Towards Knowledge Graph Construction from Entity Co-occurrence

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Nicolas Heist | Colloquium HWS 2018

06.11.2018

AGENDA

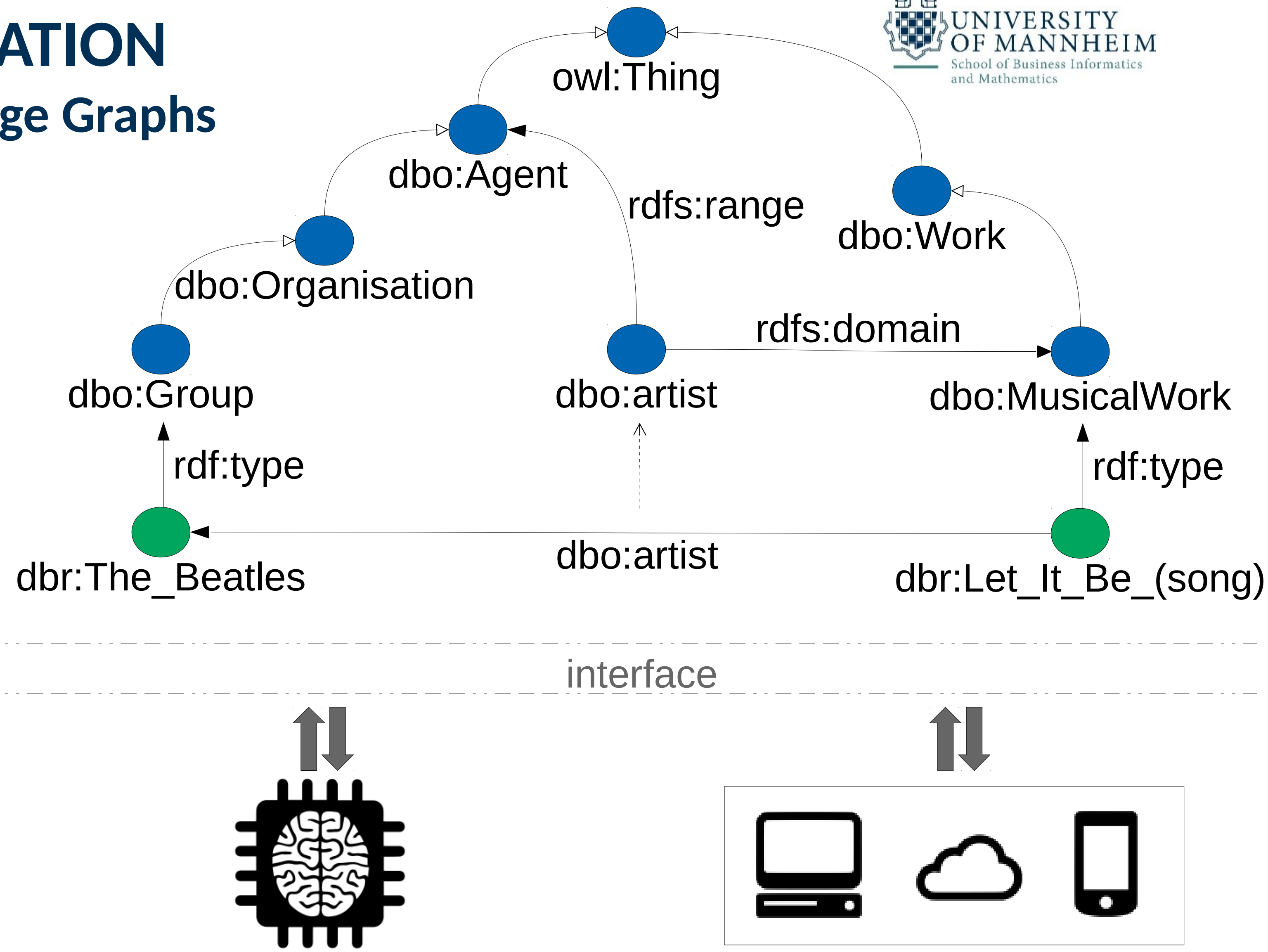
- Motivation
- Approach
- Methodology
- Experiments
- Discussion

