What's in a <unittitle>? Cross-Lingual Topic Detection & Information Retrieval in Archives Portal Europe

1. INTRODUCTION

Archives are traditionally organised not by their subject, but around the entity (person, organisation, body) that created and/or collected the documents composing the archive in the course of their activities. Finding aids in archives start by describing the records creator, to later go on describing the content of the collection's components, most often down to the single files and documents.[1] The starting point of archival research, then, is traditionally also based around the records creator; searching documents on Napoleon will most likely start with a trip to the Archives nationales in Paris, which hold the documents produced by the French state. While provenance remains an undisputed pillar of archival management, the development of online catalogues has caused historical archives operating in a digital environment to require new tools for archival research. In particular, searching by keywords, arguably the most intuitive and most often used feature of online search, has made the challenge of searching by subject, rather than simply by records creator, one of the most vivid in the development of online archival catalogues – after all, searching by keywords very often means searching with specific content in mind.[2] In this scenario, topic association becomes a fundamental feature of online archival research; and automated topic detection an essential tool for archivists and researchers working in an online environment.

This paper presents the results of a proof-of-concept project for an automated topic detection tool developed specifically for Archives Portal Europe (APE), www.archivesportaleurope.net, the portal on archives from and about Europe. APE is an aggregator that connects on a single research point the catalogues and digitised archival material from institutions in more than 30 countries, and from a variety of organisations (such as State archives, city archives, university and parish archives, private institutions, etc). It is maintained by the Archives Portal Europe Foundation, an international consortium of State archives and other archival institutions that aim to connect the archival material of single institutions into one digital catalogue, allowing researchers to access a myriad of archival institutions across Europe through a single research portal. One of the research tools made available by Archives Portal Europe are topics; however, these are currently maintained manually by the archivists, and the vast amount of archival material ingested in the portal, in more than 20 languages, makes it extremely complex to have a comprehensive approach to topics that fits the various archival traditions represented. Because of the multi-institution, multi-country, and multilingual characteristics of the portal, APE is the ideal recipient of an automated topic detection tool, which would both allow users to browse the vast amount of material in a much easier way, and the portal's curators to organise multi-institution and multi-country collections in a more consistent way. The highly challenging scenario presented by APE is also a benchmark for

[1] For a general overview of archival practices, see [22]
[2] For a study on keyword-based research in information retrieval, see [26] and [13]

assessing the current performance of Natural Language Processing (NLP) and Information Retrieval (IR) tools developed for cross-language tasks in a realistic end-user application.

Through a proof-of-concept project conducted by the Archives Portal Europe Foundation and supported by King's College London, we developed an approach for automated topic detection in a multilingual environment, where human-created archival descriptions and associated topic labels act as training data for an algorithm that discovers new relevant material yet to be topically annotated. We used this approach as part of an information retrieval prototype that allows the identification of topically-relevant materials across different languages without employing a machine translation pipeline. The development of such functions was based on supervised machine learning and cross-lingual word embeddings; domain experts contributed to the information retrieval prototype by developing topic taxonomies in different languages which, integrated with the usage of Wikipedia pages, allowed the retrieval of concepts and entities in materials' descriptions across languages. While historical documents have been at the centre of many NLP-based projects, topic detection projects working on archival catalogues (i.e., on metadata), rather than on the actual content of the documents, have not yet been at the centre of any research project, particularly in a highly multilingual scenario such as the one described.

Following a general overview of NLP in Digital Humanities, and a description of how Archives Portal Europe currently uses and produces topics, this paper focuses on outlining the methodology that has been applied for this proof-of-concept. It also presents an initial analysis of the results, which will act as the starting point for the next steps in bringing the proof-of-concept to pilot, alpha and beta phases, with the intent to provide an open source version of the tool during the upcoming years.

2. NATURAL LANGUAGE PROCESSING AND ARCHIVAL RESEARCH

During the last decade, there has been a great focus in Digital Humanities (DH) on the adoption of NLP methods for identifying topics in textual corpora. Such approaches could be divided in two main groups: supervised and unsupervised. Supervised approaches detect topics by relying upon *a)* a predefined and (supposedly) comprehensive list of topics and *b)* a (potentially large and representative) set of materials manually annotated with the relevant topic. To train an algorithm employing for instance a Support Vector Machine (SVM)[8], or a Convolutional Neural Network[9], each document is represented as a single feature-vector, which should capture the "meaning" of its content. To do so, researchers often rely on the use of pre-trained word embeddings[11]; these are vector representation of words, initially obtained from large corpora, that combined together can model longer sequences[11], such as the description of materials in APE. In recent years, research has shown that such pre-trained word embeddings can be aligned across different languages by generating a common "semantic" space[6] [7]. Having each description represented with an embedding and associated with a topic-label, the algorithm will then learn the relation between the embedding and the associated topic. While supervised approaches generally offer reliable performance for topic detection tasks (see for instance the experiments conducted by Merz et al. [15] and Glavas et al [7] on the Manifesto Corpus), in order to learn such vector-label relation they need large amounts of so-called training examples, which are often difficult to collect because they have to be sets of consistently-conducted human annotations.

For this reason, unsupervised approaches for identifying topics in texts, and Latent Dirichlet Allocation (LDA) topic modelling in particular [7], have become highly popular in DH during the last ten years [14] [20]. These approaches do not require any training data or topic labels in advance, and leverage upon similarities in the materials under study (for instance in LDA how words co-occur with each other) to identify underlying patterns in the data, which could correspond to topics. Unsupervised approaches are simple to use and very useful for initial corpus exploration; however, it is very hard to employ them for topic detection especially in cases in which the user already knows which are the topics contained in the collection [21]. This is because results of LDA topic models are often extremely hard to interpret [5], and it is not straightforward to align them with our common notion of topics.

In our research setting, we have a relatively large amount of materials, approximately 2 million documents, at our hands, that is already annotated with topics across different languages (see Table A). Even though this only represents a small part of Archives Portal Europe's dataset and bears the challenge that topics might not have been assigned consistently in all cases (see section 3), we decided to make use of this, and opted for a *supervised approach* for topic detection as a first step. Nevertheless, it is important to keep in mind that the annotations in the test dataset will most likely not cover all topics in our collection, and that our sample of annotated materials might not entirely represent the distribution of topics present in the entire APE collection. Moreover, it is important to take into consideration that employing word embeddings pre-trained on a collection (in our case, Wikipedia) will necessarily embed social, gender and ethnic biases that exist in society (and consequently in the corpus), which might have consequences in the modelling and produced output [4]. For these reasons, readers should consider our experiments only as a first attempt towards a very challenging goal, and we plan to work with an interaction of supervised and unsupervised methods in future experiments, in order to tackle these challenges in a more comprehensive way.

3. AN ARCHIVAL CATALOGUE OF ARCHIVAL CATALOGUES: SEARCHING IN ARCHIVES PORTAL EUROPE

Archives Portal Europe was set up in 2009 by a consortium of 12 national archives in Europe, and went online in 2012, with 14 million descriptive units. As of October 2020, the portal holds 282 million descriptive units, from over a thousand institutions in more than 30 countries which actively provide content to the portal.[3] The main objective of APE is to provide a single research point when looking for archival material, and to display finding aids describing individual documents or collections along with contextual information. Whenever available, the portal provides the link to the individual digital object on the web presence of an institution or another portal, e.g. a thematic or a national aggregator. The scope of the portal is to gather archival material from Europe or about Europe, without limitations on the nationality or type of institution holding the archival material. While the bulk of initial collections belonged to administrative archives based in Europe, mostly national archives, the portal is now expanding to include many different types of private and public archival institutions. The portal currently operates in 24 languages (and five different alphabets) and holds descriptions of more than 590,000 archival fonds and collections, making it the largest aggregator of its kind in the world.

