Semantische Personensuche auf Hochschulwebsites

Johannes Güttler

Albert-Ludwigs-Universität Freiburg im Breisgau

01. September 2017

Überblick

1. Motivation
2. Informationsextraktion
3. Umsetzung
4. Ergebnisse

Motivation

Ziel:

Suche nach Wissenschaftlern / Experten eines Fachgebietes (in einem bestimmten geographischem Gebiet)

Problem:

- Volltextsuche funktioniert nicht gut
- Listen i.d.R. nicht verfügbar

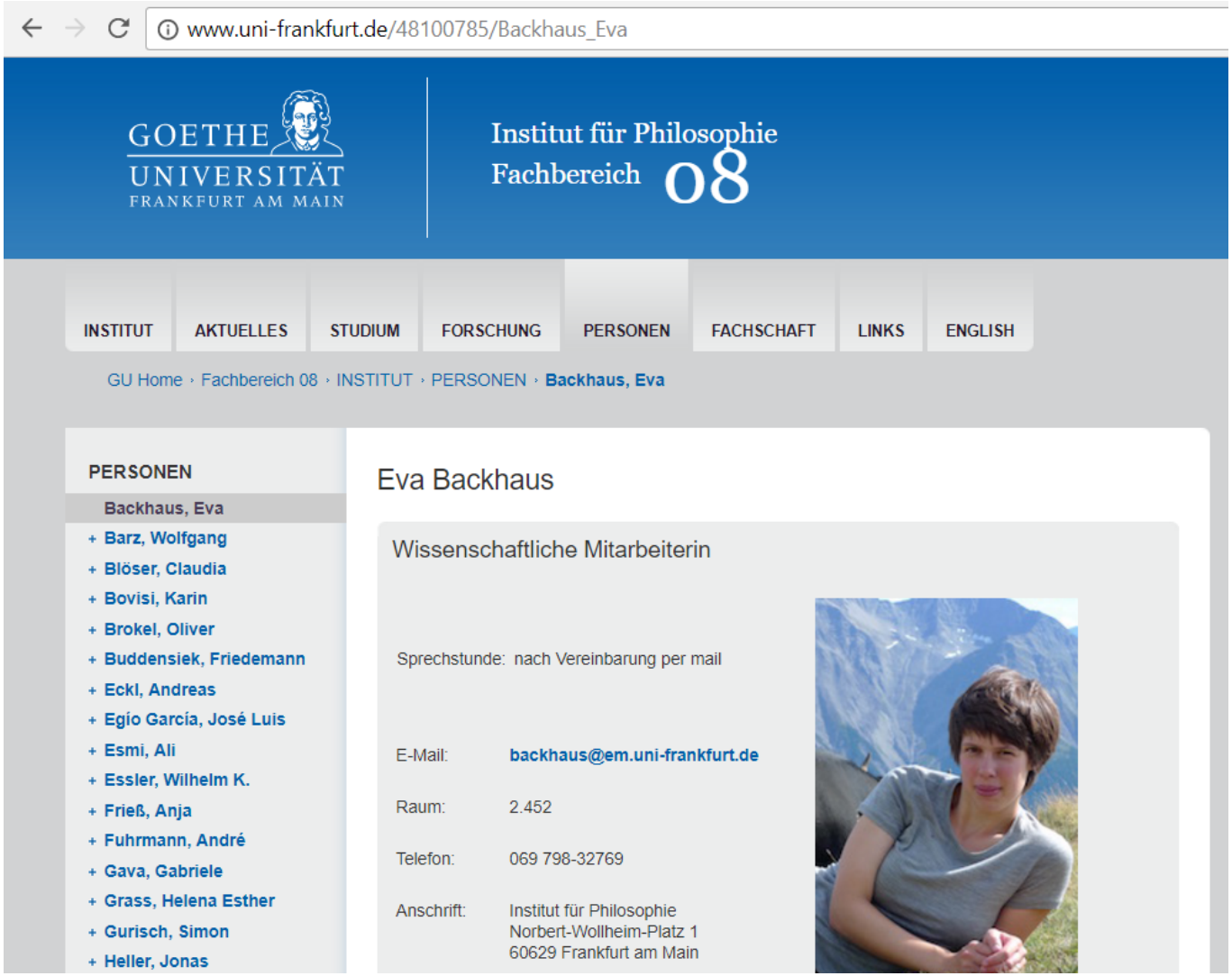
$\rightarrow$ Sehr mühsam

Demo

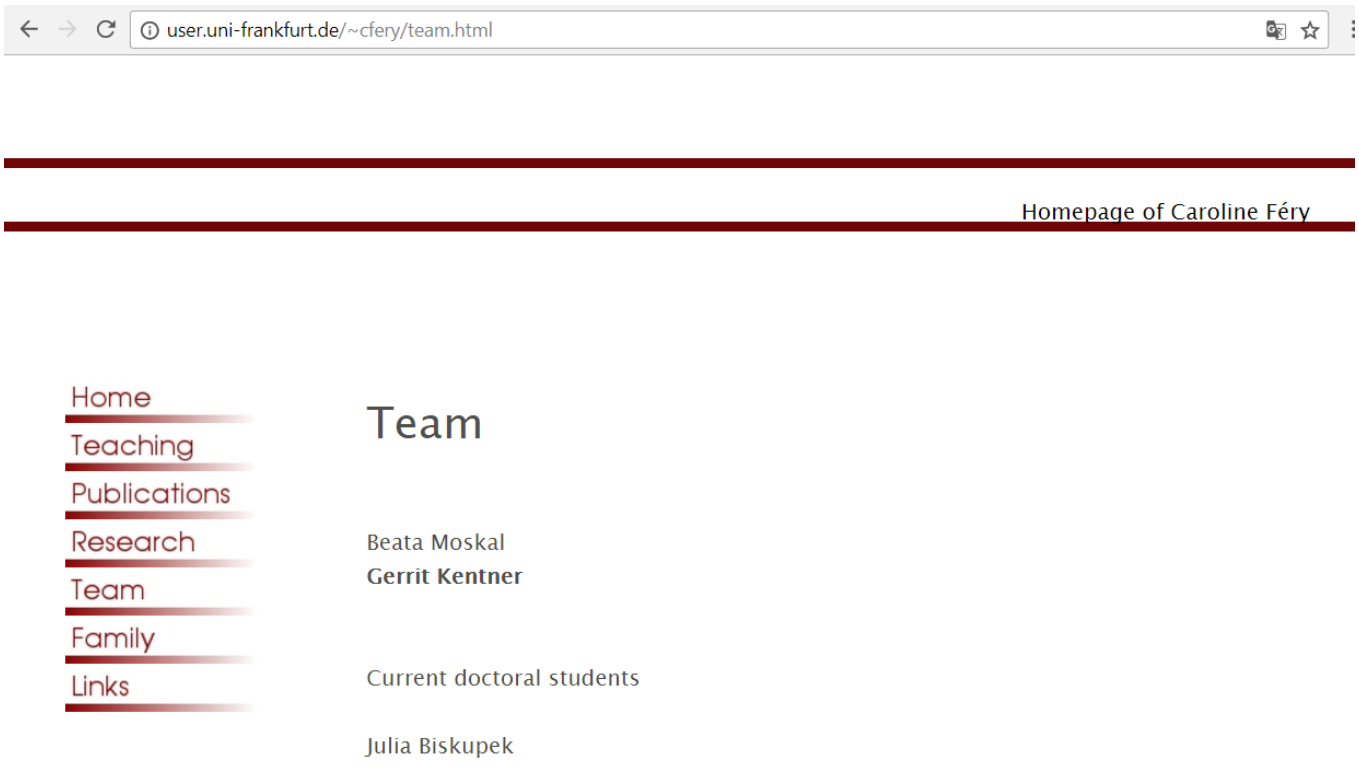
Ziel

- Datengrundlage: Websites deutscher Hochschulen
- Extraktion der Mitarbeitenden dieser Hochschulen
- Extraktion von Eigenschaften
  - Hochschule
  - Homepage und Subdomain
  - Titel (Prof., Dr.)
  - Stadt / Bundesland
  - Geschlecht
- Verfügbarkeit der Textabschnitte in denen die Person genannt wird.

Daten auf Hochschulwebsites - persönliche Informationen



Daten auf Hochschulwebsites - persönliche Informationen



Daten

- Viele Personen ohne eigene Homepage
- Titel und URLs oft nicht aussagekräftig