MOTIVATION

Knowledge Graphs

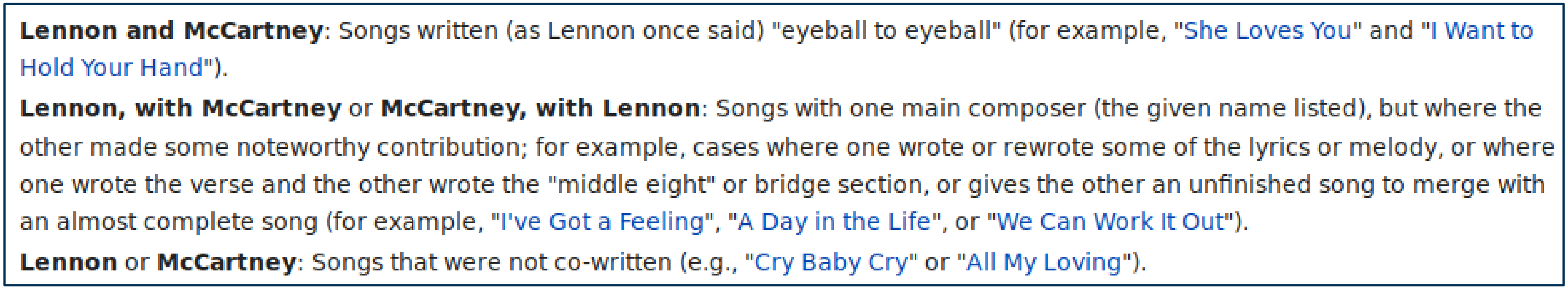
T-Box

A-Box



MOTIVATION

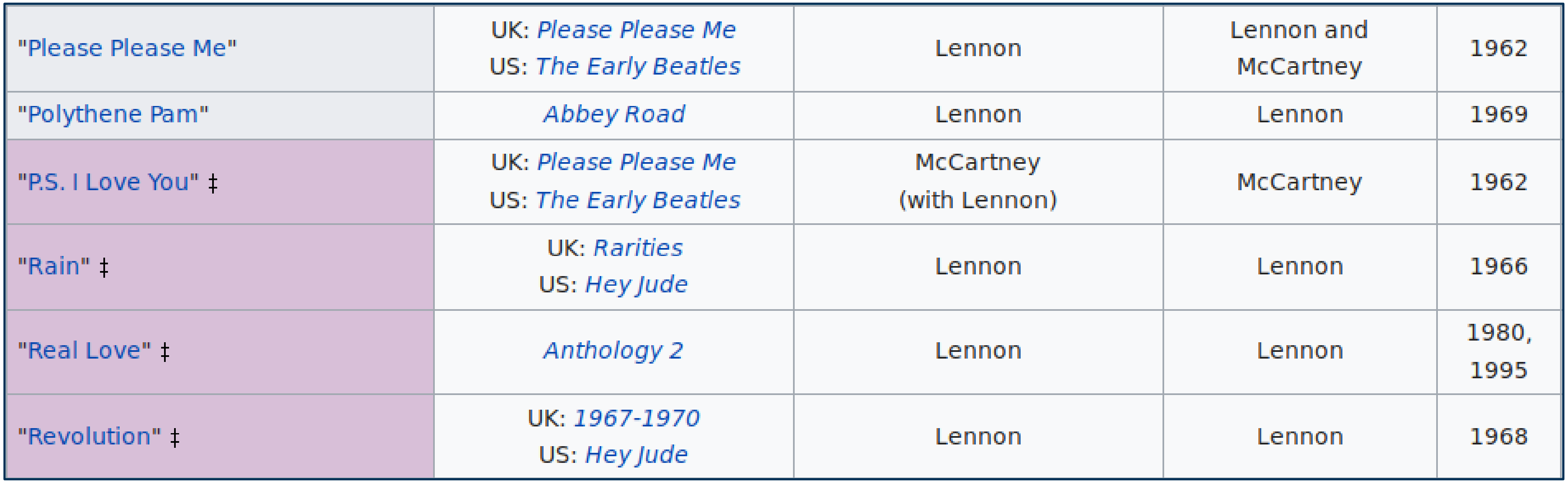
Entity Co-occurrence



Lennon and McCartney: Songs written (as Lennon once said) "eyeball to eyeball" (for example, "She Loves You" and "I Want to Hold Your Hand").

Lennon, with McCartney or **McCartney, with Lennon**: Songs with one main composer (the given name listed), but where the other made some noteworthy contribution; for example, cases where one wrote or rewrote some of the lyrics or melody, or where one wrote the verse and the other wrote the "middle eight" or bridge section, or gives the other an unfinished song to merge with an almost complete song (for example, "I've Got a Feeling", "A Day in the Life", or "We Can Work It Out").

Lennon or **McCartney**: Songs that were not co-written (e.g., "Cry Baby Cry" or "All My Loving").

