Exploratory Analysis of the Applicability of Formalised Knowledge to Personal Experience Narration

Victor Mireles, Stephanie Billib, Artem Revenko, Stefan Jänicke, Frank Uiterwaal, Pavel Pecina

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

Funded by the European Union

iDSC 2023

The project

MEMORISE: Virtualisation and Multimodal Exploration of Heritage on Nazi Persecution

Funded by **Horizon Europe** Programme of the European Commission

Funded by the European Union

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

The project

Take testimonials, diaries, etc.
by victims of national socialism

The project

Take testimonials, diaries, etc. by victims of national socialism

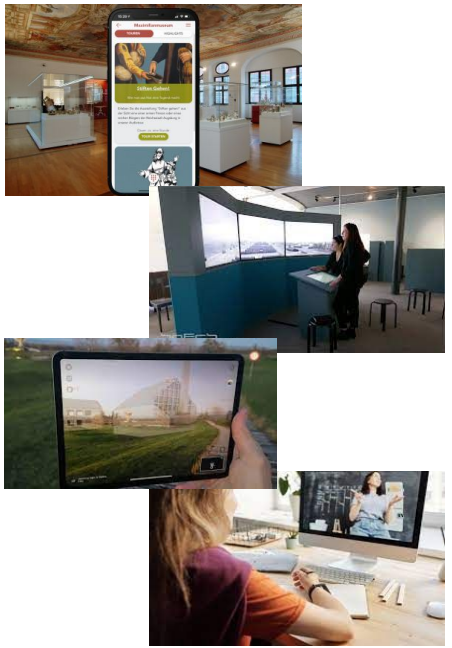
Organize them, link them and prepare narrative content

The project

Take testimonials, diaries, etc. by victims of national socialism

Organize them, link them and prepare narrative content

For visitors of memorial sites, students, researchers, etc.