→ Nach persönlichen Homepages zu suchen reicht nicht aus

→ Ca. 9 Mio. Seiten

Umsetzung - Daten und Namen

1. Extraktion der Websitetexte mittels MapReduce
2. Entfernen bekannter Bibliotheks- und Publikationsseiten
   - Verringerung der Anzahl falsch-positiver Treffer
3. Namensextraktion: Stanford NER
   - Verfahren: Conditional Random Fields
   - Liefert ungeordnete Namesliste
   - Regelbasierte Umsetzung in vollständige Namen
4. Regelbasierte Personenerstellung
5. Liste: Hochschulen, Städte, Bundesländer
6. Ontologie und Textdatenbank

Evaluation - Vorgehen

- Groundtruth
  - Praktisch keine Listen vorhanden
  - Wer soll (nicht) einbezogen werden?
- Evaluiert wird immer auch die NER

Evaluation - Ergebnisse

Nr.	# Personen	Präzision	Recall	f
1	29	0.22	0.21	0.21
2	18	0.70	0.89	0.78

1. Personen mit Fachbereich Psychologie in Schleswig-Holstein
2. Mitarbeiterinnen und Mitarbeiter am Institut für Hirnforschung Uni Bremen mit eigener Homepage

Evaluation - Ergebnisse

Gesamt: 19 Suchanfragen mit 865 Personen.