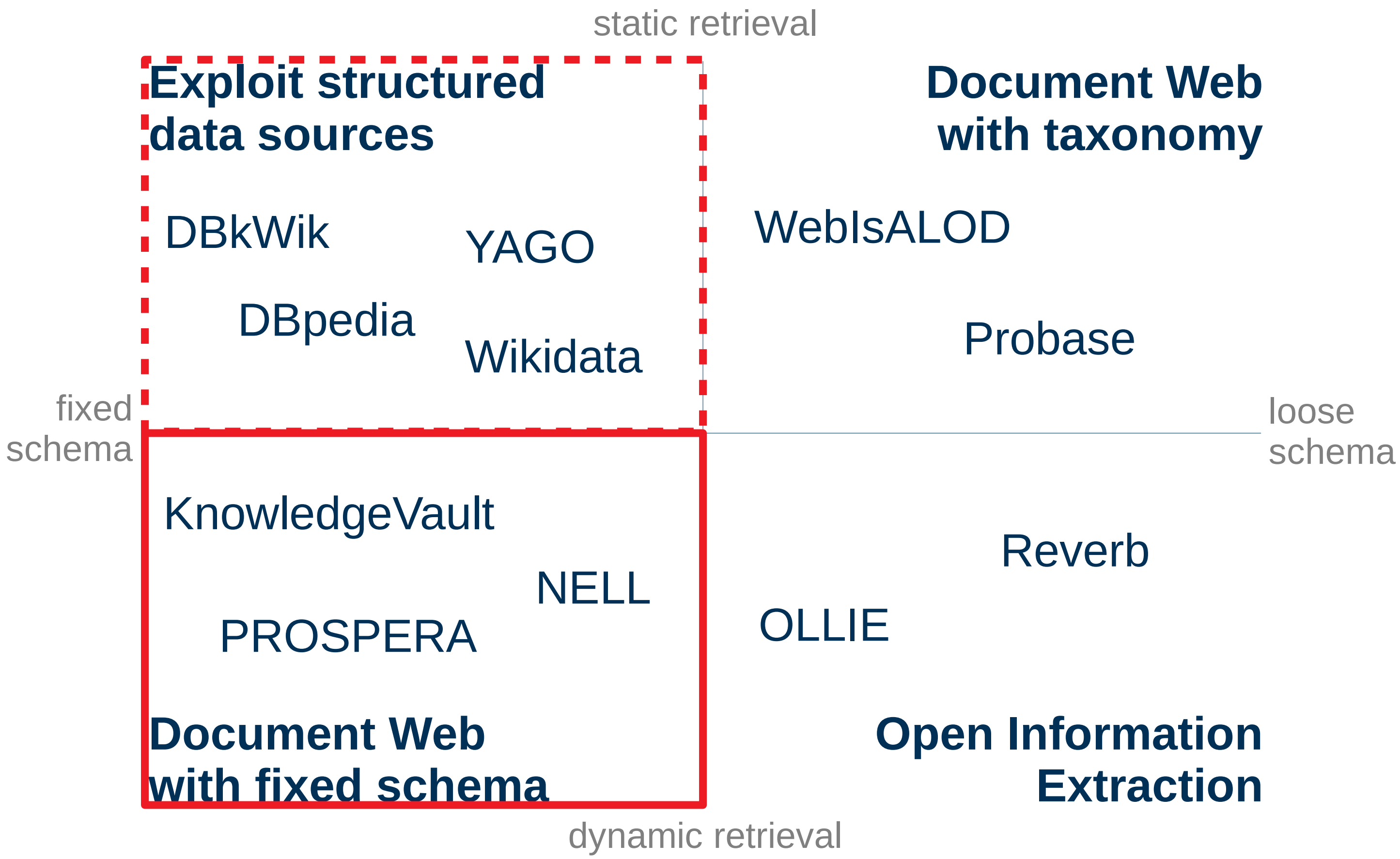


"Please Please Me"	UK: *Please Please Me* US: *The Early Beatles*	Lennon	Lennon and McCartney	1962
"Polythene Pam"	*Abbey Road*	Lennon	Lennon	1969
"P.S. I Love You" ‡	UK: *Please Please Me* US: *The Early Beatles*	McCartney (with Lennon)	McCartney	1962
"Rain" ‡	UK: *Rarities* US: *Hey Jude*	Lennon	Lennon	1966
"Real Love" ‡	*Anthology 2*	Lennon	Lennon	1980, 1995
"Revolution" ‡	UK: *1967-1970* US: *Hey Jude*	Lennon	Lennon	1968