[3] Please refer to the homepage www.archivesportaleurope.net for the latest figures

Approaching an online portal of historical archives means being at the crossroad of the old creator-based way to search an archive, and the new google-like information retrieval way, based on keywords. In order to retain all the characteristics of traditional archival research while simultaneously making use of the new forms of information retrieval enabled by digital technologies, Archives Portal Europe allows for three forms of research. First of all, it is possible to search an archival institution by starting with their holding guides (defined as overviews of the collections and fonds of one archival institution), if applicable, and/or a list of their finding aids (defined as structured descriptions of archival materials per collection or fonds, often up to item level), as it would be possible in any physical institution.

Secondly, the portal allows for keyword search, with words matching the descriptions of the document collections. While the links to the digital objects of the documents are provided whenever available, APE itself only holds the descriptive metadata of the collections, and it does not feature the possibility of text search within the digitised documents: these are held and managed by the single institutions' web presences (and they rarely exist in a fully machine-readable form). This means that the APE search engine only operates on archival descriptions, not the content of the documents. These descriptions are in all the languages represented in the portal, and will be returned as search results provided they match the spelling of the keywords used; in order to maximise the results and help researchers navigating multilingualism, Boolean operators and wildcards are in place to make searching across languages as comprehensive as possible. For example, searching for "**N?pol?* OR 'Ναπολέων Α'' OR 'ნაპოლეონ ბონაპარტი'**" will provide many more results on Napoleon Bonaparte than a standard query for "**Napoleon**", which only encompasses one form of spelling of this specific query.

The third tool of research are the *topics*, and it is the one for which Archives Portal Europe is investing the most. Assigning an archival collection to a specific topic allows to go beyond simple keyword search, to make larger semantic associations on a specific subject, expanding a query not only across different languages, but across different interpretations of a specific research. Topics group together archival collections, from different institutions and records creators, which relate to the same argument. It is possible to start with specific content already aggregated from different archival institutions and different countries, without the need to use specific keywords which may not be included in document descriptions that actually refer to a specific topic – for example, the word "slavery" (in any language that it is used) may not be contained in the descriptions of collections that are highly relevant to the subject; furthermore, even with Boolean operators and wildcards in place, it may be extremely lengthy and repetitive to search keywords related to slavery in all languages represented in the portal.

By being the single entry-point to the European archival heritage, the portal enables new types of digital archival research, freed from geographical limitations and based on cross-country comparison and multilingualism. The portal entirely respects the principles of provenance and original order, and it allows searching by records creator, through high-level holding guides, and through the finding aids of each institution that provides content to the portal[19] [12]. However, the added value of Archives Portal Europe as a new technology for archival research in a digital environment lies in the possibility of searching across multiple archives from different countries and cultures, in a comparative perspective. For this reason, APE strongly promotes searching by keyword and by topic, in order to scrape the portal in a horizontal way, making new connections between archives and the subject of a specific search, ultimately allowing a researcher to find what s/he did not know existed on a specific subject, in archives that were not originally under consideration.

4. TOPIC ASSOCIATION IN ARCHIVES PORTAL EUROPE

While the benefits of topic association are clear for researchers, the implementation of this metadata point in the repository of Archives Portal Europe carries several important obstacles and challenges, which are embedded in the general organisation of the portal. The ingestion and maintenance of the data in Archives Portal Europe is organised through a decentralised approach, in which each single institution providing material to the portal has access to a back-end dashboard to autonomously ingest the material; in the majority of cases, however, a national-level aggregator (e.g. national archives, or national-level portals such as Archives Hub in the UK)[4] take care of the ingestion of multiple institutions in their countries directly. At the moment, topics are assigned manually by the institutions holding the archival materials, or by the national-level aggregator on their behalf; they do so according to their local or national classification systems, their own ways of organising archival material, and their own sensitivity. Furthermore, this approach follows the principles of data sovereignty applied to all metadata provided to Archives Portal Europe, which not always allow for a top-level intervention on the data. At the same time, institutions and national aggregators can only assign topics from a predefined list established by the Archives Portal Europe Foundation at an admin level.

There are currently two ways in which topics can be assigned: via the association to *source guides* or based on the *subject headings* of finding aids. A source guide (or thematic guide) is a special type of holdings guide based on subject rather than on holding institution: in the portal's dashboard, the single institution or national-level aggregator can gather finding aids that refer to the same subject into a source guide. This can be as simple as only naming the titles of the connected finding aids as well as linking out to them for more details, but there are also more elaborate examples, with a content summary of the topic represented by the source guide, as well as short descriptions of each finding aid that is referenced therein. When assigning a source guide to one of the predefined topics in Archives Portal Europe, all components of all the finding aids linked in the source guide will be assigned to this topic. While this is an easy "catch-all" approach, it bears the potential risk of some materials being tagged with a topic that does not specifically apply to these documents per se, but rather to the collection that they are a part of. Furthermore, the source guide approach in itself does not connect the predefined topic terms with related terms that might be present in the archival descriptions.

The latter aspect is addressed by the second way to assign topics, through *subject headings*. Here, the predefined list of topic terms is put in relation to subject headings that are already part of the descriptive metadata of a collection. This e.g. allows to connect the general topic term "Arts" with more specific subject headings such as "Paintings", "Sculpture", "Drawings", etc. as they appear in the archival descriptions. These relationships between topic terms and subject headings can be established on an institutional as well as on a national level, and they enable content providers to connect to national vocabularies and ontologies as well as to institutional rules and guidelines for creating subject headings. With these relationships set up once, every component that includes one of the mapped terms in its subject headings will then be assigned to the central topic as defined in APE (see Figure 01). [2]

[4] https://archiveshub.jisc.ac.uk/

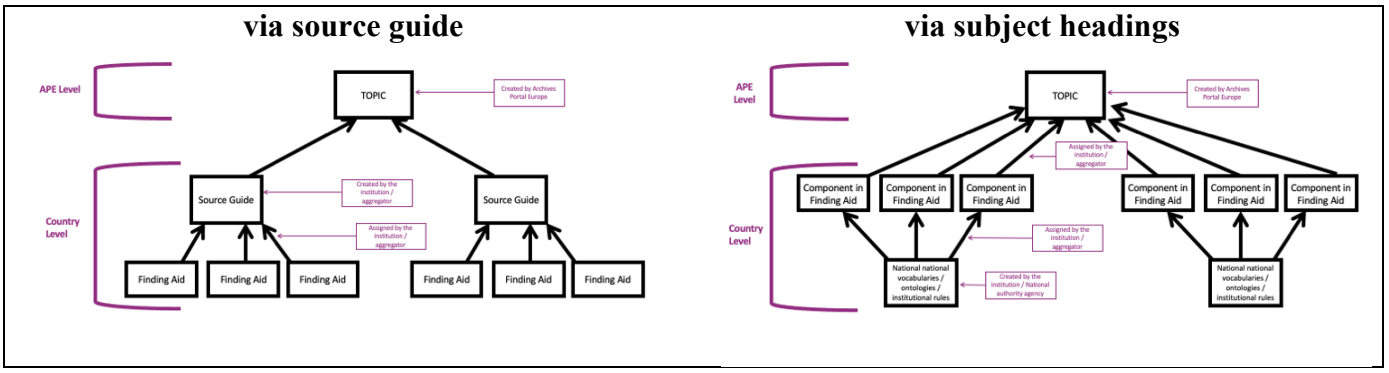

Figure 01 – topic assignment workflow

Topics are currently far from being comprehensive of what is available on Archives Portal Europe on any given subject. On the contrary, several topics are clearly not representative of the whole repository, with only a handful of documents tagged on subjects that are actually strongly present in European history (e.g. the topic "Napoléon I, Emperor of the French" is currently only associated with six descriptive units). This depends on three main factors.

