KINDEX – Automatic Subject Indexing with Knowledge Graph Services

Lisa Wenige *, Claus Stadler, Simon Bin, Lorenz Bühmann, Kurt Junghanns and Michael Martin
Institute for Applied Informatics, University Leipzig, Germany
E-mail: wenige@infai.org

Abstract. Automatic subject indexing has been a longstanding goal of digital curators to facilitate effective retrieval access to large collections of both online and offline information resources. Controlled vocabularies are often used for this purpose, as they standardise annotation practices and help users to navigate online resources through following interlinked topical concepts. However, to this date, the assignment of suitable text annotations from a controlled vocabulary is still largely done manually, or at most (semi-)automatically, even though effective machine learning tools are already in place. This is because existing procedures require a sufficient amount of training data and they have to be adapted to each vocabulary, language and application domain anew. Against the background of tight budgets and missing IT personnel in cultural heritage as well as research infrastructure institutions, adoption of automatic subject annotation tools is hindered, while manual assignment of index terms is an even greater burden on the available financial resources. In this paper, we argue that there is a third solution to subject indexing which harnesses cross-domain knowledge graphs (i.e., DBpedia and Wikidata) to facilitate cost-effective automatic descriptor assignments that can be done without any algorithm tuning and training corpora. Our KINDEX approach fuses distributed knowledge graph information from different sources. Experimental evaluation shows that the approach achieves good accuracy scores by exploiting correspondence links of publicly available knowledge graphs.

Keywords: Automatic Indexing, Named Entity Recognition, Key-phrase Extraction, Authority File

1. Introduction

Cultural heritage institutions have a need to organise and facilitate access to large collections of printed and online materials. In this regard, topic schemes are important tools that facilitate effective content searches when navigating the ever growing amount of offline and online data. These schemes are often referred to as *knowledge organisation systems* (KOS) or *controlled vocabularies*. They mediate between user information needs and the data quantity that can be found in both digital and analogue collections. In KOS, topics are represented as uniquely identified concepts, that can be characterised by different synonyms/labels [1]. Thus, it is made sure that users with identical information needs find items that are aligned with their interests, even though they conduct searches using different keywords.

*Corresponding author. E-mail: wenige@infai.org.

The process of assigning KOS keywords to resources is called subject indexing and has a long tradition in cultural heritage institutions. It is not a trivial task, since sets of descriptors have to be identified that best capture the content of library, museum or archival artifacts. In this context, Hutchins coined the term *aboutness* to underline that subject indexing is predominantly concerned with identifying an item's topics through condensed keyword descriptions [2].
In information providing institutions, trained professionals manually inspect and assign suitable concepts. Even with the growing amount of data this practice has not been replaced by automatic approaches. This is an interesting fact, considering that there are already algorithms in place that can extract and assign suitable KOS concepts from digital text. Automatic keyphrase extraction approaches can be grouped into two categories, namely ML-enabled *associative* and *lexical indexing*. The former learns a model from training data by identifying frequently occurring associations

between n-grams from input texts with the corresponding KOS descriptors. The *lexical approach* matches text snippets from the to-be-described texts with the labels of controlled vocabularies thus assigning the appropriate descriptor term by taking advantage of the synonymous labels being prevalent for each KOS concept [3].

However, even though these approaches have yielded good accuracy scores and in some cases even achieved results comparable to human indexers [4], their application is not as widespread in both cultural heritage and research infrastructure institutions as one would expect given their high performance.

The reasons for this phenomenon are manifold: First and foremost, a considerable amount of preprocessing and training effort is required to adapt the existing machine learning approaches to the training corpora (e.g., language and domain specificities) as well as the vocabulary. Besides being dependent on the availability of a large enough training corpus of annotated items, this requires both expertise in programming and NLP methods and sufficient time resources. However, many information providing institutions operate on tight budgets and cannot employ additional staff for these tasks [5]. On the other hand, manual indexing as it is still practised, is time consuming as well. Hence, a (semi-)automatic indexing tool would spare resources that could be used to speed up cataloguing routines and would lead to more items being sufficiently indexed which would in turn benefit retrieval interfaces in digital libraries and data portals. It would also enable collections that are currently not linked to any controlled vocabulary to be integrated with the distributed knowledge infrastructure of the web of data thus facilitating integration of related information resources.

For illustration purposes, consider the following example: An open data portal of a governmental agency frequently publishes data sets that might be of interest to its citizens. Data sets are described with a short abstract. An indexing tool being able to assign subject descriptors of a widespread KOS would enable users to inspect publications from a library catalogue applying the same concept annotations.

Given the high prevalence of bibliographic information on the web of data, this appears to be a promising research direction both in terms of data integration and semantics-aware retrieval [6]. Against the background of the aforementioned opportunities of the web of data and the difficulties of ML-based annotating, we present a novel method and architecture for automated subject indexing. Rather than fine-tuning existing ML and natural language processing (NLP) approaches for a given thesaurus and text collection. We present the KINDEX approach that leverages existing general purpose annotation tools, such as DBpedia Spotlight [7] in conjunction with KOS correspondence links of the data web (e.g. `owl:sameAs`, `skos:exactMatch`) in order to provide relevant keyword suggestions to information professionals. The benefits of our approach are as follows:

- **Minimum training and preprocessing effort**: The approach requires a minimum of resources from the side of cultural heritage and research infrastructure institutions as compared to state-of-the-art automatic keyword indexing systems (see Sect. 2). In KINDEX, this is realised through a coupling of existing knowledge graph web services in a unified annotation workflow.
- **Domain Independence**: KINDEX provides annotations for many potential application domains (see Sect. 3 and 4). It achieves domain independence through the following characteristics
    - *KOS independence*: The system does not need to be trained for a particular controlled vocabulary. The sole precondition is that the KOS system, from which keywords are to be generated, is published on the web of data and either linked to DBpedia or Wikidata.
    - *Multilingualism*: Annotations can be provided for various languages of the input text.
    - *Corpus Invariance*: The method does not require a training corpus of annotated texts in order to yield keyword recommendations. Hence, concept suggestions can be made for collections that are currently void of any index terms from a controlled vocabulary
- **Decent quality of accuracy scores**: Evaluation results from two real world usage scenario have shown the viability of the approach in terms of accuracy scores (see Sect. 5).