Source: https://en.wikipedia.org/wiki/List_of_songs_recorded_by_the_Beatles

MOTIVATION

Entity Co-occurrence



static retrieval

Exploit structured data sources

DBkWik

YAGO

DBpedia

Wikidata

Document Web with taxonomy

WebIsALOD

Probase

fixed schema

loose schema

KnowledgeVault

NELL

PROSPERA

Document Web with fixed schema

Reverb

OLLIE

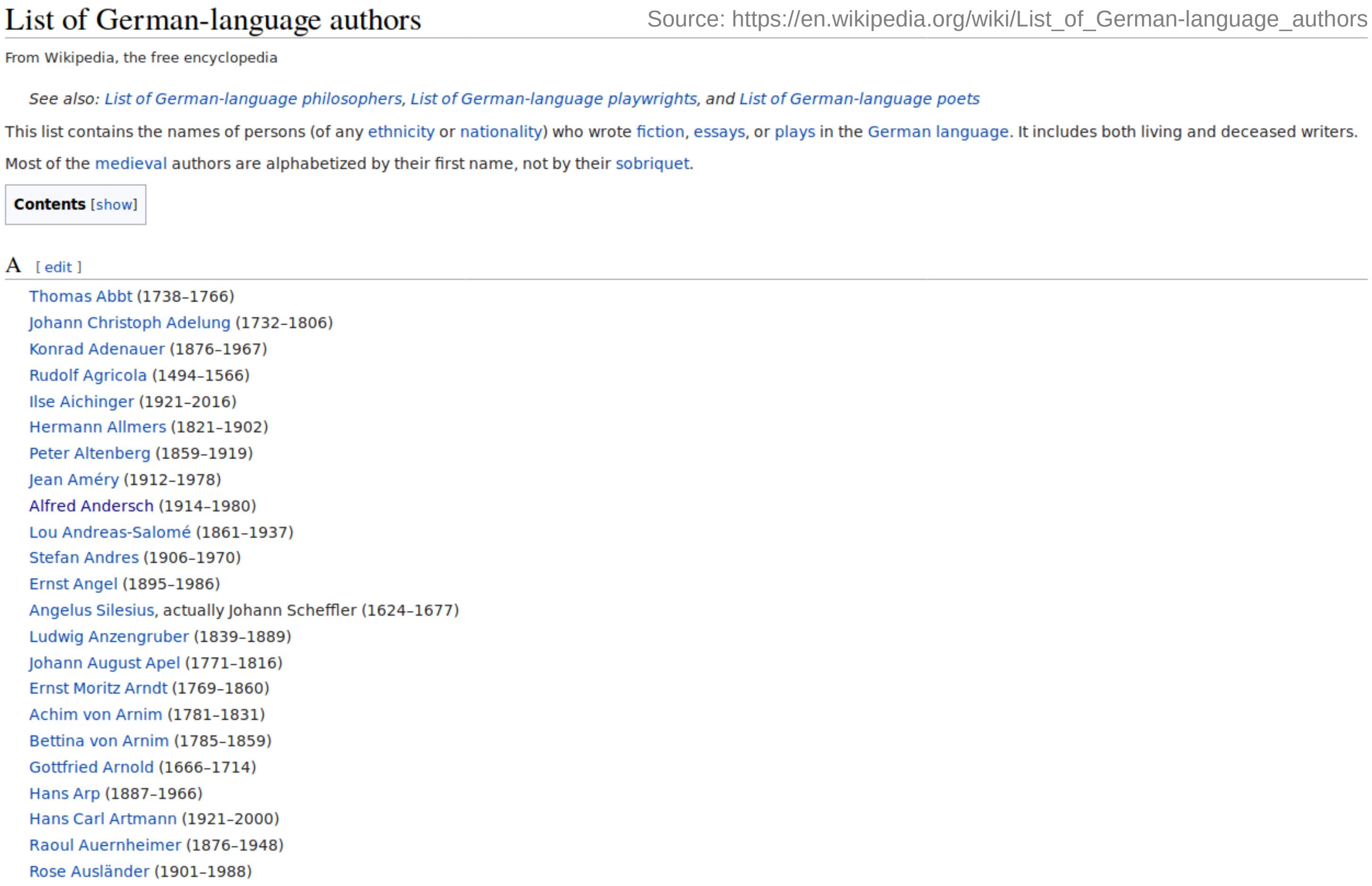
Open Information Extraction

dynamic retrieval

[Dong et al., 2014]

APPROACH

Example | Wikipedia List Page



Source: https://en.wikipedia.org/wiki/List_of_German-language_authors

List of German-language authors

From Wikipedia, the free encyclopedia

See also: List of German-language philosophers, List of German-language playwrights, and List of German-language poets

This list contains the names of persons (of any ethnicity or nationality) who wrote fiction, essays, or plays in the German language. It includes both living and deceased writers.

Most of the medieval authors are alphabetized by their first name, not by their sobriquet.

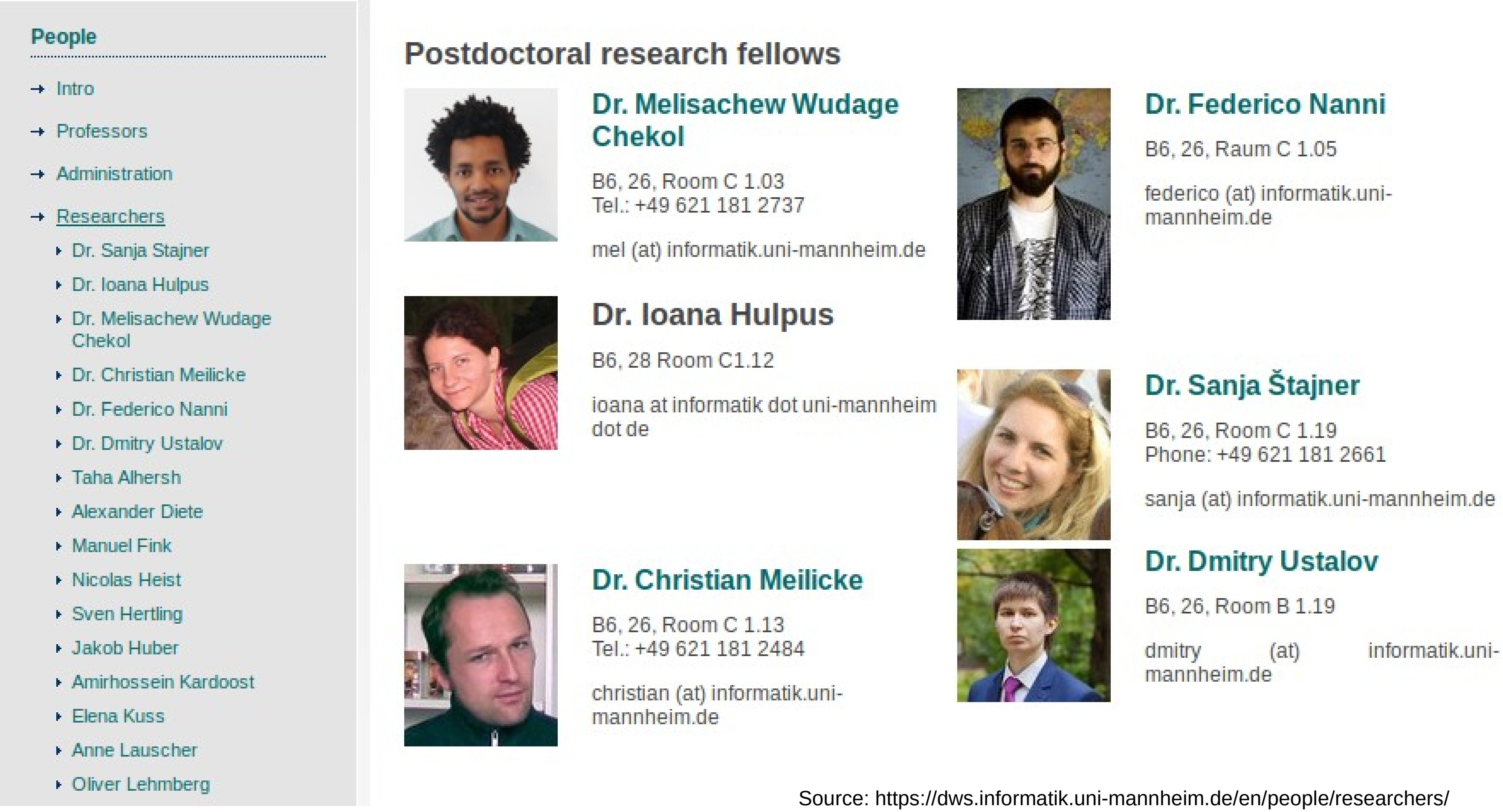
Contents [show]

A [edit]