Found	true positive	false positive	false negative
1590	**308**	**1282**	**557**

Effektivität:

Präzision	Recall	F-Maß
0.34	**0.40**	**0.34**

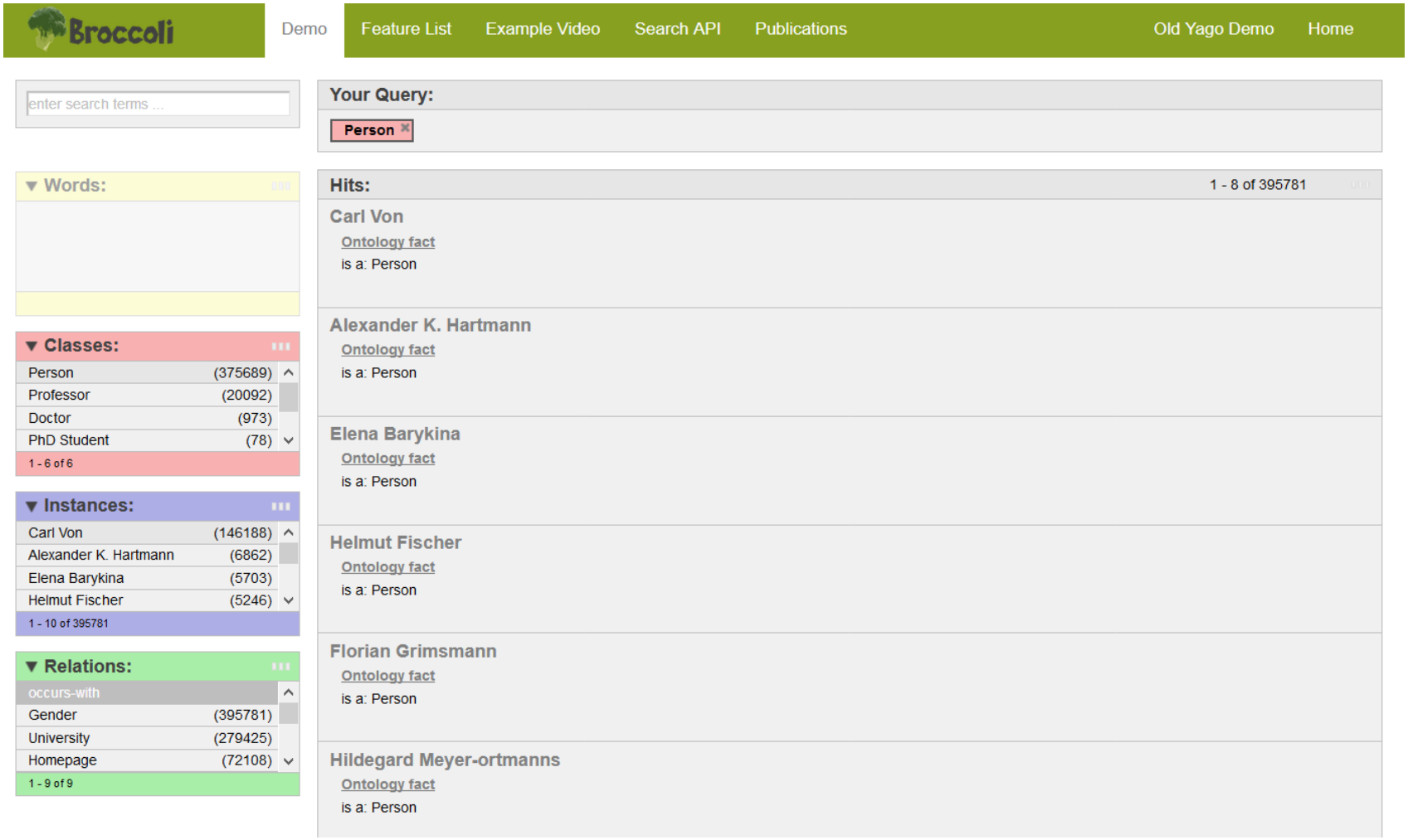
Evaluation - Allgemein

- Wie sieht die *korrekte* Groundtruth aus?
- Je kleiner der (geografische) Bereich der Anfrage desto besser das Ergebnis
- Präzision: 0.05 bis 0.86
  - ↳ hat oft aber trotzdem mit dem Thema zu tun
- Recall: 0.10 bis 0.90
  - ↳ oft nicht korrekt erkannt oder falsche HS
- Fehlende Daten: lmu.de oder uni-muenchen.de?

