Automatic Subject Indexing with Knowledge Graphs

Lisa Wenige, Claus Stadler, Simon Bin, Lorenz Bühmann, Kurt Junghanns, and Michael Martin

Institute for Applied Informatics, University Leipzig
{surname}@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. In this paper, we argue that there is a third solution to subject indexing which harnesses cross-domain knowledge graphs. 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 · Keyphrase Extraction · Authority File.

1 Introduction

Cultural heritage institutions need to facilitate access to large collections of printed and online materials. Topic schemes are important tools that enable effective content searches when navigating data. These schemes are often referred to as knowledge organisation systems (KOS) or controlled vocabularies. In KOS, topics are represented as uniquely identified concepts, that can be characterised by different synonyms [17].
In information providing institutions, trained professionals manually inspect and assign suitable concepts. This is an interesting fact, considering that there are algorithms in place that can extract and assign suitable KOS concepts from digital text. These automatic extraction approaches can be grouped into two categories, namely ML-enabled *associative* and *lexical indexing*. The former learns a model by identifying frequently occurring associations between n-grams from input texts with the corresponding KOS descriptors. The *lexical approach* matches

text snippets 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 [22].
However, even though these approaches have yielded human-level accuracy scores [13], their application is not as widespread. The reasons for this phenomenon are manifold: A considerable amount of preprocessing and training effort is required to adapt the existing ML approaches to training corpora (e.g., language and domain specificities) and the vocabulary. Besides being dependent on the availability of a large enough training corpus with annotated items, this requires expertise in programming and natural language processing (NLP) methods. However, many information providing institutions operate on tight budgets and cannot employ additional staff for these tasks [25]. On the other hand, manual indexing as it is still practised, is time consuming as well. Hence, a (semi-)automatic and efficient indexing tool based on knowledge graphs (KG) 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 information retrieval. It would also link collections with the distributed knowledge infrastructure of the web of data. In the following, we will present a novel method and architecture for automated subject indexing. We propose the KINDEX approach that leverages existing annotation tools, such as DBpedia Spotlight [16] in conjunction with KOS correspondence links of the data web (e.g. `owl:sameAs`, `skos:exactMatch`) to provide relevant keyword suggestions to information professionals. The benefits of our approach are as follows:

- **Minimum training and preprocessing effort** in comparison to existing methods for automatic subject indexing (Section 2). It will be shown how this is realized through a coupling of existing KG web services in a unified annotation workflow.
- **Domain Independence**: KINDEX provides annotations for many potential application domains (see Section 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. However, a strong precondition is that the KOS 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*: 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 scenarios have shown the viability of the approach in terms of accuracy scores (see Section 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 ana-

logue 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 [6,21]. 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 [8].
Despite the practice of manual annotation, the task of automatically identifying information 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 [18].
With ML-based *lexical indexing*, automatic learning techniques are applied to identify how features among the characteristics of the n-gram 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 [14]. KEA has been tested on a collection of agricultural documents with the AGROVOC[1] 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 [15,13]. 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 learned 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 [1,11,28]. 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 [22]. 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

[1] `http://aims.fao.org/vest-registry/vocabularies/agrovoc`

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 [25]. The majority of machine-based indexing approaches is tested under laboratory conditions, often without a follow-up productive implementation of the system [4,5]. Against this background, it seems appropriate to leverage open knowledge graphs (KG) for this task. For more than a decade, many of the existing thesauri have been made publicly available. The vocabulary *Simple Knowledge Organisation System* (SKOS) offers a schema language to describe knowledge organisation systems. SKOS provides expressions to declare cross-concordances/identity links (i.e. `skos:exactMatch`) [17,25,27]. Given the wide availability of SKOS vocabularies, which are often densely interlinked with 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 automatic indexing in conjunction with identity links has great potential to increase cataloguing efficiency [9]. However, in the context of enhanced library services, SKOS vocabularies have rarely been taken advantage of. Only a few papers studied the effect of SKOS relations in order to improve retrieval systems [7,26,27]. To the best of our knowledge, we are the first to investigate the usage of identity links for automatic subject indexing.

3 Use Cases

3.1 Use Case Selection

We showcase the KINDEX approach with two different use cases. In *Use Case 1* 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.

