Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269
Date of acceptance: 16-Sep-2023

Identification of social scientifically relevant topics in an interview repository

A natural language processing experiment

Judit Gárdos, Júlia Egyed-Gergely and Anna Horváth
Centre for Social Sciences, Budapest, Hungary

Balázs Pataki
Department of Distributed Systems, Institute for Computer Science and Control, Budapest, Hungary

Roza Vajda
Centre for Social Sciences, Budapest, Hungary

András Micsik
Department of Distributed Systems, Institute for Computer Science and Control, Budapest, Hungary

Abstract

Purpose: The present study is about generating metadata to enhance thematic transparency and to facilitate research of interview collections at the Research Documentation Centre, Centre for Social Sciences (TK KDK) in Budapest. It explores the use of artificial intelligence in producing, managing and processing social science data and its potential to generate useful metadata to describe the contents of such archives on a large scale.

Study design/methodology/approach: We combined manual and automated/semi-automated methods of metadata development and curation. We developed a suitable domain-oriented taxonomy to classify a large text corpus of semi-structured interviews. To this end, we adapted the

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

European Language Social Science Thesaurus (ELSST) to produce a concise, hierarchical structure of topics relevant in social sciences. We identified and tested the most promising NLP tools supporting the Hungarian language. The results of manual and machine coding will be presented in a user interface.

Findings: The study describes how an international social scientific taxonomy can be adapted to a specific local setting and tailored to be used by automated NLP tools. We show the potentials and limitations of existing and new NLP methods for thematic assignment. Current possibilities of multi-label classification in social scientific metadata assignment are discussed, i.e. the problem of automated selection of relevant labels from a large pool.

Originality/value: Interview materials have not yet been used for building manually annotated training datasets for automated indexing of scientifically relevant topics in a data repository. Comparing various automated indexing methods this study shows a possible implementation of a researcher tool supporting custom visualizations and the faceted search of interview collections.

Keywords: sociology, research data repository, natural language processing (NLP), thesaurus, multi-label classification, exploratory UI, text visualization

Introduction

In recent years, the various dimensions of the emerging open science framework have had a lasting impact on discussions about archival practices. Most notably, from the point of view of creating suitable and sufficient information for accessing archival documents, the FAIR guiding principles (Findable, Accessible, Interoperable and Reusable, Wilkinson et al, 2016) of scientific data management and stewardship have been gaining considerable ground. Metadata, tailored to specific needs, plays a vital role in making the FAIR principles a reality in oral history archives as well as other digital data environments. The ‘digital turn’ has transformed many aspects of scientific research too, including our ways of engagement with its processes and results.

Digitalization has brought on a new era not only in conducting sociological and historical research, but also in conserving and making such research available for the interested audiences. Thus the discipline of digital curation has emerged. Data curators and researchers, possessing accumulated knowledge in their specific field and, ideally, also in other sciences, play a vital role in providing access to research data (Chateau et al, 2012). Digital data curation and research focuses on the interplay between the professional, academic and technical dimensions of the research process in order to ensure that information created digitally remains accessible and usable (Higgins, 2018).

In this article, we explore how the thematic analysis and sharing of qualitative research data can be achieved in a multi-language setting. The European social scientific research field is a highly interconnected one through numerous international funding schemes, yet at the same time also fragmented due to limitations caused by language barriers. Most prominently, non-English qualitative research data, such as interviews, cannot be used in a way as the FAIR principles of data sharing would suggest. FAIR principles describe and prescribe technical characteristics of research

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

archives, however, the actual ways of sharing data in smaller languages is still an unresolved issue – which has been tackled recently in some projects concerning data in other small languages (like Finnish, Skenderi et al, 2021).

Our archivists at the Research Documentation Center (*Kutatási Dokumentációs Központ,* KDK[1]) at the Centre for Social Sciences in Budapest, Hungary, have been long working towards more and better metadata for the collections to serve researchers and connect with other similar enterprises worldwide. KDK is a social scientific data repository collecting, organizing and making digitally available the documents (interviews, questionnaires, survey data, etc.) generated during research conducted in the four institutes of the Centre for Social Sciences, the largest social scientific research facility of the country. Its digital holdings cover various disciplines, including political science, sociology, minority studies and law. KDK also hosts the Voices of the 20th Century Archive and Research Group[2], which collects, digitizes and curates the dispersed materials of qualitative sociological research in Hungary between 1960 and 2010 (mainly interview transcripts but also photographs, drawings, video interviews, notes, study drafts, etc.). The two online repositories together make tens of thousands of digital files available free of charge for researchers and other interested audiences. In addition to providing repository services, KDK also conducts research projects, participates in policy making and actively promotes a culture of data management.

An opportunity in testing new methods to explore what data sharing could look like for qualitative social scientific research data appeared with a nationwide EU-funded project encouraging the use of artificial intelligence in scientific research.[3] We began in 2020 by laying the groundwork for a project that would, in the coming years, grow to be a comprehensive effort to add contentual metadata to the documents and collections in our archives. The staff has worked in cooperation with the computer scientists at the Institute for Computer Science and Control (*Számítástechnikai és Automatizálási Kutatóintézet*, SZTAKI) in Budapest, Hungary, to construct an AI-based system using Natural Language Processing that would automatically create topic-related metadata for every interview in our archives. Given the huge number of interviews stored already and the prospect of our collections growing further in the future, such an endeavor, if done manually by our archivists, would be next to impossible to accomplish due to the lack of human capacity, and would yield insufficient results anyway, lacking in coherence. A kind of time lag in processing collections constitutes a general problem shared by libraries and other digital infrastructures all over the world. Instead of doing the work, many academic databases and publishers maintain a taxonomy that authors or editors reference in order to manually assign topics, research fields, or concepts to scientific publications. However, manual labeling is notoriously laborious and error-prone. To be sure, in the past two or three years there has been some work done on the machine-aided assignment of topics for scientific literature and the creation of metadata for research data collections (Toney-Wales and Dunham, 2022; Suominen et al, 2022). As opposed to our project, such innovations mostly focus on the contextual metadata of research data collections and rarely concern the very texts of primary research documents.

[1] https://kdk.tk.hu/en
[2] https://20szazadhangja.tk.hu/en
[3] Artificial Intelligence National Laboratory, https://mi.nemzetilabor.hu/

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Oral history in digital archives

Our archives at KDK contain thousands of sociological interviews conducted using a vast array of methods: narrative life stories, semistructured and structured interviews as well as collective or focus groups interviews. In this diverse landscape, oral history interviews form the core of our collections, representing a crucial 'genre' in the history of Hungarian qualitative sociology. Being the least structured of its kind, oral history interviews pose the greatest challenges in terms of indexing and systematically mapping their content. Hence, our theoretical focus lies with this form of thematically versatile accounts of life histories, available as transcriptions of records of quasi-spontaneous speech, which syntactically, or even grammatically, constitute the least coherent type of interviews, not to mention other kinds of research documents.

Due to its specific nature, oral history proves to be a cumbersome resource to use. Related issues are well spelt out by the oral historian Alessandro Portelli (Portelli, 1991, p.7), who notes that he is trying to: "*convey the sense of fluidity, of unfinishedness, of an inexhaustible work in progress, which is inherent to the fascination and frustration of oral history – floating as it does in time between the present and an ever-changing past, oscillating in the dialogue between the narrator and the interviewer, and melting and coalescing in the no man's land from orality to writing and back*." This particular complexity proves to be the crux with regard not only to analyzing oral history resources but to making them findable and accessible as well.

The way in which oral history has been represented as an entity throughout (its) history has largely been mediated by technology, so our very understanding of what oral history actually is has been changing with time, partly based on the tools at the disposal of its practitioners' (Boyd and Larson, 2014). One could assume that with improved web-based access, oral history organically becomes a resource for anyone and everyone with a viable internet connection. Digitalization has certainly brought along changes not only in its form and representation but also a major shift in the accessibility and significance of oral history as a research method and a type of archival material or resource. With technology becoming more and more affordable and accessible in recording as well as storing and analyzing interviews, oral history practice rapidly increased. Thus oral histories grew in number and also became more diverse in their form. This rapid growth left archivists grappling with the practicalities of preserving, and providing access to, materials in their collections. To be sure, some formats (like audio or video recordings) have already traditionally been more difficult and expensive to curate in an archival setting. The technological advance has provided solutions to better preserving and processing data, while the threats of incompatibility and obsolescence have only been growing. Digitalization can potentially save oral history from perishing but it takes more to protect it from the other great challenge, namely obscurity (Boyd and Larson, 2014).

The nearly instant access to digital archives and archival materials via the internet is changing the primary function of oral history archives. Once mainly responsible for storing and preserving valuable documents and providing access to sophisticated audiences, they have become vehicles that present history to a much broader set of users from almost every part of the world (Cohen, 2013). However, obscurity is something digital oral history archives need to escape to fulfill their role – even more so in case of non-English oral history sources. If a digital collection is placed online, yet the interface for accessing the interviews is not easy to use, the repository may have

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

theoretically increased the potential audience of its archival materials, however, not functionally, as the access it provides will be closer to analogue access models represented by boxes of tapes or stacks of printed transcripts (Boyd, 2013).