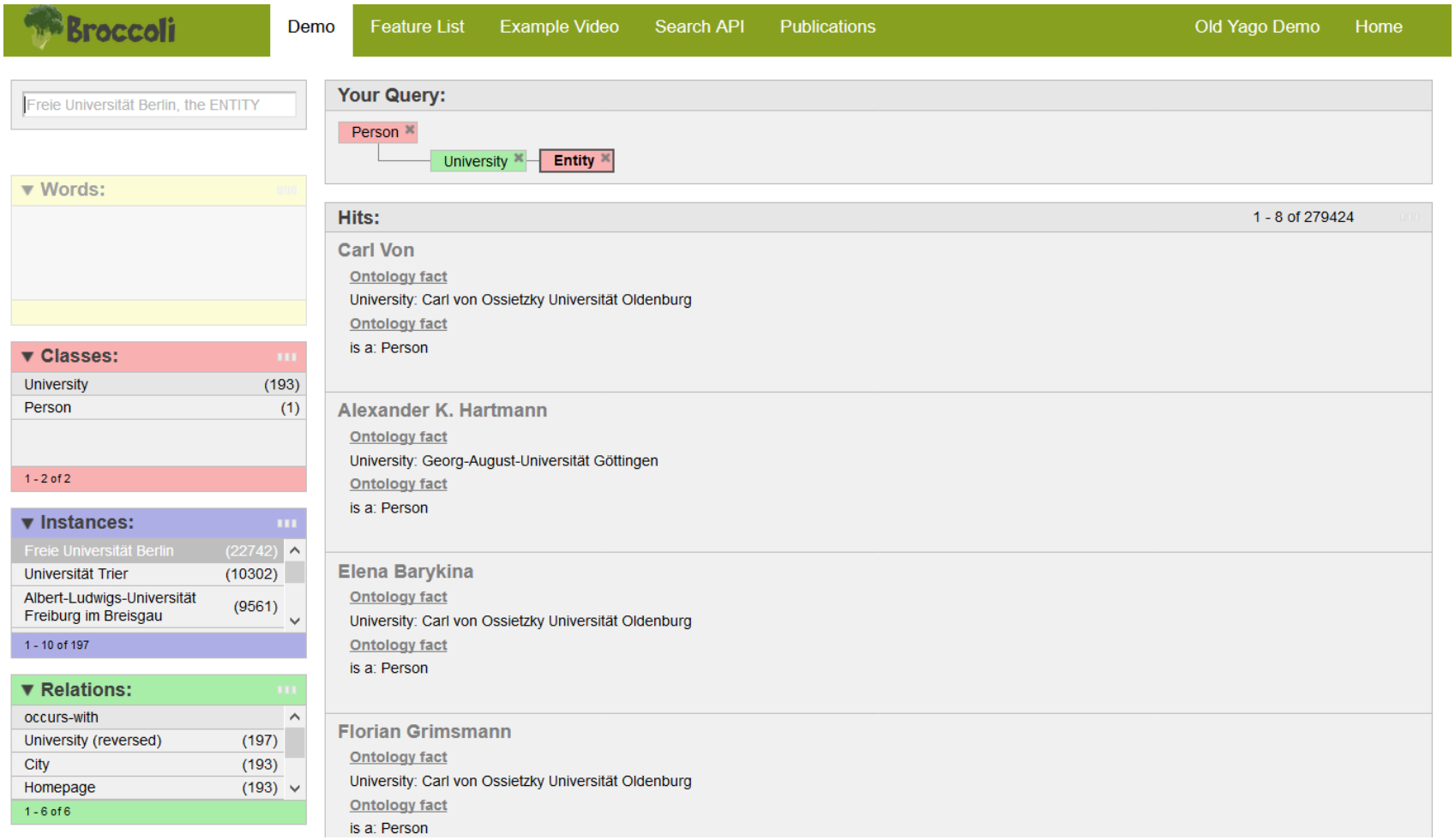
Fazit

- Struktur der Websites beeinflusst das Ergebnis stark
- Hochschulzuordnung ist entscheidend
- Publikationen müssen (besser) erkannt werden
- Name Disambiguation
- Weitere Daten wie Mail-Adresse möglich
- Ausweitung auf andere Einrichtungen

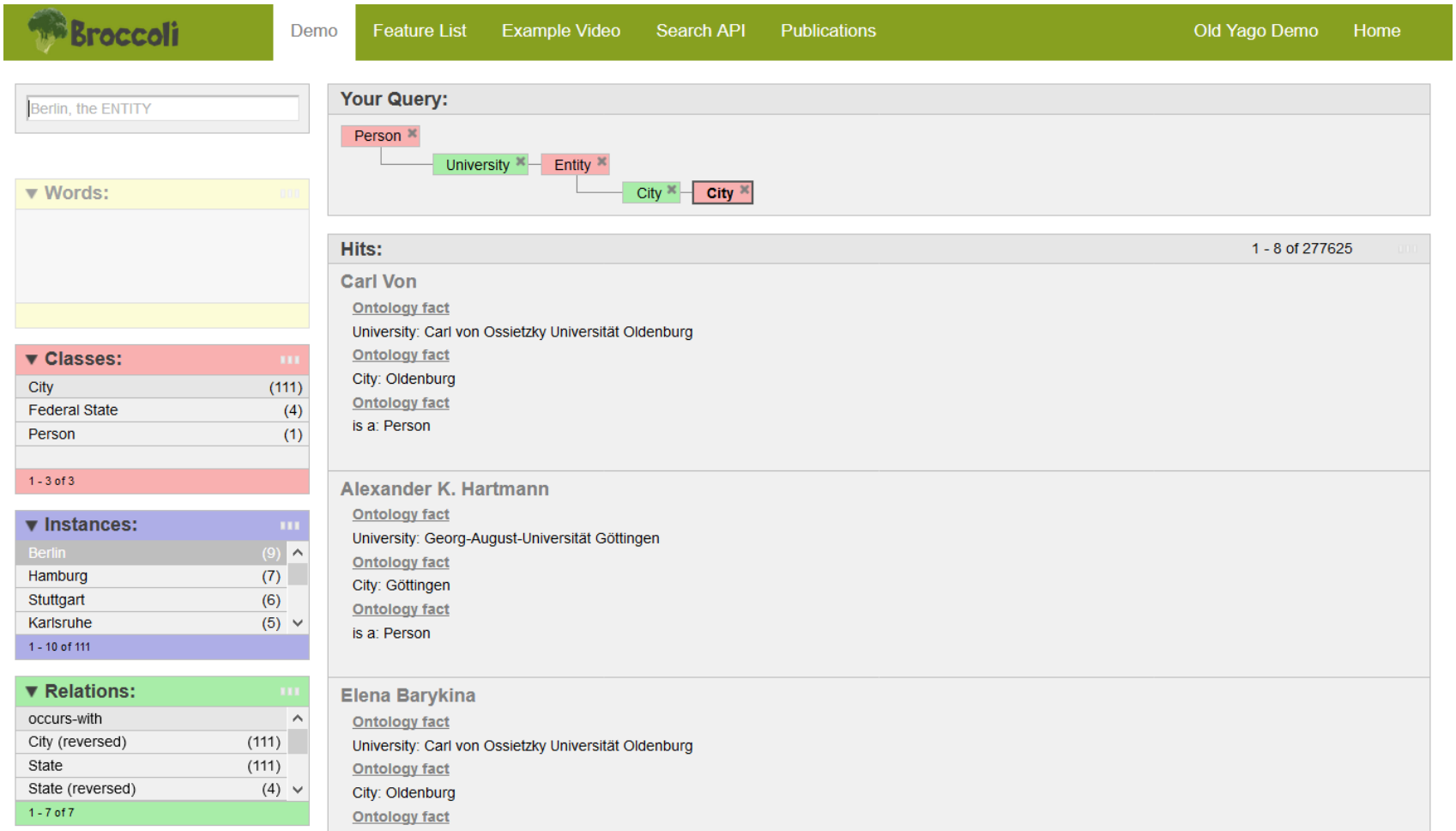
Broccoli "Demo"