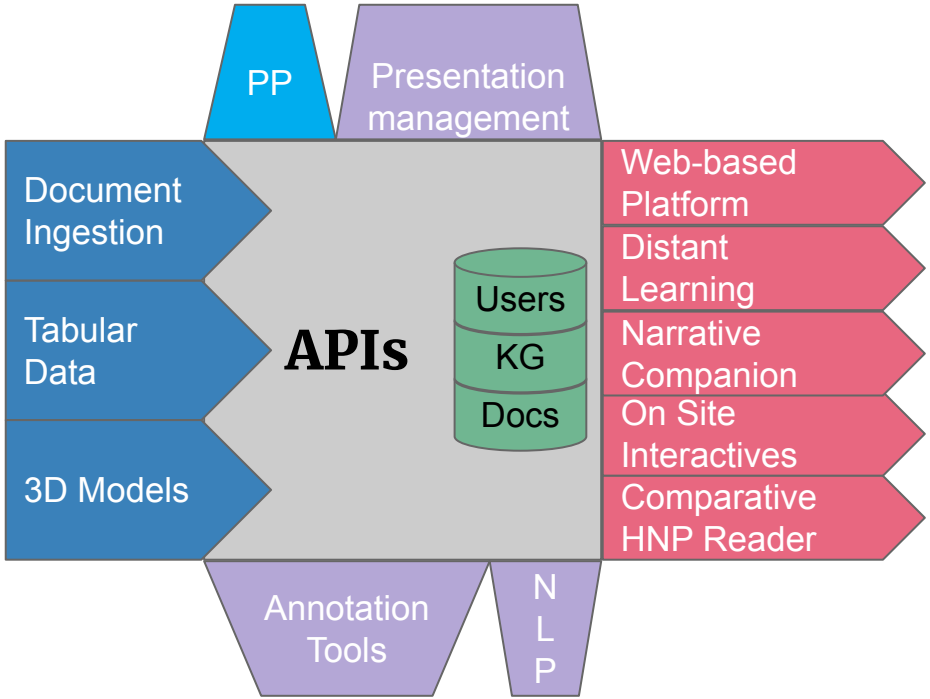


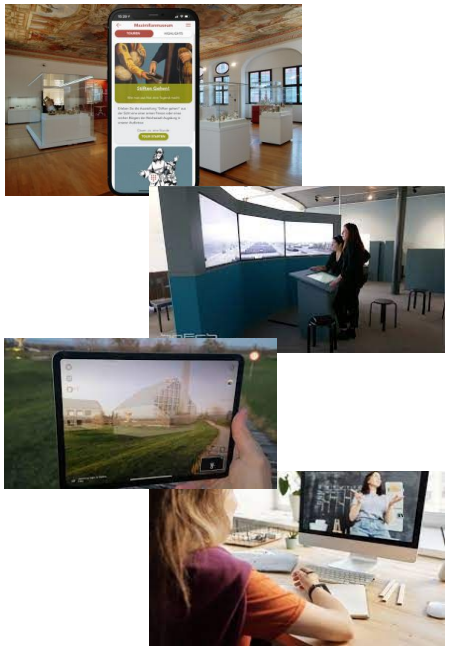
The project

Take testimonials, diaries, etc. by victims of national socialism

Organize them, link them and prepare narrative content

For visitors of memorial sites, students, researchers, etc.



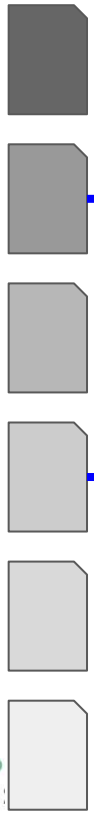


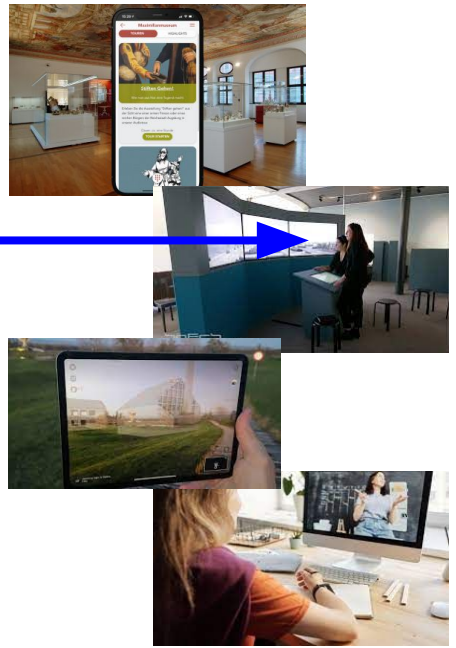
A knowledge graph

Take testimonials, diaries, etc.

Organize them, link them

For visitors of memorial sites, students, researchers, etc.



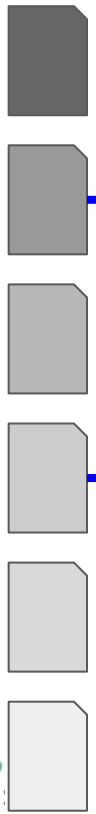


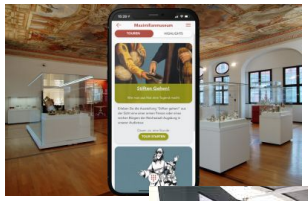
A knowledge graph

Take testimonials, diaries, etc.

Organize them, link them

For visitors of memorial sites, students, researchers, etc.