First, the organisation and definition of the topics needs levels of coordination and agreements amongst archives that have been started to be discussed only recently. At the moment, the topic list in use was created by the Archives Portal Europe staff and network, closely following the UK Archival Thesaurus (UKAT), the subject thesaurus created for the archive sector in the United Kingdom[23]. While the UKAT is based on the general UNESCO thesaurus structure[24], it is still rooted in the archival tradition of a specific country, and it does not provide a one-size-fits-all solution[5]. The APE's topic list is being currently reviewed with a more interactive approach by all country representatives, but the new structure will not enter into function for the foreseeable future.

Secondly, the semantic relations between documents and topics are still at the discretion of the single archivist conducting the association, and decisions are often made on the basis of pre-existing archival traditions that may strongly vary from country to country. Some countries have national vocabularies in place, which are used by archivists when creating subject headings, while others do not. Some institutions use national vocabularies alongside their own rules and guidelines in creating subject headings, while others have decided not to use national vocabularies when they originated from libraries and/or museums, or when there is the impression that archives' material could not be represented comprehensively. Lastly, there are institutions that specifically decided against including subject headings in the finding aids, for reasons of resources in keeping up with the task of archival description in general. This impedes the linking of finding aids to topics through semantic affinities.

The combination of these aspects leads to the current state of play, where only a handful of countries has had their metadata available in a way that would have made it easy to connect to the topics at the time of ingestion in Archives Portal Europe, and where most archival institutions do not have the means to conduct topic association manually. While the latter challenge could be overcome with the application of the source guide approach, this is only high-level, and adding the information manually to each archival collection is not feasible with human workforce only, not even via crowdsourced projects.

[5] For example, France follows a series of vocabularies set up by the French Ministry of Culture [18]

At the moment, Archives Portal Europe features 62 topics which vary from very general ones, such as "Economics" or "Education", to specific entities, such as "German Democratic Republic" or "Napoleon I." The size of topic association varies from as little as five to as many as 400 thousand descriptive units being associated with a specific topic; however, this is not reflective of the specificity level of a topic: something as general as "Statistics" includes only 5 associated records; a specific topic such as "German Democratic Republic" includes more than 100 thousands records. France is the only country that has participated significantly in topic association, contributing to 60 existing topics, strongly shaping the outlook of the existing topics architecture; in much smaller measures, Germany has contributed to six topics; Poland to two, and Finland and Latvia to one topic each. In total, less than 2 million documents have been assigned to one or more topics, over a total repository of 280 million; a mere 0.69% of the data available in Archives Portal Europe (see Table A). Manual topic association, whether carried out through crowdsourcing or by professional archivists, has clear limits to what can be achieved in the presence of such a vast and varied repository. Automated topic detection, however, could open the doors to a complete redesign of the research approach, and new accessibility, of the portal.

5. TOPIC DETECTION: BUILDING A NEW TOOL

The first prototype of an automated topic detection tool was developed on a test sample out of the whole dataset of Archives Portal Europe, consisting of a pool of nine topics and a total of 457,000 descriptive units (see Table B for a full list). The topics were selected according to the following criteria:

- Multilingualism: As the vast majority of documents currently tagged with a topic have been contributed by institutions either from France or from Germany, the first challenge was to find a balance between having languages with a big enough representation in the test data to learn from and finding topics that included documents from more than one country, and hence in more than one language, to address the multilingualism that is central in the context of Archives Portal Europe. In most cases, this meant using topics that included documents in French and German, but this criterion also led to including languages that are less tackled in NLP, such as Finnish, Latvian, and Polish.[6]
- Representation: The pool of selected topics tried to represent the variety of topics in Archives Portal Europe. Topics varied in size and scope, from the very specific ("Napoleon I"; "German Democratic Republic") to the very generic ("Economics"); from referring to a specific type of record ("Maps"), to a historical practice ("Genealogy"), to a certain group of

[6] With regard to languages used in the archival descriptions, it is important to notice that most of the times the language of description equals the official language (or languages) of the country where the institution holding the materials is located. This in turn often matches the language of the records themselves, but not necessarily: for example, a medieval charter from the Hungarian National Archives may be described in Hungarian but be written in Latin. Furthermore, there are also cases where not only the documents, but the archival descriptions of those documents are in another language than the (currently) official language of the country. For example, the city archives of Nice, France, hold many archival descriptions in Italian, for historical reasons. The metadata standard currently used in APE only allows for the identification of the language of the actual document, not the language of the description; furthermore, this element is an optional metadata field, making it difficult to rely on the metadata for detecting a document's language. In future versions of the prototype the language of each description will be automatically assessed, in order to generate a new metadata field.

records creators ("Notaries"); from being barely sketched (6 records on "Napoleon I", 765 for "Slavery") to being already a useful tool for research (40,000 on "Genealogy", from three different countries).

- Entities & Concepts: To address two of the main search approaches in archival research, persons/places and subjects/themes, the selected topics also aimed at including entity-based topics such as "Napoleon I" or "German Democratic Republic"; concept-based topics such as "Economics" or "Slavery"; and mixed topics such as "First World War", which contains both entities and concepts.

Using this sample, we developed a prototype to highlight the potential and the challenges of cross-lingual topic detection, as well as a new approach to information retrieval based on this. The tool has four main components: a supervised topic detection algorithm that works across languages; a topic taxonomy of relevant words; two information retrieval functions that allow users to browse for information across languages on a given concept or entity.

5.1 Cross-Lingual Topic Classification

The first step addressed during the development of the prototype was the training of a supervised classifier for topic detection on the collection, considering as "documents" the descriptive units within a finding aid.[7] The developed tool could be used both for enriching materials not already manually labelled and to discover entries that might be relevant to a specific topic beyond the label they are currently associated with (for instance a document under "First World War" but also relevant for the topic "Economics"). In recent years, the NLP community has intensively focused on distributional semantics and word embedding methods in order to move beyond word-frequency approaches to determine text similarity[16]. These methods have become increasingly useful also in cross-lingual scenarios[7], as they allow to capture underlying topic similarities without requiring complex machine translation systems. In our experiments we have employed Fast-Text word-embeddings from all languages present in our dataset[8] and aligned in a common cross-lingual "semantic" space by the project MUSE[6]. We represented the description of each descriptive unit in the collection as the averaged vector of all its words, therefore obtaining a single "document embedding" for each description. Then we trained a supervised topic classifier (a multi-class Support Vector Machine) in a 10-fold cross validation setting, using these document embeddings as feature-vectors, similar to what has been done already by Glavas et al.[7]. This approach achieved really high performance, identifying the correct topical label for the materials in over 90% of the cases.[9] We additionally ensured that the classifier was

[7] A descriptive unit in this context is any unit of archival description that is treated as a potential result by the current search process in Archives Portal Europe. These can be descriptions of the actual records themselves, or descriptions of higher levels, including the collection level. The tool used the Solr results in JSON format for each of these "documents", where some major parts of the archival description are captured in singular fields (e.g., the title of the unit itself or of the upper hierarchical levels that this unit is a part of). However, other parts of the archival descriptions are only included in a placeholder field of the Solr index, capturing all additional metadata that might be part of the original EAD-XML file. This is currently not part of the "document" as used by the tool.