When improving accessibility it has to be considered that the ever transforming digital world calls for a computational way of interaction and knowing (Cohen, 2013). There are formats within the archival realm which have coped better with the opportunities and barriers presented by the digital turn. Still, it is generally true that softwares designed to offer access to archived materials are not sensitive enough to specific challenges posed by oral histories (Boyd, 2013).

The purpose of oral history metadata is to ensure access to (i.e. information about) both the content and the context of oral history interviews and collections. This implies that it is vital to have metadata on multiple structural levels of an oral history archive as well as at all stages of the oral history life cycle. Metadata range from more or less detailed concepts (describing content) and information about the format of the digital object (technical metadata) to those related to the big picture (contextual data facilitating discoverability and patron use). Every community needs an approach that balances the special nature of oral history with its local mission and actual goals and resources; one that acknowledges oral history interviews to become part of a repository's or organization's collection (Kata et al, 2020:3). Therefore the question of how to create suitable and sufficient metadata for oral history documents remains a persisting problem, whose terms change with the social and the technological contexts.

In comparison with other types of personal documents or especially professional literature that are easier to be indexed and provided with precise metadata, oral history documents, including transcripts, pose a complex challenge when it comes to making them visible and accessible. Even if using perfect transcripts accessibility has limitations, as digital full text searches do not always offer all the relevant interviews. This is because keyword searching of verbatim results often fails in mapping natural language conversations, that is, arriving at meaningful and descriptive concepts. This is where social science experts, who can identify social phenomena relevant in a text, come into the picture: they may be able to uncover issues not spelt out explicitly, i.e. themes and topics expressed in a circumstantial manner, or even by way of characteristic omissions.

Researchers visiting interview archives have to be inventive and use a variety of methods in mapping collections and scrutinizing texts to select the documents they need for their work. Browsing the contents using the text search function is of course very useful when the researcher is interested in a very specific concept or a particular location. However, once someone is interested in a more general phenomenon or topic, they are forced to first "translate" it into various commonly used related terms that they assume may appear literally in the texts. In the case of a rather heterogeneous set of texts consisting mostly of unstructured interviews, it is no small task to "shoot" a topic with specific terms. For example, when looking for social mobility, one can start their search with words like 'poverty', 'money', 'education', and so on. Yet it would be pointless to expect to find all the relevant documents for each research question by blindly hitting synonyms and connoted terms related to specific concepts. By the same token, if a researcher wants to learn about the history of, say, the capital city of Hungary, they should not be satisfied with a simple search for the word 'Budapest'. Instead, they would have to try an inexhaustible list of place names,

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

including those of the different districts of the capital, its famous streets or landmarks found in Budapest – an impossible task indeed. At the same time, there may be a legitimate need to easily identify the most frequently mentioned places, persons, institutions or periods, or to find out which researchers worked on what topics, and when.

The idea of our project was inspired by our acknowledgment of the presumed need to increase the transparency of the holdings in our archives converging with the availability of technological innovations that looked promising in terms of providing a solution to this need. Given that manual annotation of scientifically relevant topics would have required enormous human resources on such a large corpus of texts, this approach was ruled out as an alternative. Although so far used for dealing with significantly simpler indexing and analysis of texts, artificial intelligence algorithms, already making their way into the social sciences, seemed a suitable tool to try for our purposes.

The task of assigning social scientifically relevant topics (called ‘subject headings’ in this article) to sociological data has been discussed recently as a case of extreme multi-label classification (XMC) (Skenderi et al, 2021). XMC consists in organizing thematic labels (subject headings) as leaves on a tree structure: there is an input text or texts, and we want the system to return the most relevant ones from a large pool of labels. The challenge constitutes a text classification problem on an industrial scale, which is currently one of the most important and fundamental topics of discussion in machine learning and natural language processing communities (Chang et al, 2020). To make things even more difficult, we set forth to analyze the body of original interview transcripts, while topic assignment is usually carried out on metadata only as opposed to the research data themselves.

In sum, the constantly growing amount of digital research data stored and the need for transparency constitute an ongoing challenge for the staff of KDK and of other similar archives (De et al, 2022). The metadata structure that records the basic data of the documents allows for only a limited overview, searchability and thematic organization of the materials. A deeper thematic analysis of all the interviews, executed by analogue methods, would be enormously time consuming, and there is always the ‘human component’ that hampers any efforts to standardize data on this large scale. Fortunately, besides posing challenges to oral history collections in the digital context, technological advancement is also a creative resource inspiring new types of solutions to existing difficulties. Computational technologies have revolutionized the archival sciences field, offering innovative approaches to process the extensive data in such collections. One relatively recent tool oral historians can benefit from in discerning and describing the contents of their collections is provided by Artificial Intelligence (AI). In particular, automatic speech recognition and natural language processing (NLP) offer unique possibilities for a thorough thematic analysis of oral history interviews (Pessanha and Salah, 2022).

Methodology

The machine annotation of topics covered by the two archives (KDK and the Voices of the 20th Century Archive and Research Group) was preceded by various preparatory activities, including the selection of samples and cleansing and formatting the texts, but also involving other tasks that evolved to become sub-projects in their own right (namely, creation of social science thesauri). Our main goal was to train the NLP algorithms using a specific set of concepts and then validate the

results. This required the creation of a vocabulary (thesaurus) of sociologically relevant topics; the assembly of a training corpus; the testing of both the vocabulary and the algorithms; and finally the production of a training dataset. These steps were carried out in iterative rounds (see Figure 3 for an overview).

Thesaurus

A structured thesaurus of sociologically relevant topics is a *sine qua non* when it comes to describing research materials in an interview archive. At the beginning of the project, no social scientific thesaurus was available in Hungarian. With regard to the specific task at hand, we needed a set of concepts that was sufficiently detailed to cover the key topics of the materials in the two archives, yet narrow enough to be used by NLP tools (Hase, 2022). Finding a balance involved compromises, on one side, and accepting hard constraints, on the other. In addition, we had to keep in mind that the structured set of concepts was intended to serve as the basis for a common search engine to scrutinize the materials in the two archives. It was, therefore, necessary to abandon the idea of a meticulously elaborated system assigning detailed topics to every segment of the texts and adapt our envisioned thesaurus to the limitations of automated coding and an apt browser.

After reviewing the general Hungarian thesaurus (OSZK Köztaurusz), the subject headings of several textbooks, as well as various international social science vocabularies, the thesaurus of CESSDA[4] (Consortium of European Social Science Data Archives) called ELSST[5] (European Language Social Science Thesaurus) was chosen as a starting point for developing our own. The rationale behind this choice was that an internationally established vocabulary in the relevant research fields seemed reliable enough to be used for producing a Hungarian social scientific vocabulary and, besides, it was expected to increase the international visibility of the two archives – a typical effort of document archives using ‘small’ languages. Due to the complexity of the work, the translation of ELSST evolved into a separate project. The Hungarian translation of the main concepts in this thesaurus was first published in September 2022 on the ELSST website, joining the other 14 language versions, and updated regularly since, according to the normal course of action of the international team.

With respect to its key characteristics, the ELSST as the multilingual thesaurus of the CESSDA Research Infrastructure forms a specific kind of intellectual support system. It is an open, constantly updated and modified dictionary, managed by social scientists from all over Europe. Its structure and terminology are constantly revised and regularly amended based on close cooperation of an international team. The Consortium maintains the vocabulary on an ongoing basis, drawing on the experience and insight of national partners responsible for translation, who agree on specific changes (like dropping and replacing irrelevant or out-of-place terms, adding new topics, etc.) and annual updates in a consensual manner. Therefore, ELSST represents an open, coherent and diversified structure that can be considered a Knowledge Organization System (KOS).[6]

[4] https://www.cessda.eu/
[5] https://thesauri.cessda.eu/elsst/en/index
[6] The open KOS is a generic network-based model that aims to allow for diverse, and even controversial, theories and viewpoints on the domain to harmonically and logically co-exist (Zhitomirsky-Geffet, 2019).

To produce our own vocabulary feeding into NLP algorithms, we first translated the ELSST thesaurus consisting of over 3,300 key terms (or 'preferred concepts') into Hungarian. This task was executed by three staff members of the Research Documentation Centre (KDK) trained in social sciences, one of them a professional translator, and supported by experts in linguistics (2 persons) and law (1 person) acting as consultants. As a start, using the SZTAKI dictionary[7] that incorporates the results of various translation services available on the internet and other online resources, all the terms were translated automatically from the English, German and French versions of ELSST into Hungarian. This was followed by extensive multi-perspectival manual corrections by our team of experts in social sciences, translation, linguistics and law. The validity and robustness of the translated terms were ensured by multiple checks and discussions, arriving at absolute consensus regarding each one of them. First, the 3-member KDK team revised the machine translation and found that about 50% of it had to be corrected. Using a wide range of tools available online and making individual searches to explore potential meanings of the terms, we picked the most suitable solution in every case, agreed by all. Second, our linguist consultants scrutinized the entire work, while our lawyer colleague double checked the legal terms, making comments and suggestions. This was followed by discussions involving the expert teams aiming at unanimously accepted solutions.

