Unlocking the Corpus: Enriching Metadata with State-of-the-Art NLP Methodology and Linked Data

Jennifer Ecker[1], **Stefan Fischer**[2], **Pia Schwarz**[1], **Thorsten Trippel**[1],
Antonina Werthmann[1], **Rebecca Wilm**[1]
[1]Leibniz Institute for the German Language (IDS), Mannheim, Germany
{ecker,schwarz,trippel,werthmann,wilm}@ids-mannheim.de
[2]Saarland University, Saarbrücken, Germany
stefan.fischer@uni-saarland.de

Abstract

In research data management, metadata are indispensable to describing data and are a key element in preparing data according to the FAIR principles. Metadata in catalogues and registries are usually recorded either by archivists or subject matter experts, i.e. researchers involved in the creation or assembling of the data, or provided in the data preparation workflow. Extracting metadata from textual research data is currently not part of most metadata workflows, even more so if a research data set can be subdivided into smaller parts, such as a newspaper corpus containing multiple newspaper articles. If we look at descriptive metadata from a large corpus of newspapers, the basic metadata may consist of information, for example, about the title, or year of publication. Our approach is to add semantic metadata on the text level to facilitate the search over data. We show how to enrich metadata with three methods: named entity recognition, keyword extraction, and topic modeling. The goal is to make it possible to search for texts that are about certain topics or described using certain keywords or to identify people, places, and organisations mentioned in texts without actually having to read them.

1 Introduction

Enriching the information extracted from corpora to find more relevant parts of a corpus for deeper analysis is the overall aim of this contribution. Newspaper corpora or other large collections contain a multitude of texts of various topics, timespans, authors, etc. Some of the properties are already available in form of metadata, hence they can be used to select partitions of these corpora. If these properties are not provided in the metadata, the corpus can basically only be used as is, which may still be a valid and useful application. To enhance usability of the data according to the FAIR principles (Wilkinson et al., 2016), rich metadata as a meaningful representation of the data are a key element. In this contribution, we explore options based on a large reference corpus to enrich the metadata not only for the whole corpus but on the level of subportions, such as articles in a newspaper corpus. For developing the processes, we select a small section of the corpus. With this approach we provide options to find parts of the corpus that may be more relevant for specific tasks, for example to create a subcorpus for specific topics, on given individuals, places or organisations.

2 Motivation

The German Reference Corpus *DeReKo* is the largest linguistically motivated collection of German language material (Kupietz & Keibel, 2009; Kupietz et al., 2010, 2018). The corpus is an example of a national corpus, with all legal restrictions of modern data. Although the sample corpus is not called a national corpus, DeReKo serves the same purposes as national corpora for other languages. It contains multiple newspapers, books, transcriptions, etc. For the purpose of this contribution, the authors received access to a number of data files in their native XML structures. For developing the methodology we focus on one file in DeReKo's native XML format with all issues of one newspaper of one year.

Specific research questions may focus on parts of such a corpus. However, there is no general criterion for substructuring a corpus, as this is highly dependent on the research questions. For someone interested in the style of specific authors, the substructuring of such a corpus would be best if all articles or contributions were clustered by their author; for someone interested in specific topics, the clustering should be by topic, for specific timeframes by dates, etc. For a reference corpus that is intended for multiple uses, no such clustering seems to be a universal way to structure a corpus, and the granularity – the archival unit – is not fixed.

Usually each archival unit receives a persistent identifier, hence there is a clear relation between the archival units and PIDs. For PIDs, criteria have been recommended to determine the granularity of archival objects. *ISO 24619:2011* (2011) recommends four different procedures for determining the granularity of objects to receive a persistent identifier which can also be used to determine the granularity of archival objects. These four options in the standard are: (1) use the granularity of an existing PID schema, if such exists, otherwise (2) the PID should be assigned if a resource is complete within one file; if this is not the case (3) it should be assigned to a unit that exists autonomously outside a larger context, and else (4) the PID should be assigned to a unit that should become citable.

For DeReKo, there are files that are used for extending the corpus, such as books represented according to the DeReKo native format I5 (Lüngen & Sperberg-McQueen, 2012), which is an XML schema defined according to recommendations of the Text Encoding Initiative (TEI P5; *TEI P5*, 2021). Most newspapers are added into the reference corpus on a yearly basis, i.e. every issue of the newspaper in its digital form is part of an archive of a specific year. Sources such as Wikipedia with their discussions and history functions are ingested based on these different functions, i.e. one file for all articles, one for the discussions and one for the history. Hence, for archiving and sustainable preservation, DeReKo currently relies on the ingest files, which are also the base for assigning PIDs, and the ingest files constitute the archival object's unit. Each of these collections is represented by a metadata file which is publicly available even if the data file itself is only available under certain restrictions. The internal structure of the original corpus data, however, also allows for other partitions, such as the identification of all newspaper articles published on a specific date, belonging to specific sections, etc. These structures can be identified by their internal unique text *sigles*, which are part of the XML representations of the data. Hence, a sigle is a unique identifier of a subpart of a corpus, either on a corpus, document, or text level, the latter for example for individual newspaper articles. Using the sigle for identifying parts of the corpus requires access to the underlying files.

For accessing linguistic structures, searches on the word, phrase or sentence level are possible with the specialised corpus query tool *KorAP* (Bański et al., 2013; Diewald & Margaretha, 2016; Diewald et al., 2016; Kupietz et al., 2017). Someone who has (legal) access to the full source file can utilise available information and create their own selection of units from the reference corpus based on for example the text sigle as well. Hence, it becomes feasible to create arbitrary subcorpora based on all available attributes, for example clustering subparts by persons with a specific role, place, date, topic, or keyword, if these attributes can be identified for a specific unit.