2. Related Work

Matching free text keywords is the number one retrieval strategy of today's search engines. However, meaningful KOS-based annotation of digital and analogue artefacts is still important for collections that serve highly specialised information needs. Several studies in the (digital) library context have shown that users were better able to find annotated documents as opposed to content without any descriptors [8, 9]. This is one of the reasons why subject indexing is a typical part of the cataloguing procedure to this date. It is also still - apart from a few exceptions - typically manually carried out by domain experts [10, 11].

Despite the practice of manual annotation, the task of automatically identifying content from text has been extensively studied under the terms *named entity recognition* and *information extraction*. Starting from hand-crafted rules, the field has moved toward applying more advanced machine learning approaches, such as support vector machines and decision trees [12]. With the advent of Web 2.0 applications and the popularity of both social bookmarking and collaborative tagging, a plethora of keyword recommendation approaches was introduced in the literature. These approaches are often quite similar to the field of automated indexing with the sole difference that in regular subject indexing, terms are part of a knowledge organisation system, while social tagging is mostly done with free text keywords [13–15]. Among the KOS-based indexing approaches, one can distinguish between ML-enabled *lexical* and *associative indexing*. The former matches n-grams from text input with the lexical entries of a controlled vocabulary. With these approaches, machine learning techniques are applied to identify how features among the characteristics of the input sequence (e.g., term-document frequency, keyphraseness or text position) should be weighted in order to make high quality predictions. Among the most prominent approaches of *lexical matching* for thesaurus-based indexing are the software applications KEA and its successor MAUI [16]. KEA has been tested on a collection of agricultural documents with the AGROVOC thesaurus and MAUI was run on Wikipedia articles. For Wikipedia articles, the authors showed that F1 scores of almost 50% are possible when automatically assigning concepts to documents [4, 17]. In contrast to *lexical indexing*, *associative indexing* learns the likelihood of associating the extracted text snippets with a corresponding index term. For both methods, a model has to be learnt from a training corpus. In the case of *associative indexing*, the document collection has to be even larger to achieve good results. This is because the method can only identify concept terms once they occur in the training data.

For instance, successful application of ML-based *associative indexing* methods was demonstrated on text collections from agricultural, medical and computer science research showing that learning association models can provide high quality automated subject assignments [18–20]. In addition to that, Toepfer et al. have shown that it is even more efficient to fuse the two approaches of *lexical* and *associative indexing* by unifying the sets of subject descriptors that were identified by the two methods. They evaluated the approach on a collection of manually annotated economics texts by testing the systems' performance regarding its ability to assign concepts from the STW Thesaurus for Economics [3]. The evaluations showed that the fusion approach significantly boosted F1 scores as compared to a sole application of either *lexical* or *associative* indexing. Further, the authors demonstrated that it is possible to achieve fairly good prediction results (with F1 scores between 30% and 40%), even though the model was exclusively trained on short texts (i.e., the titles and free-text keywords of publications). However, the above described approaches each learn models that are only applicable to a single document collection. They are therefore dependent on the idiosyncrasies of the input texts and the thesaurus that is used in that particular application domain. Hence, there is a considerable effort involved when adapting the available approaches for other resource collections and vocabularies. This is a problem, given the limited IT personnel that might not be available for algorithm fine-tuning, beside the already consuming tasks of handling day-to-day operational IT services in cultural heritage institutions [5]. Given the exponential growth of electronic resources on the one hand and the commitment of libraries and archives to provide suitable entry points for these resources on the other hand, (semi-)automated methods for subject indexing are desperately needed. The German National Library can serve as an example for this development: Its legal mandate was expanded to the collection of electronic resources being released by German publishers or on topics relevant to Germany. Thus, the library had to deal with a quickly growing amount of publications that could not be handled with the traditional procedures of manual indexing. Existing metadata from the publishers can also not be used since they often do not provide topic-specific metadata for their resources. Because of this situation, an automatic indexing strategy was implemented as a production system in the national library. The automatic strategy has proven to be feasible with mostly high quality assignments of index terms [10, 11].

However, this example of automatic indexing in an operating environment can be considered a rare exception. The majority of machine-based indexing approaches is tested under laboratory conditions, often without a follow-up productive implementation of the system [21, 22]. The low adoption rates indicate that existing automatic annotation tools are predominantly domain-specific and can not be easily transferred to other application scenarios. Against this background, it seems appropriate to investigate suitable alternative approaches for automated annotation, such as leveraging openly available data sources for this task.

The cultural heritage sector has long been an active advocate of data publishing as it has generated high quality interlinked and structured data for decades. Thus, it was a natural inclination for these institutions to join the Linked (Open) Data movement to redistribute the tax-funded collection of catalogue data to a wider public. To this date, the activities toward an open knowledge graph infrastructure of library data are predominantly concerned with the publication of catalogue data and controlled vocabularies [23]. In the context of automatic subject indexing, the available KOS vocabularies are especially relevant. For more than a decade, many of the existing thesauri have been made publicly available in SKOS format [5, 24]. Since 2009, the vocabulary *Simple Knowledge Organisation System* (SKOS) is a W3C recommendation. It offers a schema language to describe knowledge organisation systems in such a way that they can be published, exchanged and interlinked on the web of data. SKOS provides expressions to uniquely identify descriptors (i.e., `skos:Concepts`) declare hierarchical relationships (`skos:broader`) as well as cross-concordances/ identity links (i.e. `skos:exactMatch`) [1]. On the forefront of the SKOS-related activities was the publication of library authority files. Authority files standardise bibliographic control as they uniquely identify and describe persons, organisations, and subject descriptors. Among them were the Library of Congress Subject Headings (LCSH), the Integrated Authority File (GND) of the German National Library and the Virtual Integrated Authority File (VIAF) [25, 26]. Started as a joint effort between the Library of Congress and the German National Library, the VIAF registry has been joined by more than 50 institutions from the cultural heritage sector by now.[1] On the other hand, there are also many domain-specific vocabularies that have been made publicly available in RDF format. Examples are the AGROVOC thesaurus published by the Food and Agriculture Organisation of the United Nations (FAO) [27] and the STW Thesaurus for Economics by the Leibniz Information Centre for Economics [28]. The Basel Register of Thesauri, Ontologies and Classifications (BARTOC) lists hundreds of knowledge organisation systems from various fields, such as life sciences, law or history [29].[2]