The translation process raised a host of unexpected dilemmas. These included formal ones like whether to adopt the conventional plural form of English, or singularize the terms as is customary in Hungarian thesauri (the latter solution was finally adopted). Content-wise, the most important types of challenges were presented by the need to find an appropriate translation of terms not used in Hungarian, such as historically contextualized concepts ('access to countryside') including special legal terms ('common rights', 'public access rights', 'offshore employment'). Some phrases do not have a straightforward Hungarian equivalent due to different ways of demarking semantic boundaries, like the names of some policy areas and institutions ('emergency and protective services', 'architecture and building education'). Confronted with the problem of whether or not to keep certain too particular or social scientifically irrelevant terms disturbing a sense of proportion ('otitis media', 'teacher salaries', 'teacher conditions of employment') concerned not only the Hungarian version but the entire ELSST project, demanding that we address the managers directly. The same policy was adopted with regard to the need to introduce some important concepts missing from the list of preferred terms ('financial situation', 'public broadcasting', 'ethnic discrimination'), including inversions to adjust the meaning to current realities (e.g. 'Roma' to replace 'traveling people'). In the course of the work involving regular exchanges with the CESSDA staff and partners, some elements of the international vocabulary have already been rethought at our initiative to support its renewal.

Once the translation was ready, the multitude of concepts had to be severely trimmed to fit our purposes. In fact, the vocabulary of over 3,300 terms was downsized to less than its 10%. Customization involved several rounds of narrowing and refining but also complementing the original set of terms in the ELSST. Omissions concerned terms with too specific meaning, i.e. of little capture, and those covering social worlds not particularly represented in our archives (like medicine). As for additions, some terms important for social sciences (like 'social movements') or

[7] https://szotar.sztaki.hu/

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

with special relevance in the Hungarian context ('language use') were identified, and certain overarching concepts ('social inequality') and aggregate phrases ('economic and social change', 'policing and administration of justice') had to be found to stand for whole sets of more peculiar terms. Eventually, 220 terms were selected and ordered in a three-level taxonomic structure based on a principle of incorporation (i.e. where higher level concepts embrace lower level ones) to serve as the basis of the KDK Thesaurus.

Training corpus

In order to create the training set, as a first step 39 interviews (altogether 1183 pages) were handpicked from the two archives. The selected interviews, drawn from different collections, covered a variety of topics: life histories of major Hungarian sociologists, accounts of people who had spent time in prison, experiences of persons belonging to the Roma minority, family histories of Holocaust survivors, research on the memory of the state socialist era, etc. The idea was to employ a wide array of text types, in terms of both form and content, so that key topics are represented and crucial issues may come up. Out of the 39 interviews 9 were used to test the NLP algorithms and the KDK Thesaurus in several rounds, then 21 were included in the training set and 9 more will be used to improve the results so far achieved.

The interviews in the training set (see Annex 1 for the Training set of interviews and how to access them) are different in length, level of structuredness and also in terms of the applied methods: narrative interview (9 interviews, 500 pages), in-depth interview (6 interviews, 103 pages), semi-structured interview (4 interviews, 84 pages), focus group interview (2 interviews, 48 pages). The balance of the various categories reflects the distribution of the interview material in the two archives. While the Voices of the 20th Century Archive and Research Group archives mostly records of oral histories in the form of semi-structured and unstructured long narrative interviews about personal, professional and community matters, including a few group interviews, the KDK collections contain a variety of structured or semi-structured, often short or medium-sized, more focused accounts of different topics, as well as some rather long in-depth interviews and focus group discussions. The interviews contain large amounts of personal and sensitive data and therefore they are not readily available. The collections of the Voices of the 20th Century Archive and Research Group are only accessible with permission (according to the rules of the archive), while those in the KDK archive are varied in terms of accessibility: some of them are subject to registration[8], while others are available only for the researchers at the Centre for Social Sciences and still others for other, external, registered users as well[9] (also see Annex 1).

The annotation of texts was carried out by experts (scholars of different social scientific disciplines), based on pre-established guidelines. The annotators came from different social scientific backgrounds:

- 5 members of the Research Documentation Centre, sociologists and economists with a masters or a PhD degree

[8] registration: https://20szazadhangja.tk.hu/en/documents-to-download
[9] researcher access/registration: https://kdk.tk.hu/en/where-can-i-find-research-data-and-documents

- one Professor Emerita at the Institute of Sociology, Centre for Social Sciences, psychologist with a PhD degree
- one external researcher at the Institute of Sociology, Centre for Social Sciences, sociologist with a PhD degree
- one external researcher at the Institute for Legal Studies, Centre for Social Sciences, legal expert with a PhD degree
- two doctoral students, both sociologists
- one university student, sociologist.

The manual annotation of the selected interviews proved to be a complex task (Bayerl and Paul, 2011; Neuendorf, 2017). In order to build the training set (so that various algorithms of machine based harvesting of topics could be compared), the annotators marked keywords in every section of an interview text, matching each of them with a term listed in the KDK Thesaurus. The purpose of assigning terms as they occur in the texts to our set of concepts was twofold: in addition to producing the training material for machine learning, this exercise also served to validate and correct our set of subject headings, i.e. the KDK Thesaurus. By manual annotation our expert coders could determine whether our superimposed concepts were really suitable for the task of creating thematic metadata attached to the interviews. In some cases, it was necessary to modify subject headings in the KDK Thesaurus. Such modifications served the streamlining of the structure of subject headings and consisted in including missing terms, discarding redundant ones, specifying overly general concepts, creating umbrella terms by merging other, more specific ones at lower levels of the hierarchy, deciding between terms with overlapping semantic fields, etc.

Given the (relatively) large amount of subject headings in the KDK Thesaurus and the essentially unstructured texts, achieving consistency in the annotation process, or even specifying and following any principles, represented major challenges (Neuendorf, 2017). Manual annotation itself defies consistency, since prior assumptions, habits developed during coding, as well as any possible mistakes and misinterpretations all contribute to disparity (Bayerl and Paul, 2011). In order to reduce such risks, we developed coding guidelines with strict principles of annotation (Gárdos *et al*, 2023). Two annotators were assigned to code each text independently in a parallel manner – customary in other similar projects involving manual annotation – so as to increase reliability. Thus the development of guidelines and the training of the annotators served to ensure that manual coding was done methodologically and meticulously, so that its result was as consistent as possible (Bayerl and Paul, 2011; Neuendorf, 2017).

The guidelines clearly described the purpose of coding, defined the main concepts of the workflow (e.g. subject heading of the KDK Thesaurus, keyword, interview section), identified the units to be coded (the interview sections), described the methodology and the steps of annotating, specified the maximum number of subject headings per section, contained the interview schedule of annotators and listed other tasks to be performed in addition to the coding process (e.g. collecting suggestions for better coding or for amending the KDK Thesaurus). Before anything, the guidelines made it clear that the purpose of annotation was to identify the main content elements of a text, which are both important in the context of the given interview section and relevant in social science. Annotators were asked to assign the most appropriate subject headings to every section of a text with this dual aim in mind. Thus each section was treated separately, as if it were a stand-alone text,

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

which is especially important for machine learning. During the multiphase coding process, the guidelines and the policy itself was modified at some points (e.g. some details became more specified, some steps were further elaborated, etc.), based on the effectiveness of the work, the feedback from the annotators and regular discussions of the results within the team (Neuendorf, 2017).

The manual annotation was carried out using Label Studio[10], providing a user-friendly interface for this type of work (Figure 1). The preparatory process included integrating the multi-level KDK Thesaurus into the system and developing a structure and presentation suitable for matching the selected keywords with subject headings of the KDK Thesaurus. The interviews, broken down into sections, were first annotated independently by two annotators. Then, depending on the degree of agreement, either the original two annotators, or a third independent annotator determined the final selection of subject headings, constituting the training dataset. Figure 1 shows the interface with the keywords marked (in pink), which were then matched with subject headings chosen from the hierarchical list of subject headings of the KDK Thesaurus (on the left side of the figure) by the annotators.