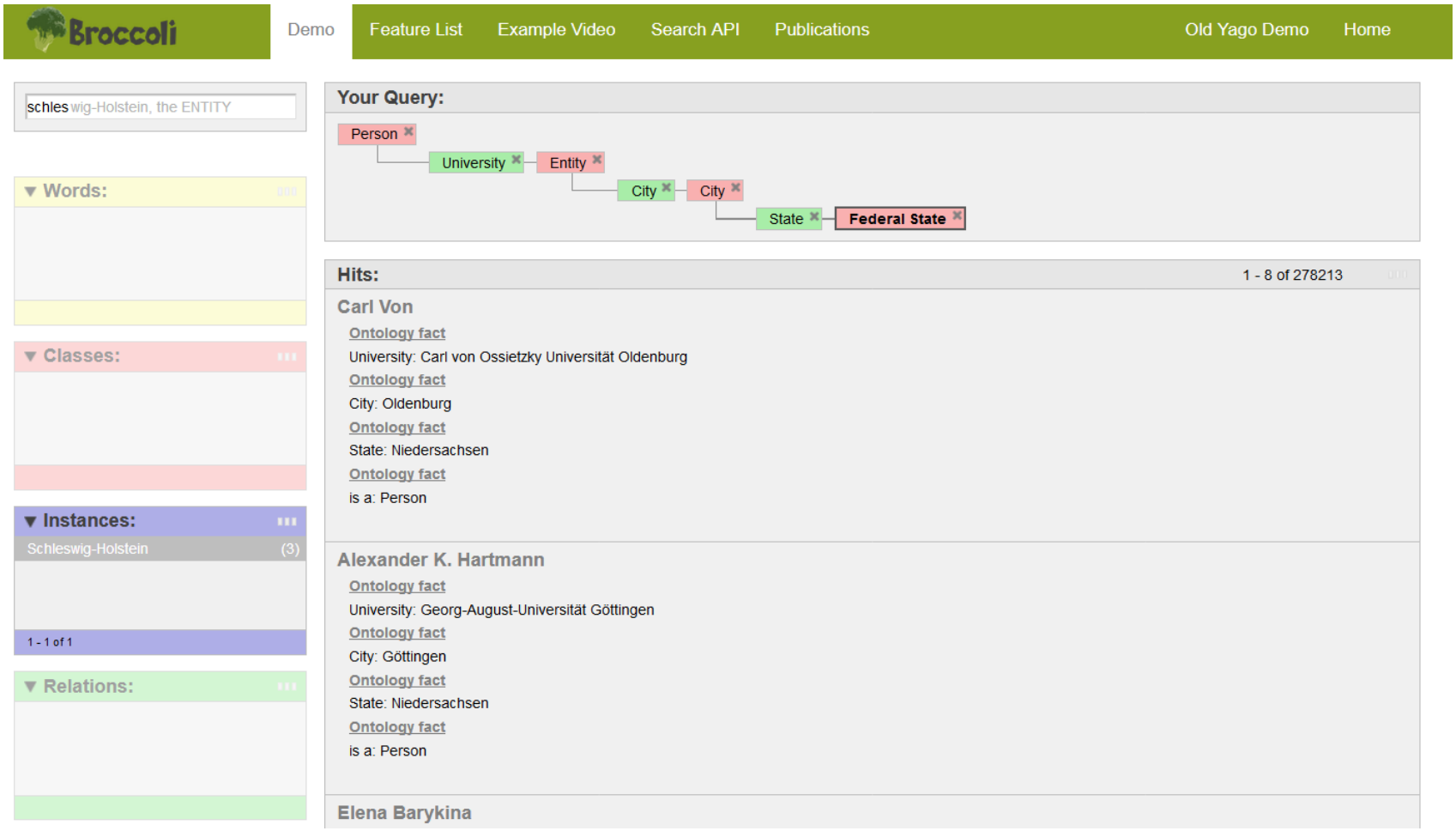
Broccoli "Demo"