In the current standard representation, there are only limited options due to sparseness of properties available in the metadata representation. Consequently, it seems very relevant to extract properties of units on various granularity levels. Due to the sheer size of the reference corpus, it will be impossible – and is indeed out of any reasonable suggestion – to add these properties manually: NLP methods need to automate the enrichment of the data.

3 Approach

As a starting point to enrich DeReKo with semantic metadata, we focus on extracting topics, keywords, and three types of named entities: academic institutions, research areas, and persons with an academic background. The experiments are described in Section 5.1 for topic modeling, Section 5.2 for keyword extraction and Section 5.3 for named entity recognition (NER). We believe that these might be useful entry points for researchers to partition the reference corpus to fit their particular research questions. Additionally, these NLP methods give important insights into the corpus and facilitate the search over data.

At the same time, applying three different NLP methods allows us to explore how we can implement a metadata extension – see Section 3.1 – which captures semantic metadata besides the existing catalog metadata. This extension also includes the possibility to record potential links between the extracted semantic information and other external data sets, further described in Section 3.2. There are other established tools for the processing of corpora provided by CLARIN consortia, such as WebLicht (Hinrichs et al., 2010) developed in CLARIN-D and a web-based natural language processing workflow (Walkowiak & Piasecki, 2015) within CLARIN-PL. Due to legal restrictions, our data must remain within our organisation during processing. For this reason, we were unable to utilise external tools that require data to be processed on other servers.

3.1 Metadata Profile Extension

Before enriching the metadata with semantic information, we considered different approaches to accomplish our aim straightforwardly. There were four main options:

1. We can enhance the metadata by adding semantic information into the metadata header provided in the I5 files. These files are based on TEI standards, rendering them not only extensible but also adaptable to our specific requirements. The advantage of this approach is that the text and the newly enriched information are stored within the same file. The major disadvantage of this approach is that it necessitates modifications to the primary I5 data. This can lead to issues if something goes wrong during the enrichment process and requires repairs or modifications. Additionally, this approach of enriching data results in the I5 files becoming larger and more complex. Any subsequent changes or extensions to the I5 data demand greater processing capacity to process the entire DeReKo data set. Furthermore, modifying the I5 files also requires adjustments to the I5 scheme, adding to the overall workload. The update here also poses an issue with regards to long term archiving, as the integrity of the archived files has to be ensured and a change in the underlying data format may change results based on the previously archived files. However, this can be addressed by providing different versions of the files, which in itself causes additional obstacles by additional memory requirements, etc.

2. Enriched metadata can be stored in separate metadata files, e.g. in CMDI[1] (Broeder et al. 2012; *ISO 24622-1:2015* 2015; *ISO 24622-2:2019* 2019) or JSON-LD[2] (*JSON-LD 1.1* 2020) format for each individual sigle of the I5 file. The advantage of this approach is that semantic information can be sequentially extended without affecting the I5 files. However, there are certain disadvantages to consider: The persistent identification for each sigle, using a PID with the sigle ID as part of the identifier, is necessary. This requires technical adjustments within the existing repository. In addition, the heightened complexity of the I5 data structure may lead to reduced clarity. A single I5 file can encompass numerous sigles, requiring the generation of corresponding semantic metadata files for each of them. This can result in increased data complexity, making the data less user-friendly and potentially more challenging to interpret.

3. We can generate a CMDI file for each individual I5 file. The advantages of this approach are: CMDI files containing descriptive metadata can be automatically generated from I5 files and then subsequently expanded with semantic metadata. Both descriptive and semantic information are stored in the same file, and they can be extended and modified at a later stage without requiring changes to the primary I5 files. Changing the CMDI schema is also possible at any time. Additionally, the CMDI files can be converted to alternative formats, such as JSON-LD or HTML, in the future. However, there are some disadvantages to consider: Enriching semantic CMDI metadata requires intricate data modeling and interpretation. Furthermore, the size of a CMDI file depends on the size of the corresponding I5 file and can become quite large. In cases with numerous sigles, the CMDI file may become increasingly complex and challenging to understand.

[1] https://www.clarin.eu/content/component-metadata

[2] https://www.w3.org/TR/json-ld11/

4. We can generate semantic information through real-time analysis. The advantages of this approach are: The I5 files remain unaltered, and there is no need to create additional metadata files. However, it is essential to take into account the significant technical and result-related disadvantages: Intensive real-time processing places a significant burden on computational resources. Such tasks require substantial processing power, leading to higher infrastructure costs and potential bottlenecks in data processing pipelines. The regular changing, modifying, and updating of the tools with the latest technology after deployment can be exceptionally expensive. Changing the licenses of the used NLP applications and models can also have unpleasant consequences because, in the worst case, we may no longer be able to use one or the other tool. Processing queries on huge text data is time-consuming, making real-time query responses practically impossible or excessively difficult. Additionally, the results of queries are reliant on the tools used. If these tools are updated or modified, replicating the results becomes challenging, if not impossible.

After careful consideration, we have chosen to pursue option (3) as this way of adding semantic standoff annotation is comparatively easy to maintain in case of any changes, whether they concern the original I5-formatted data or the semantic annotation itself. On top of that, legal restrictions of the data are not specified on the metadata. With CMDI files, we can make our findings available, even though the original data can only be accessed by specific users of the corpus. The extracted metadata can be shared under open licenses (such as CC-0 or CC-BY), offering reference to the original data and thereby promoting accessibility and transparency.