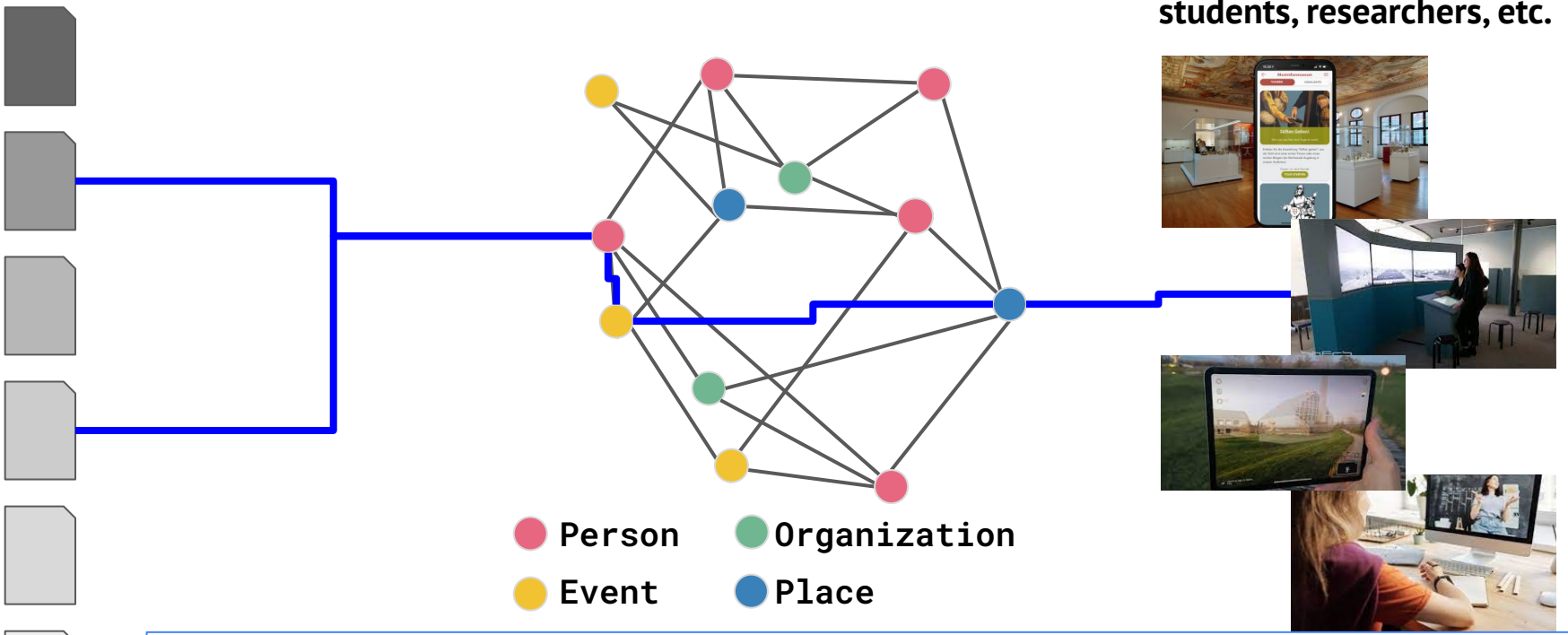
Show the user these documents because therein is mentioned a person involved in an event that occurred in the place where they are standing.

A knowledge graph

Take testimonials, diaries, etc.