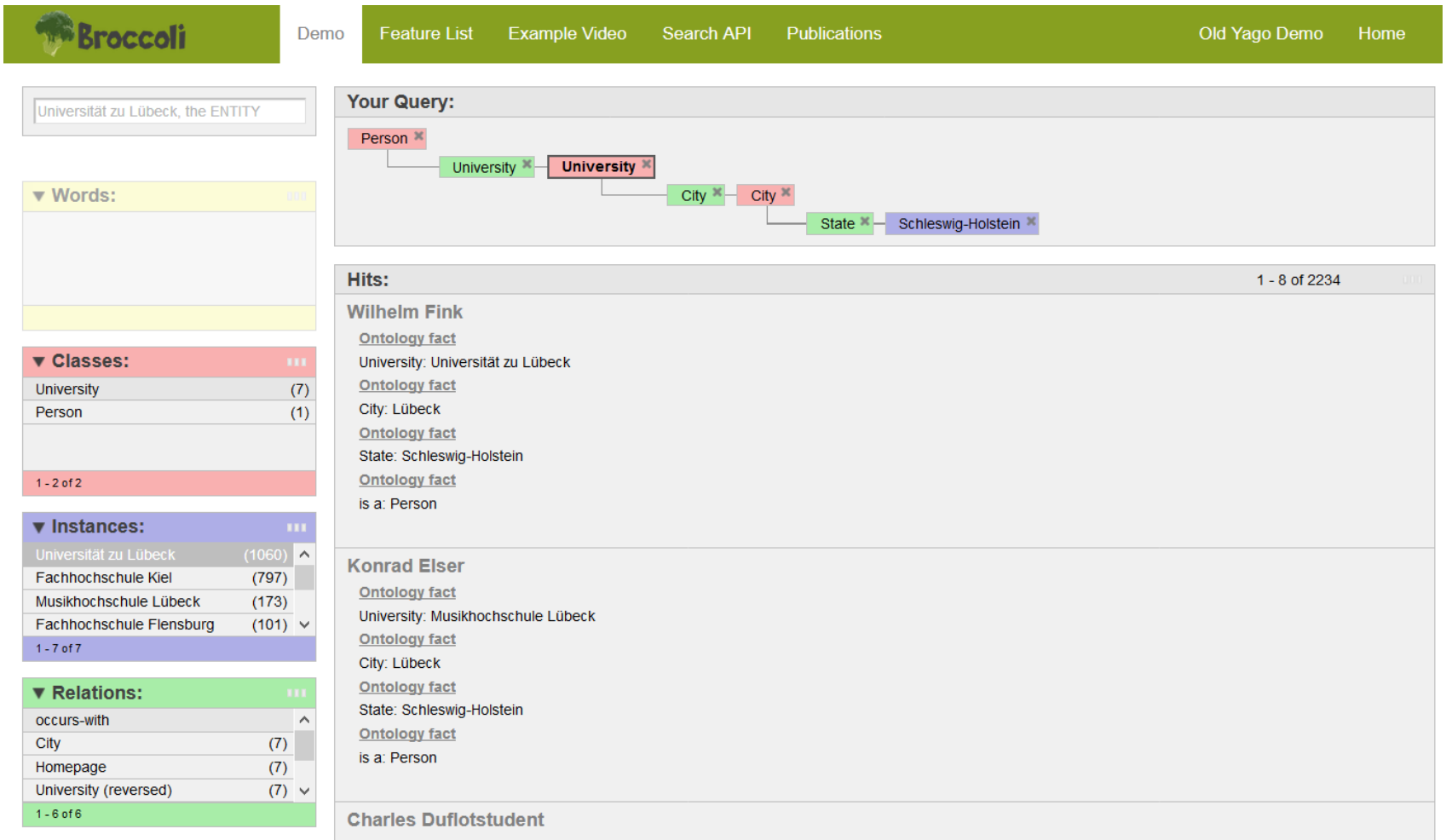
Broccoli "Demo"



