linked data and making it interoperable. Towards having high-quality data, Penteado et al. (2021) presented a unified process of publishing linked open data in a systematic way by following specific steps, like dataset selection, data cleaning, defining vocabularies, metadata specification, linking data to external sources, data serialization in a linked open data representation and data publication. Milić et al. (2020) proposed a model for linking open government datasets using Semantic Web technologies and metadata information to identify similarities between different datasets. For Greek initiatives, Chalkidis et al. (2017) developed the Nomothesia (law) ontology for publishing Greek legal documents as linked open data based on European standards and best practices for ontology modeling.

Similarly, Bratsas et al. (2021) defined an RDF vocabulary to convert data from a relational database into a knowledge graph. Then, they used this knowledge graph for facilitating ETL (Extract Transform Load) processes and trained a clustering model for identifying red flags in fiscal projects. Finally, Savvas & Bassiliades (2009) presented a process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration that provides an up-to-date and accurate legal framework for interpreting, producing and processing administrative documents. However, most of these methods proposed a new solution for future transactions without dealing with the existing information that is in an unstructured format, as this process would require extensive human effort.

In view of the above, it is evident that there is no unified framework that addresses all the aspects of knowledge graph creation.

Open government data portals comprise a typical use case of a DMS and have been implemented at national and European level (de Juana-Espinosa & Luján-Mora, 2019). However, most of the portals' DMS contain information in unstructured format leading to low data quality. DIAVGEIA is one of the most prominent efforts in Greece and has been recognized as a best practice, receiving awards at an international level

(OECD, 2011). According to national law, all decisions made by public services have to be published on the portal in order to be valid. Nevertheless, the current version is a repository of documents in PDF, indicating low data quality. The importance of improving transparency and user-friendliness in DIAVGEIA has already been addressed by Gritzalis et al. (2017). The authors conducted a survey on its impact and concluded that transparency of administrative acts and decisions could lead to citizens' trust in institutions. According to Matheus and Janssen (2020), "transparency is about creating an insight for someone who is not involved" and in the case of DIAVGEIA, for the citizens and other public institutions.

There has been limited work for improving DIAVGEIA's data quality and enhancing transparency based on information from the portal. Towards this, Vafopoulos et al. (2012) presented publicspending.net that collects metadata from the DMS to increase public awareness on public expenditures. Beris & Koubarakis (2018) proposed improved portal publishing decisions as linked open data called DiavgeiaRedefined by developing an ontology for describing information of administrative acts and decisions.

The Doc2KG Framework

In this section, we present our proposed framework for the continuous conversion of open data to a knowledge graph, exploiting existing domain ontology standards. Effectively, we are utilizing pre-existing vocabularies to convert the already existing textual information to linked data, transforming them to 4 stars open data according to Tim Berners-Lee's scheme. In addition, our framework is designed to work on an active DMS system, handling both initial conversion of pre-existing information and annotation of new documents to add them to the knowledge graph.

In comparison to previous approaches, Doc2KG offers a unified methodology to handle both the initial conversion and further enrichment. As such, it represents a consistent medium towards knowledge graph creation and maintenance. What is more, our approach uses domain-specific ontology standards to ensure that the created knowledge graph adheres to the domain's restrictions at all times. Furthermore, our IE pipeline is modularly designed to allow for easy replacement of the contribution models with better ones as they are developed.

To that end, we start with a large volume document repository (i.e., a DMS) holding a vast collection of heterogeneous documents and a domain-specific ontology. Our proposed methodology is comprised of three stages. As our document collection is heterogeneous, we make the realistic assumption that some of the documents in the repository are already annotated with domain-specific meta-data (2 or 3 stars open data) while others are not (1-star open data).

The first stage involves the pre-processing of the collection for intelligent information extraction using a variety of ML and NLP methods. In practice, we first split the documents into sentences of tokens while also disambiguating abbreviations. The required information is then extracted using pre-trained models or tools such as CoreNLP2.

In the second stage, we transform the document collection into a knowledge graph based on the descriptions of the classes, attributes and relations in the ontology. At this point, we assume that the ontology descriptions provide sufficient information for contextual representations to be created. We require descriptions in the ontology so

---

[2] https://stanfordnlp.github.io/CoreNLP/

that we can create as more accurately semantic representations of the nodes as possible for the knowledge graph creation process. In cases where class descriptions are absent, we will only use the class names for the node representations. During this stage, we present a metadata constrained approach for the documents with such pre-existing information (2 or 3 stars open data) as well as a semantically constrained approach to the already extracted information with ontological entities.

The third stage describes the continuous integration of new documents, as they are created, to the knowledge graph.
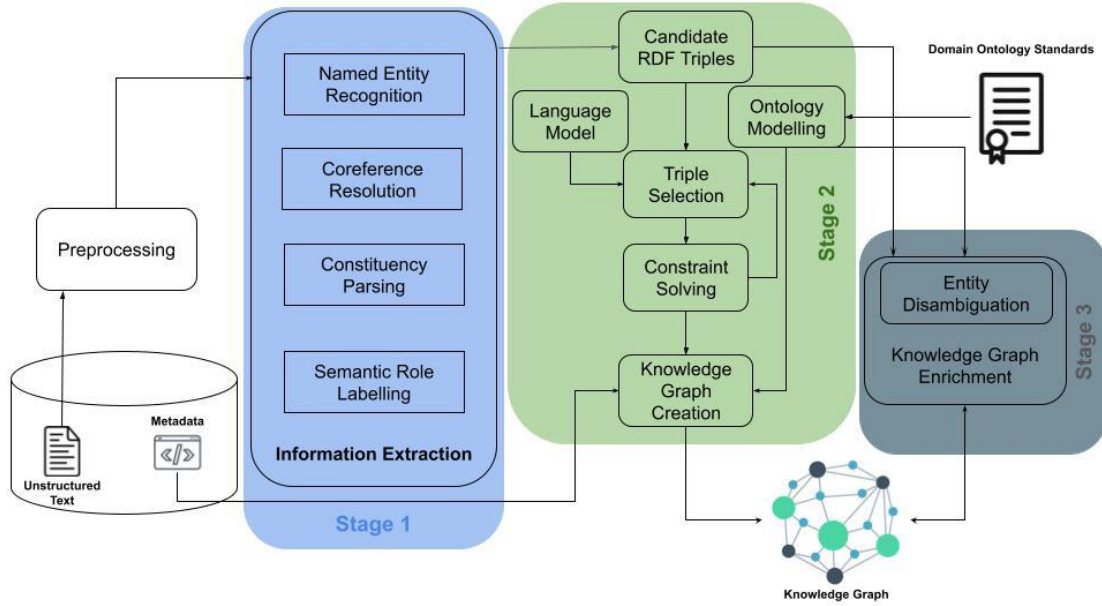

Figure 1: Doc2KG architecture overview with deployment stages colour annotated.

