

EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

www.ehri-project.eu

Multilingual Automated Subject Indexing

A comparative study of LLMs vs alternative approaches in the context of the EHRI Project

DH BENELUX 2024 6th June 2024 Leuven, Belgium

EHRL is funded by the EU

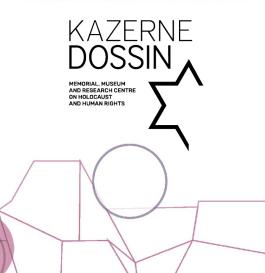
@EHRIproject

ESFRI

Maria Dermentzi (King's College London) Mike Bryant (King's College London, NIOD) Fabio Rovigo (Vienna Wiesenthal Institute for Holocaust Studies) Herminio García-González (Kazerne Dossin)



WIENER WIESENTHAL INSTITUT FÜR HOLOCAUST-STUDIEN (VWI)



♥ ± □ ♥ ● Ø Ø ≤ =

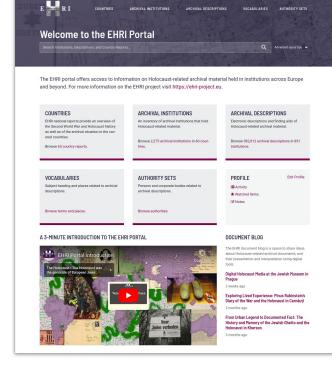
– п х



www.ehri-project.eu

EHRI's Mission

- The European Holocaust Research Infrastructure (EHRI) seeks to improve access to information about Holocaust-related archival material held by different institutions around the world
- One way it does this is the EHRI Portal, which makes collection metadata more accessible and interconnected





🗈 🗧 EHRI - Welcome to the × +

O A https://portal.ehri-project.eu

 $\leftarrow \rightarrow C$

Multilingual archival descriptions

For Research and Remembrance



Related Items **Filter Results** Q Advanced guery tips -Search Related Items.. LANGUAGE OF DESCRIPTION @ **Multilingual** English (304) **DISPLAYING ITEMS 1 TO 20 OF 309** - Relevance ID/Code Name Recently Updated Detail French (3) catalogue Give Them a Face portrait collection. Collection Hebrew (138) Kazerne Dossin; Memoriaal, Museum en Onderzoekscentrum over Holocaust en Mensenrechten / Kazerne Dossin; Mémorial, Musée et Centre de Recherche sur la Shoah et les Italian (2) **Droits Humains** KD_00017 English, Dutch 1880-1945 over 19,650 digitised images (individual photos) Updated 1 month ago Latvian (1) This collection contains over 19,650 portraits of Roma. Sinti and Jewish men, women and children from Belgium and the north of France, whom have Dutch (3) been deported from the SS-Sammellager Mecheln (Dossin barracks) to Auschwitz-Birkenau, Buchenwald, Ravensbrück, Bergen-Belsen and Vittel be-Diverse HOLDING INSTITUTION O tween August 1942 and July 1944. Kazerne Dossin: Memoriaal, Museum en

Julien Abrahamer, born in 1924; details regarding his activities in the French Jewish underground during World War II 0.89 - Collection of Personal Files of Jewish Underground Fighters in France

ארכיון יד ושם / Yad Vashem Archives 5263466 English, Hebrew xerox 4 pages/frames Updated 5 years ago

Julien Abrahamer, born in 1924; details regarding his activities in the French Jewish underground during World War II Life in Chambery. His work as assistant bookkeeper in the OSE (Children's Aid Society) office; detention with the office staff, 08 February 1944; due to his presence of mind, the bookkeeper's secretary was able to escape; deportation on Transport 69 to Auschwitz, 07 March 1944. In the file: - Photograph of Julien Abrahamer; -Summaries of his activities.

Abram Hofnung. Collection

Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaust en Mensenrechten / Kazerne Dossin: Mémorial, Musée et Centre de Recherche sur la Shoah et les Droits Humains

KD_00520 English 1942 10 digitised images (5 documents) Updated 1 month ago

This collection consists of five postcards sent by Abram Hofnung while detained at the Dossin barracks in August 1942 to his wife Bajla Taszman in Brussels.