Given the wide availability of SKOS vocabularies, which are often densely interlinked both with each other as well as the web of data, it seems promising to investigate whether these links can be leveraged for automatic subject indexing. For instance, Kempf et al. pointed out that a mixed-methods strategy combining manual and (semi-)automatic indexing in conjunction with identity links has great potential to increase cataloguing efficiency [30]. However, in the context of enhanced library services, SKOS vocabularies have been rarely taken advantage of. Only a few papers studied the effect of SKOS relations in order to improve retrieval systems [24, 31, 32]. Empirical investigations into the potential and effects of cross-concordances to leverage automatic subject indexing have not been conducted so far. This might be a hindrance for the introduction of automatic indexing tools in information providing institutions on a broader scale, since ML-based indexing methods require a considerable amount of data preprocessing, algorithm fine-tuning and are crucially dependent on a training corpus with a sufficient amount of labelled data; i.e., subjects that have already been manually assigned to a multitude of publications. Thus ML-based subject indexing is largely domain dependent and will be in many cases not an option due to the missing required high quality data sources.

However, in the web of data, there are cross-domain indexing tools available that can be applied in conjunction with identity links in order to facilitate (semi-) automatic subject indexing. In the remainder, we will investigate whether it is possible to leverage identity links with named entity recognition facilitated by DBpedia Spotlight in order to offer an alternative route for subject indexing, in cases where either the data or the available personnel resources do not allow for an ML solution.

3. Use Cases

3.1. Use Case Selection

We showcase the KINDEX approach in the context of two different use cases. *Use Case 1* is characterised

[1] https://viaf.org

[2] https://bartoc.org

by the fact that the relevant test collection has no annotations and thus belongs to the potential scenarios of (semi-)automatic subject indexing where an ML-based approach is not feasible due to missing training data. In contrast to that, *Use Case 2* is associated with a data collection that is comprised of a considerable amount of training data making it a natural test bed for evaluations regarding the performance of the identity-based approach in comparison to ML-enabled automatic subject indexing.

3.2. Use Case 1: mCLOUD - LIMBO

The mCLOUD platform is an open data portal that is maintained by the German Federal Ministry of Transport and Digital Infrastructure. It currently registers more than 1,500 traffic-related data sets (e.g., rails, roads or waterways) as well as climate and weather data. Data set owners are either the ministry or associated public agencies. The goal of the mCLOUD portal is to support data-driven research and development projects on unprecedented navigational services, smart travel and route planning as well as novel approaches towards precise weather forecasting.[3] The ministry also supports the research project *Linked Data Services for Mobility (LIMBO)*. LIMBO is concerned with semantically describing and integrating the mCLOUD data sets with the web of data in order to facilitate improved retrieval access to these resources. In the project, a metadata-catalogue in DCAT/DATAID format has already been crawled from the mCLOUD's publicly available web pages [33]. A typical record is comprised of the data set's title, its license and issue date, the publishing agency as well as a short description in text format (predominantly in German).[4] What these entries are missing, however, is a structured semantic description of the data set's content that integrates well with other resources on the web of data. For instance, it would be nice to be able to retrieve matching publications from the catalogue of the German National Library for each data set. Hence, a subject indexing approach that automatically assigns GND descriptors is required for this purpose.

3.3. Use Case 2: Econstor LOD

The second use case example concerns the Linked Open Data (LOD) collection of the Econstor repository which is one of the largest Open Access servers in the field of Economics. The German National Library of Economics published an excerpt of this collection in RDF format thus making more than 180k papers with predominantly German and English text descriptions available to a wider public [34]. A large fraction of the publications that are contained in the data set has subject annotations from the STW thesaurus. Due to the large number of labelled data, Econstor LOD is a natural candidate corpus for testing machine learning methods with regard to subject indexing. Hence, the collection has already been used in an empirical investigation into automatic indexing methods by Toepfer et al. [3]. Results of the authors' findings can serve as baseline indicators for the feasibility of subject indexing methods that harness identity links in comparison to ML-based approaches.

4. The KINDEX approach

4.1. Architecture & Workflow

Fig. 1 shows the architecture of the KINDEX approach. It can be implemented as a lightweight command-line script which harnesses existing knowledge graph technologies and web services by combining straightforward `HTTP` and `SPARQL` requests as well as `JSON` processing operations.

For this purpose, it relies on a running instance of DBpedia Spotlight [7] as well as mappings from DBpedia [35] and Wikidata [36] to Knowledge Organisation Systems. The indexing process starts with text snippets (e.g. a title and/or description of a publication, image or data set). Prior to the annotation it is often useful to apply a *blacklist filter* that suppresses annotations for sequences that are generally known to lead to faulty keywords (such as two-character sequences). After blacklist filtering, text snippets are fed into the DBpedia Spotlight web service that is tailored to a particular language. Currently, there exist 11 indices for DBpedia Spotlight [5] which facilitate multilingual annotations.
Imagine you would want to index a publication from Econstor with STW descriptors. An example publica-

[3]https://www.mcloud.de/

[4]https://gitlab.com/limbo-project/metadata-catalog/blob/master/catalog.all.ttl

[5]https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/faq
This list can possibly be extended to comprise the entirety of the 15 available DBpedia chapters https://wiki.dbpedia.org/services-resources/datasets/dbpedia-datasets