[8] Word Embeddings, pre-trained on Wikipedia, are available on GitHub at this link: <https://github.com/facebookresearch/MUSE>

[9] We obtained over 0.9 of both micro and macro F1-score, which is the harmonic mean of precision and recall. To know more see the documentation of the metrics on Scikit-learn, the library we adopted: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

correctly distinguishing between topics, and not languages, by conducting an in-depth error analysis.

5.2 Topic Taxonomies Generation

In parallel, in order to generate a first list of potentially relevant entities and concepts, we processed each document using the Spacy python library. We extracted all detected entities (people, locations, organisations, etc.) through a Named Entity Recogniser (NER) and searched for the most relevant corresponding entry in Wikipedia (i.e., the entity in Wikipedia that is most frequently mentioned with such name). [10] Then we provided subject-experts with two lists of potentially relevant topical entities: one list of overall most frequent entities for each topic and a second list of most distinctive entities (i.e., entities that are frequent in a topic but not very frequent in others - this was computed using an entity adaptation of Term Frequency - Inverse Document Frequency (TF-IDF), as in Lauscher et al.[10]). Using these lists as starting points, domain experts curated a series of topic taxonomies by adding additional keywords that they thought were relevant to search for a specific topic in the repository, either because frequently used in relation to a topic (e.g. “efficiency” in “Economics”), or because they represented entities that were clearly related to a subject topic (e.g. “Stasi” for “German Democratic Republic”). Domain experts included historians and archivists, thus combining the sensitivity around a topic from both the point of view of researchers, and of records managers; they were also selected amongst different countries, thus expanding the pool of languages present in the taxonomies by Maltese, Italian, English, Norwegian, Spanish, and Portuguese.

5.3 Concept search

Relying on both the cross-lingual classifier and the generated taxonomies, we then developed a function for browsing the collection for concepts across languages; in order to further test the functionalities, we added English and Italian as supported languages for user queries. Given a user query in one of the supported languages, for instance “traité” in French, this would be represented as a cross-lingual “document” embedding, similar to the approach employed above. Then, using the trained topic classifier, we would be able to detect the potentially relevant topics for the query (e.g. “First World War” and “Notaries”). Next, we would rank all materials in the collection by measuring the “semantic” similarity of their vectors with the vector of the query (we used cosine similarity, a common practice in information retrieval).[25] As we use cross-lingual document embeddings, results of this ranking would therefore also be in languages different from the query, but they should supposedly address the same concept; by default we showed only the first 100 retrieved materials, but this can be easily modified in the interface. Wildcards and Boolean operators are currently not supported; nevertheless, because basic algebraic operations are possible in word embedding spaces (see for instance the famous “King - man + woman = queen” in Mikolov et al. [17]) we will explore their usefulness for cross-lingual information retrieval on the APE collection in the future. We finally enriched the presentation of the results by using the constructed taxonomies and suggesting the most relevant topical words and entities, given the user query, as other possible query terms for further exploring the collection. These are selected as they appear in a topical taxonomy relevant for the user query (for instance “First World War”) and are semantically similar to the user query. We plan to

[10] For the first experiment, we focused on French, German, and Polish. A Named Entity Recogniser (NER) is a tool that automatically identifies mentions of entities (such as people, locations, organisations) in a string of text.

further improve this functionality by additionally suggesting topical words that are extracted from the content of the retrieved materials.

5.4 Entity search

As an additional feature we also offer the option to users to search for entities across languages. Instead of relying on cross-lingual embeddings our retrieval function first maps the entity inserted by the user as a query to its equivalent in Wikipedia (when present). Next, it retrieves name variations in the other languages under study,[11] and finally searches for their occurrence in the corpus. While this function is an early prototype and does not fully rely upon entity disambiguation approaches, we found it useful as an additional way of exploring the collection. As for concept search, we provide together with the ranking of potentially relevant documents, an additional list of potentially relevant associated topical words and entities from the taxonomies. In this case, however, we return results only when they contain a mention of the entity (in one of the covered languages), so they might be less than the default cut-off (set at 100 materials).

All of the steps presented in this section are in their early stages; while they will be extended in the future to include other NLP functionalities (e.g., entity disambiguation approaches, pre-trained deep language models) they already show the potential of our strategy for cross-language topic detection and information retrieval. The current status of the prototype, together with future developments, is available on APE's GitHub page[1]; once fully developed, the tool will be released as open source, following the general development principles of APE.

6. TESTING THE TOOL

While the tool was designed to be applied to the complete dataset of Archives Portal Europe, testing on the whole repository would require processing power and memory space beyond the scope of this proof-of-concept. We thus concentrated on a sample of 457,538 descriptive units, corresponding to nine selected topics.

First of all, several keyword searches were conducted for each topic, according to the following criteria:

- The keywords had to be relevant to the topic in question
- One keyword for each topic had to be an entity (ideally a person or a place)
- One keyword for each topic had to be a concept
- One search for each topic had to be a combination of two or more keywords (with the caveat that the tool is not yet working with Boolean operators to distinguish between several keywords being used as a complete text string, and several keywords being used in combination with each other);
- If a keyword could be both a concept and an entity (e.g., "Keynes" as John Maynard Keynes or as in a Keynesian policy), the search would be done both as an entity, and as a concept

[11] We pre-process name variations removing leaving life dates aside for persons or other characteristics sometimes included in brackets.

- Taking into account the language distribution of the test dataset, each search was conducted in French and German, by far the most represented languages in the sample;
- At least one of the other languages under consideration was used: Polish and Finnish, as well as Italian, English, and Slovenian.

For each query, the tool returned the 100 most relevant results or fewer, if the tool could not detect at least 100 results (this was particularly the case of search for entities, as they are narrower). Firstly, we captured how many results were already tagged with the topic that the keyword suggested, and how many were tagged with other topics. As we worked on the assumption that all the material tagged by the archivists operating in the APE dashboard was correctly assigned, we trusted that all results *already tagged* with the topic in question were relevant. We checked instead how many results *not tagged* with the topic in question were indeed relevant. As it would have been impossible to run a thorough testing by checking all the results provided, we checked up to the first 10 results for each of the other topics.[12]

Because the research focussed on topic detection, a result was considered relevant if it was clearly and beyond doubt related to a topic, whether or not the result was also related to the specific search query. For example, when searching "Keynes" with the search interest of the topic "Economics," a document was considered to be relevant when its description referred to Keynes, and furthermore to an economic context, such as a document tagged under the topic "First World War" that related to Keynes' role in the Versailles peace conference. On the other hand, a document referring to Keynes attending a performance given by his wife, Lydia Lopokova, would have been considered not to be relevant to the topic of "Economics", even though it were a match for the entity "Keynes."