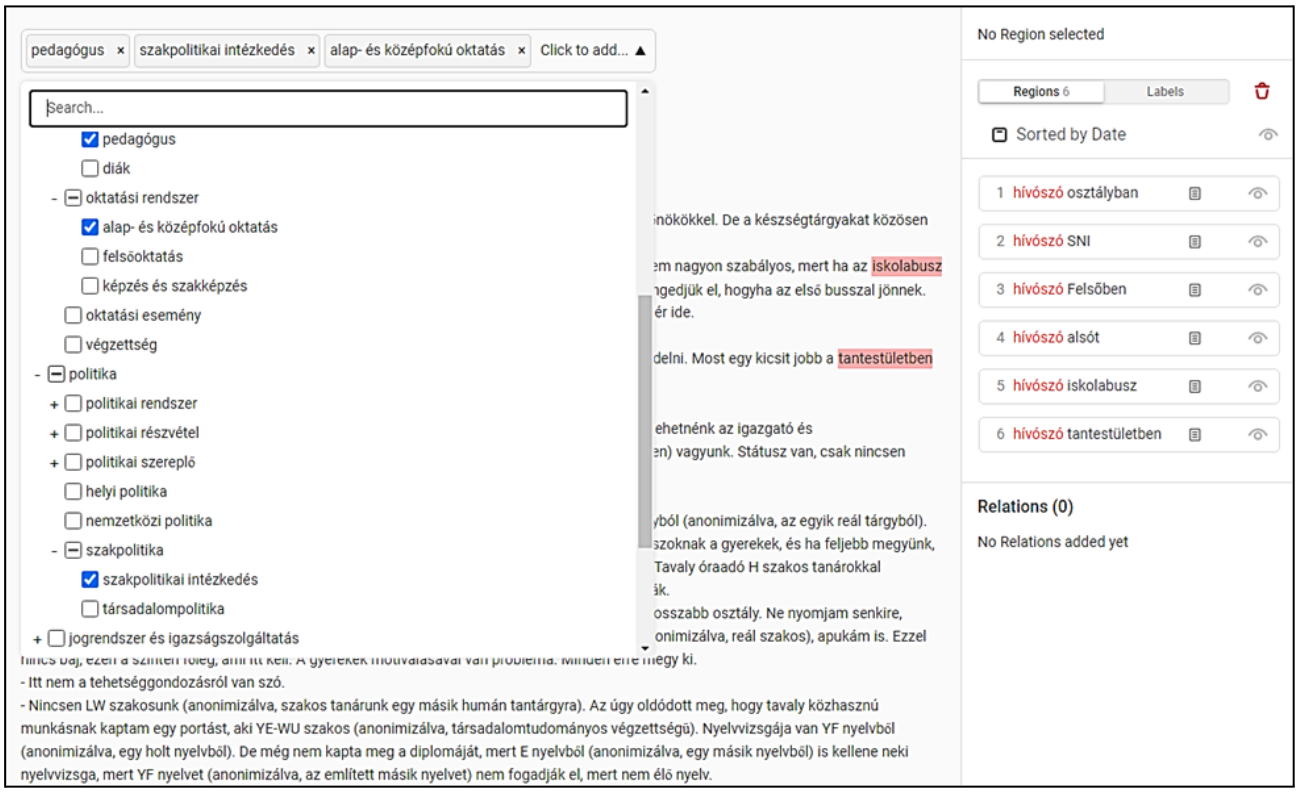

Figure 1. Label Studio interface and the hierarchical set of subject headings of the KDK Thesaurus during annotation. Figure by authors.

In accounting for the nature of the project, we developed our own principles for defining accepted subject heading matching (Bayerl and Paul, 2011; Hase, 2022). We considered it a match between two annotations if the top-level term in the hierarchy of the two chosen subject headings was identical. That is, for example, if one of the annotators added the subject heading ‘customs and

[10] https://labelstud.io/

traditions' to a section and the other one chose 'cultural event', this was accepted as a match (given that both subject headings were considered to be adequate) since both terms belong to the same thread in the hierarchy, i.e. are classified under the same overarching subject heading 'culture'. In other words, the two annotators, although choosing different subject headings, used them to refer to the same broader category.

The desired degree of matching was determined according to the specificities of the task, that is, considering the large number of variables and the open-ended nature of interpretation (i.e. identifying topics and associating them with concepts). After careful consideration, the limit of minimal coincidence was set at 30%. In each interview section, the matching of the results of the two annotators was determined by calculating the degree of similarity between the two entire sets of subject headings selected by the two annotators. For example, if one annotator attached five and the other one four subject headings to a section, out of which three were identical, this was calculated to be a 50% matching (3/[3+2+1]*100)[11]. When matching was below 30% (i.e. the degree of coincidence of subject headings selected by the two annotators for a given section of the text was less than 30%), a third annotator stepped in to decide on the final set of subject headings. When, in turn, matching was above 30%, the original two annotators together decided on the final version. As a matter of fact, 25% of the interview sections had a score below 30% (of which 4% had zero matches), while 75% had a score above 30% (including 6% with full matching).

The literature tends to expect stricter matching than this (Bayerl and Paul, 2011). However, the specificities of the project required the development of specific criteria of matching and of setting the degree of optimal results (Bayerl and Paul, 2011). Compared to "traditional" annotation tasks, the situation appeared to be special in several respects. First, the text corpus to be coded consisted in largely unstructured and thematically versatile interview transcripts close to colloquial speech. Second, the multitude of subject headings was much more difficult to handle than only a few variables normally used in thematic coding of texts aided by AI. Thirdly, on a technical note, the interviews could not be broken down into sentences, short paragraphs, or other clear-cut units easy to grasp. As a result, sections usually covered several topics of which the annotators had to identify the most relevant ones. In most of the cases, it was possible to assign several subject headings to a given section, depending on the number of relevant topics covered. At the same time, a maximum selection of five subject headings was defined in order to better grasp the key themes and make the results more comparable. Still, it was very difficult to achieve consistency, posing a major challenge both to the annotators and the workflow coordinators (Albaugh et al, 2014; Neuendorf, 2017). To be sure, the subject headings of the KDK Thesaurus are not, and should not be, mutually exclusive, as various themes appear in the texts intertwined. It is difficult, if not impossible, to standardize the method of selecting the most relevant ones among different topics appearing in the same section. Therefore, different subject headings given by the two annotators could be equally relevant, even in case of a complete lack of matching. Such a result does not even represent a real failure, given the nature of the work, i.e. the interpretation of free-flowing conversations. And yet, a minimum degree of matching had to be defined – after all, in this light, 30% is not at all that small. All these factors distinguishing our project from other exercises of thematic coding justified the development of

[11] 3 is the number of the identical subject headings, and 3+2+1 is the number of all the subject headings added to the given section: 3 identical ones, 2 and 1 differing ones.

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

specific criteria and guidelines that are more permissive than usual. Results of the annotation processes have been uploaded to the KDK repository (Gárdos *et al*, 2023).

Automated assignment of subject headings

The main goal of automated subject heading assignment was to select a few relevant subject headings out of a three-level subject heading taxonomy (the KDK Thesaurus) for text passages (paragraphs) of 2-5000 characters each. This is a rather specialized task for which traditional classification AI algorithms cannot be used due to the large number of subject headings, while topic modeling solutions are usually not suitable for such tasks, either, using a pre-given taxonomy.

Nevertheless, we made an experiment for topic modeling using gensim's LDA implementation and BERtopic with the Hungarian Spacy model. We found that topic keywords contained mostly generic words such as ‘wife’, ‘make’, ‘interview’, ‘tell’, etc. This may be caused by the nature of Hungarian language (strongly agglutinative) and the lower quality of the language model (compared to English Spacy models). In conclusion, using this method, the machine detected topics could not be matched with taxonomy entries.

Fortunately, the open source Annif software tool (Suominen et al, 2022) has been developed specifically for this purpose, allowing users to choose from ten different methods for this kind of topic assignment. The methods include learning and statistical methods, which can be combined.[12] All methods have been tested with our training dataset; the results of the best performing ones are shown in Figure 2. Among the tested methods, the best performing NN-ensemble combines the Omikuji and TF-IDF subject headings in a 3:1 ratio.

The methods present in Figure 2 are documented on the Annif wiki page[13] but some details are provided here as well. *Omikuji* is a family of tree-based machine learning algorithms, and *TF-IDF* is a baseline statistical approach for automated subject indexing based on term frequencies. The *ensemble* method combines the results of two other methods by calculating the mean of scores. The NN-*ensemble* plugin implements an intelligent combination of results of multiple approaches. The results of the other methods are re-weighted using a neural network.

In all cases, we used the same train and test sets for training by manually splitting each interview into a train and a test part in a ratio of 4:1. Subsequently, the train and test sets were converted into the corpus in TSV format required by Annif (this is a pure format conversion where no text preprocessing is included). The training and evaluation were performed using the built-in Annif commands, which yielded the precision, recall and F1 scores shown in Figure 2.