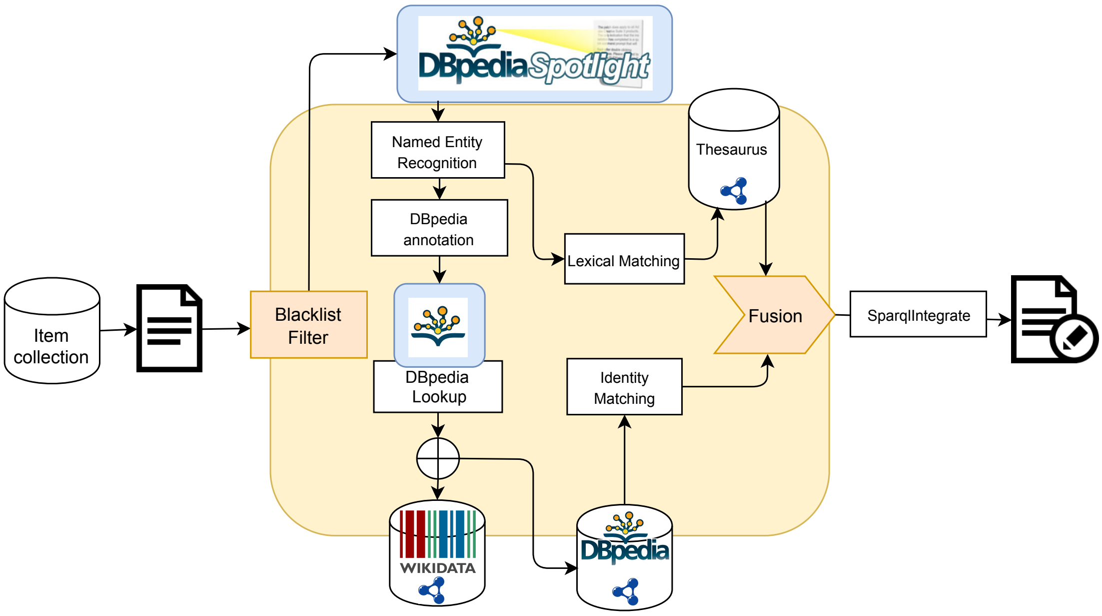

Figure 1. Conceptual overview of the KINDEX approach

tion that textually describes an economics publication on trade policy in English is shown in Listing 1 [34].

The title is sent to an English DBpedia Spotlight instance which performs *named entity recognition and disambiguation* and returns the annotated text as is shown in Fig. 2. The respective result file is comprised of the annotated text, its surface form (e.g., taxation) and the corresponding *DBpedia annotation* as a URI (e.g., http://dbpedia.org/resource/Tax). For each of the identified entities, a custom workflow is then invoked which combines both lexical as well as identity matching. In this particular example, the most suitable strategy is to first match the surface form as it was determined by the spotlight index with the STW thesaurus' preferred or alternative labels. In case no descriptors can be found, the engine leverages the identity links (i.e., `skos:exactMatch` or `owl:sameAs`) that are present in the web of data starting with the DBpedia URI. For instance, there exist cross-concordance links between the STW thesaurus and the GND, the DBpedia and Wikidata. For each of these knowledge graphs or thesauri, the KINDEX tool tests if there exist any identity links to the STW thesaurus that can be utilised for annotation. In this context, correspondences can be determined by different means depending on the quality and quantity of the existing mappings. The following lookup strategies are possible:

– *DBpedia Lookup*: The same-thing lookup service has been developed as part of the fusion of knowledge graphs from different DBpedia chapters as well as Wikidata with the FlexiFusion approach presented by Frey et al. The web API[6] serves as a registry to resolve identity links to manage resources of the largest publicly available cross-domain knowledge graphs. For instance, for a given language-specific DBpedia URI the web API returns the corresponding Wikidata and DBpedia URIs that are linked to the input URI via the `owl:sameAs` property. Even though, the public DBpedia SPARQL endpoint contains some of these identity links, the same-thing lookup service represents the most comprehensive mapping registry to identify correspondence links between DBpedia and Wikidata [37].
– *DBpedia*: The public DBpedia SPARQL endpoint is also offering other mapping links for identity resolution to thesauri, such as the GND or BBC Things. When hosting a keyword indexing service in the own institution, these links can be inserted into a local triple store or – in cases of small to medium-sized mapping collections – they can even be accessed through querying an in-memory

[6]https://global.dbpedia.org/same-thing/lookup/

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix econstor: <http://linkeddata.econstor.eu/beta/resource/publications/> .
@prefix econ-auth: <http://linkeddata.econstor.eu/beta/resource/authors/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