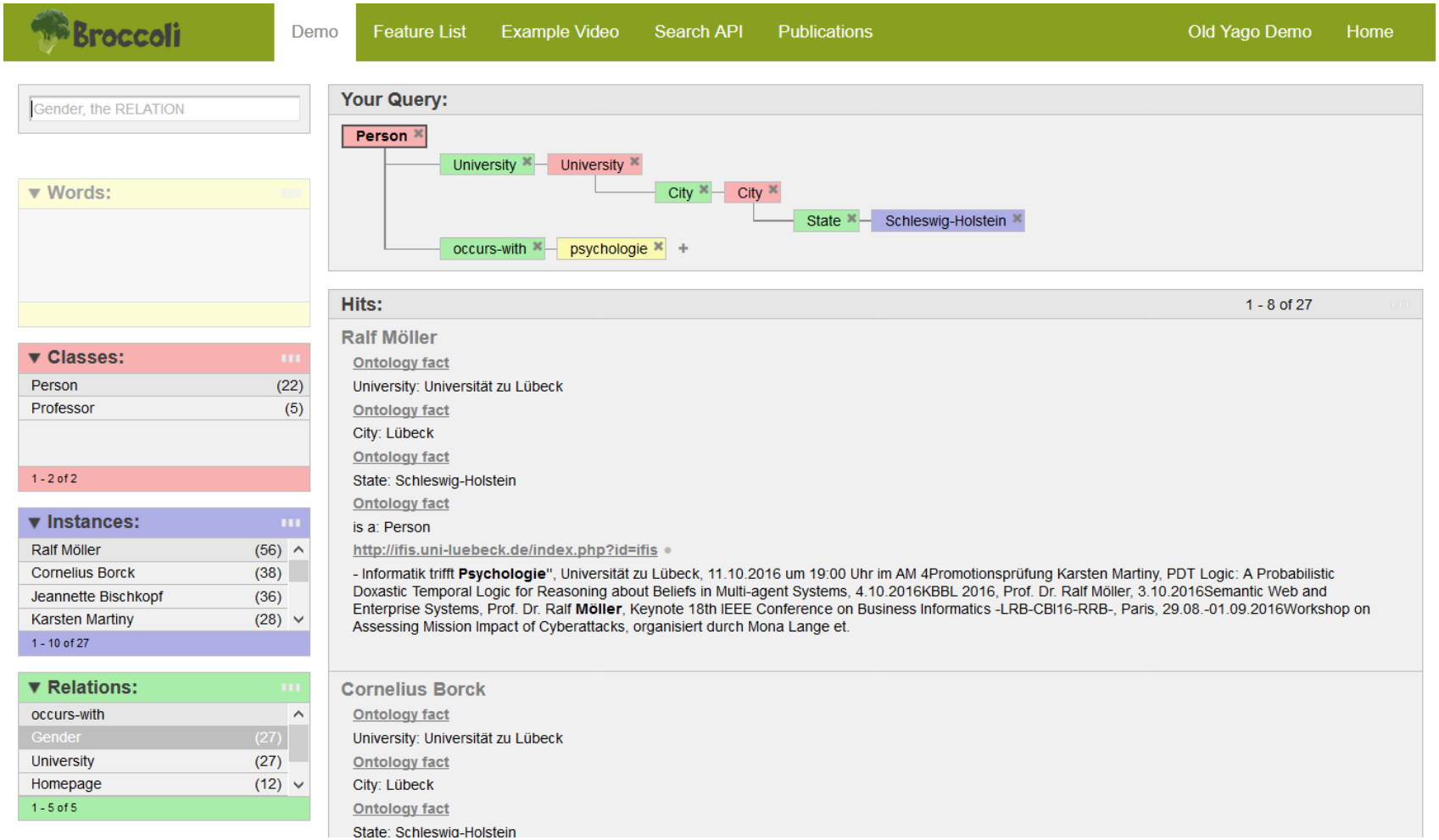
Broccoli "Demo"



Broccoli "Demo"



Broccoli "Demo"



Daten auf Hochschulwebsites - persönliche URLs

cecad.uni-koeln.de/Prof-Dr-Adam-Antebi.244.0.html

↓

Prof. Dr. Adam Antebi

VS.

gko.uni-leipzig.de/index.php?id=63

↓?

Prof. Dr. Inge Baxmann

CommonCrawl

- Non-Profit Organisation die einen WebCrawler betreibt
- Daten öffentlich verfügbar
- Verschiedene Formate:

  Rohdaten - mit Metadaten - reiner Text
- Unsortiert
- Index mit URLs, durchsuchbar nach Domain

Datengewinnung aus einem unsortierten Webcrawl

- Struktur: Segmente als GZip-Dateien
- Vollständigkeit
  - Überschneidung einzelner Crawls ca 50%
  - Einbeziehen mehrerer Monate
  - Aktualität?
- Bei 401 Hochschulen und 5 Crawls:
  - 8,9 Mio URLs
  - 545 000 Segmente
  - Ca. 70 TB
- Verarbeitung „lokal“ in einem Rechencluster

Informationsextraktion

- Natürliche Sprachverarbeitung
  - Analyse der Satzstruktur
  - POS-Tagging
  - Erkennen von Entitäten
  - Immer abhängig von der Sprache
- Entitätenextraktion: Named Entity Recognition
  - Regelbasiert
  - Maschinelles Lernen
- Relationsextraktion

Named Entity Recognition

- Erkennen von Personen, Organisationen, geografischen Entitäten, Geldbeträgen
- Basiert auf Großschreibung, Grammatik, Wörterbüchern, ...
- Regelbasiert
  - Aus (semi-) strukturierten Daten
  - Einfach anzupassen
  - Keine Lernphase
- Maschinelles Lernen
  - Erkennung auf Grundlage gelernter Satzstruktur
  - Labelset und Trainingsdaten (supervised learning)
  - Hohe Genauigkeit
  - Anpassung für neue Label / Daten notwendig
- Verfahren: SVM, HMM, MEMM, CRF