3.2 Linking Data to External Knowledge Bases

There are a number of available knowledge bases that the named entities, topics, and keywords that we extracted from our corpus could be linked to. These resources differ with respect to size, domain, and annotation quality. In part, this can be explained by their different creation processes: While some resources are created manually, others are generated automatically. In addition, resources can be subject to strict curation processes, or they might be limited to a specific domain.

Wikidata[3] (Vrandečić & Krötzsch, 2014) is a multilingual knowledge graph that is part of the Wikimedia project. Like Wikipedia, it is created by volunteers and available under a permissive license. Each Wikidata item has a persistent identifier (QID) and is explained by a label and short description. Additional information is provided in structured key–value statements. As the knowledge graph is based on the collaborative editing effort of volunteers, records might be inaccurate or not up-to-date. With more than 100,000,000 items, which are not restricted by topic, Wikidata provides large coverage and can serve as a “hub” that connects identifiers from different authority files. Wikidata has been used for the development of named entity linkers (Delpeuch, 2020; Möller et al., 2022; Sakor et al., 2020).

The Integrated Authority File[4] (Gemeinsame Normdatei; GND; Behrens-Neumann & Pfeifer, 2011) provides a catalogue of entities with unique and reliable identifiers. The GND covers persons, corporate bodies, events, place names, subject headings as well as cultural and academic works. It comprises roughly 10,000,000 records, which can only be edited by participating organisations. In comparison to Wikidata, the GND provides high-quality entries from an authoritative source at the cost of lower coverage.

Although the aforementioned resources cover a broad range of domains, they might not provide sufficient coverage for rather specific use cases, e.g. academic named entity recognition (see Section 5.3). However, there are domain-specific databases providing information about both research organisations and individual researchers: the Research Organization Registry[5] (ROR; Lammey, 2020) and the Open Researcher and Contributor ID[6] (ORCID; Haak et al., 2012). The ROR is a community effort to provide persistent identifiers and metadata about research organisations. The ROR contains entries for more than 100,000 international organisations. The registry is searchable via an API and can be downloaded freely.

[3] https://www.wikidata.org

[4] https://explore.gnd.network

[5] https://ror.org

[6] https://orcid.org

ORCID is a platform that assigns persistent identifiers to participating researchers. After registration, a researcher can provide information about their publications, affiliations or grants. The ORCID database can be downloaded freely and the records are also searchable via an API.

GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) is a lexical-semantic network for German. In this network, lexical units are organised into concepts (*synsets*) whose members share the same meaning. Both lexical units and synsets are identified by unique IDs. GermaNet provides lexical-semantic information and a conceptual network for adjectives, adverbs, nouns, and verbs. However, named entities referring to persons are not part of its database. In our use case, the hierarchical relations between synsets might be of interest. For example, one could determine (co-)hyponyms of a detected named entity in order to connect texts that cover the same concept but use different surface forms. The GermaNet database can be used freely for academic research.

In order to link extracted topics, keywords, and named entities, the respective identifiers (QIDs, GND IDs, IDs from ROR and ORCID, GermaNet IDs) of corresponding items from all these external knowledge bases can be encoded per text sigle in the semantic stand-off annotation files. So far, we have not implemented this linking of the data, partly because the problem of word sense disambiguation between the extracted items in our corpus and possible counterparts in the different knowledge bases has not yet been solved.

4 Data

For our experiments we selected a data sample from DeReKo, the 2020 volume of the newspaper corpus Mannheimer Morgen (M20), published under the QAO-NC license (Kupietz & Lüngen, 2014) allowing for query-and-analysis only academic and non-commercial use. According to DeReKo's structure, M20 is a single I5-formatted XML file containing several individual newspaper articles each identified by their text sigle (e.g. M20/APR.00192), which consists of the corpus identifier M20, the document identifier, corresponding to the month in which the article was released, and a five-digit text identifier. In total the M20 subcorpus comprises 44,383 texts.

The selection of a subset for developing the process has not been chosen arbitrarily. We chose the subset based on the following criteria:

- The size of the data set is substantial. Running metadata extraction processes on a test set should provide sufficient information on runtime and hardware use to evaluate whether these processes could be scaled to the full data set, i.e. some thousand additional subcorpora as input.

- The size of the data is small enough to allow experimenting with the data in short periods of time and with limited computing power. The overall goal was to make sure that individual processes on such a corpus would not take too long. The duration of such a process influences the development, as restarts and new tries may be required frequently. In addition, the CO_2 footprint of failed experiments would not be too large.

- The data and its TEIish serialisation are prototypical for many other data sets that are part of the corpus, e.g. other newspapers and magazines, but also books and Wikipedia articles.

As such a metadata enrichment process is not a real-time application, initial processes have to be fast enough and modest in hardware requirements to scale in a way that the full process could be run without HPC environments in a reasonable amount of time.[7] The M20 subcorpus had the right size and format for experimenting.

5 Experiments

Using NLP for keyword extraction, named entity recognition and topic modeling is a rather typical task in the development of these methods. Enriching metadata based on a large set of linguistically annotated

[7] In the current development state, we experience runtimes without specific optimisations that could easily lead to durations in the magnitude of years on the full data set of DeReKo.