Formally, we consider our document repository to consist of a collection of documents $D = (d_1, d_2, …, d_l)$, where l is the number of documents in the collection, a subset of which will be annotated with metadata $M = \{p_1:v_1, p_2:v_2, …, p_q:v_q\}$ such that each metadata is described by a label p_q and a value v_q, where q is the number of available metadata. All documents utilize the same set of metadata, where the labels are predefined properties and the values are excerpts of text from the respective document when populated.

Stage 1: Pre-processing and Information Extraction

During the first stage of our framework, we initially split the document into sentences (s) and the sentences in tokens (t), such that $d_i = \{s_1, s_2, ..., s_n\}$ and $s_j = \{t_1, t_2, …, t_m\}$, where d_i and s_j are a randomly selected document and a randomly selected sentence from that document respectively, and m and n are the number of tokens in sentence s_j and document d_i. Next, we use a combination of Regular Expressions (RegEx) and Part-of-speech (POS) tagging to extract and resolve abbreviations in text and replace them with their original forms (*AbreviationResolver*). This step allows for better performance in the following steps of our proposed methodology to better identify similar entities and link them together. The resulting pre-processed document will have n number of

sentences, each with at least **m** tokens, as resolving any abbreviations occurring in the sentence will result in more tokens.

With the documents processed, the next step involves a series of information extraction tasks that aim to identify and annotate specific types of information present in the text. Specifically, we apply a combination of NER, Coreference Resolution (CR), Constituency Parsing (CP) and Semantic Role Labelling (SRL) in order to identify entity mentions in text, to link them together, as well as to identify the syntactic dependencies of each sentence. The NER process, given a sequence of tokens **t** (i.e., a sentence), produces a second equal length sequence (**e**) which annotates each token to a predefined entity such that $e_j = (e_{j1}, e_{j2}, \ldots, e_{jm})$. The set of entities identified by a NER process can change depending on the domain (e.g., in a biomedical domain, we are more interested in identifying all diseases mentioned in a document than locations and organization names). The CR process, on the other hand, given a document **d** identifies which tokens refer to the same entity and returns a series of coreference chains (**c**) such that $c_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$, where k is the number of identified real-world entities in d, where each chain has to be comprised by at least one token.

Similarly, we utilize a CP process to transform each sentence in each document into a constituent parse tree (cp_i). A CP tree splits each sentence based on its grammar categories, in which the root of the tree is the sentence, traversing to phrase type (e.g., noun phrases and verb phrases), part-of-speech tag and ending with the tokens as leaves.

The SRL process is used to identify the arguments of all the verbs in a sentence and their relations, represented as triplets where each triplet (**sr**) contains the two arguments related to the verb. The SRL process makes use of further resources like VerbNet (Schuler, 2005) or PropBank (Kingsbury & Palmer, 2002) to annotate the verbs with role names, as they are defined in the respective corpus. These resources help identify the roles of the attributes of the verbs and propositions, respectively, to *Agent* and *Patient*, allowing for better SRL performance.

Having extracted all the required information from our documents, we focus on the creation of candidate RDF triples. Each triple is in the form of **tr = (s, p, o)**, where **s**, **p**, **o** is the subject, predicate and object, respectively. As a result, for each document, a representative mention from each coreference chain c is selected and replaces all entity mentions in the document (*CoreferenceMentionResolution*). We opt towards using proper nouns or identified named entities when that is possible. Then, using the identified named entities as candidates for subject or object, we find all the instances where they have been identified as arguments to a verb. All the triples identified in **sr,** which contain entities, are candidate RDF triples, with the verbs playing the role of predicates. Based on the identified verb semantics, Agents are identified as subject and Patients as an object. If this information is not available, we built two triples with each element being a subject or an object and resolve this in Stage 2 of the proposed framework through restriction compliance checking. With this approach, we maximize the information throughput from Stage 1 to Stage 2 to ensure that we do not mistakenly discard identified nodes.

ALGORITHM 1 PRE-PROCESSING AND INFORMATION EXTRACTION PROCESS

Input: D ← *document,*
NER ← *Named Entity Recognition Model,*
CR ← *Coreference Model,*
CP ← *Constituency Parsing Model,*
SRL ← *Semantic Role Labelling Model*

Output: *Candidate RDF triples*

```
1    d ← Tokenize(SentenceSplit(D))
2    for i ←1 … n do
3        tokenizedSentence ← di
4        di ← AbreviationResolver(tokenizedSentence)
5    end for
6    e ← NER(d) //entities
7    c ← CR(d) //coreference chains
8    d ← CoreferenceMentionResolution(d, c) //replace all mentions with a candidate
9    cp ← CP(d) //constituent parse tree
10   sr ← SRL(d, VerbNet, cp) //verb semantic role labelling
11   tr ← TripleConstructor(d, e, sr) //candidate RDF triples
```

Consequently, the set of candidate RDF triples extracted from each document, along with the metadata of each document, when these are available, are used in the next stage of our framework toward the creation of a knowledge graph using a domain-specific ontology.

This stage's performance is depended on the individual performance of the IE components that comprise it. Therefore, each task has its own set of evaluation criteria and metrics, which can be used to verify a model's performance given a set of labelled data. As such, the NER model, the CR model and the CP model are all independent as they are applied directly to the source data. The effects of propagated errors from the CP model in the SRL model are based on the noise resilience and generalizing ability of the SRL model and are case depended. The triple construction approach is designed to maximize the number of candidate triples from this process and its performance is directly dependent on the NER and SRL models.