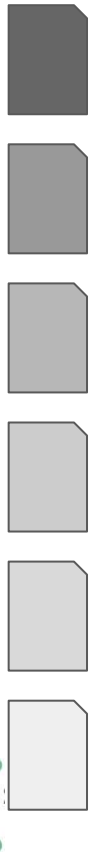
Organize them, link them

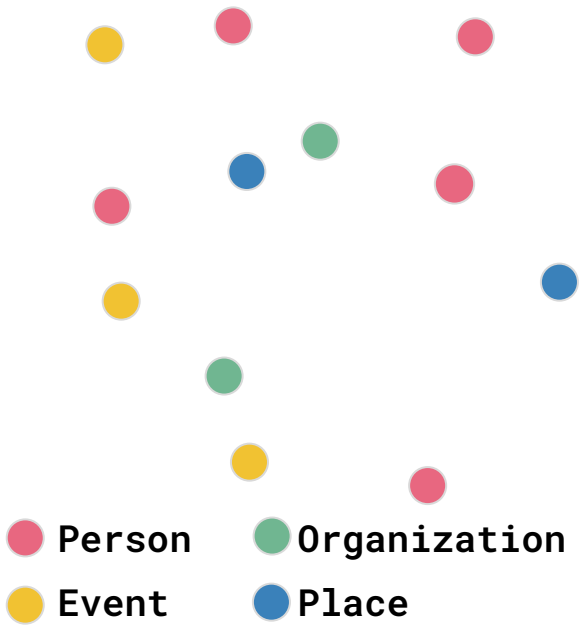
For visitors of memorial sites, students, researchers, etc.



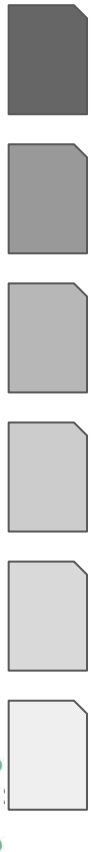
Show the user these documents because therein is mentioned a person involved in an event that occurred in the place where they are standing.

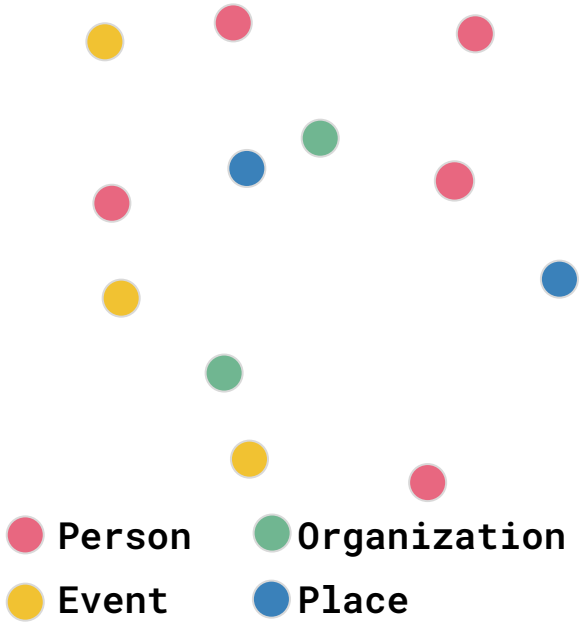
What are the actual entities?





What are the actual entities?





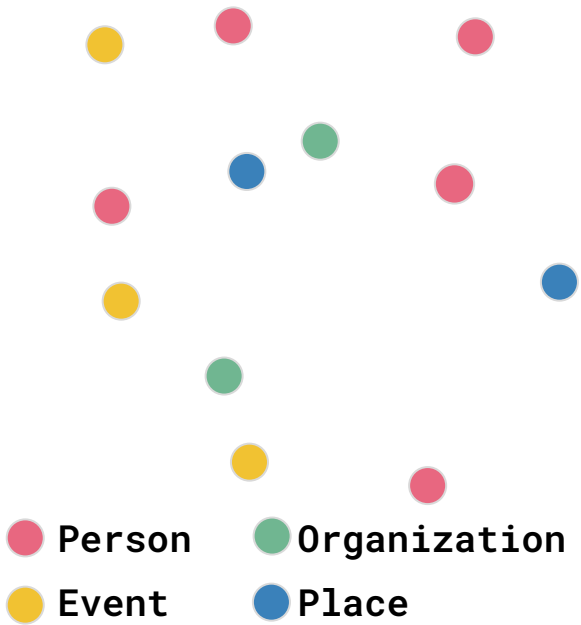
They must be:

- Interesting to the user.
- Distinguishable among themselves.
- Enrichable with as much extra data as possible.
- (Automatically) easy to find.

Approaches for choosing entities

What documents mention:

- Based on syntactic patterns
- No need for curation



What experts say:

- Curated list of entities
- Knowledge from many sources integrated.