Slightly better results were achieved by our in-house developed method (named SZTAKI in Figure 2, Micsik and Kukucska, 2023), based on the use of keywords assigned to subject headings during the construction of the training set. Initially during corpus building, keywords were collected for each topic. This collection of keywords was revised and compiled into a keyword-to-topic mapping

[12] https://annif.org/
[13] https://github.com/NatLibFi/Annif/wiki

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

dictionary. Here, we used textacy[14] and the latest HuSpaCy (Orosz et al, 2022) to collect key terms for the text set. The main steps of the "SZTAKI" annotation approach are as follows:

1. Remove special audio transcript notations (e.g. for hawk),
2. Process the text using HuSpaCy,
3. Extract keywords as lemmatized noun chunks (word sequences of nouns, proper nouns and adjectives),
4. Find out if extracted keywords are mapped to any topic in the taxonomy,
5. Generate a score for each topic found, based on the frequency and weight of keywords,
6. Select topics with a score above a preset threshold.

The best values for weights and threshold were selected using a heuristic parameter sweep with the F1 score as fitness value.

Based on the statistics of the subject headings associated with the keywords, KDK Thesaurus was further refined. For example, topics with very low and very high usage were reconsidered for splitting or merging.

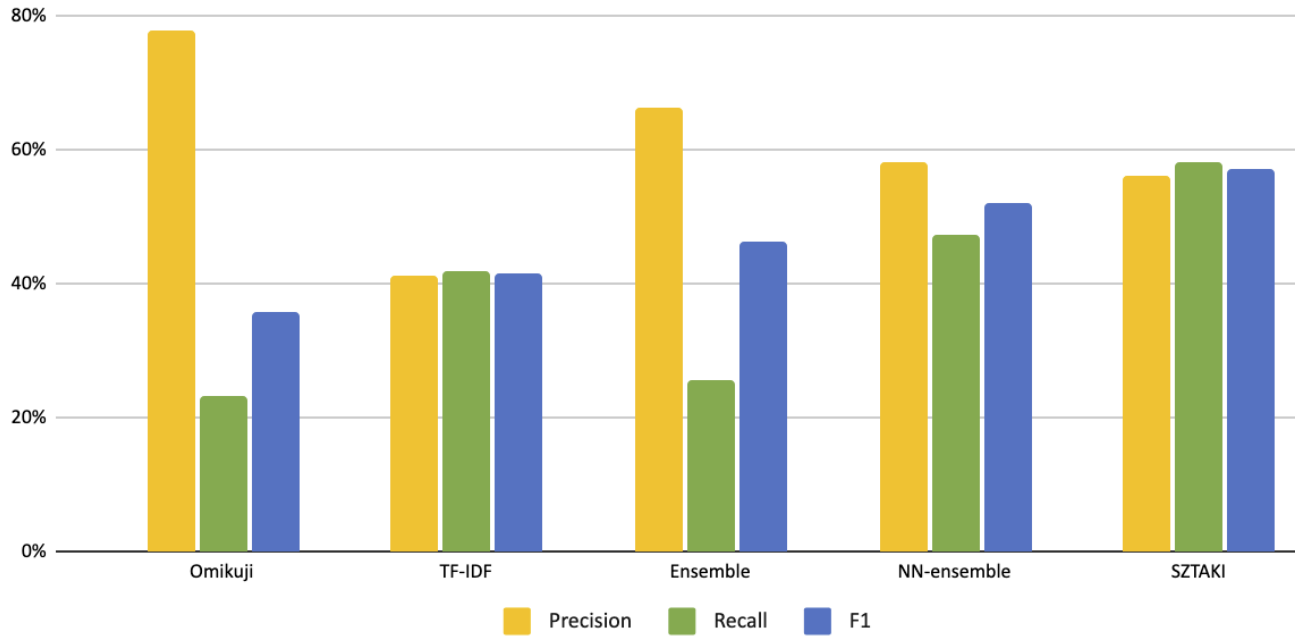

Figure 2. Comparison of the results of the different methods. Figure by authors.

The sustainability of the methods ranked first and second (NN-ensemble and SZTAKI) should be examined here. For example, what happens if you need to change the subject vocabulary? In the case of the SZTAKI method, the list of associated keywords would need to be reconfigured for the changed KDK Thesaurus and validated on a small sample of the developed set. Then it would be necessary to re-calculate the keyword and subject annotations for the interviews (or part of them), which can be slightly time-consuming as the text needs to be linguistically analyzed (word order, word genre definition, etc.). The situation is more complicated with respect to NN-ensemble: the method would need to be re-taught, and the learning set expanded with a sufficient number of

[14] https://github.com/chartbeat-labs/textacy

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

instances of the changed subject words. This is an iterative process whereby, if the results are good enough, the automatic assignment of subject headings can be rerun in the interviews.

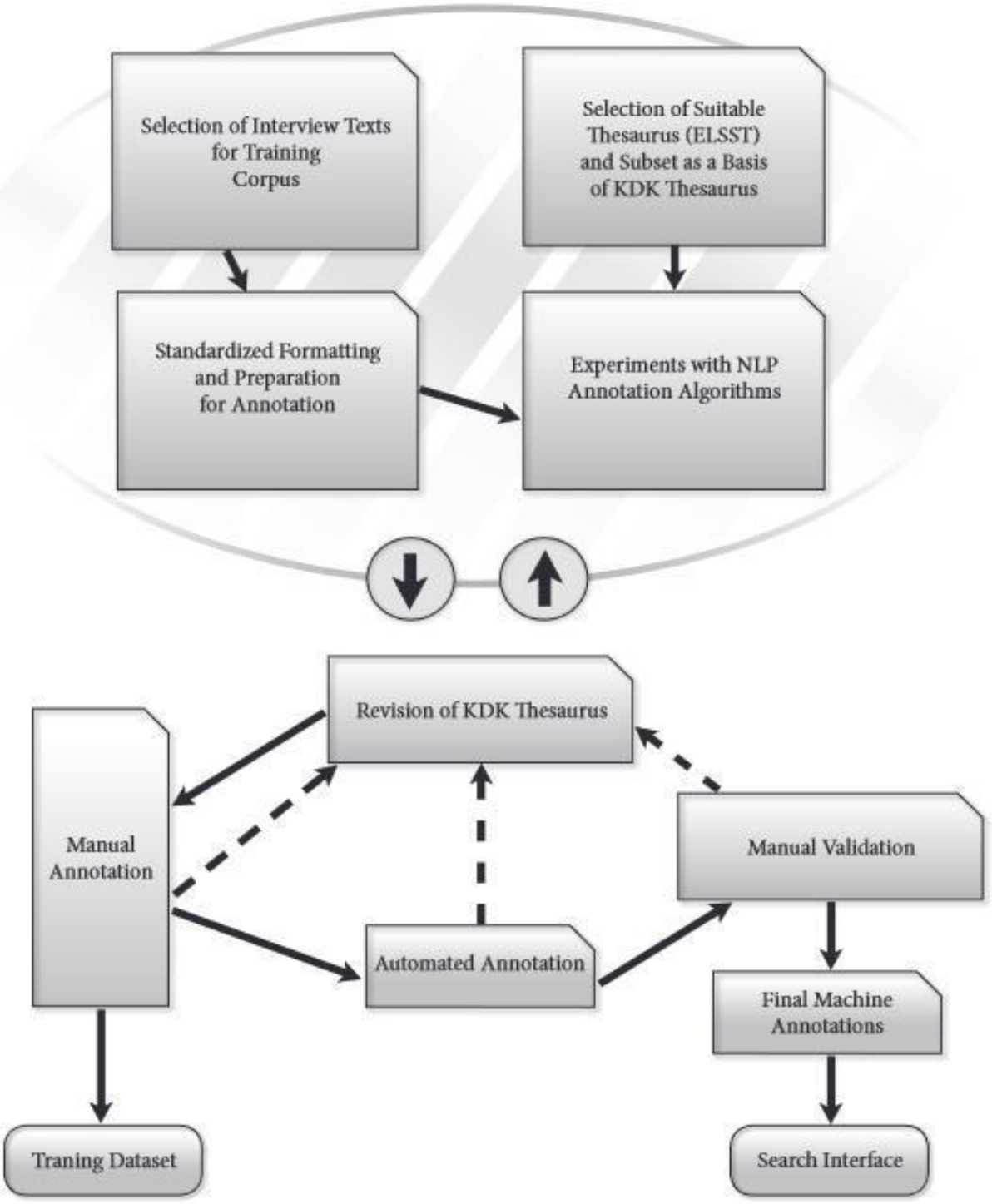

Figure 3. An overview of the production of the KDK Thesaurus, the training dataset and the search interface. Figure by authors.