Number	Generated Name	Topic Words
0	Musik ‘music’	Songs, Album, Sound, Pop, Song, Musiker, Hits, Gitarren, Schlagzeuger, Gitarrist, ... ‘songs’, ‘music album’, ‘sound’, ‘pop’, ‘song’, ‘musician’, ‘hits’, ‘guitars’, ‘drummer’, ‘guitarist’
1	Religion ‘religion’	Kirche, Gläubigen, Pfarrer, Gottesdienst, Kirchen, Gottesdienste, Gebet, Andacht, Gläubige, beten, ... ‘church’, ‘the faithful’, ‘pastor’, ‘service’, ‘churches’, ‘services’, ‘prayer’, ‘devotions’, ‘the faithful’, ‘to pray’
88	Entfesselung ‘unleashing’	schlank, Zeugen, Jacke, trug, Hinweise, Hose, Täter, bekleidet, Fahndung, flüchtete, ... ‘slim’, ‘witnesses’, ‘jacket’, ‘wore’, ‘evidence’, ‘trousers’, ‘offender’, ‘dressed’, ‘tracing’, ‘escaped’

Table 1: Topics named with the help of a Llama model, two of them correctly labeled (number 0 and number 1) and one of them incorrectly labeled (number 88).

corpora has not been applied to the DeReKo corpus. As the authors are not the maintainers of the corpus, it is obvious that the representation of the added information will have to follow stand-off procedures, such that the archived corpus may not be modified or tampered with. Hence, our approach is a standard procedure in NLP by using independent processes on the data to be investigated in order to extract the wanted set of information and to see if these methods can beneficially be applied to this corpus, as well as to figure out how the extracted information could best be represented in a productive system. In the following, we present current work from topic modeling, keyword extraction and named entity recognition.

5.1 Topic Modeling

One exception for existing metadata of the DeReKo corpus is that each text sigle is annotated with a topic as described in Weiß (2005): First, a human annotator constructs a thematic taxonomy based on specific guidelines and the training data. These guidelines include clusters generated by a document clusterer and an external ontology (i.e., the Open Directory Project). Afterwards, corpus texts are automatically classified using the training data. However, this method has significant deficiencies: the taxonomy, developed partly bottom-up through clustering, no longer covers all the domains needed, the granularity is inadequate, the base taxonomy from the Open Directory Project is no longer in use, and the classifier is outdated due to its almost 20-year-old training data. Therefore, a new approach of assigning topics is required.

For our data set, we employ topic modeling to group the articles into categories. The categories are not predefined with the help of an ontology, but learned by a topic model. We use the Top2Vec model[8] (Angelov, 2020) to assign topics to the articles. Before deciding to use Top2Vec, we also experimented with other topic modeling approaches like Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and BERTopic (Grootendorst, 2022). After manually reviewing the results, we found that the results from Top2Vec best met our requirements. Furthermore, Top2Vec proved to be the easiest approach to implement. With Top2Vec, we selected the doc2vec embedding model for application to our dataset, configuring the speed parameter to deeplearn while using the default settings for all other parameters. Top2Vec divides the articles automatically into 348 topics and assigns a similarity to every article. The cosine similarity between the article vector and the topic vector depicts this semantic similarity. After a manual inspection, we use hierarchical topic reduction to

[8] https://github.com/ddangelov/Top2Vec

reduce the topics from 348 to 150 to circumvent that about half the topics are semantically too close to each other.

The output of the topic model is a number of unnamed topics and corresponding topic words for each topic. Unnamed is meant in the sense that there is no word that describes or sums up the topic words. Instead, numbers are assigned to differentiate between the topics. There are different approaches available for labeling the topics. While some take the highest ranked topic word as the name, others only provide the first n words as a description. Another alternative is to use graph-based labeling with the help of a knowledge resource to automatically label the topics (Ecker, 2024). Manually labeling the topics is also an option, but we decide to prompt a large language model[9] to do the labeling. The used model is based on Meta's Llama 2 model[10]. The top twenty topic words calculated and ranked by the model form the basis for computing the label (see Table 1 for an example with a few topic words). The topic name can be one of the topic words or a word derived from them. If the generated label is not suitable to describe the topic words, we define a label manually (see Section 6 for examples).

5.2 Keyword Extraction

Up to ten uni- or bigram keywords are extracted for each article by combining YAKE! (Campos et al., 2018a, 2018b, 2020), a state-of-the-art unsupervised keyword extraction method that assigns a score to each possible keyword based on statistical features, with a filter based on spaCy (Honnibal et al., 2020) part-of-speech tags[11] to exclude any parts of speech other than nouns and proper nouns. In order to avoid inflected forms such as 'Bundesfinanzministeriums' ('Federal Ministry of Finance's'), the resulting keywords are lemmatised using spaCy.

5.3 Academic NER

In order to recognise academic named entities, we fine-tune a German BERT$_{\text{BASE}}$ model as, to our knowledge, no German NER model with the required entity types academic person, academic organisation, and research area exists. Sentences were filtered out of 10,000 randomly selected texts from DeReKo (mainly newspaper articles and press releases) with the help of an off-the-shelf NER model from the Stanza NLP package (Qi et al., 2020) and word lists containing prototypical mentions for each of the entity types. The word list for the entity type academic person (PER-RES) lists academic titles such that a string match with an item from the list combined with a detected person entity from the Stanza NER model results in a candidate entity. For academic institutions (ORG-RES) and research areas (AREA-RES), the word lists were created using official lists from the German Research Foundation[12] and the Federal Government[13] to detect candidate entities using simple string matching. During post-processing, candidate entities were manually reviewed, which resulted in a data set of 4,928 sentences with a total of 7,199 tags. The data was split into a training, development, and test set with a ratio of 70/20/10. Table 2 provides an overview of the entity type distribution across data splits.