They must be:

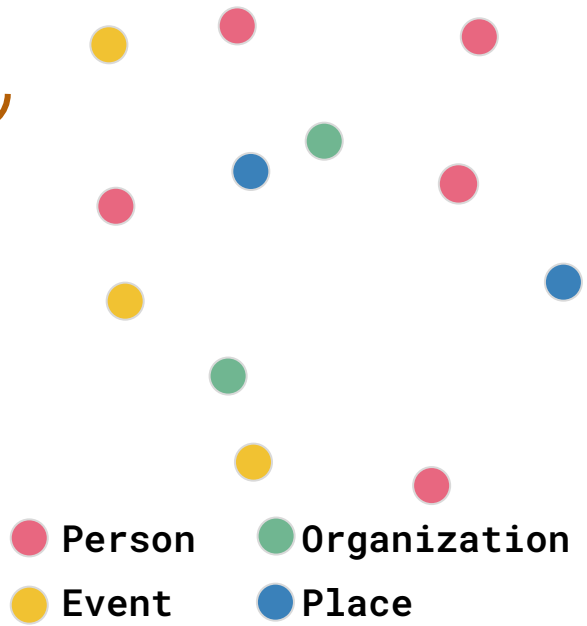
- Interesting to the user.
- Distinguishable among themselves.
- Enrichable with as much extra data as possible.
- (Automatically) easy to find.

Approaches for choosing entities

What documents mention:

- Based on syntactic patterns
- No need for curation

Named Entity Recognition



What experts say:

- Curated list of entities
- Knowledge from many sources integrated.

Entity Extraction

They must be:

- Interesting to the user.
- Distinguishable among themselves.
- Enrichable with as much extra data as possible.
- (Automatically) easy to find.

Approaches for choosing entities

Entity Extraction

NIOD WW2 Thesaurus

- Created by the **Dutch Institute for War, Holocaust and Genocide Studies** over the past decade.
  https://www.niod.nl/en/collections/ww2-thesaurus
- It has been used by the KNAW to catalogue over a million historical sources
- 6135 Entities
  - Organisations: 1431
  - Events: 1489
  - Internment Camps: 1629
  - Places: 433
  - General Concepts: 5676
- Multi-hierarchical up to 8 levels deep.
- 17646 SKOS-XL labels in Dutch, German and English

Approaches for choosing entities

Named Entity Recognition

German BERT fine-tuned on GermEval14

https://huggingface.co/fhswf/bert_de_ner

- F_1 of 86.89 on GermEval 2014
- F_1 of 85.52 on CoNLL-2003 test data sets.
- We use the four high-level classes:
  - LOC
  - PER
  - ORG
  - OTH

Data

Diaries from six inmates of the Bergen-Belsen concentration camp.