- **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 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.[2] The ministry also supports the research project *Linked Data Services for Mobility (LIMBO)*.[3]

[2] `https://www.mcloud.de/`

[3] `https://www.limbo-project.org/`

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 [19].

– **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.[4] 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 [10]. A large fraction of the publications that are contained in the data set has subject annotations from the STW thesaurus [22].

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 KG technologies and web services by combining `HTTP` and `SPARQL` requests as well as `JSON` processing operations.

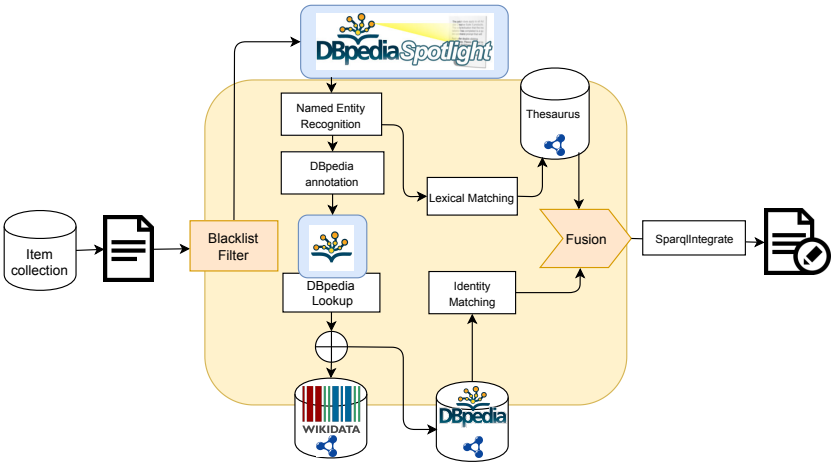

Figure 1. Conceptual overview of the KINDEX approach

It relies on a running instance of DBpedia Spotlight [16] as well as mappings from DBpedia [12] and Wikidata [24] to KOS. The indexing process starts with

[4] `https://www.econstor.eu`

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]
Imagine you would want to index the Econstor publication `econstor:62535` on trade policy with STW descriptors [10]: The publication title is sent to an English DBpedia Spotlight instance which performs *named entity recognition and disambiguation* and returns the annotated text. The respective result file is comprised of the assigned descriptor(s), their surface form (e.g., taxation) and the corresponding *DBpedia annotation* as URI (e.g., `dbr: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 identity the surface form via n-gram matching 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 integrated authority file of the German National Library (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 KGs from different DBpedia chapters and Wikidata with the FlexiFusion approach presented by Frey et al. [3]. 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 [3].
- *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

[5] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/faq`
[6] `https://global.dbpedia.org/same-thing/lookup/`

querying an in-memory 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.

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 `dbr:Tax` is not linked to a Wikidata resource, the corresponding STW descriptor (`stw: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 exists between the GND and the STW can be obtained by querying the STW-to-GND mapping.[7] from an in-memory model that is accessed with the help of SparqlIntegrate[8].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 descriptors as triple statements 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. In the example of the publication `econstor:62535`, the tool would assign the subject keywords `stw:12072-1` (Enterprise), `stw:11547-6` (Tax), `stw:11322-2` (Wages).
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 sent to either Wikidata or DBpedia in order to identify 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 [20].[9] An example implementation of the KINDEX approach has been made publicly

[7] `http://zbw.eu/stw/version/latest/download/about.de.html`

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

[9] `https://lobid.org/`

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.

4.2 SparqlIntegrate

SparqlIntegrate[10] 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 parameter 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. 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 additional data transformations during a KINDEX pipeline.[11] 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 with a simple command.

[10] `https://github.com/SmartDataAnalytics/SparqlIntegrate`
[11] `https://stedolan.github.io/jq/`

5 Experiments

Prior to conducting the final performance evaluations, we carried out a few 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. The tests indicate that lower to medium-level confidence values produce higher accuracy scores. This might be explained by 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. 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 (no. of relevant and predicted index terms/no. of predicted index terms), recall (no. of relevant and predicted index terms/no. of relevant index terms) and F1 scores (harmonic mean of precision and recall). Tabs. 1-2 show the final evaluation results of the KINDEX approach in the different use cases.