- Thomas Abbt (1738–1766)
- Johann Christoph Adelung (1732–1806)
- Konrad Adenauer (1876–1967)
- Rudolf Agricola (1494–1566)
- Ilse Aichinger (1921–2016)
- Hermann Allmers (1821–1902)
- Peter Altenberg (1859–1919)
- Jean Améry (1912–1978)
- Alfred Andersch (1914–1980)
- Lou Andreas-Salomé (1861–1937)
- Stefan Andres (1906–1970)
- Ernst Angel (1895–1986)
- Angelus Silesius, actually Johann Scheffler (1624–1677)
- Ludwig Anzengruber (1839–1889)
- Johann August Apel (1771–1816)
- Ernst Moritz Arndt (1769–1860)
- Achim von Arnim (1781–1831)
- Bettina von Arnim (1785–1859)
- Gottfried Arnold (1666–1714)
- Hans Arp (1887–1966)
- Hans Carl Artmann (1921–2000)
- Raoul Auernheimer (1876–1948)
- Rose Ausländer (1901–1988)

APPROACH

Example | DWS Researchers



People

- → Intro
- → Professors
- → Administration
- → Researchers
  - Dr. Sanja Stajner
  - Dr. Ioana Hulpus
  - Dr. Melisachew Wudage Chekol
  - Dr. Christian Meilicke
  - Dr. Federico Nanni
  - Dr. Dmitry Ustalov
  - Taha Alhersh
  - Alexander Diete
  - Manuel Fink
  - Nicolas Heist
  - Sven Hertling
  - Jakob Huber
  - Amirhossein Kardoost
  - Elena Kuss
  - Anne Lauscher
  - Oliver Lehmberg

Postdoctoral research fellows

Dr. Melisachew Wudage Chekol

B6, 26, Room C 1.03
Tel.: +49 621 181 2737

mel (at) informatik.uni-mannheim.de

Dr. Ioana Hulpus

B6, 28 Room C1.12

ioana at informatik dot uni-mannheim dot de

Dr. Christian Meilicke

B6, 26, Room C 1.13
Tel.: +49 621 181 2484

christian (at) informatik.uni-mannheim.de

Dr. Federico Nanni

B6, 26, Raum C 1.05

federico (at) informatik.uni-mannheim.de

Dr. Sanja Štajner

B6, 26, Room C 1.19
Phone: +49 621 181 2661

sanja (at) informatik.uni-mannheim.de

Dr. Dmitry Ustalov

B6, 26, Room B 1.19

dmitry (at) informatik.uni-mannheim.de

Source: https://dws.informatik.uni-mannheim.de/en/people/researchers/

APPROACH

Problem Description

D	document corpus
E	entities in D
r_{sur}	a *surface relationship* between two entities
r_{sem}	a *semantic relationship* between two entities

For every *d* ∈ *D*: Find set of patterns P_d with

$$P_d = \left\{ p_{d, r_{sur}, r_{sem}} \middle| \forall e_1, e_2 \in E_p : e_1, e_2 \in E_d \wedge r_{sur}(e_1, e_2) \wedge r_{sem}(e_1, e_2) \right\}$$

then fuse individual P_d to generalized set of patterns *P*

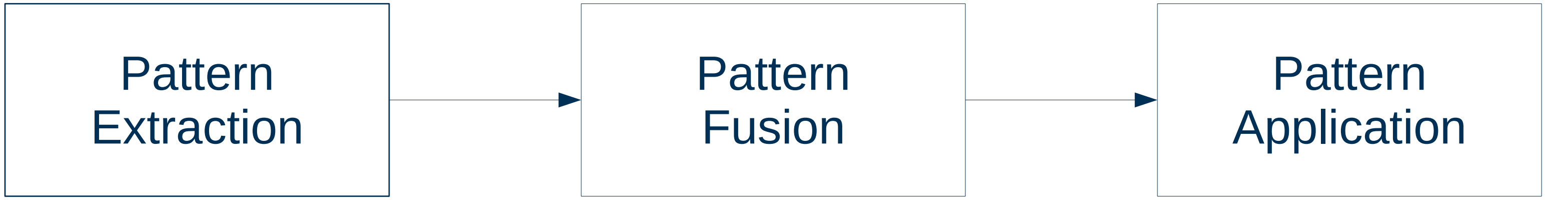
APPROACH

Research Questions

1) Is it possible to discover arbitrary entity co-occurrence patterns locally (i.e. within a bounded context like Wikipedia) as well as globally (on the Web)?
2) Can co-occurrence patterns be grouped into different types of patterns and if so, how do these groups differ in their performance?
3) How well can (groups of) co-occurrence patterns be generalized so that they can be applied to arbitrary web documents?

METHODOLOGY

Pipeline



- Three-phased approach
- Inputs:

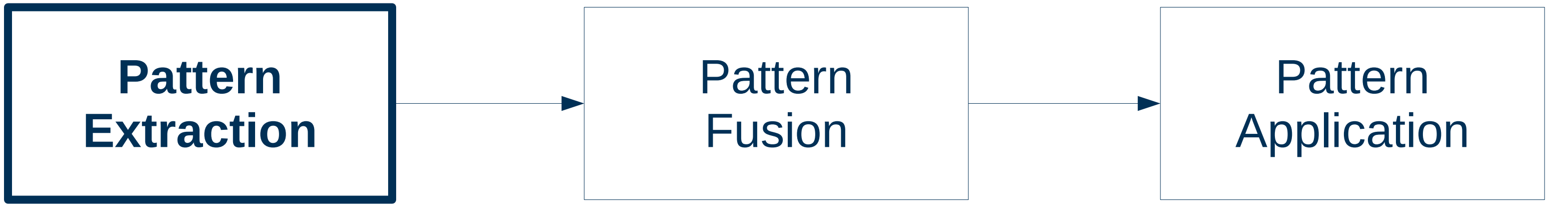
  KG_s seed knowledge graph

  D_s seed corpus of web documents (containing entities described in KG_s)