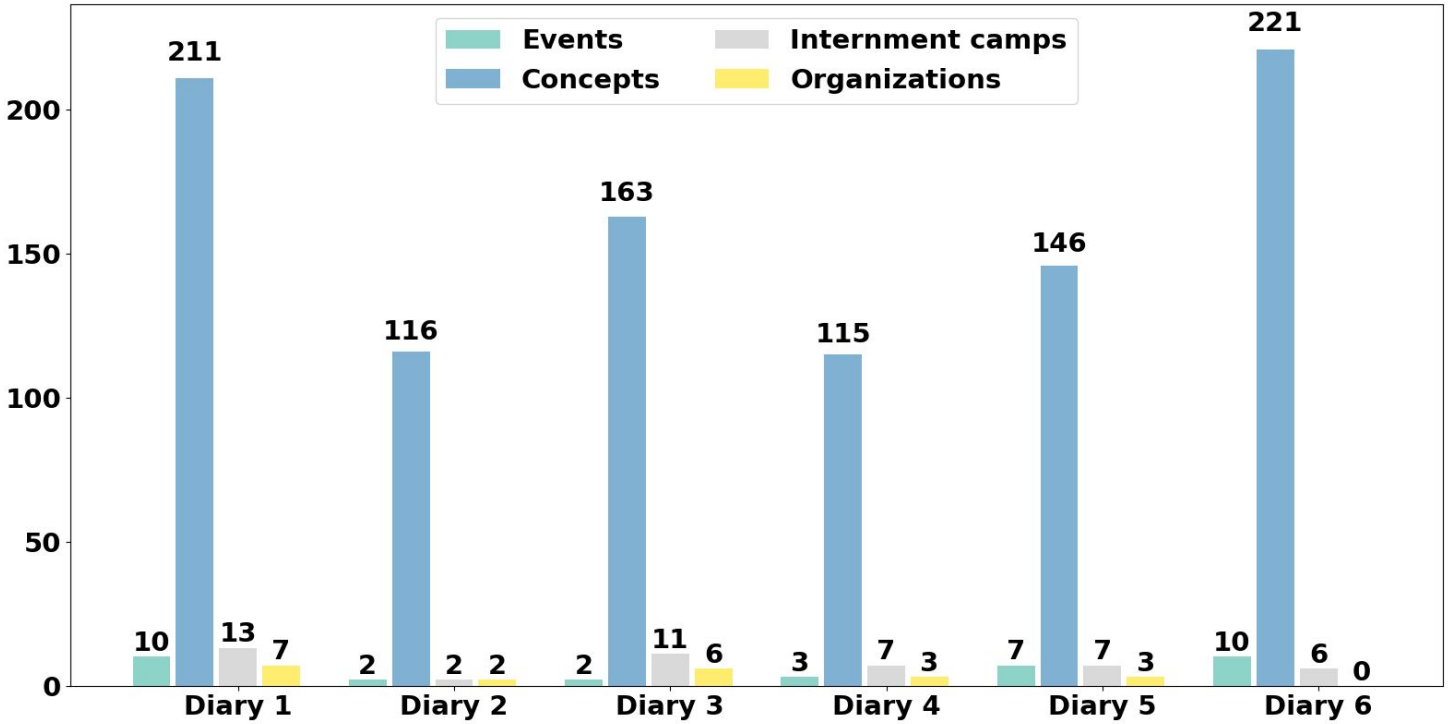
- Two female, four male authors
- From 4 to 30 months duration
- Informal vocabulary, with variations between authors.
- A common topic in all the diaries is hunger and the lack of food as well as the social behaviour in the authors' surrounding.

- Human-translated into German (except one, originally in DE)
- In total 228245 tokens