Entity Type	No. of Tags in Train / Dev / Test	P	R	F1
PER-RES	2,942 / 858 / 423	93.68	97.19	95.4
ORG-RES	1,624 / 484 / 192	89.58	88.66	89.12
AREA-RES	450 / 147 / 79	89.47	80.0	84.47
Overall	5,016 / 1,489 / 694	92.12	92.78	92.45

Table 2: Tag distribution of the entity types and resulting model performance measured in precision, recall, and F1-score in percent. Overall scores are micro-averaged.

[9] https://huggingface.co/TheBloke/Llama-2-13B-chat-GGML

[10] https://huggingface.co/meta-llama

[11] The 'de_core_news_lg' model is used for spaCy part-of-speech tagging and lemmatisation.

[12] https://www.dfg.de/dfg_profil/gremien/fachkollegien/faecher/index.jsp

[13] https://www.bundesbericht-forschung-innovation.de/de/Liste-der-Einrichtungen-1790.html

Using the spaCy transformer library[14], we fine-tune the model *de_dep_news_trf*[15] on a single Tesla P4 GPU. With respect to the hyperparameters for model training, spaCy's default settings were used: a batch size of 128, a dropout rate of 0.1, the Adam optimiser with an initial learning rate of 10^{-5}, and early stopping based on the F1 score. Model evaluation was done on the 489 sentences of the test split, yielding an overall micro-F1 score of 92.45%. One observation that can be made already are that the individual F1 scores for the entity types increase according to the amount of tags per type (see Table 2). However, these numbers do not necessarily need to correlate, and there might be other reasons why for the model some entity types are more difficult to recognize than others. For more details about the NER model refer to Schwarz (2024), which also includes experiments with LLMs compared to the BERT approach: The evaluation on the test data shows that the fine-tuned BERT model yields better results than any of the LLMs. However, given the very dynamic developments of LLMs, we can certainly not exclude that later models can keep up with the fine-tuned BERT model. The approach of fine-tuning an LLM with the given NER task was not tested and might be subject of further experiments.

6 Results and Discussion

Enriched metadata is crucial for usability, accessibility, and overall value of corpora for users and must therefore be of high standards. Information about the whole corpus can be accessed without the need to read each text and therefore users can quickly locate pertinent information about the topic of the text, relevant keywords, and people, places, and organisations named in the text. In this section, we present the results of the application of the three methods and we discuss which improvements could be made.

6.1 Topic Modeling

For the method of topic labeling, the majority of the generated names for the topics are suitable. Nevertheless, we changed the label in 39 out of 150 cases (26 percent). Table 1 includes an example of an incorrectly labeled topic (number 88). The German word 'Entfesselung' ('unleashing') is not fitting, and a more appropriate name for the topic would be 'Fahndung' ('tracing'), which is also included in the topic words. Besides, there are other mistakes, such as pseudowords, which are similar to a German word that is suitable as the name of the topic but contain a wrong letter, or English words, which we then translated from English to German, because the generated English word is generally suitable. An example for a pseudoword is 'Fahrverböt' instead of 'Fahrverbot' ('suspended driving licence'), and an example for an English name is 'Digital Life' instead of 'digitales Leben'. To improve the topic words, one option is to apply lemmatisation before or after the topic model. In this case, 'Gottesdienst' ('service') and 'Gottesdienste' ('services') would only be listed once as 'Gottesdienst' (see Table 1 for topic number 1). The list of topic words (in topic number 88) would also not include conjugated verbs such as 'trug' ('wore'), which would then be represented with the infinitive form 'tragen' ('to wear'). With lemmatisation, the number of topics correctly labeled by the Llama model could increase, but again, incorrectly lemmatised words may impede this. Furthermore, experiments with other Llama models should be carried out to compare the outcome of different models.

6.2 Keyword Extraction

While no quantitative evaluation is performed for keyword extraction due to a lack of gold-standard data, table 3 presents the results obtained for three different articles, showcasing some remaining difficulties. For all three articles, the keywords make it easy to guess what they may be about (in combination, at least), thus fulfilling their general purpose. However, some of the keywords are not ideal due to part-of-speech tagging and lemmatisation mistakes. Keywords such as 'jugendliche' ('young') and 'Ungewöhnlichst Buchtitel' ('most unusual book title') were selected although all constituent tokens were supposed to be restricted to be nouns or proper nouns. 'Gedankengänge' ('trains of thought') was lemmatised to 'Gedankengäng' rather than the correct 'Gedankengang', a form that does not actually exist in German and would be unlikely to be searched for. It may be possible to improve the results by experimenting with

[14] https://spacy.io/api/transformer

[15] https://github.com/explosion/spacy-models/releases/tag/de_dep_news_trf-3.7.2

other part-of-speech taggers and lemmatisers. Most notably, the respective annotations that are already available for DeReKo seem worth exploring.

Further difficulties arise from the parameters chosen for YAKE!. As demonstrated by the keywords ‘Saturday Night’ and ‘Night Live’, which would be more adequately represented by the trigram keyword ‘Saturday Night Live’, there are cases in which the approach of only taking unigram and bigram keywords into consideration falls short. With no gold keywords available, however, it is difficult to determine whether the advantages of increasing the value of n would outweigh the drawbacks. The same holds true for the number of keywords extracted per article. For the text sigle M20/APR.00002, the surname ‘Sträter’ appears in three of the extracted keywords, but the combination of the comedian’s first name and surname, ‘Torsten Sträter’ (arguably an especially important keyword since it is one that users would seem likely to search for), which does appear in the original text, is missing. While increasing the number of keywords extracted per article would lead to more relevant keywords being found, more irrelevant keywords would also be extracted. It is not clear what the ideal number or cutoff value in terms of YAKE! score would be.