Current state-of-the-art approaches achieve very high-performance scores in the independent components as they are heavily based on pre-trained LMs and Transfer Learning methodologies (Li et al., 2020; Stylianou & Vlahavas, 2021). As a result, the propagated error from this stage will be minimal. Considering the domain, the NER model described by (Baevski et al., 2019) achieves 93.5% F1-score, the CR model by (Khosla & Rose, 2020) achieves an 85.8% F1-score, the CP model by (Papay et al., 2021) reaches a 93.8% F1-score and the SRL model by (Papay et al., 2021) attains a % F1-score. In addition, all these model designs satisfy our scheme requirements and are capable of producing the required outputs for the completion of first stage of the framework. Moreover, all these components are independent models that can be replaced with better performing counterparts are NLP research progresses in these fields.

Stage 2: Ontology-Constrained Knowledge Graph creation

ALGORITHM 2 ONTOLOGY-CONSTRAINED KNOWLEDGE GRAPH CREATION

Input:
X ← Predicted RDF candidate and metadata (tr, m) for each document in D
H ← heuristic mapping,
O ← Domain-specific Ontology
PLM ← Pre-trained LM
k ← number of maximum closest terms
t ← Compliance threshold
Output:
Knowledge Graph (KG) from the document collection

```
1    O* ← PLM(O) //contextualized representations of ontology labels and descriptions
2    KG ← ( ) // Knowledge Graph triples initialization
3    for x in X do
4        tr, m ← x
```

5	{*s, p, o*} ← *PLM(tr) //contextualized representations of triple terms*
6	*s$_s$, p$_s$, o$_s$* ← *SimilarityScoring(O*, {s, p, o})*
7	*s*$_s$, p*$_s$, o*$_s$* ← *TopKSampling(k, {s$_s$, p$_s$, o$_s$})*
8	*sl, pl, ol* ← *OntologyLabelling(H, {s*$_s$, p*$_s$, o*$_s$}, m) //label for each type*
9	**while** *sl, pl, ol are not None and (s*$_s$, p*$_s$, o*$_s$) is not empty* **do:**
10	**for** *s, p, o in (s*$_s$, p*$_s$, o*$_s$)* **do:**
11	*sl, pl, ol* ← *RestrictionsCompliance((s, p, o), t)*
12	*Remove s, p, o from s*$_s$, p*$_s$, o*$_s$*
13	**end for**
14	**end while**
15	**Add** (*sl, pl, ol*) *to KG if not None in (sl, pl, ol)*
16	**end for**

In this stage, each document will be represented by its metadata and the set of candidate RDF triples. This information is used to create restrictions and construct the final knowledge graph. In doing so, a heuristic mapping (**H**) between the annotated metadata labels and the ontology has to be created. This process has to be done manually for each domain based on the type of information the metadata holds. While this process can require a varying amount of human labour, the majority of the connections and rules are provided by domain-specific ontologies, which are considered a given. What is more, this process only takes place once, during the KG creation stage and licenses the KG creation process to have a standardized format.

Using the created candidate RDF triples from the previously described process, we can match their entities to the ontology. While approaches such as string-matching (strict or fuzzy) with the described classes, attributes and relationships can be successful assuming a similar vocabulary is used to describe the terms, in our approach, we propose a comparison of semantic representation of the candidate RDF triples' terms with the ontology entities.

With pre-trained LMs, which have been trained on millions of documents, we can extract contextualized representations of each word in their vocabulary, which are designed to semantically represent the words. As a result, even without a direct match, we can identify semantically close concepts and roles, making use of both the ontology labels and their descriptions (e.g., *rdfs:comment*), leading to better performance than that of strict or fuzzy matching when comparing terms. Using LMs has the inherited advantage of not requiring any annotated or hard to create resources such as lexicons (Inan, 2020). However, as LMs are very volatile to the pre-trained domain, the LM choice can have a significant performance impact in this process (Chalkidis et al., 2020; Lee et al., 2020). When using domain-specific pre-trained LMs is not possible, domain adaptation techniques should be considered (Rietzler et al., 2020).

Effectively, for each entity in the triple, which can be a sequence of tokens, we extract vectors $T \in \mathbb{R}^{1 \cdot d}$, which is the pooled contextualized representation of the tokens and compare them with the similarly extracted representations of the respective classes, attributes, and relationships in the ontology (*PLM*). Constructively, we are performing a probabilistic classification in the case of concepts and a mapping in the case of roles using the similarity scores (*SimilarityScoring*). For the comparison, we are using cosine similarity and maintain a list of the similarity scores of each triple entity with all the classes, attributes and relationships. From this step, we create a similarity score for each triple element with all the ontology elements.

With the use of a top-k subsampling process (*TopKSampling*), where **k** is a user-defined parameter that can vary based on the application domain, we select only the **k** closest classes, attributes and relations from the list (i.e., the ones with the highest simi-

larity score), leading to a three-dimensional vector of size **k**. For documents that are accompanied with metadata, based on this previously created heuristic mapping (**H**), we impose the respective ontology label to a span of text regardless of their score to other ontology labels and its relative part in the RDF triple.

For the remaining of the elements of the triple, starting with the semantically closest predicate candidate, we apply a restriction compliance checking that attempt to verify the predicate's relation with the subject and the object based on the domain-specific ontology, i.e., the domains/range compatibilities (*RestrictionsCompliance*). Specifically, for each predicate, we calculate the probability that the respective subject and object candidates are suitable options, based on the ontology restrictions. As a result, for each triple, we calculate a conditional probability based on the previously calculated probabilities and chose the one with the highest probability, if that is higher than a predefined threshold. The threshold (**t**) can vary based on the domain of application and the quality of the data and is subject to experimentation and fine-tuning.

If we cannot solve a candidate RDF triple to specific ontology labels, in the cases where the predicate is identified in the metadata, we endorse that label and assume that both subject and object adhere to the ontology's definitions (namely domain and range restrictions). As a result, candidate RDF triples are assigned with the relevant ontology label based on this process. If a triple fails to match with any label and its predicate is not in the metadata, we discard it.

Having obtained the triples and linked the entities from the triples to the elements of the selected ontology, we create a local Uniform Resource Identifier (URI) for each term in the triple to be used as a unique identifier.

This stage's evaluation is two-fold as both the individual components that comprise it and the created knowledge graph can be evaluated independently. Starting with the individual components, the heuristic mapping is a human-driven mapping process that only needs to be conducted once, and is not expected to have errors given that the semantic connections between the metadata are provided by the domain-specific ontology. For the term comparison, the performance is tied with the ability of the LM to accurately capture the data and is evaluated differently based on the LM type (Liu et al., 2020). However, given a large sampling number (**k**), the propagated error can be mitigated in the expense of computational time. Lastly, the restriction compliance checking cannot be evaluated directly and is only evaluated indirectly through the created knowledge graph as described below.