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Results

The KDK Thesaurus and the Training Dataset

As a result of trimming and adjustments, a vocabulary of 220 items (the "KDK Thesaurus") was produced: a concise and customized structure of concepts (Albaugh et al, 2014), about 7% in size compared with ELSST (see Annex 2 for the complete KDK Thesaurus and here: Gárdos *et al*, 2023[15]). The list of terms was developed with regard to *relevance* (based on a selection of concepts that are meaningful in social sciences, taking into account the main topics and key formulations of Hungarian sociology), *coverage* (ensuring representation of all the important fields and topics in social sciences), *proportionality* (including adjustment of the structure of vocabulary to the thematic foci represented in our archival collections), *disjunctivity* (by the selection of well-differentiated concepts) and *quantity* (limiting the number of concepts so that they can be easily accommodated by machine algorithms). Thus, in addition to the drastic reduction of the original list, new terms had to be added, as noted above, to capture the important themes characterizing Hungarian sociology, recent Hungarian history and the contents of our archives. It has also already been discussed how the terms in the final thesaurus (KDK Thesaurus) are organized in a threefold hierarchical structure, providing transparency and thus supporting manual and machine-based coding. The layout also serves as the blueprint of the concept tree to be used in the future search engine shared by the two archives.

The process of negotiations and its outcomes, characteristic of an open KOS, is well illustrated by the emergence of our KDK Thesaurus. Based mainly on ELSST entries, it also includes other kinds of input. The underlying goal being to enhance the quality of our thematic mapping, the structure of concepts was developed considering many factors simultaneously. As a result, KDK Thesaurus meets important quality standards expected from thesauri, like allowing for semantic navigation and providing different search options (Pinto, 2008).

Out of the 220 terms in the KDK Thesaurus, 172 are straightforward or approximate translations of ELSST concepts, while there are 48 that could not be reasonably matched with any of the concepts therein (see Annex 2). As for the latter, a host of reasons interplayed in the decisions of making such additions and alterations. We have already pointed out the general goals behind customizing the thesaurus used in our project. Here the more specific considerations and types of adjustments are explained.

Even though it contains fifteen times more items, the list of 'preferred terms' in ELSST does not always cover the themes we wanted to capture. Thus we made additions like 'division of labor in the family', 'deprivation of rights' or 'non-state institutions'. Occasionally, the concept we were looking for could be found in ELSST, however, only as a secondary one (not as a suggested term but as one to be replaced by a 'preferred term'). Such synonyms called 'entry terms', when deemed important or better than the associated 'preferred term', were upgraded to become part of our main list of concepts. Thus we have 'segregation' (not only 'integration'), and 'financial situation' (instead of 'financial resources') in the KDK Thesaurus.

[15] with a Creative Commons license BY-NC-SA [Attribution-NonCommercial-ShareAlike]

In some of the cases, generic terms were sought when the term or terms in ELSST reflected more particular usage, representing selected fields of application or distinct types of a phenomenon. Thus ‘economy’, ‘media’, ‘healthcare’, ‘origin’, ‘violence’ and ‘crisis’ were added to cover all kinds of occurrences of these concepts. The underlying reason was to find overarching concepts that help accommodate many others, partly to delineate important fields, and partly to limit the vocabulary. Similar considerations related to the lack of space and the intention to customize the vocabulary gave rise to attempts to draw together various concepts under a single umbrella term. For this reason we have ‘institutions and organizations’, ‘educational event’, ‘primary and secondary education’, ‘policy’, ‘political actor’ or ‘sex and gender’. These general terms seemed sufficient for our purposes, i.e. no further differentiation of meaning was needed for the level of content mapping aimed in our project.

By contrast, at other times, we needed more specific terms. Original terms were altered for reasons of accuracy, occasionally designating a smaller set of concepts or more specific usage. Hence, ‘financial situation’ came to replace ‘life circumstances’, ‘natural disaster’ was specified as a kind of ‘catastrophe’ as such, ‘civil activism’ was introduced to expand on ‘civil and political rights’, and ‘lower class’ was used in place of ‘working class’. Such clarifications were needed to better denote social phenomena as appearing in life stories and reflect common language use. There were occasions when two competing translations had to be considered. Examples include ‘travel’ that, although part of the ELSST, has a narrower and a broader equivalent in Hungarian, and the latter (closer to ‘movement’) was preferred. The same is true for the case of ‘integration’, which is directly adopted to Hungarian from the Latin *integratio*, primarily to refer to policy, while its other translation we used instead (*beilleszkedés*) expresses the process from the perspective of its subjects. Similar situations arose related to the Hungarian noun generally referring to the idea of parenting, literally meaning ‘having children’ (*gyermekvállalás*), which was used instead of the more restrictive phrase ‘giving birth’, or with respect to ‘scientific work’, a more comprehensive term as compared with ‘scientific research’.

Additions were sometimes made necessary to accommodate the vocabulary in the Hungarian social and historical context. Here, for instance, ‘language competence’ (not part of ELSST) is often considered a crucial skill. Also, there are good reasons to underline the importance of ‘economic processes’ and their particular instances like ‘privatization’, ‘nationalization’ or ‘reform’. Specifics of Hungarian history may, at times, alter or, so to speak, usurp the meaning of certain terms (common nouns) restricting them to refer to particular events or phenomena, rather than standing for one of a kind. For instance, it is because of the overwhelming importance of the end of World War II that the term ‘liberation’ in Hungarian really refers to the victory of the invading Soviet army over the German troops. By the same token, ‘regime change’ denotes the democratic transition after 1989, and ‘restitution’ is appropriated to signify the distribution of vouchers in the early 1990s to compensate for lost property due to nationalization by the socialist state. Such problems were resolved sometimes by introducing a more abstract category to be the generic term, while reserving the term tainted by Hungarian history to denote the particular phenomenon. Hence, ‘political change’ was added to refer to the general idea of a new system replacing the old, keeping ‘regime change’ to name the particular historical instance, that is, the democratic transition after 1989.

The thematic characteristics of our archives, obviously reflecting key issues in Hungarian history and society, while also embodying particular scholarly perspectives, made it necessary to introduce or prioritize certain terms over others. For instance, additions like ‘foreign country’, ‘creation’, or ‘scientific event’ are owing to our heavy focus on scientific work and the scientific community as such on account of our collection of interviews with Hungarian social scientists that can be read as professional histories adding up to a history of the profession. Other inclusions like ‘trauma’, ‘forced relocation’, ‘suppression’, ‘remembrance’ or ‘Romaphobia’ were needed because of our collections of life histories and family histories of Roma and Jewish Holocaust survivors. Ethnonyms like ‘Jews’ and ‘Roma’ were introduced for the same reason. The inclusion of ‘pandemic’, in turn, was due to its current significance in society and social research. Finally, there were instances, like in the case of ‘region’, ‘public policy’ or ‘advocacy’ which terms were missing in ELSST and the reason for omission could not be detected. These additions were made simply to account for important categories.

The adjusted vocabulary served as the basis for processing the texts providing a kind of topic analysis. As a result of manual annotation using KDK Thesaurus, a training dataset of social scientific interviews was produced with sociologically relevant subject headings. Used as a standard for measuring the success of machine assisted coding (i.e. validation), this corpus remains to be used and amended in the future for further refining the KDK Thesaurus and developing methods of operation for the search engine serving our archives.

Browsing platform

In order to make the interviews in the archives readily available for researchers, we wanted to offer other tools in addition to full text search. Thus based on the linguistic analysis of the interview texts, producing a large database comprising documents in both archives, a faceted search engine was compiled from open source components, with different filtering options (Figure 4). In the middle of the figure, the two active filtering conditions are shown in the blue boxes in the frame, and the filtered items are marked with a red background.