ארכיון יד ושם / Yad Vashem Archives (140)

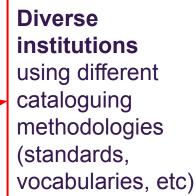
Holokauszt Emlékközpont (3)

Archives Nationales de France (2)

Biblioteca Archivio "Renato Maestro" (2)

Latvijas Nacionālā arhīva, Latvijas Valsts arhīvs (1)

Mémorial de la Shoah Fondation (1)







EHRI is funded by the EU





EUROPEAN HOLOCAUSI RESEARCH INFRASTRUCTUR

www.enri-project.eu

The EHRI Terms Vocabulary

- Hierarchically organised, multilingual set of subject headings
- At the time of writing consists of **913 terms** translated in **12 languages**
- Used by EHRI for:
 - Manually cataloguing descriptions in the EHRI Portal
 - Aligning with subject headings given on descriptions shared by EHRI partner institutions
- Available in the Simple Knowledge Organisation System (SKOS) RDF format

https://portal.ehri-project.eu/vocabularies/ehri_terms



EHRI Terms: usage example

For Research and Remembrance



EHRI Terms / Deportations / People		
Deportees	KEYWORDS	
Deportees	LANGUAGE	
	English - Deportees 🔻	
	ONLINE HISTORY	
	5 years ago Link created Show full history	
Related Items	O Advanced query line - Filter Results	
Search Related Items	Q Advanced query tips - Filter Results	
DISPLAYING ITEMS 1 TO 20 OF 277	LANGUAGE OF DESCRIPTION English (272)	
DISPLAYING ITEMS 1 TO 20 OF 277 Give Them a Face portrait collection. Collection	Relevance ID/Code Name Recently Updated Detail English (272) French (3)	
Give Them a Face portrait collection. Collection Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaus Droits Humains	English (272) French (3) Hebrew (138) Italian (2)	
Give Them a Face portrait collection. Collection Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaus Droits Humains K0_00017 English, Dutch 1880-1945 over 19,650 digitised images (individ This collection contains over 19,650 portraits of Roma, Sinti and J	Relevance ID/Code Name Recently Updated Detail rench (3) Hebrew (138) Italian (2) Iativan (1) Dutch (4)	
Give Them a Face portrait collection. Collection Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaus Droits Humains K0_00017 English, Dutch 1880-1945 over 19,650 digitised images (individ This collection contains over 19,650 portraits of Roma, Sinti and J	EANGUAGE OF DESCRIPTION Relevance ID/Code Name Recently Updated Detail English (272) French (3) Hebrew (138) Italian (2) Italian (2) Latvian (1)	
Give Them a Face portrait collection. Collection Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaus Droits Humains KD_00017 English, Dutch 1880-1945 over 19,650 digitised images (individ This collection contains over 19,650 portraits of Roma, Sinti and been deported from the SS-Sammellager Mecheln (Dossin barrac	English (272) French (3) Hebrew (138) Italian (2) Latvian (1) withish men, women and children from Belgium and the north of France, whom have s) to Auschwitz-Birkenau, Buchenwald, Ravensbrück, Bergen-Belsen and Vittel be-	ves (140)
Give Them a Face portrait collection. Collection Kazerne Dossin: Memoriaal, Museum en Onderzoekscentrum over Holocaus Droits Humains KD_00017 English, Dutch 1880-1945 over 19,650 digitised images (individ This collection contains over 19,650 portraits of Roma, Sinti and J been deported from the SS-Sammellager Mecheln (Dossin barrac tween August 1942 and July 1944.	EANBUAGE OF DESCRIPTION Relevance ID/Code Name Recently Updated Detail Relevance ID/Code Name Recently Updated Detail English (272) French (3) Hebrew (138) Italian (2) Latvian (1) Dutch (4) bi to Auschwitz-Birkenau, Buchenwald, Ravensbrück, Bergen-Belsen and Vittel be- HOLDING INSTITUTION O HOLDING INSTITUTION O HOL	eum en



Example of Subject Access Points

For Research and Remembrance



	EXISTENCE AND LOCATION OF Originals	Directorate-General War Victims, Brussels. The list of transport IV was lost and has been recreated using the index cards of the Sicherheitspolizei- Sicherheitsdienst (Sipo-SD).
	EXISTENCE AND LOCATION OF Copies	Directorate-general War Victims, Brussels ; USHMM, Washington DC
ſ	SUBJECTS	Transit camps Sondertransporte Deportees
	PLACES	Mechelen