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**

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 Tables 1 and 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 [2].
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 adaptation for Econstor) and only slightly weaker than meta-learning strategies that fuse the results of different base learners [22].[12]

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

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

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 [23]. 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 to be used by a larger audience. Other future research directions in regard to the KINDEX method will be investigations into the scalablity of the approach and the handling of performance bottlenecks.

References

1. Chung, Y.M., Pottenger, W.M., Schatz, B.R.: Automatic subject indexing using an associative neural network. In: ACM DL. pp. 59–68 (1998)
2. Färber, M., Rettinger, A.: Which knowledge graph is best for me? arXiv preprint arXiv:1809.11099 (2018)
3. Frey, J., Hofer, M., Obraczka, D., Lehmann, J., Hellmann, S.: Dbpedia flexifusion the best of wikipedia¿ wikidata¿ your data
4. Golub, K.: Automatic subject indexing of text (2019)
5. Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., Hiom, D.: A framework for evaluating automatic indexing or classification in the context of retrieval. Journal of the Association for Information Science and Technology **67**(1), 3–16 (2016)
6. Gross, T., Taylor, A.G.: What have we got to lose? the effect of controlled vocabulary on keyword searching results. College & Research Libraries (2005)
7. Hajra, A., Tochtermann, K.: Linking science: approaches for linking scientific publications across different lod repositories. International Journal of Metadata, Semantics and Ontologies **12**(2-3), 124–141 (2017)
8. Junger, U.: Quo vadis inhaltserschliessung der deutschen nationalbibliothek? herausforderungen und perspektiven. o-bib. Das offene Bibliotheksjournal/Herausgeber VDB **2**(1), 15–26 (2015)
9. Kempf, A.O., Rebholz, T.: 'mixed methods' indexing: Building-up a multi-level infrastructure for subject indexing (2017)
10. Latif, A., Borst, T., Tochtermann, K.: Exposing data from an open access repository for economics as linked data. D-Lib magazine **20**(9/10) (2014)

11. Lauser, B., Hotho, A.: Automatic multi-label subject indexing in a multilingual environment. In: International Conference on Theory and Practice of Digital Libraries. pp. 140–151. Springer (2003)
12. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
13. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. pp. 1318–1327. Association for Computational Linguistics (2009)
14. Medelyan, O., Perrone, V., Witten, I.H.: Subject metadata support powered by maui. In: JCDL'10. pp. 407–408. ACM (2010)
15. Medelyan, O., Witten, I.H.: Thesaurus based automatic keyphrase indexing. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06). pp. 296–297. IEEE (2006)
16. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
17. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. W3C recommendation **18**, W3C (2009)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
19. Stadler, C., Wenige, L., Tramp, S., Junghanns, K., Martin, M.: Rdf-based deployment pipelining for efficient dataset release management (2019)
20. Steeg, F., Pohl, A., Christoph, P.: lobid-gnd–eine schnittstelle zur gemeinsamen normdatei für mensch und maschine. Informationspraxis **5**(1) (2019)
21. Taylor, A.G.: On the subject of subjects. The journal of academic librarianship **21**(6), 484–491 (1995)
22. Toepfer, M., Seifert, C.: Descriptor-invariant fusion architectures for automatic subject indexing. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 1–10. IEEE (2017)
23. Vahdati, S., Arndt, N., Auer, S., Lange, C.: Openresearch: collaborative management of scholarly communication metadata. In: European Knowledge Acquisition Workshop. pp. 778–793. Springer (2016)
24. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base (2014)
25. Wenige, L.: The application of linked data resources for library recommender systems. Theorie, Semantik und Organisation von Wissen **13**, 212 (2017)
26. Wenige, L., Berger, G., Ruhland, J.: Skos-based concept expansion for lod-enabled recommender systems. In: Research Conference on Metadata and Semantics Research. pp. 101–112. Springer (2018)
27. Wenige, L., Ruhland, J.: Retrieval by recommendation: using lod technologies to improve digital library search. International Journal on Digital Libraries **19**(2-3), 253–269 (2018)
28. Wilbur, W.J., Kim, W.: Stochastic gradient descent and the prediction of mesh for pubmed records. In: AMIA Annual Symposium Proceedings. vol. 2014, p. 1198. American Medical Informatics Association (2014)