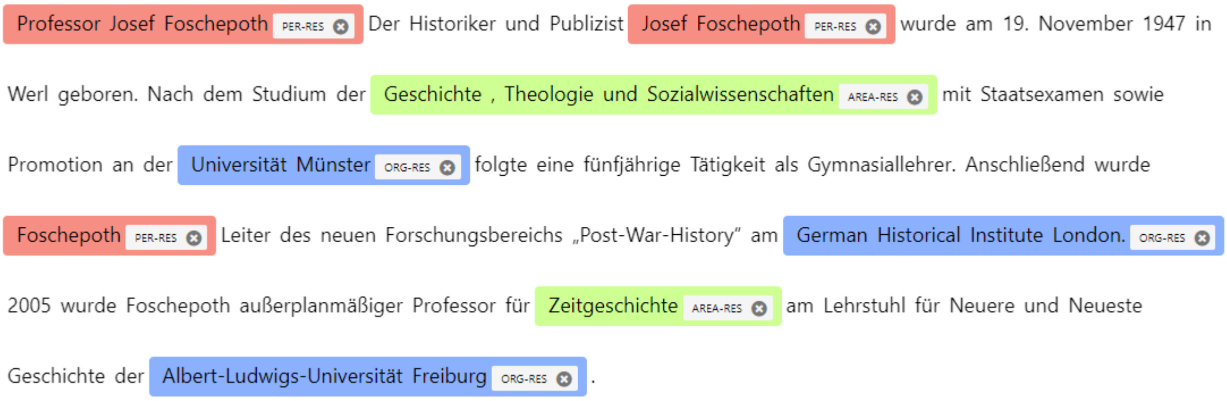
Finally, it may be worth exploring more modern, LLM-based approaches to keyword extraction. Unfortunately, given that no gold keywords are available for the given articles, it is difficult to compare the performance of different models on the given task. We decided against manually annotating the data with gold keywords since the task would be very time-consuming and highly subjective.

Article	Keywords
M20/JAN.00004	Jugendförderung, Zeltlager, Vorbereitung, Toskana, Möglichkeit, Viernheim, jugendliche, Stadtteilbüro Ost, Ferienfreizeit, Anmeldeformular ‘youth empowerment’, ‘camp’, ‘preparation’, ‘Tuscany’, ‘possibility’, ‘Viernheim’, ‘young, ‘district office East’, ‘holiday camp’, ‘registration form’
M20/JAN.00060	Saturday Night, Night Live, Sender NBC, Howard Shore, – Aviator, Panic Room, Kanada, Bild, Ton, Howard ‘Saturday Night’, ‘Night Live’, ‘channel NBC’, ‘Howard Shore’, ‘– Aviator’, ‘Panic Room’, ‘Canada’, ‘picture’, ‘sound’, ‘Howard’
M20/APR.00002	Sträter Gedankengäng, Gedankengäng, Zuschauer, lachen, Luft, Bühne, Sträter, Ungewöhnlichst Buchtitel, Thalia Buch, Sträter Verhältnis ‘Sträter train of thought’, ‘train of thought’, ‘spectator’, ‘laughing’, ‘air’, ‘stage’, ‘Sträter’, ‘most unusual book title’, ‘Thalia book’, ‘Sträter relationship’

Table 3: Keyword extraction results for three articles.

6.3 Academic NER

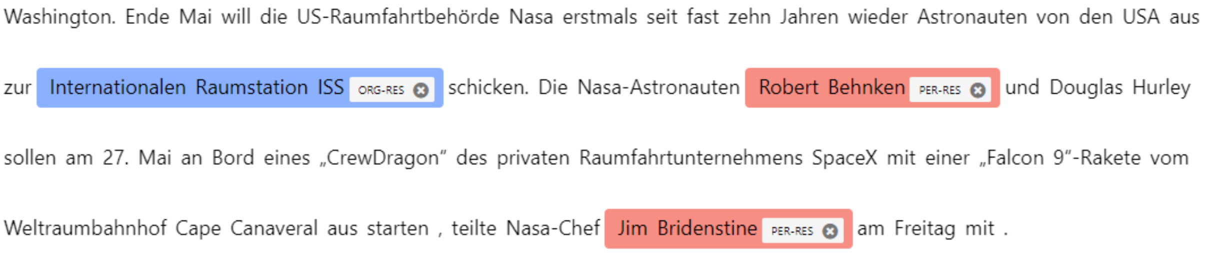
When applying the fine-tuned NER model to the M20 subcorpus, at least one academic named entity is tagged in almost 40 percent of the 44,383 newspaper articles. Most of the tags, over 20,000, fall upon the type PER-RES, almost 10,000 items are tagged with ORG-RES, and a bit more than 3,000 with the entity type AREA-RES. Sentences from two randomly selected articles illustrate good and bad example output of the NER model. In Figure 1, the person ‘Josef Foschepoth’ is detected as a researcher three times. In the first occurrence, the preceding academic title makes it an obvious choice. For the second and third occurrence, however, no such title is present, but the model is still able to correctly assign a PER-RES tag based on the context, unlike in the last sentence. Although through the context it is obvious for a human reader to identify ‘Foschepoth’ as an person with academic background, the model leaves this occurrence untagged. Regarding the detected entities of type AREA-RES, the first one contains three research areas for which it would have been preferable to have each research field tagged as an individual entity. Whereas the second detected entity of the type AREA-RES is fine, the model fails to tag the sequence ‘Neuere und Neueste Geschichte’ (‘recent and modern history’) in the last sentence. The entities of the type ORG-RES are all tagged correctly. This is not the case for the sentences in Figure 2, where the International Space Station ISS is erroneously tagged as ORG-RES and two NASA members