The created knowledge graph can also be used as a means of evaluating both Stage 1 and Stage 2. In order to evaluate the graph itself, the existence of graph from the original data is required. Given a pre-existing graph, graph sampling techniques can be applied to ensure static evaluation of efficiency and high-quality accuracy in the knowledge graph (Gao et al., 2018). The extracted sample can be given to human annotators in order to estimate the accuracy of the knowledge graph with metrics like precision, recall and F1-score but also the annotation cost. This process is performed iteratively in order to ensure the quality satisfies required expectations. Furthermore, in the case of semantic requirements in the system that will define a set of conditions, these rules can be applied with the SHACL Shapes Constraint Language to capture quality problems with specific constraints (Farid, 2020; Holger Knublauch & Dimitris Kontokostas, 2017). However, this requires manual annotation from domain experts of all the constraints which becomes infeasible in large datasets.

To that end, evaluation can be formulated as a ML task (Mihindukulasooriya, 2020). For identifying constraints, ML can be used to automatically construct SHACL Shapes by extracting different instance-based features and predicting cardinality and

range constraints. Furthermore, ML can help at predicting inconsistent mappings. This approach is based on the notion that if an entity has the same object value for two different predicates, then it is very probable that there is mapping inconsistency. Lastly, a classifier can be obtained to predict if an instance belongs to a specific class. These approaches provide a data-driven assessment of quality for the created knowledge graph.

Stage 3: Knowledge Graph enrichment

ALGORITHM 3 KNOWLEDGE GRAPH ENRICHMENT

Input:
$D \leftarrow document,$
$Algorithm1 \leftarrow Pre\text{-}processing\ and\ Information\ Extraction\ process$
$Algorithm2 \leftarrow Ontology\text{-}Constrained\ Knowledge\ Graph\ creation$
$KG \leftarrow Created\ knowledge\ graph$
$l \leftarrow fixed\ path\ length\ for\ co\text{-}occurrences\ count$
$t \leftarrow Compliance\ threshold$
Output:
Enriched Knowledge Graph (KG)

1	$Ent, W \leftarrow Training\ entity\ and\ word\ embeddings\ models\ using\ KG$
2	$Co \leftarrow CoOccurrenceCount(KG, l)$
3	$P(k_i|e) \leftarrow Normalize(Co,\ range=[0,1])$
4	**for** d in D **do:**
5	$\quad E \leftarrow NER(d)$ // Apply named entity recognition in document d
6	$\quad t_r, m \leftarrow d$ //
7	$\quad$ **for** e in E **do:** // Entity linking
8	$\quad\quad Con \leftarrow Context(e)$
9	$\quad\quad Con^* \leftarrow W(Con)$ // calculate word embeddings of the context of entity e
10	$\quad\quad C_e \leftarrow TopKCandidateSelection(e, P(k_i|e))$
11	$\quad\quad$ **for** k_e in C_e **do:**
12	$\quad\quad\quad k_e^* \leftarrow Ent(k_e)$
13	$\quad\quad\quad simScore \leftarrow SimilarityScore(k_e^*, Con^*)$
14	$\quad\quad$ **end for**
15	$\quad\quad k_{max} \leftarrow argmax(simScore)$ // link the entity with the highest similarity
16	$\quad$ **end for**
17	$\quad \{s, p, o\} \leftarrow Algorithm1(tr)$ // triple extraction with linked entities according to Algorithm 1
18	$\quad$ **for** (s^*_s, p_s, o^*_s) in $\{s, p, o\}$ **do:**
20	$\quad\quad p^*_s \leftarrow TopKSampling(k, p_s)$
21	$\quad\quad (s_l, p_l, o_l) \leftarrow Algorithm2.RestrictionsCompliance\ ((s^*_s, p^*_s, o^*_s), t)$
22	$\quad\quad$ **Add** (s_l, p_l, o_l) to KG if not None in (s_l, p_l, o_l)
23	$\quad$ **end for**
24	**end for**

Having obtained a large knowledge graph with a huge number of nodes and edges, for new documents, we can perform more advanced entity linking on an instance level based on the context. Therefore, we reuse the process described in the Pre-processing and Information Extraction section to extract triples and perform NER. However, for entity linking we follow the approach proposed by van Hulst et al. (Van Hulst et al., 2020), which performed state-of-the-art results with a 83.3% F1-score. The process is based on a general architecture for entity linking that consists of three steps: mention detection, candidate selection and entity disambiguation (Balog, 2018). Mention detection is already performed through the extracted triples and NER tool.

Since we have identified the mentioned entities, we want to select a subset of the top-k candidate entities to be linked, as it is intractable to check each entity pair in the whole knowledge graph (TopKCandidateSelection). The candidate entities are selected based on the entities with the higher prior conditional probability $p(k_e|e)$, where k_e is an entity in the knowledge graph and e is the mentioned entity in the document. To calculate this probability, we initially count co-occurrences between all entity pairs on a fixed path length in the knowledge graph. Subsequently, we normalize the values to create probability scores.

Entity disambiguation, which is the most important step of the entity linking process, is the process of linking a mention to one of the candidate entities in the graph. In this step, for each candidate entity, we calculate the similarity between the mentioned entity and the candidate entity $k_e^T \sum t_m, t_m \in C$, where k_e and t_m are the embeddings of the candidate entity and the word m of the context of the mentioned entity, respectively. The entity and word embeddings are trained according to the approach by Ganea & Hofmann (2017). While approaches for word sense disambiguation can also be leveraged for entity linking (Moro et al., 2014), we opted them out as entities can be disambiguated with knowledge graph information and would increase the complexity of the framework. If an entity cannot be linked due to a low similarity score but has relations with other linked entities in the text, we create a new entity in the graph. For linking predicates to the ontology, we reuse the methodology described in the ontology-constrained creation process to create semantic representations and apply restrictions compliance checking in order to validate the connections between the respective ontology entities.

Similar to stages 1 and 2, each component can be evaluated independently with its own set of evaluation criteria and metrics. The difference in this stage is that dynamic evaluation is performed in order to ensure incremental evaluation as the knowledge graph is continuously evolving. To that end, different sampling techniques like reservoir and stratified sampling can be leveraged for incremental evaluation (Gao et al., 2018).