Harmonising subject term usage in the EHRI Portal:

- 70% of descriptions with access points do not align with EHRI Terms common vocabulary
- 25% of collection-level descriptions have no access points at all





Research Question

Can **automated subject indexing** make archival descriptions in the EHRI Portal **more discoverable at scale** via:

- 1. Out-of-the-box, general-purpose tools, or...
- 2. Training custom tools on the Portal's existing metadata





Typical Subject Indexing Process



Determine what its subject content is Decide what subjects should be indexed Translate subject(s) into specific index terms





RESEARCH INFRASTRUCTURE

Automated Subject Indexing (ASI)

Automated subject indexing is the use of machine-based methods to perform the subject indexing steps typically performed by human indexers (Golub, 2021).



ASI: Process





Depending on the **institution's strategy**, resulting terms may be added to the record:

- Automatically (ASI)
- Semi-automatically after expert verification (semi-ASI)





ASI: Approaches

Two prevailing approaches (Suominen et al., 2022, Toepfer and Seifert, 2020):

- Statistical associative approach (multilabel classification)
- Lexical approaches (string-matching to match terms in a controlled vocabulary with words in the text of the record's description) (Golub, 2021)

PLUS:

- Fusion approaches that combine the two (ensemble models)
- **Transfer learning using "zero-shot" classification** (Zhang et al., 2023). LLMs used out-of-the-box to predict suitable terms from a controlled vocabulary.





ASI & EHRI

EHRI has been experimenting with most of the aforementioned approaches.

- The text and the subject terms of already ingested EHRI Portal archival descriptions are used as a dataset for training ML models.
- These models predict suitable EHRI Terms for any text we pass through them based on what terms were assigned to similar descriptions in the past.



ASI: Example of Archival Description Text

	Astivity People		Help + Advan Site 🛃 +	
	United States / United States	: Holocaust Memorial Museum		
	Oral history in	terview with Jakob Lapides	Addenius, pescarimonia Addenius Page	
	LANDJAGE OF DESCRIPTION	English	🗘 Wedch Item	
	KT. DOMIFICHS	2005.603.4 RG-50.637.0004	ARCHINAL CONTEXT	
	LEVEL OF DESCRIPTION	ltem	United States 4- United States Holocaust Memorial	
	LAMPUAGES	Hebrew	Maseum	
	SOURCE	EHRI Partner	EUFORT MEXAANA ▲ EAD 2002 XML	
	WEB SOURCE	https://collections.ushmm.org/search/catalog/im44161	▲ EAD-3 100L	
	DOTEST AND HEIRUP	1 videocassette (DVCAM), sound, color ; 1/4 in.	OWLINE INSTORY	
			11 months ago Incres imported	
_	DIEATORIES	Polity, Ms. Teresa A.	Show full history	
	BOGAAPHICAL HISTORY	Associate Curator Teresa Polin conducted the interview with Jakos March 27, 2005 for the United States Holocaust Memorial Museum Voices from the Lodz Ghetto' exhibition.		
	ARCHIVE, HISTORY	United States Holocaust Memorial Museum		
	CONDITIONS OF VERMIN ACCESS	No restrictions on access		
	PEPLE	Lapóles, Jakob Polite, Ma: Teresa A. Rumikowski, Mordecal Hayim. Polite, Ma: Teresa A. Lapóles, Jakob,		
	CORPORATE BODIES	Lapites, Jakob. Betar. Crestochowa (Concentration carro)		
	5/8-/C75	Czestochowa (Concentration camp) Jewish youth-Poland-Societies and clubs.	T	
		Men-Personal nerrotives. Lódz (Poland)	Associate Curator 16	resa Pollin conducted the interview with Jakob Lapides in Israel on
		Bar mitzveh.		
		Czestochowa (Poland) Jewish orphans-Poland-Lódž.	March 27 2005 for t	ne United States Holocaust Memorial Museum's "Give Me Your Children:
		Jews-Paland-Lódź. Zlotists.	Warch 27, 2005 101 t	le officer states holocaust Meriorial Museurits. Give Me rour children.
		Polend-History-Occupation, 1939-1945. Holocaust, Jewith (1939-1945)-Personal nerratives.		
		World War, 1939-1945-Conscript labor. Israel-Emigration and immigration.	Voices from the Lod	z Ghetto" exhibition
		Jewish oplice officers-Poland. Jewish police officers-Poland.	voices nom the Lou	Conecto exhibition.
		Holoceust survivors-Israel-Interviews.		
		Rationing-Poland. World War, 1939-1945-Deportations from Poland.		
		Jewish ghettos-Polend-Lód2. Forced labor.		
	GENRE	Onal History		
	This description is derived should be	inro structured data provided to EVM by a partner institution. This collection building it	national considerate this	
	description as an accurate reflection	rare amazzando data provided to zirve by a partner instituzion. I ha conection oursing a 1 of the archival baldings to which it refers at the moment of data transfer.		
	Notes on this description			
	C) ADDINON			
	Keep up with EHRI news		English Français Deutsch Polisi Españal	
	Subscribe to our mailing list			
	fracebook ¥ twitten			
	0 Copyright ID-40 Consortium 2023 Th	e 10-99 Project is supported by the Suropean Commission		
	Terms & Conditions Data Folioy	API Emountly Asked Questions	FRIARS	