  D_t target corpus of web documents (to gather new knowledge from)

METHODOLOGY

Pipeline | Pattern Extraction

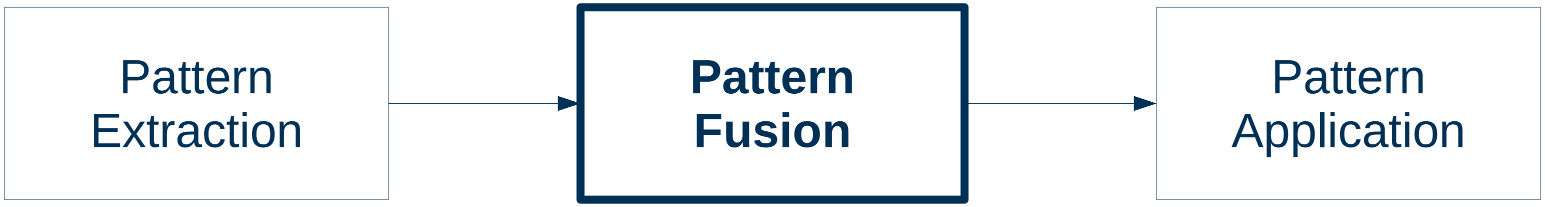


Pattern Extraction → Pattern Fusion → Pattern Application

- locate and and link entities in D_s to KG_s
- gather data for extraction of patterns
  - use *distant supervision* and *local closed world assumption*
- use structural features of the document to extract patterns
  - e.g. *DOM-tree paths* (generic) or *Wiki markup* (specific)
- output: (possibly empty) set of patterns P_d for every $d \in D$

METHODOLOGY

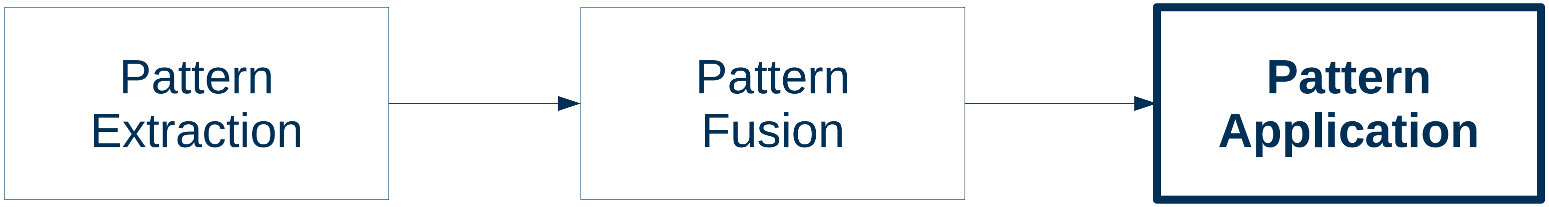
Pipeline | Pattern Fusion



- fuse patterns p_1 and p_2 if their respective relations r_{sur} and r_{sem}:
  - ➢ are equal
  - ➢ subsume one another
  - ➢ can be subsumed by a more general relation
- output: generalized set of patterns *P*

METHODOLOGY

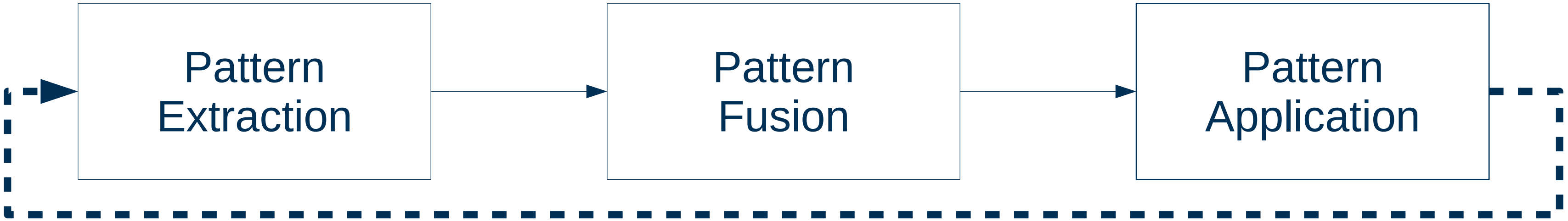
Pipeline | Pattern Application



- apply patterns in P to D_t
  - depending on the generality of patterns in P, we might set $D_t = D_s$
- output: set of (novel) entities and facts
  - new entities can be discovered with arbitrary pattern p
  - new facts are extracted in the context of r_{sem} of a pattern p

METHODOLOGY

Pipeline | Iterative Extension



- possible extension (as used by NELL):
  - extract only high-quality patterns *P*
  - iteratively run the pipeline with *P* as seed patterns