econstor:62535 a swc:Paper, sioc:Item, foaf:Document;
   dc:title "Business taxation and wages: Evidence from individual panel data"^^
        xsd:string;
   dc:creator econ-auth:38331874, econ-auth:38331875; econ-auth:38331876;
   dc:issued "2012"^^xsd:gYear;
   dc:language "eng"^^xsd:string .
```

Listing 1 Example bibliographic description in RDF from Econstor

Figure 2. DBpedia Spotlight results

model by using the tool SparqlIntegrate (see Section 4.2).

- *Wikidata* In cases, where DBpedia does not provide the required identity link, additional mappings can be determined from Wikidata, which contains mappings to a multitude of thesauri, such as the GND, VIAF, the Library of Congress authority files or MeSH. The advantage of utilising Wikidata for identity resolution is that there exists a separate property for each target thesaurus. For instance, the mappings to the GND descriptors are accessible via the property <http://www.wikidata.org/prop/direct-normalized/P227>, whereas mappings to VIAF can be determined from <https://www.wikidata.org/wiki/Property:P214> thus enabling faster access to the target KOS. As is the case with using DBpedia links, the KOS-specific mappings are best loaded into a local triple store or into an in-memory model.
- *Additional mappings* There might be cases in which there neither exist sufficient mappings in the DBpedia nor in Wikidata to the required target KOS. In these cases, it is often beneficial to utilise a larger mapping collection as an intermediate step. Thus the matching of descriptors is enabled through traversing identity paths. For instance, even though DBpedia contains mappings to the STW thesaurus, there currently exist only 915 `skos:exactMatch` links between the two knowledge graphs.[7] In contrast to that, the existing identity links between DBpedia and the GND (>80,000 `owl:sameAs` links) as well as GND and the STW (>4,800 `skos:exactMatch` links) is sufficiently higher[8]. Thus, the chances of finding an identity link by navigating from DBpedia to the GND and then to the STW is even higher then directly navigating from DBpedia to the STW.

Hence, link-enabled keyword indexing is based on trying to find the descriptor from the target KOS, which matches the DBpedia URI as identified by Spotlight. In the given example, although the resource http://dbpedia.org/resource/Tax is not linked to a Wikidata resource, the corresponding STW descriptor (http://zbw.eu/stw/descriptor/11547-6) can be determined from following the identity links that exist between DBpedia and the GND (e.g., via the public DBpedia SPARQL endpoint matching the property `owl:sameAs`).

Afterward, the cross-concordance that exist between the GND and the STW can be obained by querying the STW-to-GND mapping `stw_gnd_mapping.ttl`.[9]

[7] http://zbw.eu/stw/versions/latest/mapping/dbpedia/about.de.html
[8] https://lod-cloud.net/dataset/dnb-gemeinsame-normdatei
[9] http://zbw.eu/stw/version/latest/download/about.de.html

from an in-memory model that is accessed with the help of SparqlIntegrate[10].

The same procedure is carried out for each of the Spotlight annotations. It stops as soon as the respective STW descriptor is found. It is assumed that through the combination of lexical as well as identity matching, more relevant high quality keywords can be identified for subject indexing. Once the set of relevant KOS descriptors has been obtained, they can be assigned to a metadata catalogue in RDF format by applying a SPARQL UPDATE command that is customised with the command-line tool SparqlIntegrate (see Sect. 4.2).

The execution of the UPDATE request adds the triple statements of Listing 2 to the publication catalogue thus enabling a seamless integration of keyword annotation with existing bibliographic records. To ensure that only relevant keywords are added to the catalogue, the KINDEX approach can be implemented as an interactive command-line script. Thus, the tool asks the subject indexer whether a keyword is relevant for a particular publication, each time before a subject is added to the catalogue. Listing 2 shows the automatically generated and manually verified keywords for the Econstor publication.

The automatic annotation of data set descriptions in the context of the LIMBO project (*Use Case 1*) only slightly differs from the previously outlined processing steps: Subject descriptors from the GND are determined by matching the text snippets of a data set description from the LIMBO metadata-catalogue with DBpedia URIs. DBpedia URIs are then send to either Wikidata or DBpedia in order to identify identity links to the GND thesaurus. Alternatively, relevant text snippets (surface forms as determined by DBpedia Spotlight) are matched with the preferred or alternative labels of the GND. For this purpose, we queried the LOBID API of the University Library Centre of North Rhine-Westphalia (HBZ) as it is offering a public interface to determine GND descriptors [38].[11] An example implementation of the KINDEX approach has been made publicly available and can be downloaded from: https://gitlab.com/limbo-project/keyword-indexing. As SparqlIntegrate plays a major role during the indexing procedures of both use cases as a tool to perform the RDF triple transformations, it will be now explained in more detail.

[10] https://gitlab.com/limbo-project/keyword-indexing/blob/master/queries/mapping.sparql

[11] https://lobid.org/

4.2. SparqlIntegrate

SparqlIntegrate[12] is a tool that leverages SPARQL together with extension functions as the lingua franca for RDFisation and integration of the common heterogeneous data formats XML, CSV, JSON and of course RDF itself. Furthermore, it supports interfacing with scripting environments by allowing for passing environment variables to SPARQL statements as well serialization of result sets in a JSON representation that is suitable for immediate consumption by several existing JSON processors. Thus, it is possible to seamlessly integrate multiple data transformation steps that occur during automatic indexing into an efficient pipeline by means of lightweight command-line processing. Effectively, triple statements are transformed and/or newly generated based on the variables that were passed to the SPARQL interface (see Sect. 4.1). During KINDEX processing, SparqlIntegrate transformations are typically handy, when triple statements are to be altered or inserted into a small to medium-sized data collection that can be easily loaded into the main memory of the host machine. For instance, in order to determine identity links for different keywords and mapping collections, SparqlIntegrate offers convenient query interface options. Depending on the flag option that is set during execution, different mapping collections can be flexibly queried and corresponding results (e.g. keyword descriptors from the specified KOS) can be immediately consumed.
For example, when combined with a scripting language, SparqlIntegrate could first query the DBpedia same-thing lookup service for an identity link to Wikidata.

When there exists a mapping, a `getDescriptor` function determines the corresponding STW descriptor of the STW-to-Wikidata mapping collection (i.e. `stw_wikidata_maping.ttl`). In case, there is no such link, the procedure sends a `HTTP` request to the DBpedia SPARQL endpoint determining whether there exist any mappings to the GND. Upon obtaining a match, the `getDescriptor` function is called, which this time queries the STW-to-GND mapping collection for a suitable STW descriptor.

Due to the built-in feature of SparqlIntegrate to provide SPARQL results in JSON format they can be conveniently accessed by standard command-line processors, such as `jq` in order to be available for

[12] https://github.com/SmartDataAnalytics/SparqlIntegrate

```
@prefix econstor: <http://linkeddata.econstor.eu/beta/resource/publications/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix stw: <http://zbw.eu/stw/descriptor/>