Advanced searching functionality

The new framework extends the current functionalities of the document repository with the use of the knowledge graph for advanced searching. Given our heterogenous document collection, in order to retrieve documents, the users had to rely on strict or fuzzy term matching inside the document and in the metadata when these were available. While this functionality remains unchanged as it presents a trivial approach towards document retrieval, we increase the searching abilities of the enhanced repository with knowledge graph label searching using SPARQL queries.

For documents that previously had no metadata (1-star documents), this enables semantic searching to specific classes or attributes as defined in the domain-specific ontology. What is more, we can now search for a specific relation between the attributes or classes that are identified in the document, which was a previously unattainable functionality. As a result, we enable semantic searching, limiting the results of our queries to only the ones that contain relevant documents, some of which would previously be unretrievable or returned as a small subset of a vast collection of results. This enhanced functionality, when applied to open data repositories, increases the transparency and accessibility of the repository.

Case Study

We investigate the application and effects of Doc2KG on the DIAVGEIA's DMS. We reuse the European Union (EU) ISA^2 (Interoperability Solutions for European Public Administrations) Core Vocabularies, European Legislation Identifier (ELI), schema.org and Dublin Core Metadata, that are W3C recommendations, for ontology modelling related to the defined concepts on DIAVGEIA and showcase the implementation process of the Knowledge Graph for DIAVGEIA by using Doc2KG.

DIAVGEIA makes a perfect candidate for a case study due to the large document volume coming from 3,677 public authorities, while on average, around 28,000 decisions are uploaded every working day. Furthermore, it is an open document repository that includes basic metadata and a PDF document for each decision. Therefore, and as already discussed, it exhibits big data characteristics with high volume, velocity and ill-structured format.

DIAVGEIA

DIAVGEIA is, first and foremost, an openness project designed with the goal to increase transparency in government operations and provide enhanced accountability for citizens (transparency by design). The portal was initiated in October 2010, forcing all public administration bodies to publish their administrative acts and decisions in an open manner. Its name in Greek means transparency, denoting the goal for the implementation of this initiative. The documents are uniquely identified with the Internet Uploading Number (IUN) which is attached to each document. Based on the Greek law 4210/2013 of the Ministry of Administrative Reform and e-Governance, administrative decisions and acts are valid only after publication at the portal.

Administrative acts and decisions are an umbrella term for various types of decisions, like laws, acts, executive orders, circulars, budgets and expenses, instruments of appointment, calls for job positions, reserve lists, acts of acceptance of donations, acts of funding, public contract awarding and more public proceedings.

The administrative acts and decisions are uploaded in PDF format and include basic metadata information, including the protocol number of the document, date of publication, email of the publisher, subject of the document, type of act, organization of the signee, signee name, thematic area. Depending on the type of administrative act, additionally, there is analogous information provided on the portal's DMS. For example, when a decision is modified, it includes the IUN of the previous decision. When the act refers to expenses, it includes tax information of the institution and of the subcontractor, description of the subject, amount of the expense, etc. Similarly, if the act is about budget or funding, it includes information on the amount, the type of the budget and tax information of the institution receiving the funding. Finally, for administrative acts on contracts, it includes the tax information of the contractor.

Currently, DIAVGEIA allows citizens and public servants to perform two types of basic search functionalities: simple keyword search and advanced search by using multiple criteria based on the metadata information like date range or type of administrative act. Registered users (any citizen or public servant) also have the ability to provide feedback regarding errors in the published decisions. Furthermore, the portal allows users to visualize the organizational structure of a public institution and the number of published decisions by this body. Finally, any user can check the validity of any

uploaded PDF file that includes a UIN. Developers can access information on DI-AVGEIA by using the OpenDataAPI[3] that constitutes an endpoint for making requests and queries over the metadata.

Domain-specific ontology standards

In the case of DIAVGEIA, we first have to identify the specific requirements in order to tie the portal to the domain ontology standards. In our case, the requirements are described through the current OpenDataAPI, which provides access to the metadata available at the portal. Our work reuses and extends the DIAVGEIA ontology proposed by Beris and Koubarakis (2018). Although it consists of persistent URIs and reuses the ELI ontology for describing a decision as a LegalResource concept (ELI Task Force, 2016), it represents named entities like person and government institution without using existing EU standards. For this purpose, we focus on two main EU domain standards that match our use-case, the ISA[2] Core Vocabularies and the ELI ontology, both designed by the European Commission for describing important entities and legal resources. Furthermore, for properties that cannot be aligned with the ISA[2] Core Vocabularies, we consider the existing standards schema.org and Dublin Core that are W3C recommendations and provide general data types and properties to be reused.

Having identified all required components, we proceed with a heuristic alignment between the DIAVGEIA ontology and existing standards. This allows us to reuse the standards for our use case and also provides us with a detailed framework to build our knowledge graph. We represented the signer (*dvg:signer*) and present (*dvg:present*) concepts as object properties having as range the Person class from the Core Person Vocabulary. We also considered all the government institution information as part of the PublicOrganisation class from the Core Public Organisation Vocabulary. This allows richer information about entities with higher connectivity within the knowledge graph due to the existence of more properties related to the subclasses of the ISA core vocabulary. Furthermore, we map all *xsd:date* attributes to the date data type from the Dublin Core metadata (*dc:date*).

Error! Reference source not found. presents the UML of the extended version of the core of DIAVGEIA ontology. Gray color indicates the original classes and yellow color the extended ones.

Figure 2: UML model of the extended DIAVGEIA ontology. Gray color indicates the classes of DIAVGEIA ontology and the classes with yellow color are based on the ISA[2] Core Vocabularies.

Doc2KG on DIAVGEIA

The implementation of the Doc2KG Framework to DIAVGEIA's DMS will enrich it with further functionalities, significantly improving the portal's information transparency and accessibility over its current state. The current portal supports two types of users, government employees and citizens, and two access points, the web portal and the Application Programming Interface (API). Each user-type will be equally accommodated by the future changes to the portal, under different scenarios, regardless of access method.