Other elements from the results that we captured were the language of the search results; the list of the 10 most relevant topical words suggested by the tool, based on the taxonomies; and how many of the suggested topical words were indeed pertinent to the topic under consideration. In total, 153 keyword queries were conducted.[13]

7. RESULTS OF THE FIRST TESTING

The testing of the tool in this proof-of-concept phase aimed to 1) confirm that the tool does what it is expected to do, and 2) evaluate our approaches to the human input in the process of assigning topics and in the creation of the taxonomies. The results gathered should not be treated as conclusive, as the test dataset only represented a very small sample of the repository of Archives Portal Europe (25% of all documents tagged with a topic, and 0.16% overall); however, they aim at identifying patterns, evaluating the performance of the tool, and establishing a workflow in view of the future scaling up of the project.

---

[12] For example, if a search query related to "Economics" returned 100 results of which 36 were already tagged with "Economics", 34 with "World War I", 26 with "German Democratic Republic", and 4 with "Notaries", we checked the first 10 results from the results' list that were tagged as "World War I", the first 10 results tagged as "German Democratic Republic", and all the results tagged as "Notaries"

[13] Appendix A available online at this link: https://docs.google.com/spreadsheets/d/1MWXJkC6EQjPW8wtf9DSmWnlXorTGJMMMz-AWNDWVL1k/edit?usp=sharing

7.1 Cross-lingual topics classification

An important function of the tool is the identification of documents related to a keyword subject and topic independent from the language. Initial results show that the tool does perform this task, in spite of several hindrances. One of the problematic aspects of the dataset (which is however representative of the overall current portion of tagged documents in the portal), is that the majority of the tagged documents are either in French (37.5%) or in German (32.9%), with the final 30% divided between Polish, Latvian, and Finnish. This meant that the tool was mainly trained on the basis of documents in French or German, which we addressed with the application of Fast-Text word embeddings, and their alignment in a cross-lingual "semantic" space, in order to broaden the tool's functionalities to the other three languages represented, plus two additional ones that are usually part of such multilingual settings (we chose English and Italian, as both languages are strongly represented in Archives Portal Europe amongst the non-tagged documents). In terms of testing, we have followed the breakdown of languages when choosing our search terms, but have aimed at bringing the percentages down, at least to a certain extent, by including further search terms from other languages. Despite these efforts, the result cap at 100 may have also lowered the chances of a document from the "underrepresented" languages being included in the list. Given these premises, it should not be surprising that the results of the searches were mostly in French or German (see Table E).

It is interesting however that there was a higher number of cases where the language of the search term differed from the predominant language of the search results (56.8% of all searches) rather than remaining in the same language context (43.2% of all searches). This occurs more often in entity searches (62.3% return results in other languages) than in concept searches (51.9% returns results in other languages). While more testing will be needed, we hypothesise that this is due to archival descriptions largely re-using the administrative language of the records creators, which is not "standard" nor "natural", and therefore presents additional challenges in NLP.

A second interesting observation is that in terms of changes from one language to another, language families play a role, e.g. English search terms result more often in German search results than in French ones, while Italian search terms result more often in French search results than in German ones. Finnish on the other hand, from the North-Eurasian Uralic languages family, generates search results in a larger variety of languages (there are instances in which a search in Finnish gives twice the results in Italian than it does in Finnish or German).

A third important observation is that connecting documents via semantically similar entities and concepts allowed to retrieve results in other languages than the official one of the country where the institutions are located, for example archival descriptions in Italian held in French institutions, which can be most likely explained with the changing borders and administrations of several territories.

7.2 Most relevant topical words

The findings with regard to the word lists, which the tool currently uses as a list of suggested terms that could also be of interest based on the search that has been conducted, provided mixed results which should be investigated further. The assumption is that this is less to do with the initial extraction of entities and the definition of the most frequent and the most distinctive entities for each topic, but more with the extension of these lists. This step mainly relied on the availability and engagement of the partners of the Archives Portal Europe network and of domain experts to provide additional entities to enhance these lists; while this has provided lists in many different languages, it has also led to the status quo that English, which is not

represented as a language in the documents of the test dataset, makes up nearly 40% of the total 1,654 taxonomy entries. English furthermore is the language with the highest representation in the taxonomies for five out of the nine topics of the test dataset, followed by French and Portuguese as the languages with the highest representation in the taxonomies for three respectively one of the topics. German, on the other hand, only gets to 5% of the taxonomy entries over all, being ranked second for the topics of "First World War" and "German Democratic Republic", for which German also is the predominant language represented in the documents, but only contributing between 0% and 2.3% of the taxonomy entries of the seven remaining topics.
To illustrate this, below are two representations of the topic "First World War". Figure 02 looks at the languages in relation to:

- The topic itself - documents tagged with this topic are predominantly in German (blue), with some smaller part being in French (red);
- The search term used - mainly English (black), French and German, with some searches in Italian (purple) and one in Finnish (orange);
- The most relevant topical words from the taxonomies.

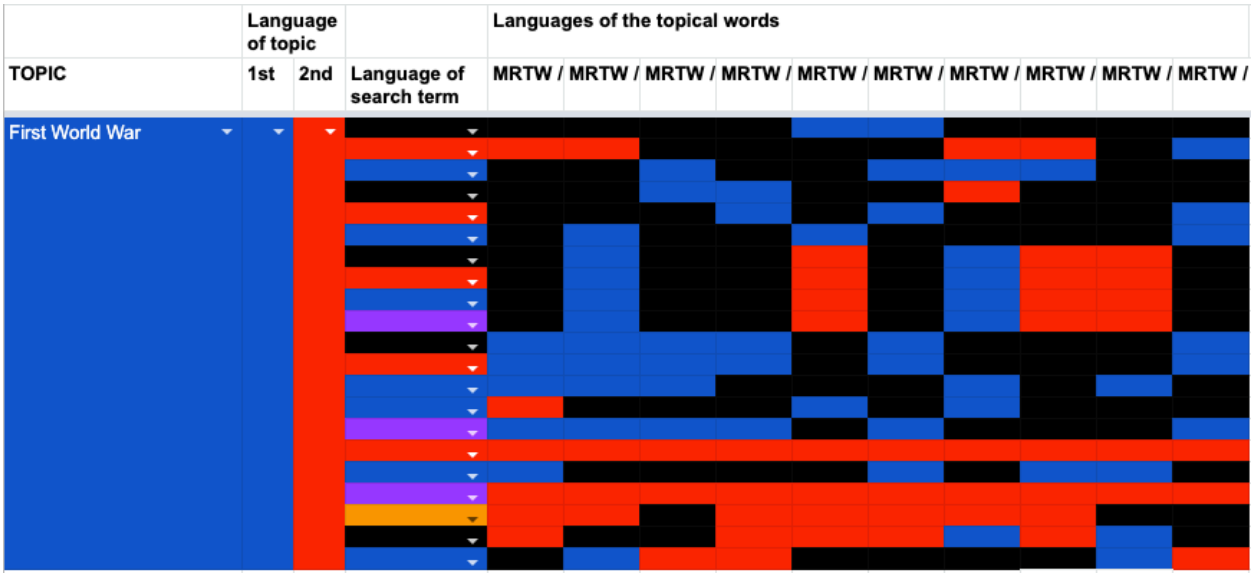

Figure 02 - A visual representation of languages related to the topic "First World War"

While German and French are the languages of the documents and have been used as often as English as the language of the search term, the colour-coded presentation shows a picture where English seems to dominate the topical words. That being said, Figure 03, which looks at the topics represented by the most relevant topical words shows a vastly homogeneous picture for terms related to "First World War" (blue) with the second predominant topic being "German Democratic Republic" (green), possibly fostered by the fact that this topic also has German as its main language. This again seems to speak positively to the cross-lingual topic classification that the tool provides, despite the improvement potential with regard to the taxonomies themselves.