Professor Josef Foschepoth PER-RES Der Historiker und Publizist Josef Foschepoth PER-RES wurde am 19. November 1947 in Werl geboren. Nach dem Studium der Geschichte , Theologie und Sozialwissenschaften AREA-RES mit Staatsexamen sowie Promotion an der Universität Münster ORG-RES folgte eine fünfjährige Tätigkeit als Gymnasiallehrer. Anschließend wurde Foschepoth PER-RES Leiter des neuen Forschungsbereichs „Post-War-History" am German Historical Institute London. ORG-RES 2005 wurde Foschepoth außerplanmäßiger Professor für Zeitgeschichte AREA-RES am Lehrstuhl für Neuere und Neueste Geschichte der Albert-Ludwigs-Universität Freiburg ORG-RES .

Figure 1: Mostly correct predictions of named entities in sentences from article M20/APR.00264 tagged with the fine-tuned NER model (with tags PER-RES in red, AREA-RES in green, and ORG-RES in blue).

are tagged as PER-RES although there is no evidence in the text for their academic background. Due to the sheer size of the subcorpus, no detailed qualitative analysis of the output is made, but it might be worth to invest some effort into finding error patterns in the results and test if targeted additional training data would diminish the amount of incorrectly tagged data. A certainly helpful feature for the NER model would be some kind of score accompanying the tags to indicate the model's confidence in the respective tags.



Washington. Ende Mai will die US-Raumfahrtbehörde Nasa erstmals seit fast zehn Jahren wieder Astronauten von den USA aus zur Internationalen Raumstation ISS ORG-RES schicken. Die Nasa-Astronauten Robert Behnken PER-RES und Douglas Hurley sollen am 27. Mai an Bord eines „CrewDragon" des privaten Raumfahrtunternehmens SpaceX mit einer „Falcon 9"-Rakete vom Weltraumbahnhof Cape Canaveral aus starten , teilte Nasa-Chef Jim Bridenstine PER-RES am Freitag mit .

Figure 2: Incorrect predictions of named entities in sentences from article M20/APR.02952 tagged with the fine-tuned NER model (with tags PER-RES in red and ORG-RES in blue).

6.4 Technical Challenges

It applies to all three methods that there are still improvements to be done regarding run-time and memory consumption. For the selected M20 subcorpus, the three processes were run individually and took between six hours and three days. When processing DeReKo as a whole, this would be multiplied by several thousand times and needed to be optimised, e.g. through parallel computing. The same issue holds for temporary files created during the preprocessing of large I5 files before the three NLP methods can be applied.

7 Conclusion and Future Work

Our experiments show that extracting metadata from linguistically motivated large corpora is possible. The usefulness of this metadata will have to be proven in the future based on possible tools (e.g. for corpus analysis) making use of these bits of information to identify subcorpora.

The results can be applied to all types of national corpora and other large corpora including those that have legal restrictions. The sample corpus is not called a national corpus, but DeReKo serves similar purposes as national corpora for other languages. The method described and implemented here will be usable across all languages and corpora with a similar structure, e.g. containing newspaper texts, articles etc. Though the method was only applied to a defined substantial subset of the corpus, it was created to finally analyse the whole corpus. To apply this to other national corpora, a requirement is the existence of appropriate models for the respective languages and that the corpora are available or can at least be preprocessed to obtain the appropriate structures suitable to the task (topic modeling, NER, keyword extraction), e.g. a segmentation into individual articles, sentences etc.

With the application of this methodology to a corpus such as DeReKo and the perspective of applying it to national corpora, there is also potential for its use within other areas of international research data infrastructures such as CLARIN. Enriching the metadata for textual corpora allows for additional functionalities, including preparing linked data applications based on the metadata involved.

One issue left open so far is the integration of linked data sources for the identified topics, keywords and named entities. In the future we will use GermaNet for words and concepts, ontologies for topics, and authority files for named entities. This will allow us to connect the metadata to the Semantic Web and other forms of knowledge graphs.

The described methods were applied to only a subset of the corpus. One of the next steps is to scale these experiments to the full data set. Running the processes several thousand times will require stable processes and fully automated metadata enrichment. The initial tests on the limited data set were successful.

Another set of intended experiments will look at different types of corpora. For example, a corpus of endangered languages such as DOBES[16] contains many different languages lacking existing NLP models tailored to their specific linguistic characteristics. However, multi-tier annotations, for example for spoken language, often contain a gloss in another language such as English, German or French, for which such models are available. Additionally, metadata often contain a textual description of the data which should also be valuable sources for NLP processes for topic modeling, keyword extraction, and named entity recognition. Hence, we will explore how multi-tier annotated corpora can utilise the same technique for enriching metadata as well.