econstor:62535 dc:subject stw:12072-1 (Enterprise),
                          stw:11547-6 (Tax),
                          stw:11322-2 (Wages) .
```

Listing 2 Automatically generated triple statements

additional data transformations during a KINDEX pipeline.[13] Thus, once a matching STW descriptor has been identified for the input DBpedia URL (and preferably verified by a professional subject indexer), it can be assigned to the respective publication and inserted into the Econstor catalogue (`catalog.ttl`) with a simple command.

5. Experiments

5.1. Experimental Settings

The performance of the KINDEX approach was evaluated in experiments, in which the two use cases and their associated data catalogues served as empirical test beds. For *LIMBO*, a baseline corpus of manually assigned GND descriptors was set up prior to conducting the evaluations. For this purpose, two test annotators assigned keywords to a random selection of 100 example records in German that were extracted from the LIMBO metadata catalogue. The annotators were presented with the metadata descriptions of the data sets and received some initial recommendations for GND descriptors that were determined according to DBpedia Spotlight, Wikidata and the LOBID API. The annotators were asked to judge the relevance of the descriptors (with either 1 = "relevant" or 0 = "not relevant") upon being presented with the data set description. Additionally, the indexers stated further keywords that they considered vital to characterise the data set. Afterwards, the relevance judgements of the two annotators were condensed in a gold standard annotation set, where each data set description was enhanced with the fitting index terms. This was done by considering each GND descriptor as relevant, whenever the test persons had consensus, i.e. both either marked the concept with 1 or stated the same keyword as being additionally relevant. The test persons reached agreement in approx. 63% of the assignments.

With the Econstor LOD use case, an annotated corpus was already available. From this corpus, we randomly selected a test collection of 250 English publications that were annotated with STW descriptors.

For both use cases, we tested whether the KINDEX approach correctly predicted the baseline annotations from the gold standard corpora. For this purpose, precision and recall values were calculated by applying the sample-based average method commonly used in text categorisation tasks [39]. Average F1 scores were determined after the summarised values for precision and recall had already been calculated. The collections for both use cases as well as the evaluation script are publicly available.[14]

5.2. Results

Prior to conducting the final performance evaluations, we carried out a few preliminary parameter tuning experiments, in which we determined the best configuration with regard to the confidence score; i.e. the degree to which the matched text snippet is deemed to refer to the correct entity by DBpedia Spotlight. Additionally, the best suited type of text information to be used by the respective Spotlight instances was also determined. In case of the LIMBO catalogue, the *title* (`dct:title`), description (abbr. *desc*, `dcat:description`) and the union of the two (*title+desc*) was tested. For Econstor publications, the *title* (`dct:title`), free text keywords (*key*, `dc:keyword`) and a combination of the two text snippets were evaluated during the tests. Abstracts were not considered, since the baseline paper by Toepfer et al. only reported on ML-based accuracy scores related to short text descriptions (e.g. title and string keywords) [3]. Since we wanted to evaluate how the identity-based

[13] https://stedolan.github.io/jq/

[14] https://gitlab.com/limbo-project/keyword-indexing/tree/master/evaluation

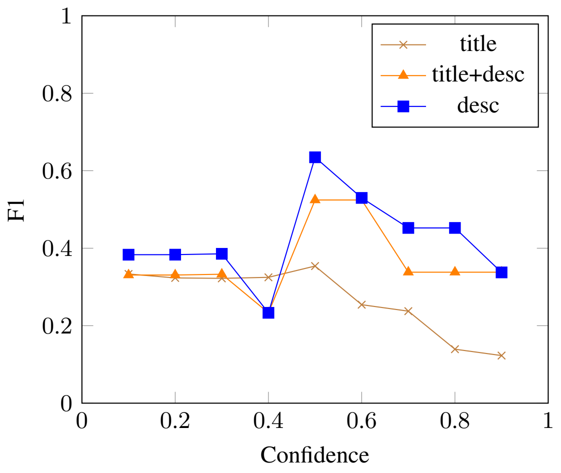

Figure 3. KINDEX parameter tuning (LIMBO)

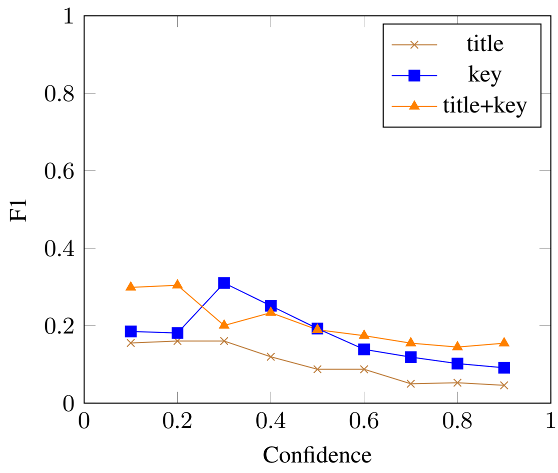

Figure 4. KINDEX parameter tuning (Econstor)

approach performs in comparison to ML-based indexing, we focused on these metadata features. Figures 5.2 and 5.2 depict the results of the parameter tuning tests. Surprisingly, the diagrams indicate that lower to medium-level confidence values produce higher accuracy scores. This might be explained with the fact that both of the evaluated catalogues are rather domain-specific and thus require narrowly defined topic descriptions that can have different meanings in other domains. For the LIMBO catalogue, it also seems to be the case that more input data (*title+desc*) leads to better F1 scores. This is why this specific parameter configuration has been used in subsequent tests. For Econstor, on the other hand, the text-based keywords (*key*) achieved the best performance results for varying confidence parameters.

Upon having determined the best parameter setting, we conducted a final analysis in which we evaluated how the KINDEX approach performs in comparison to a naive lexical matching of Spotlight surface forms as measured by precision, recall and F1 scores. Tabs. 5.2-2 show the final evaluation results of the KINDEX approach in the different use cases. Figures in bold mark the best performing score for each metric and indexing approach. Overall the evaluations show that a combination of lexical and identity matching always achieves better results than simply identifying subject descriptors based on their labels. We performed subsequent statistical tests to seek further validation for this hypothesis. For the LIMBO data, paired t-tests confirmed a significant improvement through KINDEX in comparison to naive lexical matching for each performance metric ($p < 0.001$), while for Econstor the same could only be proven for recall scores ($p < 0.07$).