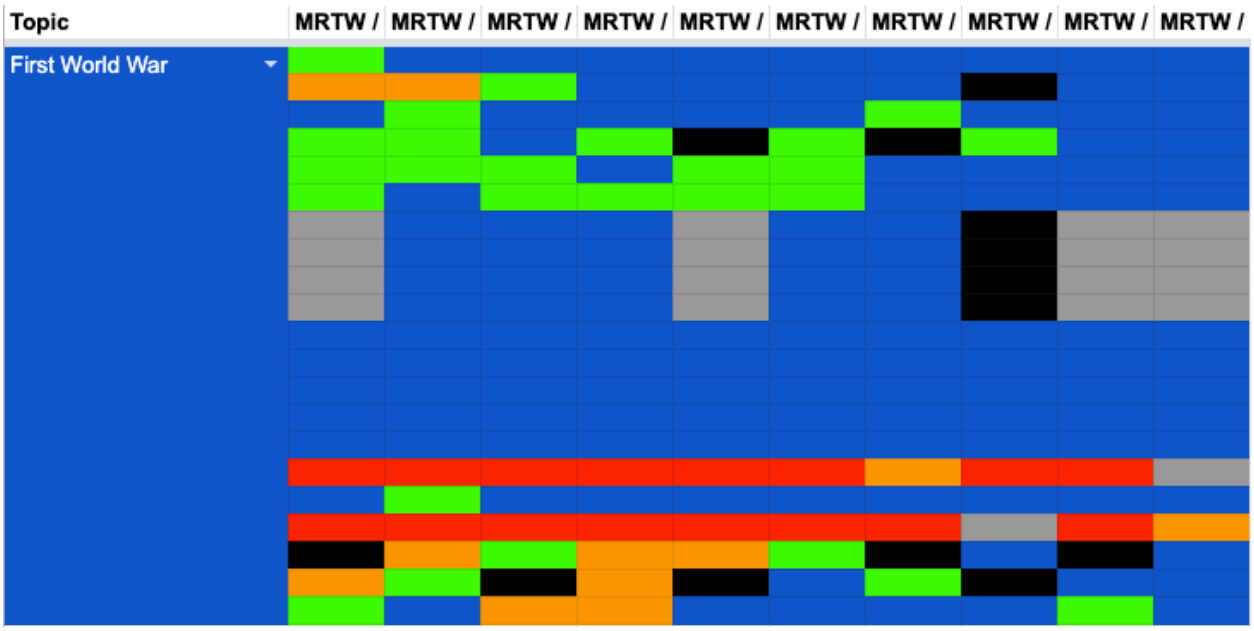

Figure 03 - Topics represented by the most relevant topical words

A second interesting observation with regard to the taxonomies and the most relevant topical words is that there seems to be an inverse relation between the percentage of relevant results in a query, and the number of relevant topical words suggested. As shown in Table C, the higher the number of new relevant results detected by the tool, the lower the number of relevant topical words suggested, and vice versa. While the picture across all nine topics of the test dataset is not entirely conclusive, this is an area to investigate further. It could especially be interesting to study 1) if there is a certain correlation between the type of topic (general vs. specific, entity vs. concept, etc.) and these results: at the moment, the two topics that show the highest percentage of relevant taxonomy entries are both entity-related: “German Democratic Republic” as a geographic entity as well as a collection of organisational and personal entities and “Notaries” as a combination of a functional entity as well as a personal entity; 2) if this inverse relation continues to hold when enlarging the test dataset; 3) if this, again, depends on the way in which the taxonomies are created.

7.3 Concept searches and entity searches

The testing included a relatively even number of concept (78) and entity (75) searches. Entity searches had an average success rate (as in, number of relevant new results) of 11.26%; concept searches the slightly lower value of 10.56% average success rate, but with much lower variance. However, the lack of Boolean operators and wildcards in the tool is currently affecting these results, as a multiple terms search does not allow the tool to consider the search as a combination of words, or as a complete string. For example, when searching for the Treaty of Versailles in the context of “First World War”, many irrelevant results belonged to the topic “Notaries”, because focussed on treaties of various kind, rather than the specific peace treaty in question. The next phases of the project will include the implementation of Boolean operators and wildcards to allow for the combination of several keywords and to compensate for different spellings.

One unexpected result related to entity searches. The initial assumption was that when searching for an entity, the results offered by the tool would be the same, independent from the language. As shown in Table D, this was often not the case. One hypothesis to explain this is the

the discrepancies in Wikipedia pages in different languages (e.g., a person might have their life dates included in their Wikipedia name in one language, but not in the other, or might have academic or other honorary titles added in one case, but not in all). Throughout this proof-of-concept phase we tried to address some of these discrepancies, but others may be playing an important role as well. Furthermore, different entities with the same or very similar names present in Wikipedia, are not disambiguated yet by the tool. This is for example the case with searching for "Wilhelm von Preußen" (DE), which returned one result more compared to searches for "Wilhelm, German Crown Prince" (EN), for "Guillaume de Prusse" (FR), or for "Guglielmo di Prussia" (IT), because in the German language the tool currently also included results relating to his great-grandfather "Wilhelm I. (Deutsches Reich)".

8. CONCLUSIONS

This first round of testing was aimed at checking the performance of the tool in its proof-of-concept phase, identifying areas of further investigation, and establishing a methodology for developing and testing the tool further. Our initial findings gave promising result; in spite of the unevenness of the sample (both in terms of topic representation and of languages included), and the vastness of the possible keyword searches around each topic, a 10% success rate (evenly distributed between concepts and entities) is a generally positive result, which allowed to identify relevant documents related to one topic well beyond the narrowness of direct keyword matching. In spite of the noise to sift through, even one extra document can be useful for research in this multilingual environment. Initial findings have also shown that good results levels come also from not very largely annotated topics (e.g. "Catholicism" only has 1,500 tagged documents, but a success rate of 31.5%), which opens the door for smaller scale projects to extend existing topics and potentially create new ones. This success rate also confirmed the feasibility of the project, and the overall soundness of the tool architecture.

With the proof-of-concept confirmed, a closer look to the results allowed us to elaborate on the next steps of the project. The first line of action for the next phase will be to enlarge the sample both in terms of topics under consideration, and of available languages. Secondly, taxonomies will be improved by combining machine-based extraction and human-based input more interactively and more iteratively. Thirdly, new functionalities will be added to the tool, starting from features that improve retrievability:

- Boolean operators;
- Wildcards;
- Disambiguation of entities;
- The inclusion of other data from the documents, e.g. dates, that might be useful in determining whether a search result is relevant or not.