Currently, government employees are the people responsible for populating the portal with new documents (Acts), and providing the metadata for each document. The citizens use DIAVGEIA to discover documents, searching through its database, using only the metadata. Specifically, government employees have access to upload data under their accounts, while citizens do not need an account. Furthermore, a citizen can have different needs depending on his role. As an example, a citizen can search through the DMS to identify specific government decisions, or s/he can search for a certain cause, e.g., to identify all spending acts of a government branch. Regardless of the roles, the privileges, functionalities and UI remain the same.

Due to the vast number of daily documents published in DIAVGEIA, the effort and costs to manually annotate them with metadata are unmanageable. Assuming, for example, 5 minutes average annotation time per document, it would have required 2000 person-years for the current collection of DIAVGEIA documents to be annotated. Automating this process would result in massive cost-saving benefits. Even in the optimistic

scenario of a reduced average annotation time per document, the overall cost would have been prohibitive. Moreover, and as discussed, this leads to ill-structured metadata that can cause some documents to be practically "hidden" in the repository. The implementation of Doc2KG will not only resolve these issues but will also further enhance the functionality of DIAVGEIA.

Doc2KG has the ability to automate the information extraction process of identifying all the previously required metadata information, thus reducing the workload of the government employees and allowing more time for them to focus on other aspects of their job. Moreover, by accurately identifying all the entities and their relations in a document, it will ensure that all documents will be adequately annotated, making them discoverable to everyone and resolving issues of missing metadata. This will not only affect future documents but also will enhance the current document collection. Furthermore, by linking the data through a knowledge graph, the retrieval process will also be improved as users will be able to search for specific entities and find all documents where that entity appears. What is more, using the knowledge graph, the document retrieval scope can be further limited to identify a specific type of connections between search terms using the graph's labelled vertices and edges, further refining searches.

As a direct effect, the functionality of DIAVGEIA will be both improved and extended and also become a powerful tool for both government employees and citizens alike.

For the government, DIAVGEIA could be utilized as both a cost-cutting mechanism and as a tool for an independent authority focused on anti-corruption. Doc2KG will instill the ability to identify all spending actions of a specific organization, making it easier to gain insight into its procured items, expenditure patterns and finally draw conclusions. For example, it will have the ability to refine the searches to filter transactions with specific vendors, simplifying the process of identifying biases and abnormal purchases.

For citizens, the previously described functionality will also hold, allowing Non-Governmental Organizations (NGOs), journalists and citizens to identify misconduct and hold the government accountable.

Finally, DIAVGEIA will be transformed into a tool for researchers and start-ups to acquire access to an extremely rich knowledge base for various use cases and needs (Goel, Kazemi, Brubaker, & Poupart, 2020; Zhang et al., 2020). Last, by providing access to its knowledge graph through its API, it will constitute a massive open-data benchmark corpus that will promote research in areas such as graph completion (Rosso, Yang, & Cudré-Mauroux, 2020), link prediction (Rossi, Barbosa, Firmani, Matinata, & Merialdo, 2021), and entity linking (Mulang' et al., 2020).

Discussion

In this paper, we proposed Doc2KG, a novel framework that handles both initial conversion of a DMS to a knowledge graph and supports the perpetual population of the created knowledge graph with new documents. This is done with a combination of NLP techniques to facilitate Information Extraction and constrained solving techniques for knowledge graph creation and manipulation. Furthermore, we exhaustively discussed the potential effects of implementing Doc2KG on DIAVGEIA with respect to accessibility, transparency and added functionality. To that, we created a heuristic mapping between the domain ontology standards suitable for the portal's DMS and its metadata scheme, which enables the framework to operate with DIAVGEIA. We postulate that the implementation of Doc2KG will bring significant functionality improvements in

both public and private sector organizations alike. Our design principle of re-using concepts (ontologies) as domain standards adds further value to previous research in the fields and highlights their importance.

The framework's performance is analogous to the performance of the individual components that comprise it. As a result, potential errors from the individual NLP extraction tools, inaccuracies during the restriction complied creation of the knowledge graph and miss-identifications of entities during knowledge graph enrichment will affect the final performance of the framework.

In our approach, we mitigate the propagated error from the NLP tools via restrictions compliance checking. By applying domain ontology standards, we discard semantically inaccurate RDF candidates, lowering the probability of false nodes being introduced to the knowledge graph. However, this is not true for false negatives (i.e., failure to identify entities of interest), as we will have no way to introduce them in the knowledge graph. During our knowledge graph creation approach, we only consider the most probable matches and the routine's performance is heavily based on the user-defined threshold. A very high threshold, meaning a low number of top candidates (**k**) can result in the elimination of a correct RDF triple, while a very low threshold will be computationally expensive. What is more, this parameter varies depending on the domain of application and the quality of the existing domain-specific ontology standards during the similarity score calculation process. Finally, miss-aligning entities found in novel text with entities in the knowledge graph will create inconsistencies to the final graph, degrading the abilities to ensue functionalities such as accurate document retrieval.

Further limitations apply depending on the domain of application and the available resources for it. Domains that have specialized vocabularies (e.g., medical, law, etc.) depend on specialized models that can identify entities of interest. Moreover, in cases such as DIAVGEIA, the scarcity of available language resources introduces further limitations. As a result, the performance of the respective NLP tools is based on the ability to train models for the respective language. In the case of DIAVGEIA, GreekBERT (Koutsikakis et al., 2020) provides a perfect resource for this domain along with Spacy's pre-trained NER and CP models. For the remaining tasks, utilizing multilingual models provide the desired functionality (Cruz et al., 2020; De Cao et al., 2021). Overall, multilingual models have shown significant improvement in the latest research (Baumann, 2019; Gromann, 2020).

In addition, there is no single performance measure approach that can capture the success of a created knowledge graph directly. Careful consideration needs to be taken on the selection of metrics to be used for evaluation based on the knowledge graph domain and the ontology. Quality measures and characteristics should be clearly defined to achieve a robust quality assessment. What is more, using ML can facilitate the fast and automated process of linked data evaluation without need of human intervention. However, it is highly dependent on the information provided in the extracted training features. It is important to have a rich linked data profile in the knowledge graph that will allow high quality features to train a ML classifier. Finally, approaches such as graph sampling and SHACL can be utilized to evaluate specific areas of the created graph, however significant attention needs to be taken on the amount of SHACL constraints as a high number of constraints could severely affect recall of the knowledge graph creation.