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

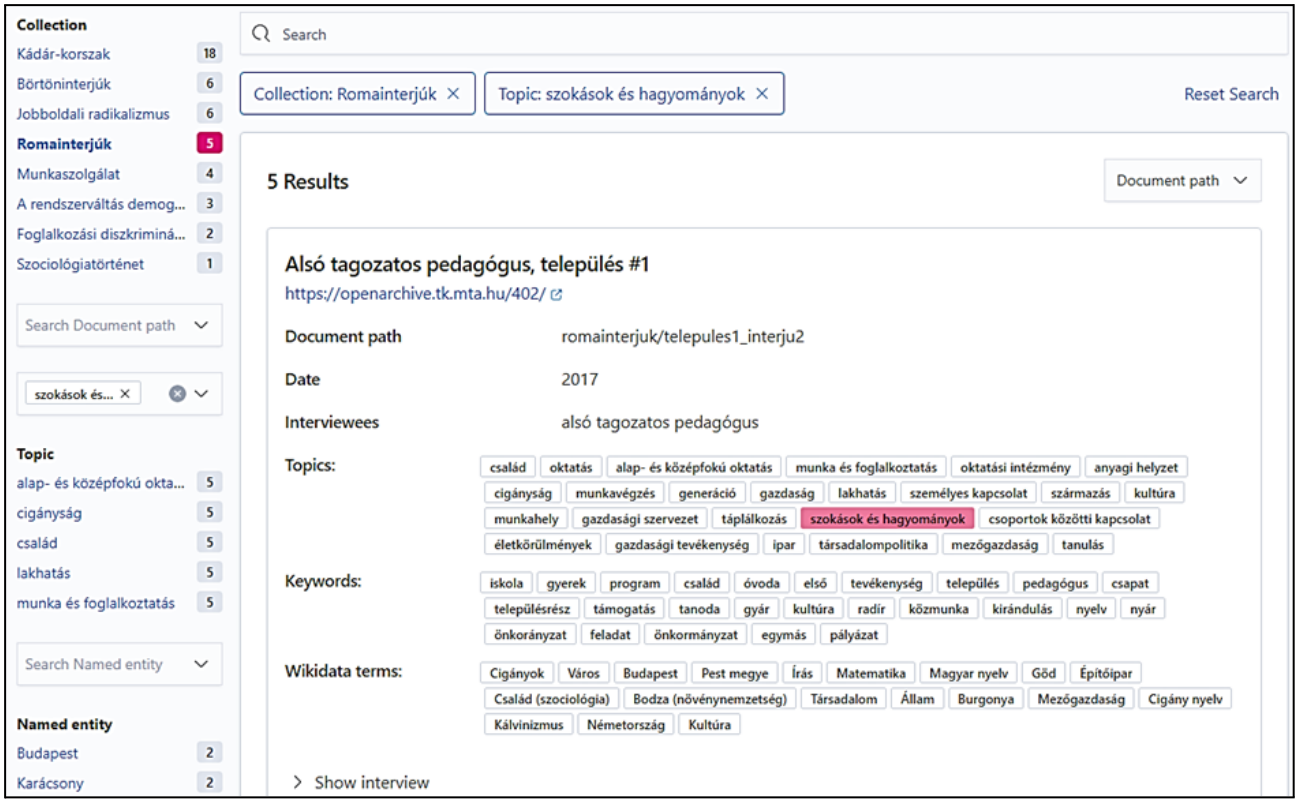

Figure 4. Search interface for interview collections. Figure by authors.

The result list includes the main metadata of the selected interviews, as well as the main subject headings of the KDK Thesaurus associated with them, furthermore keywords and Wikidata pages aggregated by the automatic analysis per document. The full text of the interview can also be opened here for reading, where a small information panel with links to the corresponding Wikidata, Geonames, VIAF, etc. pages appears next to the name terms identified in the text.

Another important aim in the project was to extract the geographical, personal and other kinds of proper names mentioned in the interview texts. This activity is only briefly reported here, as it would deserve a separate discussion. We experimented with three Named Entity Recognition (NER) tools and also searched for the names found in Wikidata in order to be able to offer links to the mentions of every name. Wikidata aggregates the identifiers of many other registers, so that most of the links can be obtained from Wikidata (e.g. VIAF, Geonames). Finally, we applied named entities together with their set of database links to add further annotations to the interview texts.

For the creation of custom visualizations, a Kibana[16] interface has been integrated alongside the search engine, allowing the display of the selected data content in a variety of graphical formats. As an example, Figure 5 shows a map visualization that presents the geographical locations mentioned in the interviews, also showing in which of them the mention occurs.

[16] https://www.elastic.co/kibana/

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

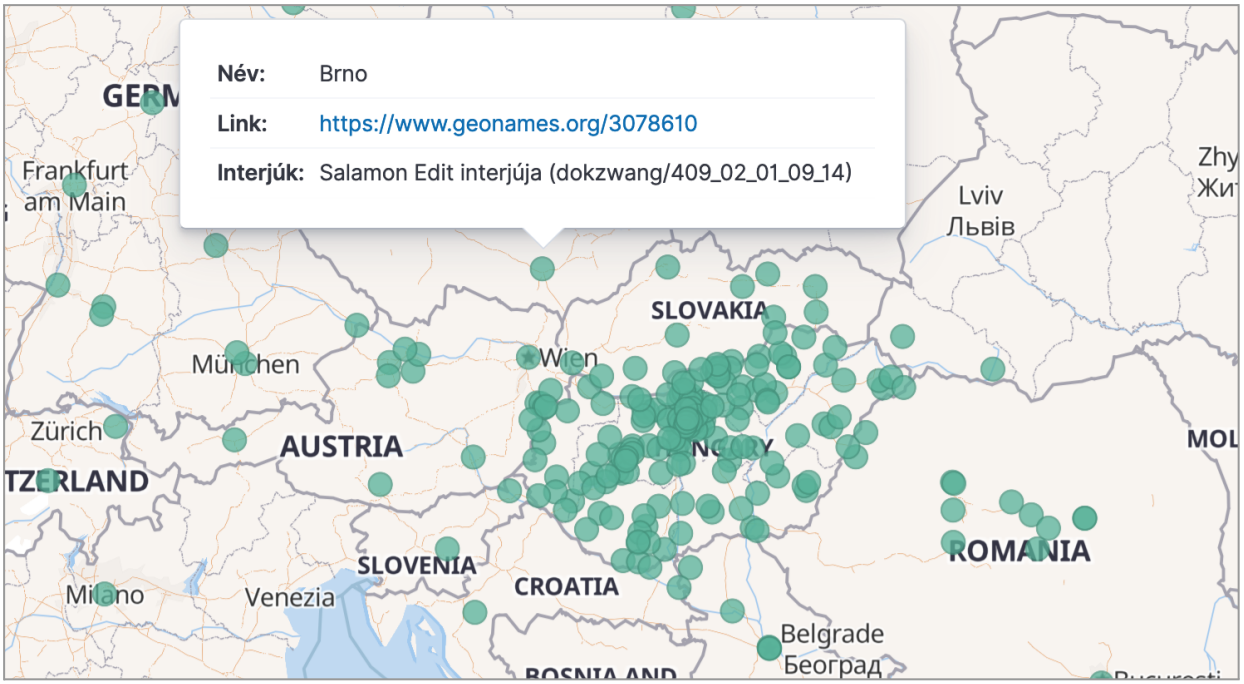

Figure 5. Map of the geographical names mentioned in the interviews. Figure by authors.

Conclusion

NLP tools provide a promising area for enhancing searchability of large textual research data archives, as demonstrated in this article. With open access software available, it is relatively easy to create rich exploratory research interfaces. This is necessary because there is a large amount of valuable research materials, generated in the form of texts or audio recordings, which cannot be reviewed and scrutinized without some form of prior processing. In the course of our project, the most popular and recent Hungarian as well as language-independent language processing tools were tested for their viability for the specific research purpose of enhancing the transparency of a social scientific interview repository by creating useful metadata. Some kind of benchmark is essential, yet it is quite resource-intensive to produce – we have made the first steps in this direction by creating a training dataset. It should be noted that word construction and noun recognition in Hungarian still suffer from a level of error that is disturbing in a research environment and may require posterior manual correction. Another key issue is selecting the most suitable subject vocabulary system for mass automatic categorisation. We opted for a domestic adaptation of a European sociological taxonomy, which, in fact, had no plausible alternatives. The process of the creation of the KDK Thesaurus with all its compromises and challenges that we have described in this article might prove insightful for other data archives and repositories with similar aims.

Funding

The project presented in this publication, implemented by TK KDK and SZTAKI DSD, was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Bibliography

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J., Walgrave, S., 2014. Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding.

Bayerl, P.S., Paul, K.I., 2011. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. Computational Linguistics 37, 699–725. https://doi.org/10.1162/COLI_a_00074

Boyd, D., 2013. OHMS: Enhancing Access to Oral History for Free. The Oral History Review 40, 95–106.

Boyd, D.A., Larson, M.A. (Eds.), 2014. Oral History and Digital Humanities: Voice, Access, and Engagement. Palgrave Macmillan US, New York. https://doi.org/10.1057/9781137322029

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., Dhillon, I.S., 2020. Taming Pretrained Transformers for Extreme Multi-label Text Classification, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Presented at the KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Virtual Event CA USA, pp. 3163–3171. https://doi.org/10.1145/3394486.3403368

Château, S. du, Boulanger, D., Mercier-Laurent, E., 2012. Managing the domain knowledge: application to cultural patrimony. Knowledge Management Research & Practice 10, 312–325. https://doi.org/10.1057/kmrp.2012.22

Cohen, S., 2013. Shifting Questions: New Paradigms for Oral History in a Digital World. The Oral History Review 40, 154–167.