Furthermore, we will include features that prepare for the full integration of the tool within the functionalities of the portal:

- For each result, the provision of a link to the full description in Archives Portal Europe, so that a user could check all details of a result in order to decide on its relevance;
- Allowing to flag in the tool the result as being part of a topic, when relevant;
- Designing a more user-friendly graphical user interface.

Finally, the next stage of the project will start to reflect on various user scenarios for the tool: firstly, the possibility of lowering or increasing the accuracy of retrieval according to the different purposes for which the tool can be used. A user doing historical research in APE may be more inclined to see a different or additional perspective on her/his research topic and will be more willing to accept higher noise levels in the results. Content providers wanting to assign document collections to a specific topic will want to be more precise in order to minimise the possibility of making mistakes in the tagging.

In second instance, and with a more long-term research approach to the project, combining the tool with Linked Data approaches, such as adding URIs to the current literals, may open to whole new lines of research, and of usefulness of Archives Portal Europe as a tool of archival research in a digital environment.

9. BIBLIOGRAPHY

[1] Archives Portal Europe Foundation. 2020. *ArchivesPortalEuropeFoundation/Topic-Detection*. Archives Portal Europe Foundation. Retrieved October 29, 2020 from https://github.com/ArchivesPortalEuropeFoundation/Topic-Detection

[2] Archives Portal Europe Foundation. How to use topics - Archives Portal Europe Wiki. Retrieved October 29, 2020 from http://wiki.archivesportaleurope.net/index.php/How_to_use_topics

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]* (May 2020). Retrieved October 28, 2020 from http://arxiv.org/abs/2005.14050

[5] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (eds.). Curran Associates, Inc., 288–296. Retrieved October 28, 2020 from http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf

[6] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. *arXiv:1710.04087 [cs]* (January 2018). Retrieved October 28, 2020 from http://arxiv.org/abs/1710.04087

[7] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-Lingual Classification of Topics in Political Texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, Association for Computational Linguistics, Vancouver, Canada, 42–46. DOI:https://doi.org/10.18653/v1/W17-2906

[8] Thorsten Joachims. 2002. *Learning to Classify Text Using Support Vector Machines*. Springer US. DOI:https://doi.org/10.1007/978-1-4615-0907-3

[9] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 1746–1751. DOI:https://doi.org/10.3115/v1/D14-1181

[10] Anne Lauscher, Pablo Ruiz Fabo, Federico Nanni, and Simone Paolo Ponzetto. Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability. 2, 2 , 22.

[11] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]* (May 2014). Retrieved October 28, 2020 from http://arxiv.org/abs/1405.4053

[12] Ana María López Cuadrado. 2018. Archives Portal Europe: los trabajos de normalización archivística en el ámbito europeo y su influencia en el acceso e intercambio de información. *TRIA* 22, (2018), 49–65.

[13] Thomas Mann. 2008. Will Google's Keyword Searching Eliminate the Need for LC Cataloging and Classification? *Journal of Library Metadata* 8, 2 (June 2008), 159–168. DOI:https://doi.org/10.1080/10911360802087366

[14] E. Meeks and S.B. Weingart. 2012. The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities* 2, 1 (2012), 1–6.

[15] Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics* 3, 2 (April 2016), 2053168016643346. DOI:https://doi.org/10.1177/2053168016643346

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (eds.). Curran Associates, Inc., 3111–3119. Retrieved October 28, 2020 from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[17] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Atlanta, Georgia, 746–751. Retrieved October 28, 2020 from https://www.aclweb.org/anthology/N13-1090

[18] Ministère de la Culture et de la Communication. Les Vocabulaires du Ministère de la Culture et de la Communication. Retrieved October 29, 2020 from http://data.culture.fr/thesaurus/static/en-savoir-plus

[19] Marta Musso and Arnold Kerstin. 2020. An Archival Repository of Archival Repositories: integrating metadata in Archives Portal Europe. *Moderna arhivistika* III, 1 (2020), 120–139.

[20] Federico Nanni, H. Kümper, and S.P. Ponzetto. 2016. Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations. *International Journal of Humanities and Arts Computing* (2016).

[21] Trevor Owens. 2012. Discovery and Justification are Different: Notes on Science-ing the Humanities. *Trevor Owens*. Retrieved October 28, 2020 from http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/

[22] The National Archives (UK). 2016. Archive Principles and Practice: an introduction to archives for non-archivists. Retrieved October 29, 2020 from https://www.nationalarchives.gov.uk/documents/archives/archive-principles-and-practice-an-introduction-to-archives-for-non-archivists.pdf

[23] UK Archval Thesaurus. UKAT - UK Archival Thesaurus | Home. Retrieved October 29, 2020 from https://ukat.aim25.com/

[24] UNESCO. UNESCO Thesaurus. Retrieved October 29, 2020 from http://vocabularies.unesco.org/browser/thesaurus/en/

[25] Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th Annual ACM SIGIR Conference on Research and Development in Information Retrieval - Full Papers (SIGIR 2015)*, ACM; New York, NY, 363–372. Retrieved from $$Uhttps://lirias.kuleuven.be/retrieve/322606$$DVulicMoensSIGIR2015Final.pdf [Available for KU Leuven users]

[26] Caroline Williams. 2006. 2 - Principles and purposes of records and archives. In *Managing Archives*, Caroline Williams (ed.). Chandos Publishing, 3–33. DOI:https://doi.org/10.1016/B978-1-84334-112-3.50002-9

Table A – Topics in Archives Portal Europe as of October 2020

Topic	N. of tagged documents	Countries
Agriculture	35,677	France
Architecture	78,145	France; Germany
Armed forces	23,068	France
Arts	4,093	France
Buildings	88,178	France
Catholicism	1,499	France
Charity	321	France
Charters	3,331	France; Germany
Church records and registers	2,056	France
Churches	721	France
Colonialism	1,130	France
Communism	433	France
Concentration camp	43,016	France; Germany
Crime	29,970	France
Culture	89,248	France
Democracy	29,829	France; Germany
Early modern period	58	France
Economics	144,157	France; Germany
Education	94,914	France
European Union	15,277	France
First World War (1914-1918)	57,445	France; Germany
French Revolution (1789-1799)	615	France
GDR (German Democratic Republic)	117,268	Germany
GDR parties and trade unions	23,029	Germany
Genealogy	43,792	France; Poland; Latvia
Genealogy archives	13,763	France
Health	53,966	France
Heresy	6	France
Industrialisation	56,793	France
Justice	91,334	France
Lifestyle	28,588	France
Maps	57,119	France; Finland
Medical sciences	21,637	France
Medieval period	3,447	France
Monasteries	204	France
Municipal government	27,088	France
Music	11,172	France
Napoléon I, Emperor of the French, 1769-1821	6	France

Napoléon III, Emperor of the French, 1808-1873	4,641	France
National administration	45,395	France
Notaries	35,487	France; Poland
Photography	149,697	France
Politics	41,764	France
Population cencuses	629	France
Poverty	13,059	France
Protestantism	15	France
Religion	7,545	France
Revolutions of 1848	6	France
Royalty	658	France
Schools	73,112	France
Science	94,465	France
Second World War (1939-1945)	32,169	France
Slavery	765	France
Social history	1,093	France
Socialism	15	France
Statistics	11	France
Taxation	30,621	France
Trade unions	21,980	France
Transport	97,417	France
Universities	11,682	France
Wars (events)	10,270	France
Women	6,390	France