Results

Results - Entity Extraction

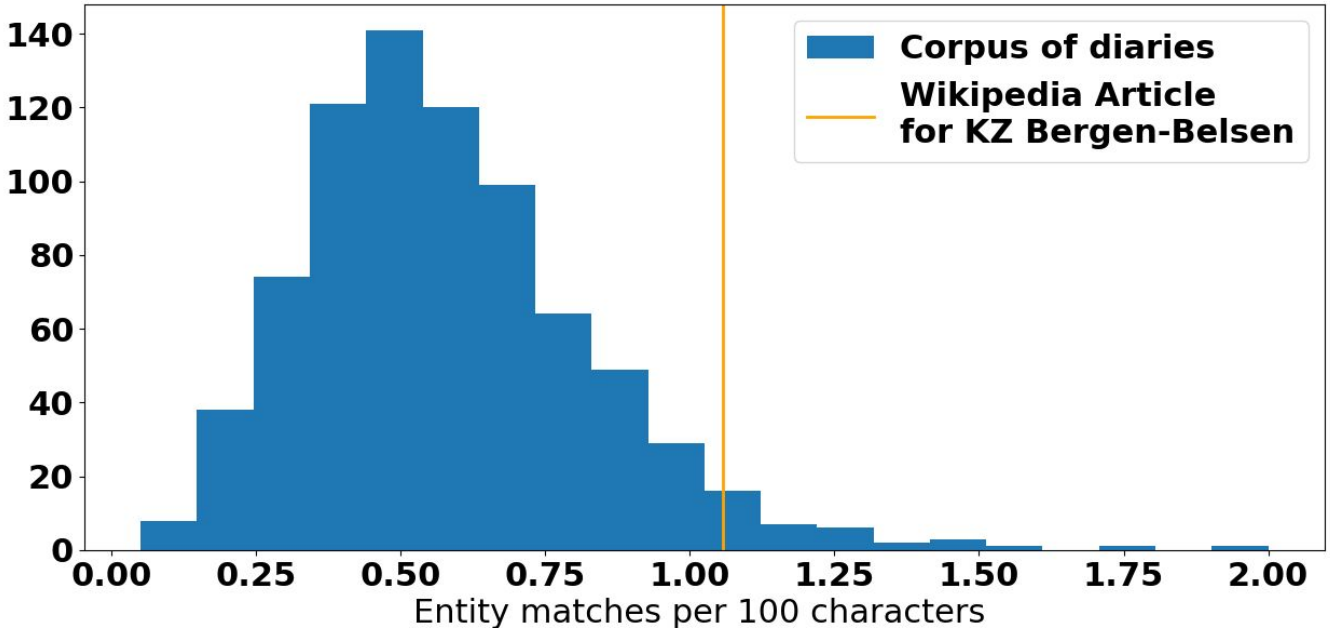
8428 concept matches in total.