Conclusively, in order to create an accurate knowledge graph representation of a document collection with the versatility to handle all the domain requirements, domain ontology standards are required. These standards will ensure the correct structure of the classes, labels and relations between the created knowledge graph nodes. Similarly, the

performance is also heavily based on the existence of NLP resources in the form of either pre-trained models or training data to create models. In low-resource languages and specific domains, multilingual approaches and domain adaption techniques can serve as a substitute to provide the desired functionality (Baumann, 2019; Rietzler et al., 2020).

Apart from DIAVGEIA, Doc2KG can be applied to other large document repository collections, public and private. In the same domain, potential applications include the Federal Depository Library Program[4] (FDLP) in United States. Similarly, National Libraries, such as the National Depository Library in Finland, can use Doc2KG to increase their functionality and make their systems more accessible and manageable. Going further, our framework is robust and can be also used in the private sector, given the existence of domain standards for the respective company's sector.

Conclusion

In this paper, we proposed a framework that handles both the initial conversion of a DMS to a knowledge graph and supports the further enrichment of the created knowledge graph, which is achieved by reusing pre-existing ontology standards. More specifically, we proposed a combination of NLP tools to extract semantic information from documents and build RDF triples, a restriction compliance approach based on domain-specific ontology standards and a knowledge graph enrichment methodology to handle the addition and automatic indexing of new documents as they become available. Consequently, Doc2KG enables the implementation of intelligent features and the modernization of DMS, increasing their functionality.

By reusing pre-existing domain ontology standards, we built upon previous research, proving the value of such ontologies by domain experts. The implementation of the Doc2KG framework can be useful in both the private and public sectors. Currently, the domain-specific sampling number (k), required during the Ontology-constrained Knowledge Graph creating is manually defined as it heavily depends on the quality of the given ontology. As such, further research is required on effective methodologies of automatically selecting k candidates during the sampling process as well as evaluating its impact on the created knowledge graph.

References

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., & Auli, M. (2019). Cloze-driven Pretraining of Self-attention Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5360–5369. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1539

Balog, K. (2018). *Entity-Oriented Search*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-93935-3

Baumann, A. (2019). Multilingual Language Models for Named Entity Recognition in German and English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 21–27. Varna, Bulgaria: INCOMA Ltd.

---

[4] https://www.fdlp.gov/

https://doi.org/10.26615/issn.2603-2821.2019_004

Beris, T., & Koubarakis, M. (2018). Modeling and preserving Greek government decisions using semantic web technologies and permissionless blockchains. *European Semantic Web Conference*, *10843 LNCS*, 81–96. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-93417-4_6

Berners-lee, T. (2010). Linked Data - Design Issues. Retrieved June 3, 2021, from Design Issues website: https://www.w3.org/DesignIssues/LinkedData.html#fivestar

Bratsas, C., Chondrokostas, E., Koupidis, K., & Antoniou, I. (2021). The Use of National Strategic Reference Framework Data in Knowledge Graphs and Data Mining to Identify Red Flags. *Data*, *6*(1), 2. https://doi.org/10.3390/data6010002

Buscaldi, D., Dessì, D., Motta, E., Osborne, F., & Reforgiato Recupero, D. (2019). Mining Scholarly Publications for Scientific Knowledge Graph Construction. *European Semantic Web Conference*, *11762 LNCS*, 8–12. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-32327-1_2

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020, November 29). *LEGAL-BERT: The Muppets straight out of Law School*. 2898–2904. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chalkidis, I., Nikolaou, C., Soursos, P., & Koubarakis, M. (2017). Modeling and querying Greek legislation using semantic web technologies. *European Semantic Web Conference*, *10249 LNCS*, 591–606. Cham, Switzerland: Springer Verlag. https://doi.org/10.1007/978-3-319-58068-5_36

Cruz, A. F., Rocha, G., & Cardoso, H. L. (2020). Coreference resolution: Toward end-to-end and cross-lingual systems. *Information (Switzerland)*, *11*(2), 74. https://doi.org/10.3390/info11020074

De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., … Petroni, F. (2021). *Multilingual Autoregressive Entity Linking*. Retrieved from http://arxiv.org/abs/2103.12528

de Juana-Espinosa, S., & Luján-Mora, S. (2019). Open government data portals in the European Union: Considerations, development, and expectations. *Technological Forecasting and Social Change*, *149*, 119769. https://doi.org/10.1016/j.techfore.2019.119769

ELI Task Force. (2016). *ELI implementation methodology - Good practices and guidelines*. Publications Office of the EU. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/a8367080-bdad-11e5-bfdd-01aa75ed71a1/language-en

Farid, M. (2020). *Extracting and Cleaning RDF Data*. University of Waterloo.

Ganea, O.-E., & Hofmann, T. (2017). Deep Joint Entity Disambiguation with Local Neural Attention. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2619–2629. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1277

Gao, J., Li, X., Xu, Y. E., Sisman, B., Dong, X. L., & Yang, J. (2018). Efficient knowledge graph accuracy evaluation. *Proceedings of the VLDB Endowment*, *12*(11), 1679–1691. VLDB Endowment. https://doi.org/10.14778/3342263.3342642

Garijo, D., & Poveda-Villalón, M. (2020). *Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web*. Retrieved from http://arxiv.org/abs/2003.13084

Goel, R., Kazemi, S. M., Brubaker, M., & Poupart, P. (2020). Diachronic embedding

for temporal knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 3988–3995.

Gritzalis, A., Tsohou, A., & Lambrinoudakis, C. (2017). Transparency-enabling systems for open governance: Their impact on citizens' trust and the role of information privacy. *Communications in Computer and Information Science*, *792*, 47–63. Springer Verlag. https://doi.org/10.1007/978-3-319-71117-1_4

Gromann, D. (2020). Neural language models for the multilingual, transcultural, and multimodal Semantic Web. *Semantic Web*, *11*(1), 29–39. https://doi.org/10.3233/SW-190373

Gutierrez, C., & Sequeda, J. F. (2021). Knowledge graphs. *Communications of the ACM*, *64*(3), 96–104. https://doi.org/10.1145/3418294

Hasnain, A., & Rebholz-Schuhmann, D. (2018). Assessing FAIR data principles against the 5-star open data principles. *European Semantic Web Conference*, *11155 LNCS*, 469–477. Cham: Springer. https://doi.org/10.1007/978-3-319-98192-5_60