ASI: The goal

Associate Curator Teresa Pollin conducted the interview with Jakob Lapides in Israel on March 27, 2005 for the United States Holocaust Memorial Museum's "Give Me Your Children: Voices from the Lodz Ghetto" exhibition.

Given some text in any language

ASI: The goal

Associate Curator Teresa Pollin conducted the interview with Jakob Lapides in Israel on March 27, 2005 for the United States Holocaust Memorial Museum's "Give Me Your Children: Voices from the Lodz Ghetto" exhibition.

- Given some text in any language
 Predict suitable EHRI
 - Terms





EHRI Experiments

• Hugging Face:

- Fine-tuning for MLC: BERT-base Multilingual Cased (Devlin et al., 2019)
- Zero-shot classification using the pre-trained mDeBERTa-v3-base-mnli-xnli (Laurer et al., 2023)
- Annif (Suominen et al. 2022):
 - Framework for *training*, *evaluating* and *combining* various non-LLM based ASI tools
 - TF-IDF, MLLM, fastText, Omikuji, NN Ensemble backends



EHRI Experiments

Key considerations in selecting the Transformer models:

- pre-trained on languages relevant to EHRI's material
- being open source and easily accessible
- transparent about training data used
- ability to quantitatively evaluate their performance using our in-house compute infrastructure

EHRI Experiments: Dataset

Associate Curator Teresa Pollin conducted the interview with Jakob Lapides in Israel on March 27, 2005 for the United States Holocaust Memorial Museum's "Give Me Your Children: Voices from the Lodz Ghetto" exhibition.

Text* and associated subject terms (linked with the EHRI Terms vocabulary) of around 37,000 archival descriptions in 15 languages.

*The fields selected were ISAD(G) elements 3.1.2. Title; 3.2.2. Administrative / Biographical History; 3.2.3. Archival History; and 3.3.1. Scope and Content.

SUBJECTS

Jewish youth--Poland--Societies and clubs. Men-Personal narratives. Łódź (Poland) Bar mitzvah. Czestochowa (Poland) Jewish orphans--Poland--Łódź. Jews--Poland--Łódź. **Zionists** Poland--History--Occupation, 1939-1945. Holocaust, Jewish (1939-1945)--Personal narratives. World War, 1939-1945--Conscript labor. Israel-Emigration and immigration. Jewish orphanages--Poland. Jewish police officers--Poland--Łódź. Holocaust survivors--Israel--Interviews Rationing--Poland. World War, 1939-1945--Deportations from Poland. Jewish ghettos--Poland--Łódź. Forced labor.

Experiments: Dataset

- To obtain our training and test sets we used iterative multilabel stratification (Sechidis et al., 2011) as implemented in the Python package iterative-stratification (Bradberry, 2018/2021).
- We did this to increase the chances that all labels (and language codes) would have positive examples in the train set as well as in the test set.