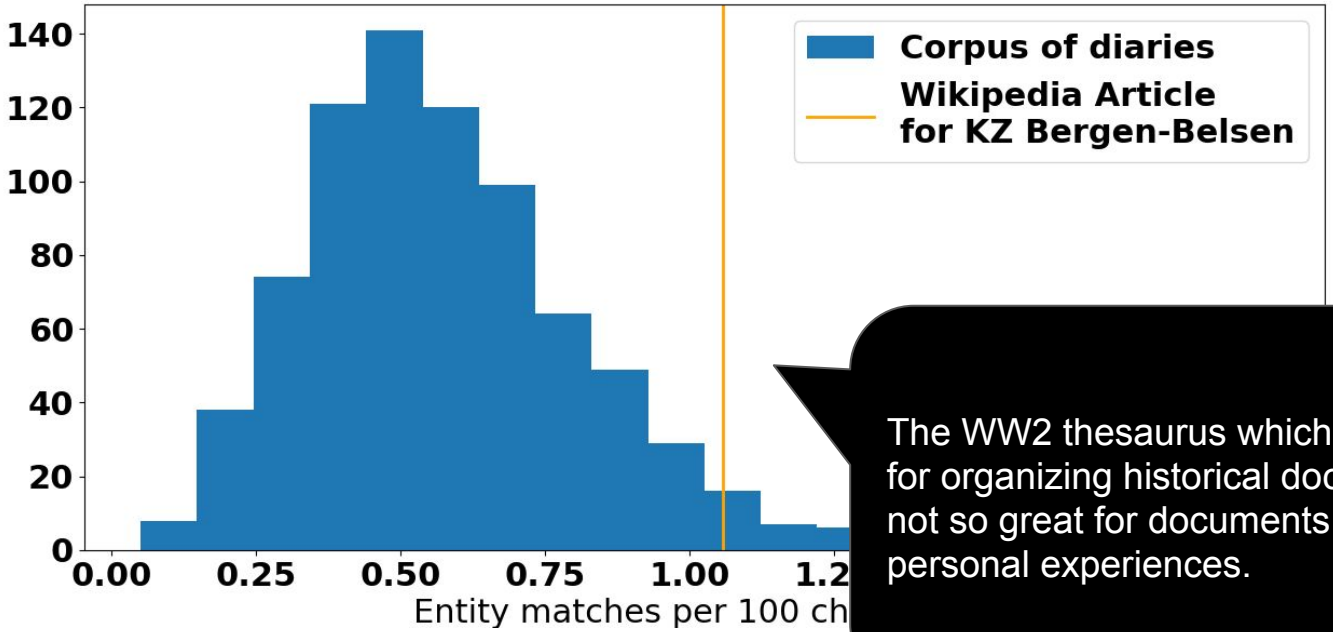
Results - Entity Extraction

Organisations	Concepts	Events	Internment camps
Durchgangslager Westerbork	Ostern	Warschauer Aufstand Beschusse	Lieben
Trawniki	Weihnachten	Ostfront	Vittel
Jewish Agency for Palestine	General-gouvernement	Schlacht	Ku
Sonderkommando	Tauschhandel	Abmarsch	Grünberg
SS	Desinfektion	Offens	Fossoli
American Jewish Joint Distribution Committee	Entlausung	Luftkrieg	Theresienstadt
World Jewish Congress	Hygiene	Luftangriff	Drancy
Waffen-SS	Täter	Landungen	Durchgangslager Westerbork
Rex	Nachschub	Invasion	Dachau
Reichssicherheitshauptamt	Hunde	Machtübernahme	Oranienburg
Luftwaffe	Helfer	Luftschutzkeller	Treblinka
Germanen	Ghetto	Attentat	Sachsenhausen
	Arbeitslager	Abwerfen	Poniatowa
	Kapo	Abzug	Neuengamme

Results - Entity Extraction



Results - Entity Extraction



The WW2 thesaurus which is great for organizing historical documents, is not so great for documents describing personal experiences.

Results - NER

2355 different strings recognized as NEs
2486 different NEs
7217 NER matches in total.

Document	LOC	ORG	OTH	PER	total
Diary 1	280	105	141	429	**955**
Diary 2	45	3	4	22	**74**
Diary 3	152	28	32	169	**381**
Diary 4	93	11	14	217	**335**
Diary 5	192	21	17	63	**293**
Diary 6	245	16	23	598	**882**
Whole Corpus	688	166	213	1419	2486

Results - Compared

2355 different strings recognized as NEs
Only 45 of them are actually in the thesaurus.
Many errors:

E.g. B*rande, Poniatowa* and *Drancy*, (names of internment camps), labelled PER

The thesaurus has a concept scheme with 1431 organisations, but only 3 out of the 166 found by NER

Named Entity Recognition

Entity Extraction

	LOC	ORG	OTH	PER
Concepts	19	3	6	6
Internment camps	18	0	1	3
Organisations	2	3	0	1
Events	0	0	0	0

Results - Compared

2355 different strings recognized as NEs
Only 45 of them are actually in the thesaurus.
Many errors:

E.g. B*rande, Poniatowa* and *Drancy*, (names of internment camps), labelled PER

The thesaurus has a concept scheme with 1431 organisations, but only 3 out of the 166 found by NER

The thesaurus is missing many concepts that are actually used by people.

Named Entity Recognition

Entity Extraction

	LOC	ORG	OTH	PER
s	19	3	6	6
internment camps	18	0	1	3
Organisations	2	3	0	1
Events	0	0	0	0

Conclusions

- We need to augment the thesaurus and other sources of formalized knowledge if we want to organized person-centric documents.
- NER can play a role in this
- But NER needs curation

The thesaurus is missing many concepts that are actually used by people.

The WW2 thesaurus which is great for organizing historical documents, is not so great for documents describing personal experiences.

SEMANTIC WEB COMPANY
linking data to knowledge

Memorise

Gedenkstätte
Bergen-Belsen

Stiftung
niedersächsische
Gedenkstätten

SDU

niod

CHARLES
UNIVERSITY

Questions ?

Funded by
the European Union

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.