During the simulation runs, it was also investigated in which sequence the individual lookup steps (i.e., lexical, Wikidata- or DBpedia-based identity lookup) should be preferably processed by the KINDEX engine. The columns with the heading *Indexing Approach* in Tabs. 5.2-2 list the different priority rules, where the order of the lookup steps (i.e, identity or lexical and Wikidata or DBpedia) denotes the corresponding processing sequence. Given the evaluation results, it seems to be the case that indexing approaches function best when they are tailored to the specific use case. While for the LIMBO catalogue, identity matching should have priority over lexical matching and Wikidata-based identity links should be detected prior to DBpedia links, the opposite is true for the Econstor use case. In the latter scenario lexical matching and DBpedia look-ups are to be processed first in order to boost accuracy scores. The reasons for these differences might be that two DBpedia Spotlight instances were applied (a German Spotlight instance for *LIMBO* and an English instance for *Econstor*) and that the topical domains might be covered differently by the two large-scale cross-domain knowledge graphs DBpedia and Wikidata [40].

Additionally, KINDEX achieved varying levels of accuracy in the two scenarios. While *LIMBO* results on average reached fairly high accuracy scores, the performance was not as good for the *Econstor* scenario. For the latter use case, however, it has to be noted that the scores were mostly as good as the best performing ML-based lexical matching approach (i.e., a MAUI

Table 1
mCLOUD - Evaluation results

Indexing Approach	**Precision**	**Recall**	**F1**
Naive Lexical	0.375	0.669	0.480
Identity (DBpedia)	0.442	0.505	0.471
Identity (Wikidata)	0.509	0.740	0.603
Identity+Lexical (Wikidata)	**0.523**	**0.806**	**0.635**
Identity+Lexical (DBpedia)	0.503	0.752	0.603
Identity+Lexical (Wikidata+DBpedia)	0.510	0.792	0.620
Lexical+Identity (DBpedia)	0.430	0.489	0.457
Lexical+Identity (Wikidata)	0.418	0.654	0.510
Lexical+Identity (Wikidata+DBpedia)	0.423	0.660	0.516

Table 2
Econstor - Evaluation results

Indexing Approach	**Precision**	**Recall**	**F1**
Naive Lexical	0.357	0.242	0.288
Identity (DBpedia)	0.145	0.083	0.105
Identity (Wikidata)	0.194	0.070	0.103
Identity+Lexical (Wikidata)	0.360	0.258	0.300
Identity+Lexical (DBpedia)	0.336	0.275	0.302
Identity+Lexical (Wikidata+DBpedia)	0.338	0.276	0.304
Lexical+Identity (Wikidata)	0.306	0.224	0.259
Lexical+Identity (DBpedia)	0.353	0.273	0.307
Lexical+Identity (DBpedia+Wikidata)	**0.363**	**0.279**	**0.315**

adaptation for Econstor) and only slightly weaker than meta-learning strategies that fuse the results of different base learners [3].[15]

6. Conclusion

Given the growing number of digital and analogue content in cultural heritage institutions, high quality metadata descriptions are more important than ever to facilitate personalised retrieval access to valuable resources. However, because of the content overload and the limited personnel in information providing institutions, manual indexing will often not be feasible. Hence, investigations into methods for automatic generation of KOS descriptor annotations are required. To this date, most of the few existing approaches focus on the application of machine learning techniques. While this is an important route for further investigations, we argue that cultural heritage institutions might also profit from harnessing the already available cross-concordance links for automatic subject indexing. Our method generates KOS annotations by combining lexical and identity matching, which is facilitated by the web of data. The evaluation results demonstrate that our KINDEX approach reaches accuracy scores that are competitive with some state-of-the-art ML-enabled methods. Hence, it can serve as a base method whose results are fused with the results of other indexing approaches. Additionally, KINDEX can be applied as a stand-alone tool that offers a viable alternative method for automated subject indexing when the application of ML approaches is not feasible due to missing data, hardware infrastructures or human resources. While it is true that there is also some performance tuning involved in using our method, KINDEX is multilingual and applicable to a large number of knowledge organisation systems almost out-of-the-box, while being independent of training data at the same time. Thus, in addition to cultural heritage institutions, it might also be an interesting tool for researchers to help them annotate their publications with KOS descriptors in order to facilitate an open research infrastructure that relies on rich metadata descriptions [41]. To this end, we plan to offer a web service in the near future that annotates text from multiple languages with descriptors from various thesauri thus leveraging identity links for subject indexing in such a manner that it can be used by a larger audience.

[15] Please note that these findings give only an indication, since the evaluations could not be run on the same sample.

References

[1] A. Miles and S. Bechhofer, SKOS simple knowledge organization system reference, *W3C recommendation* **18** (2009), W3C.

[2] W.J. Hutchins, The concept of ‘aboutness’ in subject indexing, in: *Aslib proceedings*, Vol. 30, MCB UP Ltd, 1978, pp. 172–181.

[3] M. Toepfer and C. Seifert, Descriptor-invariant fusion architectures for automatic subject indexing, in: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, IEEE, 2017, pp. 1–10.

[4] O. Medelyan, E. Frank and I.H. Witten, Human-competitive tagging using automatic keyphrase extraction, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, Association for Computational Linguistics, 2009, pp. 1318–1327.

[5] L. Wenige, The application of Linked Data resources for Library Recommender Systems, *Theorie, Semantik und Organisation von Wissen* **13** (2017), 212.

[6] M. Schmachtenberg, C. Bizer and H. Paulheim, Adoption of the linked data best practices in different topical domains, in: *International Semantic Web Conference*, Springer, 2014, pp. 245–260.

[7] P.N. Mendes, M. Jakob, A. García-Silva and C. Bizer, DBpedia spotlight: shedding light on the web of documents, in: *Proceedings of the 7th international conference on semantic systems*, ACM, 2011, pp. 1–8.

[8] T. Gross and A.G. Taylor, What have we got to lose? The effect of controlled vocabulary on keyword searching results, *College & Research Libraries* (2005).

[9] A.G. Taylor, On the subject of subjects, *The journal of academic librarianship* **21**(6) (1995), 484–491.

[10] U. Junger, Quo vadis Inhaltserschliessung der Deutschen Nationalbibliothek? Herausforderungen und Perspektiven, *o-bib. Das offene Bibliotheksjournal/Herausgeber VDB* **2**(1) (2015), 15–26.

[11] U. Junger, Automation first–the subject cataloguing policy of the Deutsche Nationalbibliothek (2017).