Umsetzung - Personen

- Mindestens: Vorname Nachname
- Optional: akademische Titel / Grade, zweiter Name
- Website-Set
  - ↳ Hochschule
  - ↳ Persönliche Homepage
  - ↳ Subdomain
- Geschlecht

Evaluation - Vorgehen

- Groundtruth
  - Praktisch keine Listen vorhanden
  - Wer soll (nicht) einbezogen werden?
- Evaluiert wird immer auch die NER

Nr.	Suchanfrage	*Relation* Entität
1	Alle Professorinnen und Professoren der Informatik an der Uni Freiburg	*University* Universität Freiburg *is-a* Professor *Subdomain* Informatik
2	Personen mit Fachbereich Psychologie in Schleswig-Holstein	*University* $\rightarrow$ *City* $\rightarrow$ *State* Schleswig-Holstein *occurs-with* Psychologie
3	Mitarbeiterinnen und Mitarbeiter im Exzellenzcluster „Hören für alle“/ “hearing4all“	*occurs-with* hearing4all
4	Mitarbeiterinnen und Mitarbeiter am Institut für Hirnforschung Uni Bremen mit eigener Homepage	*University* Universität Bremen *Subdomain* Brain

Beispiele für evaluierte Suchanfragen

Evaluation - Ergebnisse

Nr.	GT	Found	TP	FP	FN	p	Recall	f
2	29	27	6	21	23	0.22	0.21	0.21
4	18	23	16	7	2	0.70	0.89	0.78

Nr.	p@5	p@10	AP
2	0.20	0.20	0.06
4	1.00	0.90	0.83

2 Personen mit Fachbereich Psychologie in Schleswig-Holstein

4 Mitarbeiterinnen und Mitarbeiter am Institut für Hirnforschung Uni Bremen mit eigener Homepage

Gesamt Ranking: p@5: 0.57; p@10: 0.52; MAP: 0.29

Evaluation - Details Anfrage 4

Auswertung der 7 falsch-positiven Treffer

- 1 x zweiter Vorname als Nachname erkannt
- 2 x veraltete Daten (Schreibfehler, ehemaliges Mitglied)
- 1 x nicht in der GT, da nicht in der Mitarbeiterliste + geänderter Nachname
- 2 x „Freunde des Lehrstuhls“, Falsche Universität
- 1 x „Theoretical Neurobiology“

Evaluation - Details Anfrage 2

Auswertung der 21 falsch-positiven Treffer

- 7 x Personen anderer Lehrstühle mit Themenfeld Psychologie
- 2 x Mitarbeiter einer fachgebietsübergreifenden Arbeitsgruppe.
- 3 x korrektes Fachgebiet, falsche Hochschule
- 5 x zufälliges Vorkommen mit dem Stichwort
- 1 x Namesüberschneidung
- 1 x erwähnter Buchautor
- 2 x falsche Namenserkennung

enter search terms ...

▶ Words:

▼ Classes:

Person	(114)
Professor	(28)

1 - 2 of 2

▼ Instances:

Isabelle Deflers	(258)
Christian Kühner	(214)
Arndt Schreiber	(198)

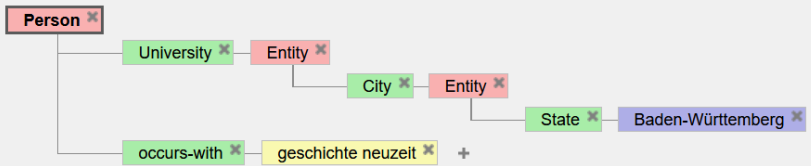
1 - 20 of 142

▼ Relations:

occurs-with	
Gender	(142)
University	(142)
Homepage	(75)
Subdomain	(75)

1 - 5 of 5

Your Query:



Hits:

1 - 8 of 142

Isabelle Deflers

Ontology fact

University: Albert-Ludwigs-Universität Freiburg im Breisgau

Ontology fact

City: Freiburg

Ontology fact

State: Baden-Württemberg

Ontology fact

is a: Person

http://fnz.geschichte.uni-freiburg.de/Personen/personen

Tagungen Lehrstuhl für **Geschichte** der Frühen **Neuzeit** Albert-Ludwigs-Universität Freiburg Sie sind hier: Startseite ' Personen Artikelaktionen Drucken Info Personen Inhaber der Professur für **Geschichte** der Frühen **Neuzeit**/ Vizesprecher des SFB 948: Professor Dr. Ronald G. Asch SS 17: Forschungsfreisemester Raum: KG IV, 4436 Telefon: 0761 / 203 3448 Email: ronald.g.asch@geschichte.uni-freiburg.de Sprechstunde: nach Vereinbarung Lebenslauf: deutsch / CV: english Veröffentlichungen Sekretariat: Sibylle Rupp Raum: KG IV, 4435 Telefon: 0761 / 203 3429 Email: sibylle.rupp@geschichte.uni-freiburg.de Bürozeiten: Mo-Do: 10.00 - 12.00 Uhr u. Di: 14.00 -16.00 Uhr Akademische Rätin für Frühe **Neuzeit**: PD Dr. Isabelle **Deflers** SS 17: Lehrstuhlvertretung von Professor Asch