Acknowledgements

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

References

Angelov, D. (2020). Top2Vec: Distributed representations of topics. https://doi.org/10.48550/arXiv.2008.09470

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C., & Witt, A. (2013). KorAP: The new corpus analysis platform at IDS Mannheim. In Z. Ventulani & H. Uszkoreit (Eds.), *Proceedings of the 6th Language and Technology Conference (LTC'13)* (pp. 586–587). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3261/file/Banski_KorAP_2013.pdf
The following values have no corresponding Zotero field:tertiary-authors: Vetulani, Z., and H. Uszkoreittertiary-title: Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conferencepages: 586-587.

[16] https://dobes.mpi.nl/

Behrens-Neumann, R., & Pfeifer, B. (2011). Die Gemeinsame Normdatei - ein Kooperationsprojekt (Deutsche Nationalbibliothek, Ed.). *Dialog mit Bibliotheken*, *23*(1), 37–40.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., & Trippel, T. (2012, May). CMDI: A component metadata infrastructure. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini, & T. Trippel (Eds.), *Proceedings of the workshop describing language resources with metadata: Towards flexibility and interoperability in the documentation of language resources (LREC'12)* (pp. 1–4). European Language Resources Association. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-108677

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018a). A text feature based automatic keyword extraction method for single documents. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 684–691). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_63

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018b). Yake! collection-independent automatic keyword extractor. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 806–810). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_80

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Delpeuch, A. (2020, November). OpenTapioca: Lightweight entity linking for Wikidata (short paper). In L.-A. Kaffee, O. Tifrea-Marciuska, E. Simperl, & D. Vrandečić (Eds.), *Proceedings of the 1st Wikidata Workshop (Wikidata 2020)*. CEUR-WS.org.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., & Witt, A. (2016). KorAP Architecture - Diving in the Deep Sea of Corpus Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 3586–3591). Paris: European Language Resources Association (ELRA) 2016.

Diewald, N., & Margaretha, E. (2016). Krill: KorAP search and analysis engine (M. Kupietz & A. Geyken, Eds.). *Corpus Linguistic Software Tools. Journal for Language Technology and Computational Linguistics (JLCL)*, *31*(1), 73–90.

Ecker, J. (2024, May). Labeling results of topic models: Word sense disambiguation as key method for automatic topic labeling with GermaNet. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 10014–10022). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.875/

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. https://doi.org/10.1087/20120404

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. https://aclanthology.org/W97-0802

Henrich, V., & Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2228–2235. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf

Hinrichs, M., Zastrow, T., & Hinrichs, E. (2010, May). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA). https://aclanthology.org/L10-1184/

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). SpaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

Language resource management – Persistent identification and sustainable access (PISA) (International Standard). (2011, May). International Organization for Standardization (ISO). Genf. https://www.iso.org/standard/37333.html

Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model (International Standard). (2015, January). International Organization for Standardization (ISO). Geneva.

Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language (International Standard). (2019, July). International Organization for Standardization (ISO). Geneva. https://www.iso.org/standard/64579.html

JSON-LD 1.1 – A JSON-based Serialization for Linked Data (W3C Recommendation 16 July 2020). (2020, July). World Wide Web Consortium (W3C). https://www.w3.org/TR/json-ld/

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (pp. 1848–1854). European Language Resources Association (ELRA) 2010.

Kupietz, M., Diewald, N., Hanl, M., & Margaretha, E. (2017). Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In M. Konopka (Ed.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung. Jahrbuch des Instituts für Deutsche Sprache 2016* (pp. 319–329). de Gruyter.

Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi (Ed.), *Workings Papers in Corpus-based Linguistics and Language Education* (pp. 53–59, Vol. 3). Tokyo University of Foreign Studies 2009.

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2378–2385. http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018, May). The German Reference Corpus DeReKo: New Developments – New Opportunities. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)* (pp. 4353–4360). European Language Resources Association (ELRA).

Lammey, R. (2020). Solutions for identification problems: A look at the Research Organization Registry. *Science Editing*, *7*(1), 65–69. https://doi.org/10.6087/kcse.192

Lüngen, H., & Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, *Issue 3*. https://doi.org/10.4000/jtei.508

Möller, C., Lehmann, J., & Usbeck, R. (2022). Survey on English entity linking on Wikidata: Datasets and approaches. *Semantic Web*, *13*(6), 925–966. https://doi.org/10.3233/SW-212865

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Sakor, A., Singh, K., Patel, A., & Vidal, M.-E. (2020). Falcon 2.0: An entity and relation linking tool over Wikidata. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3141–3148. https://doi.org/10.1145/3340531.3412777

Schwarz, P. (2024). Semiautomatic data generation for academic named entity recognition in German text corpora. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, & M. Wiegand (Eds.), *Proceedings of the 20th conference on natural language processing (konvens 2024)* (pp. 173–181). Association for Computational Linguistics. https://aclanthology.org/2024.konvens-main.20

P5: Guidelines for electronic text encoding and interchange (version 4.2.1. last updated on 1st march 2021, revision 654a5c551) (tech. rep.). (2021). Text Encoding Initiative. Retrieved October 9, 2023, from https://guidelines.tei-c.de/en/html/index.html

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Walkowiak, T., & Piasecki, M. (2015). Web-based natural language processing workflows for the research infrastructure in humanities. *5th Conference of the Japanese Association for Digital Humanities*, 61–63.

Weiß, C. (2005). Die thematische Erschließung von Sprachkorpora. *Mannheim: Institut für Deutsche Sprache. OPAL-Online publizierte Arbeiten zur Linguistik, 1/2005*. https://pub.ids-mannheim.de/laufend/opal/pdf/opal2005-1.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*. https://doi.org/https://doi.org/10.1038/sdata.2016.18