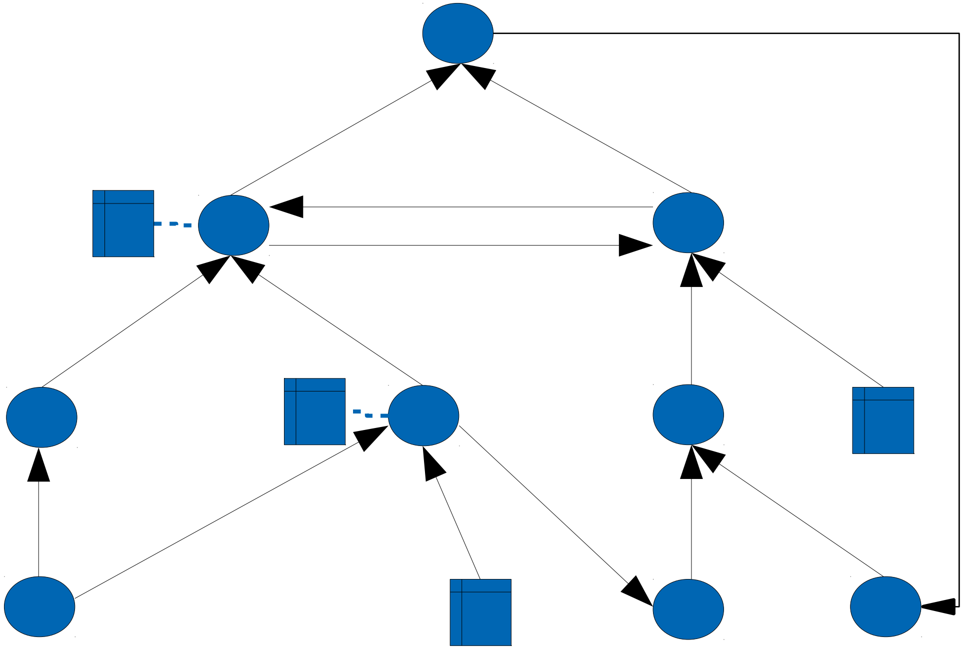
METHODOLOGY

Evaluation

- intrinsic
  - metrics like size, coverage, accuracy (with regard to LCWA)
- extrinsic
  - evaluate quality of KG with help of human judges
- task-based
  - compare with other KGs when used for specific applications

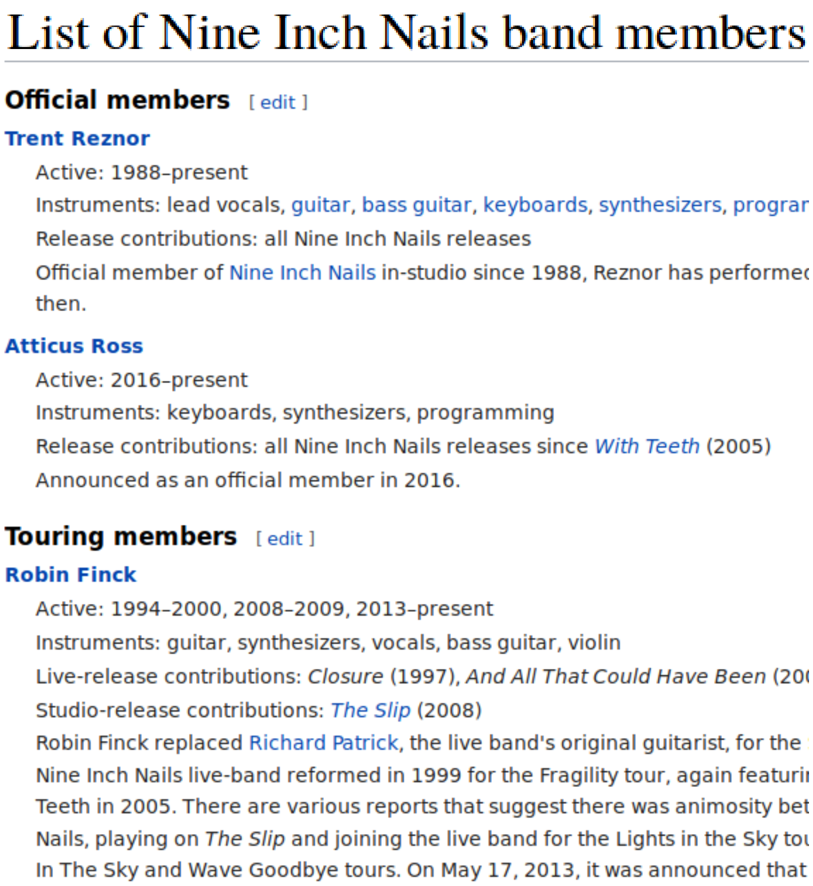
EXPERIMENTS

Concept | Wikipedia Categories and Lists



EXPERIMENTS

Example| Wikipedia Categories and Lists



List of Nine Inch Nails band members

Official members [edit]

Trent Reznor

Active: 1988–present

Instruments: lead vocals, guitar, bass guitar, keyboards, synthesizers, prograr

Release contributions: all Nine Inch Nails releases

Official member of Nine Inch Nails in-studio since 1988, Reznor has performec then.

Atticus Ross

Active: 2016–present

Instruments: keyboards, synthesizers, programming

Release contributions: all Nine Inch Nails releases since *With Teeth* (2005)

Announced as an official member in 2016.

Touring members [edit]

Robin Finck

Active: 1994–2000, 2008–2009, 2013–present

Instruments: guitar, synthesizers, vocals, bass guitar, violin

Live-release contributions: *Closure* (1997), *And All That Could Have Been* (20(

Studio-release contributions: *The Slip* (2008)

Robin Finck replaced Richard Patrick, the live band's original guitarist, for the Nine Inch Nails live-band reformed in 1999 for the Fragility tour, again featuri Teeth in 2005. There are various reports that suggest there was animosity bet Nails, playing on *The Slip* and joining the live band for the Lights in the Sky tou In The Sky and Wave Goodbye tours. On May 17, 2013, it was announced that

ENTITIES

- lists 29 band members (5 of them unlinked)

TYPES

dbo:Agent	(21)
dbo:Person	(23)
dbo:Artist	(20)
dbo:MusicalArtist	(22)

PROPERTIES

dbo:genre = dbr:Alternative_rock	(15)
dbo:genre = dbr:Industrial_rock	(10)
dbo:bandMember of dbr:Nine_Inch_Nails	(4)

possible discovery of:

5 new entities + 88 facts

EXPERIMENTS

Concept | Patterns from the Web

Common Crawl

document corpus

schema.org

microdata,
RDFa, ...