De, S., Moss, H., Johnson, J., Li, J., Pereira, H., Jabbari, S., 2022. Engineering a machine learning pipeline for automating metadata extraction from longitudinal survey questionnaires. IQ 46. https://doi.org/10.29173/iq1023

Hase, V., 2023. Automated Content Analysis, in: Oehmer-Pedrazzi, F., Kessler, S.H., Humprecht, E., Sommer, K., Castro, L. (Eds.), Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research: Ein Handbuch - A Handbook. Springer Fachmedien, Wiesbaden, pp. 23–36. https://doi.org/10.1007/978-3-658-36179-2_3

Higgins, S., 2018. Digital curation: the development of a discipline within information science. JD 74, 1318–1338. https://doi.org/10.1108/JD-02-2018-0024

Micsik, A., Kukucska, Á., 2023. A multilabel classifier for Hungarian social science interviews. https://github.com/dsd-sztaki-hu/huSocC

Neuendorf, K.A., 2017. The Content Analysis Guidebook. SAGE Publications, Inc, 2455 Teller Road, Thousand Oaks California 91320. https://doi.org/10.4135/9781071802878

Oehmer-Pedrazzi, F., Kessler, S.H., Humprecht, E., Sommer, K., Castro, L. (Eds.), 2023. Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research: Ein Handbuch - A Handbook. Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-36179-2

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., Farkas, R., 2022. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. https://doi.org/10.48550/arXiv.2201.01956

Pessanha, F., Salah, A.A., 2022. A Computational Look at Oral History Archives. J. Comput. Cult. Herit. 15, 1–16. https://doi.org/10.1145/3477605

Pinto, M., 2008. A user view of the factors affecting quality of thesauri in social science databases. Library & Information Science Research 30, 216–221. https://doi.org/10.1016/j.lisr.2007.12.003

Portelli, A., 1991. The death of Luigi Trastulli, and other stories: form and meaning in oral history, SUNY series in oral and public history. State University of New York Press, Albany, N.Y.

Skenderi, E., Huhtamäki, J., Stefanidis, K., 2021. Multi-Keyword Classification: A Case Study in Finnish Social Sciences Data Archive. Information 12, 491. https://doi.org/10.3390/info12120491

Suominen, O., Lehtinen, M., Inkinen, J., 2022. Annif and Finto AI : Developing and Implementing Automated Subject Indexing. JLIS. https://doi.org/10.4403/jlis.it-12740

Toney, A., Dunham, J., 2022. Multi-label Classification of Scientific Research Documents Across Domains and Languages, in: Proceedings of the Third Workshop on Scholarly Document Processing. Presented at the sdp 2022, Association for Computational Linguistics, Gyeongju, Republic of Korea, pp. 105–114.

Gárdos, J., Egyed-Gergely, J., Horváth, A., Micsik, A., Kovács, L., Martin, D., Marx, A., Meiszterics, E., Pataki, B., Siket, M., Vajda, R., 2023. A thematic exploration of textual research resources in CSS data repositories. Data Collection. DOI: 10.17203/KDK598

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Zhitomirsky-Geffet, M., 2019. Towards a diversified knowledge organization system: An open network of inter-linked subsystems with multiple validity scopes. JD 75, 1124–1138. https://doi.org/10.1108/JD-10-2018-0163

Online sources:

ELSST - European Language Social Science Thesaurus. https://elsst.cessda.eu/

Kata, Lauren, Milbrodt, Natalie and Sielaff, Steven and Vos, Jaycie, Oral History Metadata and Description: A Survey of Practices - Survey Report (Oral History Association, 2020). https://www.oralhistory.org/wp-content/uploads/2021/01/OHA-MTF-White-Paper_2020.pdf

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Annex 1. Training set interviews and accesses

Interview	URL	Access*
bortoninterjuk/101	https://openarchive.tk.mta.hu/185/	researcher access/registration
bortoninterjuk/120	https://openarchive.tk.mta.hu/185/	researcher access/registration
feischmidt/INT 1-177	https://openarchive.tk.mta.hu/177/	researcher access/registration
feischmidt/INT 3-177	https://openarchive.tk.mta.hu/177/	researcher access/registration
interjuk2/341 interju	https://openarchive.tk.mta.hu/198/	researcher access/registration
interjuk2/WesselyJanosTrans_arch	https://voices.tk.mta.hu/handle/123456789/70	registration
interjuk2/interju_4-221	https://openarchive.tk.mta.hu/221/	researcher access/registration
kadar/409_04_02_01_1	https://voices.tk.mta.hu/handle/123456789/210	registration
kadar/409_04_02_02_1	https://voices.tk.mta.hu/handle/123456789/187	registration
mabisz/409_39_17_02	https://voices.tk.mta.hu/handle/123456789/12209	registration
mabisz/409_39_1_02	https://voices.tk.mta.hu/handle/123456789/9366	registration
romainterjuk/telepules3_interju8	https://openarchive.tk.mta.hu/402/	researcher access/registration
romainterjuk/telepules8_interju3	https://openarchive.tk.mta.hu/402/	researcher access/registration
szoctort/409_08_01_25_1	https://voices.tk.mta.hu/handle/123456789/1180	registration
szoctort/409_08_01_27_1	https://voices.tk.mta.hu/handle/123456789/1178	registration
dokzwang/409_02_01_08_4	https://voices.tk.mta.hu/handle/123456789/40	registration
dokzwang/409_02_01_11_3	https://voices.tk.mta.hu/handle/123456789/41	registration
kdk198/11610 interju	https://openarchive.tk.mta.hu/198/	researcher access/registration
kdk198/4248 interju	https://openarchive.tk.mta.hu/198/	researcher access/registration

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

kdk221/interju_3-221	https://openarchive.tk.mta.hu/221/	researcher access/registration
kdk221/interju_9-221	https://openarchive.tk.mta.hu/221/	researcher access/registration

* Types of access:
researcher access/registration: https://kdk.tk.hu/en/where-can-i-find-research-data-and-documents
registration: https://20szazadhangja.tk.hu/en/documents-to-download

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

Annex 2. The three-level KDK Thesaurus

TERM LEVEL 1	NOT IN ELSST	TERM LEVEL 2	NOT IN ELSST	TERM LEVEL 3	NOT IN ELSST
families		origin			
		division of domestic labor	x		
health care		health professionals			
		patients			
		health care facilities			
		medical care			
		pandemic	x		
church		religious personnel			
		religious institutions			
life events		having children	x		
		weddings			
		divorce			
		accidents			
		suicide			
		mortality			
		trauma	x		
living conditions		financial situation		poverty	
		health status		diseases	
				addiction	
				mental health	
		housing		homelessness	
		nutrition			
		sport			
		everyday life			
		leisure time			
		view of future	x		
stages of life	x	child and childhood			
		youth			
		old age			

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

values		attitudes			
		behaviour			
		identity			
		sexuality	x		
		ideologies			
		religion			
		emotion			
		prejudice		anti-semitism	
				Romaphobia	x
				homophobia	
				racism	
		radicalism			
economy		economic organisation	x		
		economic activity		industries	
				trade	
				agriculture	
				service industries	
		labour market			
		consumption			
		finance and financing			
		economic process	x	privatization	
				nationalization	
				reform	
communi-cations		information	x	suppression	x
		interpersonal communication			
		communication skills		literacy	
		language competencies	x		
		media	x	internet	
				social media	
		source		document	x
				remembrance	x
environment		location		built environment	x

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

				human settlement	
				foreign country	x
				country	x
				region	x
				geographical mobility	x
		natural environment		climate change	
				pollution	
				animals	
				natural disasters	x
culture		socio-cultural clubs			
		cultural events			
		creation	x		
		arts			
		customs and traditions			
labour and employment		workplace			
		work			
		employment			
		unpaid labor			
		working conditions		commuting	
		unemployment			
		occupations			
		manual workers			
		white collar worker			
education		educational institutions			
		persons participating in education	x	teacher	
				students	
		education events	x		
		learning			
		educational background			
		primary and secondary education			
		tertiary education			

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

		(vocational) training			
politics		political actor	x	government	
				political parties	
		public policy	x	social policy	
		international politics			
		political participation			
		political systems		democracy	
				dictatorship	
				communism	
law		human rights			
		international law			
		legislation and regulations			
		offences		corruption	
				prostitution	
		crimes against humanity		forced relocation	x
				deportation	
				deprivation of rights	x
				genocide	
		implementation of rights	x	restitution	x
		administration of justice		punishment	
				prison	x
		armed forces			
institutions and organizations		state institutions and organizations			
		public administration		central government	
				local government	
		non-state institutions and organizations	x	advocacy organizations	x
				private voluntary organizations	
		management			
		decision making			
groups		sex and gender			
		minority groups		Roma	

				Jews	
				disabled persons	
				nationality	
				LGBTQI+	
		communities			
		generations (age)			
		foreigners		immigrants	
				refugees	
social phenomenon	x	population dynamics		migration	
		integration		assimilation	
		social exclusion		discrimination	
				segregation	
		social participation		civil activism	x
		globalization			
social systems		social structure		working class	
				middle class	
				upper class	
				social inequality	
		capitalism			
		socialism			
relations		interpersonal relations		partnerships (personal)	
		intergroup relations			
		international relations			
		opponent	x		
		family socialization			
		conflict			
		violence			
history		historical period and event	x	war	
				holocaust	
				liberation	x
				revolution	x

Journal of Documentation (forthcoming), DOI: 10.1108/JD-12-2022-0269. Date of acceptance: 16-Sep-2023.

				Horthy-era	x
				Rákosi-era	x
				Kádár-era	x
				regime change	x
				period after regime change	x
science		sociology			
		scientific work	x	educational visit	
		scientific event	x		
		scientific publication	x		
crisis					
technology and innovation					