Experiments: Evaluation

- Create Annif backends for LLM-based MLCs
 - Off-the-shelf zero-shot models
 - Fine-tuned EHRI model
- Compare zero-shot, fine-tuned and Annif models
- Perform quantitative and qualitative evaluation of ~160 examples





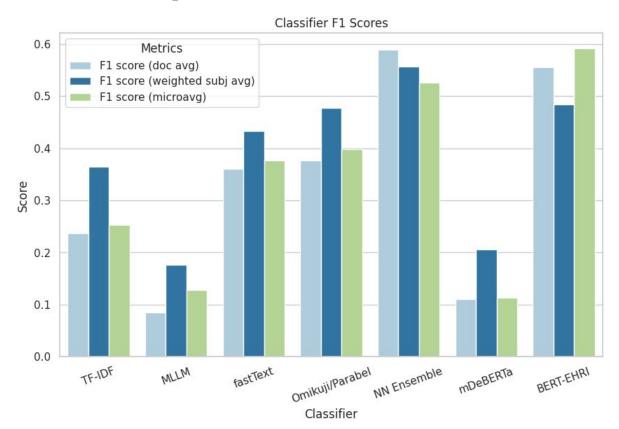
ROPEAN HOLOCAUST SEARCH INFRASTRUCTURE

Experiment: quantitative evaluation

For quantitative evaluation, we **assume** that the terms assigned to EHRI-ingested records are complete and correct (gold standard)

- Computed F1 scores:
 - Micro-averaged (taking into account total true positives, false negatives and false positives)
 - Weighted macro-averaged (computing the metrics per label weighted by each label's number of true instances)
 - Averaged per document
- All of these scores range from zero to one, with one being the best possible score.

Experiment: quantitative evaluation







Experiment: human evaluation

Why?

Quantitative evaluation doesn't necessarily tell the full story

- Subject indexing with large controlled vocabularies is a very **subjective task**—small overlap between human cataloguers
- Hard to develop a truly comprehensive gold standard dataset
- Our training dataset is derived from heterogeneous data, leading to inconsistencies





Experiment: human evaluation

Qualitative evaluation of four distinct label sets:

- Original labels assigned by EHRI staff or co-referenced with data from EHRI partner institutions
- NN Ensemble classifier (best performing Annif backend according to our quantitative evaluation)
- Fine-tuned BERT-based model
- mDeBERTa-based zero-shot model



EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

www.ehri-project.eu

Experiment: human evaluation

С	D	E	F	G H	1	J	K L	M	N	O P	Q	
	Tool 1			Tool 2			Tool 3			Tool 4	Tool 4	
fext	Labels	Good	Poor	Labels	Good	Poor	Labels	Good	Poor	Labels	Good	
4759	Healthcare			People			Diseases			Healthcare		
lemocný v nemocnici v Terezíně	Diseases			Healthcare			Healthcare			Food and eating		
atient in a hospital in Terezin				Prohibitions and bans						Diseases		
				Medical						Daily life		
			-	Pamphlets						Judenrat		
inkuva photographs	Photographs			Prewar period			Photographs			Photographs		
lder				Jewish population								
				Photographs								
nited States Holocaust Memorial Museum				Jewish culture								
nited States Holocaust Memorial Museum				Holocaust								
onsists of eight photographs of pre-war Jewish life in nkuva, Lithuania. Includes photographs of homes, the				memory								
Feissel Florette ילידת Mulhouse 1922 צרפת, פרטים על פעילותה במחתרת היהודית בצרפת בזמן מלחמת העולם				Imprisonment			Deportees			Photographs		
	Help to Jews			Ghetto life			Detentions			Personal documentation		
קסרוקס pages/frames 6	Relief and			Wartime period			Personal documentation			Deportees		
Feissel Florette ילידת Mulhouse 1922 צרפת, פרטים על פעילותה במחתרת היהודית בצרפת בזמן מלחמת העולם				Deportation to camps			Camps			Detentions		
השנייה בריחה עם היהודים ל-Lyon ביולי 1940; החיים עם המשפחה במחנה Chaudanne בסוף 1940; קבלת סיוע מארו Aide Sociale aux Refugie: קבלת עבודה בתוו				Arbeitseinsatz			Photographs			Commemoratio n		
מזכירה בארגון הודות להמלצת הרב Hirschler Rene (רבה	1											



EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

www.ehri-project.eu

Experiment: human evaluation

	GS Labels			NN Ensemble			BERT-FT			mDeBERTa		
	F	N	U	F	N	U	F	N	U	F	N	U
	221	69	4	267	252	37	233	179	28	340	312	183
J1	75%	23%	1%	48%	45%	7%	53%	41%	6%	41%	37%	22%
	210	83	1	239	288	29	197	213	30	342	444	49
J2	71%	28%	0%	43%	52%	5%	45%	48%	7%	41%	53%	6%
	191	99	4	238	303	15	217	209	14	344	460	31
J3	65%	34%	1%	43%	55%	3%	49%	48%	3%	41%	55%	4%

Human judges rated many of the zero-shot predictions as favourable, although the model also achieved some of the highest neutral and unfavourable scores.





ROPEAN HOLOCAUST SEARCH INFRASTRUCTURE

Limitations, Concerns, Discussion

- Qualitative evaluation was very limited (only meant as a preliminary check)
- Google Translate was used for descriptions in languages that authors did not understand
- No access to the actual archival material but only to metadata
- Limited metadata with insufficiently descriptive text
 - Many labels therefore judged "neutral", inconclusive
 - This could contribute to relatively low performance of all classifiers





RESEARCH INFRASTRUCTURE

Limitations, Concerns, Discussion

- Signs of **spurious correlations**
- Index terms correlating with oral testimony interviewees' names
- Index terms correlating with institution or project name
 - "The USHMM, Oral History Branch, in cooperation with Watchtower Bible and Tract Society, Inc" => "Jehovah's Witnesses"
- Several cases produced labels which, in context, would be deemed inappropriate
 - e.g. phrase "produced in cooperation with" => "collaborators"





ROPEAN HOLOCAUST

Limitations, Concerns, Discussion

• Hierarchical vocabulary resulted in some redundant, overlapping terms

e.g. "Political activities", "Politics", "Political organisations" and "Political movements"

 The model used for zero-shot classification was trained on a CommonCrawl dataset which might have already "seen" some of the descriptions in our dataset



Based on our experiments:

- ASI tools tested here only usable in a semi-ASI scenario
- Ability of LLM-based zero-shot classification to correctly predict more specific ("narrower") labels potentially useful as part of a data augmentation effort
- NN Ensemble model on a par with fine-tuned BERT (if not better)
- Future work: fine-tuning multilingual LLMs with a more balanced dataset, experimenting with more SOTA models





EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

www.ehri-project.eu

Thank you!

Abstract:







References

Bradberry, T. J. (2021). Iterative-stratification [Python]. https://github.com/trent-b/iterative-stratification (Original work published 2018)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. <u>https://doi.org/10.18653/v1/N19-1423</u>

Golub, K. (2021). Automated Subject Indexing: An Overview. Cataloging & Classification Quarterly, 59(8), Article 8. <u>https://doi.org/10.1080/01639374.2021.2012311</u>

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2023). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. Political Analysis, 1–17. https://doi.org/10.1017/pan.2023.20





References

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the Stratification of Multi-label Data. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 145–158). Springer. https://doi.org/10.1007/978-3-642-23808-6_10

Suominen, O., Inkinen, J., & Lehtinen, M. (2022). *Annif and Finto AI: Developing and Implementing Automated Subject Indexing.* JLIS.It, 13(1), Article 1. <u>https://doi.org/10.4403/jlis.it-12740</u>

Toepfer, M., & Seifert, C. (2020). Fusion architectures for automatic subject indexing under concept drift. International Journal on Digital Libraries, 21(2), Article 2. <u>https://doi.org/10.1007/s00799-018-0240-3</u>

EHRL is funded by the EU

@EHRIproject

Zhang, S., Wu, M., & Zhang, X. (2023). Utilising a Large Language Model to Annotate Subject Metadata: A Case Study in an Australian National Research Data Catalogue (arXiv:2310.11318). arXiv. <u>http://arxiv.org/abs/2310.11318</u>

ESFRI



EUROPEAN **HOLOCAUST** RESEARCH INFRASTRUCTURE

For Research and Remembrance

www.ehri-project.eu