Total	1,792,908
Total archival descriptions (at 13 October 2020)	282,110,269

Table B – Topics selected for the first round of testing

Topic	N. of tagged documents	Country(ies)	N. of institutions
Catholicism	1,499	France	1
Economics	144,157	France	7
First World War (1914-1918)	57,445	France; Germany	7
Genealogy	43,792	France; Poland; Latvia	7
GDR (German Democratic Republic)	117,268	Germany	1
Maps	57,119	France; Finland	8
Napoléon I, Emperor of the French, 1769-1821	6	France	2
Notaries	35,487	France; Poland	7
Slavery	765	France	1
Total	457,538		

Table C – Overview of the results

Topic	N. of tagged documents	Countries	N. of institutions	N. of results (total)	N. of results (checked)	New relevant results	Already tagged results	N. of topic words (total)	Relevant topic words
Catholicism	1,499	France	1	1,674	540	31.48%	1.79%	220	1.36%
Economics	144,157	France	7	1,656	584	24.32%	18.60%	180	6.67%
First World War (1914-1918)	57,445	France; Germany	7	1,278	223	6.73%	40.53%	210	60.00%
Genealogy	43,792	France; Poland; Latvia	7	1,000	285	16.6%	1.90%	100	6.00%
GDR (German Democratic Republic)	117,268	Germany	1	1,226	63	0.00%	81.40%	160	79.38%
Maps	57,119	France; Finland	8	705	101	87.13%	70.64%	90	51.11%
Napoléon I, Emperor of the French, 1769-1821	6	France	2	1,655	531	7.72%	0.00%	190	13.68%
Notaries	35,487	France; Poland	7	903	93	1.08%	70.10%	120	78.33%
Slavery	765	France	1	903	370	3.51%	37.21%	120	45.00%
TOTAL	457,538								

Table D – Summary of the keyword searches by Topic

Topic: "Catholicism"	
Number of keyword queries	22
Searched for the following entities / concepts (translated to English)	Solidarność, nicean, pope, Marian, Hyperdulia, Holy Inquisition witches
Number of total results	1,674
Number of retrieved results already tagged as "Catholicism"	30
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	170 on 540 checked (31.5 % of the sample)
Number of relevant topical words	3 over 220 (1.36%)
Number of times that entity search in different languages did not give the same results	2 out of 4 (Solidarność; Holy Inquisition witches)
Topic: "Economics"	
Number of keyword queries	20
Searched for the following entities / concepts (translated to English)	Keynes, Bank of France, Marxist, Spanish GDP
Number of total results	1,656
Number of retrieved results already tagged as "Economics"	308
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	142 on 584 checked (24.3% of the sample)
Number of relevant topical words	12 over 180 (6.7%)
Number of times that entity search in different languages did not give the same results	2 out of 3 (Keynes, Spanish GDP)
Topic: "First World War"	
Number of keyword queries	21
Searched for the following entities / concepts (translated to English)	Great War, Liège, Triple Alliance, Wilhelm German Crown Prince, Treaty of Versailles, mustard gas
Number of total results	1,278
Number of retrieved results already tagged as "First World War"	518

Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	15 out of 223 checked (6.7% of the sample)
Number of relevant topical words	126 out of 210 (60%)
Number of times that entity search in different languages did not give the same results	1 out of 2 (Wilhelm German Crown Prince)
Topic: "Genealogy"	
Number of keyword queries	10
Searched for the following entities / concepts (translated to English)	Registry Office, family tree, father
Number of total results	1,000
Number of retrieved results already tagged as "Genealogy"	19
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	47 on 285 checked (16.6 % of the sample)
Number of relevant topical words	6 over 100 (6%)
Number of times that entity search in different languages did not give the same results	1 out of 1 (Father)
Topic: "German Democratic Republic (GDR)"	
Number of keyword queries	17
Searched for the following entities / concepts (translated to English)	Erich Honecker, Schabowski, Hohenschönhausen, Fall of the Berlin Wall, Stasi Records Agency
Number of total results	1,226
Number of retrieved results already tagged as "German Democratic Republic (GDR)"	998
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	0 out of 63 checked (0% of the sample)
Number of relevant topical words	127 out of 160 (79.4%)
Number of times that entity search in different languages did not give the same results	1 out of 3 (Hohenschönhausen)
Topic: "Maps"	
Number of keyword queries	17
Searched for the following entities / concepts (translated to English)	Ptolemy, Gerardus Mercator, (only) Mercator, topographical map, map AND town

Number of total results	705
Number of retrieved results already tagged as “Maps”	498
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	88 out of 101 checked (87.1% of the sample)
Number of relevant topical words	46 out of 90
Number of times that entity search in different languages did not give the same results	2 out of 3 (Ptolemy and Mercator)
Topic: Napoléon I, Emperor of the French, 1769-1821	
Number of keyword queries	22
Searched for the following entities / concepts (translated to English)	Napoleon, Napoleon and France, Napoleon Russia, Empress Joséphine Martinique, Saint Helena, Waterloo battle, Nouveau Régime, Bonapartian
Number of total results	1,655
Number of retrieved results already tagged as “Napoléon I, Emperor of the French, 1769-1821”	0
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	41 on 531 checked (7.7% of the sample)
Number of relevant topical words	26 over 190 (13.7%)
Number of times that entity search in different languages did not give the same results	2 out of 6 (Napoleon; Saint Helena)
Topic: “Notaries”	
Number of keyword queries	12
Searched for the following entities / concepts (translated to English)	Rue Saint-Honoré, Notary, Notary AND testament, authentication
Number of total results	903
Number of retrieved results already tagged as “Notaries”	633
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	1 out of 93 checked (1.1% of the sample)
Number of relevant topical words	94 out of 120
Number of times that entity search in different languages did not give the same results	0 out of 1

Topic: “Slavery”	
Number of keyword queries	12
Searched for the following entities / concepts (translated to English)	Spartacus, encomienda, slave, Slave traffic port
Number of total results	903
Number of retrieved results already tagged as “Slavery”	336
Number of new relevant results in the checked sample (that it, relevant to the topic but tagged under other topics)	13 on 370 checked (3.5% of the sample)
Number of relevant topical words	54 over 120 (45%)
Number of times that entity search in different languages did not give the same results	0 out of 1

Results are shown as aggregated for each topic. For a complete overview of each single keyword search, please refer to Appendix A, available online <https://docs.google.com/spreadsheets/d/1MWXJkC6EQjPW8wtf9DSmWnlXorTGJMMMz-AWNDWVL1k/edit?usp=sharing>

Table E – Results by language of search

Language of the search	Number of searches	% of overall searches	Language of the majority of search results	Number of times when language dominant in search results	Percentage
English	17	11.11%	English	0	0.00%
Finnish	8	5.23%	Finnish	2	1.31%
French	49	32.03%	French	66	43.14%
German	48	31.37%	German	66	43.14%
Italian	14	9.15%	Italian	4	2.61%
Polish	13	8.50%	Polish	0	0.00%
Slovenian	4	2.61%	Slovenian	0	0.00%
Total	153		Total	138	
			(search without results)	15	n/a