Holger Knublauch, & Dimitris Kontokostas. (2017, July 20). Shapes Constraint Language (SHACL). Retrieved June 5, 2021, from W3C Recommendation website: https://www.w3.org/TR/shacl/

Hu, B., & Svensson, G. (2010). A case study of linked enterprise data. *International Semantic Web Conference*, *6497 LNCS*(PART 2), 129–144. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17749-1_9

Inan, E. (2020). SimiT: A Text Similarity Method Using Lexicon and Dependency Representations. *New Generation Computing*, *38*(3), 509–530. https://doi.org/10.1007/s00354-020-00099-8

Kertkeidkachorn, N., & Ichise, R. (2018). An automatic knowledge graph creation framework from natural language text. *IEICE Transactions on Information and Systems*, *E101D*(1), 90–98. Institute of Electronics, Information and Communication, Engineers, IEICE. https://doi.org/10.1587/transinf.2017SWP0006

Khosla, S., & Rose, C. (2020). Using Type Information to Improve Entity Coreference Resolution. *Proceedings of the First Workshop on Computational Approaches to Discourse*, 20–31. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.codi-1.3

Kingsbury, P., & Palmer, M. (2002). From TreeBank to PropBank. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). GREEK-BERT: The Greeks Visiting Sesame Street. *11th Hellenic Conference on Artificial Intelligence*, 110–117. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3411408.3411440

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, J., Sun, A., Han, J., & Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2020.2981314

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). *A Survey on Contextual Embeddings*. Retrieved from http://arxiv.org/abs/2003.07278

Martinez-Rodriguez, J. L., Lopez-Arevalo, I., & Rios-Alvarado, A. B. (2018). OpenIE-

based approach for Knowledge Graph construction from text. *Expert Systems with Applications*, *113*, 339–355. https://doi.org/10.1016/j.eswa.2018.07.017

Matheus, R., & Janssen, M. (2020). A Systematic Literature Study to Unravel Transparency Enabled by Open Government Data: The Window Theory. *Public Performance & Management Review*, *43*(3), 503–534. https://doi.org/10.1080/15309576.2019.1691025

Mihindukulasooriya, N. (2020). *A Framework for Linked Data Quality based on Data Profiling and RDF Shape Induction*. PhD Thesis, Polytechnic University of Madrid.

Mohamed, M., Pillutla, S., & Tomasi, S. (2020). Extraction of knowledge from open government data: The knowledge iterative value network framework. *VINE Journal of Information and Knowledge Management Systems*, *50*(3), 495–511. https://doi.org/10.1108/VJIKMS-05-2019-0065

Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*, *2*, 231–244. https://doi.org/10.1162/tacl_a_00179

Mulang', I. O., Singh, K., Vyas, A., Shekarpour, S., Vidal, M. E., & Auer, S. (2020). Encoding Knowledge Graph Entity Aliases in Attentive Neural Network for Wikidata Entity Linking. *International Conference on Web Information Systems Engineering*, *12342 LNCS*, 328–342. Cham, Switzerland: Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-62005-9_24

OECD. (2011). The call for innovative and open government: An overview of country initiatives. In *The Call for Innovative and Open Government: An Overview of Country Initiatives* (Vol. 9789264107052). Organisation for Economic Cooperation and Development (OECD). https://doi.org/10.1787/9789264107052-en

Papay, S., Klinger, R., & Padó, S. (2021). Constraining Linear-chain CRFs to Regular Languages. *CoRR*, *abs/2106.0*. Retrieved from https://arxiv.org/abs/2106.07306

Penteado, B. E., Maldonado, C., & Isotani, S. (2021). Methodologies for publishing linked open government data on the Web: a systematic mapping and a unified process model. *Semantic Web Journal*.

Raiman, J., & Raiman, O. (2018). DeepType: Multilingual Entity Linking by Neural Type System Evolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved from www.aaai.org

Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or Get Left Behind: Domain Adaptation through {BERT} Language Model Finetuning for Aspect-Target Sentiment Classification. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4933–4941. Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.607

Rossanez, A., Dos Reis, J. C., Torres, R. da S., & de Ribaupierre, H. (2020). KGen: a knowledge graph generator from biomedical scientific literature. *BMC Medical Informatics and Decision Making*, *20*(S4), 314. https://doi.org/10.1186/s12911-020-01341-5

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(2), 1–49.

Rosso, P., Yang, D., & Cudré-Mauroux, P. (2020). Beyond Triplets: Hyper-Relational Knowledge Graph Embedding for Link Prediction. *The Web Conference 2020 -*

Proceedings of the World Wide Web Conference, WWW 2020, 1885–1896. New York, NY, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/3366423.3380257

Savvas, I., & Bassiliades, N. (2009). A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration. *Expert Systems with Applications*, *36*(3), 4467–4478. https://doi.org/10.1016/j.eswa.2008.05.022

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. Retrieved from https://repository.upenn.edu/dissertations/AAI3179808

Stylianou, N., & Vlahavas, I. (2021). A neural Entity Coreference Resolution review. *Expert Systems with Applications*, *168*, 114466. https://doi.org/10.1016/J.ESWA.2020.114466

Vafopoulos, M., Meimaris, M., Papantoniou, A., Anagnostopoulos, I., Xidias, I., Alexiou, G., … Loumos, V. (2012). Intelligent and semantic real-time process of the Greek LOD for enhancing citizen awareness in public expenditures. *International Conference on Web Information Systems Engineering*, *7651 LNCS*, 808–811. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35063-4_72

Van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., & De Vries, A. P. (2020). REL: An Entity Linker Standing on the Shoulders of Giants. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2197–2200. New York, NY, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/3397271.3401416

Veljković, N., Milić, P., Stoimenov, L., & Kuk, K. (2020). Production of linked government datasets using enhanced lire architecture. *Computer Science and Information Systems*, *17*(2), 599–617. https://doi.org/10.2298/CSIS190420001M

Yu, H., Li, H., Mao, D., & Cai, Q. (2020). A domain knowledge graph construction method based on Wikipedia. *Journal of Information Science*, 016555152093251. https://doi.org/10.1177/0165551520932510

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), 63–93. IOS Press. https://doi.org/10.3233/SW-150175

Zhang, C., Yao, H., Huang, C., Jiang, M., Li, Z., & Chawla, N. V. (2020). Few-shot knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 3041–3048.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17–29. https://doi.org/10.1016/j.giq.2013.04.003