[12] D. Nadeau and S. Sekine, A survey of named entity recognition and classification, *Lingvisticae Investigationes* **30**(1) (2007), 3–26.

[13] J. Voss, Tagging, folksonomy & co-renaissance of manual indexing?, *arXiv preprint cs/0701072* (2007).

[14] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme and G. Stumme, Tag recommendations in folksonomies, in: *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 2007, pp. 506–514.

[15] B. Sigurbjörnsson and R. Van Zwol, Flickr tag recommendation based on collective knowledge, in: *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 327–336.

[16] O. Medelyan, V. Perrone and I.H. Witten, Subject metadata support powered by Maui, in: *JCDL'10*, ACM, 2010, pp. 407–408.

[17] O. Medelyan and I.H. Witten, Thesaurus based automatic keyphrase indexing, in: *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06)*, IEEE, 2006, pp. 296–297.

[18] Y.-M. Chung, W.M. Pottenger and B.R. Schatz, Automatic subject indexing using an associative neural network, in: *ACM DL*, 1998, pp. 59–68.

[19] B. Lauser and A. Hotho, Automatic multi-label subject indexing in a multilingual environment, in: *International Conference on Theory and Practice of Digital Libraries*, Springer, 2003, pp. 140–151.

[20] W.J. Wilbur and W. Kim, Stochastic gradient descent and the prediction of MeSH for PubMed records, in: *AMIA Annual Symposium Proceedings*, Vol. 2014, American Medical Informatics Association, 2014, p. 1198.

[21] K. Golub, Automatic subject indexing of text (2019).

[22] K. Golub, D. Soergel, G. Buchanan, D. Tudhope, M. Lykke and D. Hiom, A framework for evaluating automatic indexing or classification in the context of retrieval, *Journal of the Association for Information Science and Technology* **67**(1) (2016), 3–16.

[23] J. Neubert and K. Tochtermann, Linked library data: offering a backbone for the semantic web, in: *Knowledge Technology Week*, Springer, 2011, pp. 37–45.

[24] L. Wenige and J. Ruhland, Retrieval by recommendation: using LOD technologies to improve digital library search, *International Journal on Digital Libraries* **19**(2–3) (2018), 253–269.

[25] E. Summers, A. Isaac, C. Redding and D. Krech, LCSH, SKOS and linked data (2008).

[26] R. Bennett, C. Hengel-Dittrich, E.T. O'Neill and B.B. Tillett, Viaf (virtual international authority file): Linking die deutsche bibliothek and library of congress name authority files, in: *World library and information congress: 72nd IFLA general conference and council*, 2006.

[27] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques and J. Keizer, The AGROVOC linked dataset, *Semantic Web* **4**(3) (2013), 341–348.

[28] J. Neubert, Bringing the" Thesaurus for Economics" on to the Web of Linked Data., *LDOW* **25964** (2009).

[29] J. Waeber and A. Ledl, A Semantic Web SKOS Vocabulary Service for Open Knowledge Organization Systems, in: *Research Conference on Metadata and Semantics Research*, Springer, 2018, pp. 3–12.

[30] A.O. Kempf and T. Rebholz, 'Mixed Methods' Indexing: Building-Up a Multi-Level Infrastructure for Subject Indexing (2017).

[31] L. Wenige, G. Berger and J. Ruhland, SKOS-based concept expansion for LOD-enabled recommender systems, in: *Research Conference on Metadata and Semantics Research*, Springer, 2018, pp. 101–112.

[32] A. Hajra and K. Tochtermann, Linking science: approaches for linking scientific publications across different LOD repositories, *International Journal of Metadata, Semantics and Ontologies* **12**(2–3) (2017), 124–141.

[33] C. Stadler, L. Wenige, S. Tramp, K. Junghanns and M. Martin, RDF-based Deployment Pipelining for Efficient Dataset Release Management (2019).

[34] A. Latif, T. Borst and K. Tochtermann, Exposing data from an open access repository for economics as linked data, *D-Lib magazine* **20**(9/10) (2014).

[35] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer et al., DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web* **6**(2) (2015), 167–195.

[36] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledge base (2014).

[37] J. Frey, M. Hofer, D. Obraczka, J. Lehmann and S. Hellmann, DBpedia FlexiFusion The Best of Wikipedia> Wikidata> Your Data.

[38] F. Steeg, A. Pohl and P. Christoph, lobid-gnd–Eine Schnittstelle zur Gemeinsamen Normdatei für Mensch und Maschine, *Informationspraxis* **5**(1) (2019).

[39] F. Sebastiani, Machine learning in automated text categorization, *ACM computing surveys (CSUR)* **34**(1) (2002), 1–47.

[40] M. Färber and A. Rettinger, Which Knowledge Graph Is Best for Me?, *arXiv preprint arXiv:1809.11099* (2018).

[41] S. Vahdati, N. Arndt, S. Auer and C. Lange, OpenResearch: collaborative management of scholarly communication metadata, in: *European Knowledge Acquisition Workshop*, Springer, 2016, pp. 778–793.

[42] M. Tanaka, S. Nakazono, H. Matsuno, H. Tsujimoto, Y. Kitamura and S. Miyano, Intelligent system for topic survey in MEDLINE by keyword recommendation and learning text characteristics, *Genome Informatics* **11** (2000), 73–82.

[43] K.S. Hasan and V. Ng, Automatic keyphrase extraction: A survey of the state of the art, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1262–1273.

[44] P.D. Turney, Learning algorithms for keyphrase extraction, *Information retrieval* **2**(4) (2000), 303–336.

[45] Q. Zhang, Y. Wang, Y. Gong and X. Huang, Keyphrase extraction using deep recurrent neural networks on twitter, in: *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 836–845.

[46] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin and C.G. Nevill-Manning, Kea: Practical automated keyphrase extraction, in: *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, IGI Global, 2005, pp. 129–152.

[47] J. Hannemann and J. Kett, Linked data for libraries, in: *Proc of the world library and information congress of the Int'l Federation of Library Associations and Institutions (IFLA)*, 2010.

[48] Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten and D. Fensel, Goldenbullet: Automated classification of product data in e-commerce, in: *Proceedings of the 5th international conference on business information systems*, 2002.

[49] A. Latif, A. Scherp and K. Tochtermann, LOD for library science: benefits of applying linked open data in the digital library setting, *KI-Kunstliche Intelligenz* **30**(2) (2016), 149–157.