WDC

entity recognition

yago
select knowledge

?

WIKIDATA

DBpedia

knowledge graph

DISCUSSION

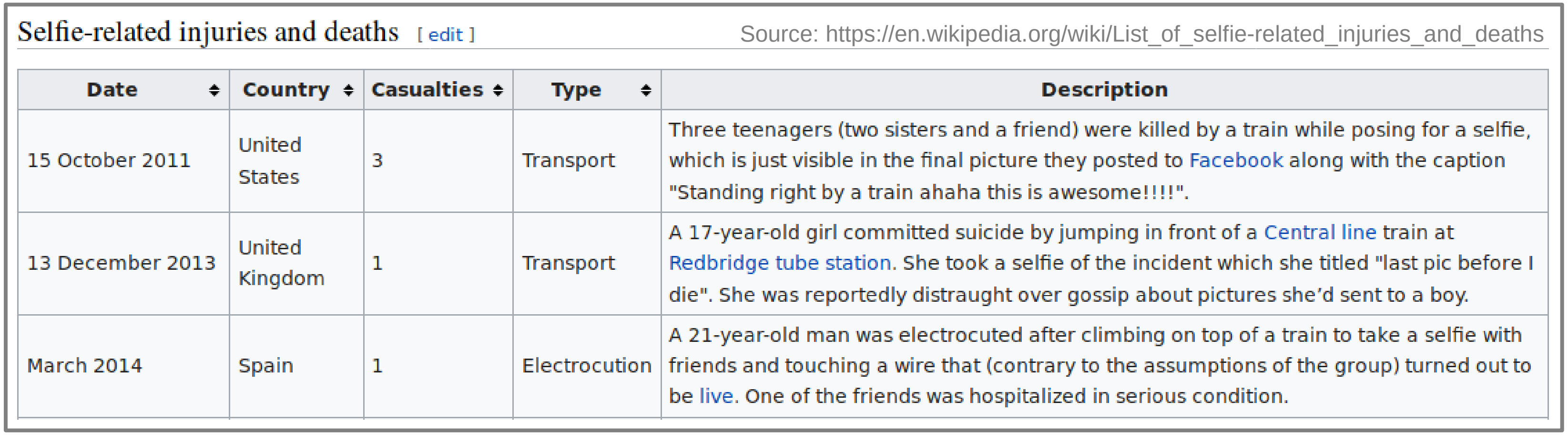
Potentials & Limitations

+ complement existing approaches

+ discover entities / facts with very little local evidence

- generalized patterns might need additional context

- no extension of T-Box / schema not flexible enough



Selfie-related injuries and deaths [edit]

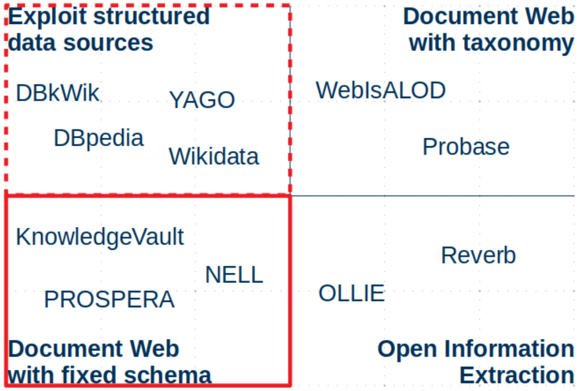
Source: https://en.wikipedia.org/wiki/List_of_selfie-related_injuries_and_deaths

Date	Country	Casualties	Type	Description
15 October 2011	United States	3	Transport	Three teenagers (two sisters and a friend) were killed by a train while posing for a selfie, which is just visible in the final picture they posted to Facebook along with the caption "Standing right by a train ahaha this is awesome!!!!".
13 December 2013	United Kingdom	1	Transport	A 17-year-old girl committed suicide by jumping in front of a Central line train at Redbridge tube station. She took a selfie of the incident which she titled "last pic before I die". She was reportedly distraught over gossip about pictures she'd sent to a boy.
March 2014	Spain	1	Electrocution	A 21-year-old man was electrocuted after climbing on top of a train to take a selfie with friends and touching a wire that (contrary to the assumptions of the group) turned out to be live. One of the friends was hospitalized in serious condition.

SOURCES

DBpedia	Lehmann et al.: DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2), 167–195 (2015)
YAGO	Suchanek et al.: YAGO: A core of semantic knowledge. In: 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
Wikidata	Vrandečić et al.: Wikidata: a free collaborative knowledgebase. In: Communications of the ACM, pp. 78-85 (2014)
DBkWik	Hofmann et al.: Dbkwik: Towards knowledge graph creation from thousands of wikis. In: International Semantic Web Conference (Posters and Demos) (2017)
Reverb	Fader et al.: Identifying relations for open information extraction. In EMNLP, 2011
OLLIE	Mausam et al.: Open language learning for information extraction. In EMNLP, 2012.
KnowledgeVault / LCWA	Dong et al.: Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. In: 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610. ACM (2014)
NELL	Carlson et al.: Toward an architecture for never-ending language learning. In: AAAI. vol. 5, p. 3. Atlanta (2010)
PROSPERA	Nakashole et al.: Scalable knowledge harvesting with high precision and high recall. In WSDM, pages 227–236, 2011.
WebIsALOD	Hertling, S., Paulheim, H.: WebIsALOD: providing hypernymy relations extracted from the web as linked open data. In: International Semantic Web Conference. pp. 111–119. Springer (2017)
Probase	Wu et al.: Probase: A probabilistic taxonomy for text understanding. In: 2012 ACM SIGMOD International Conference on Management of Data. pp. 481–492. ACM (2012)
Distant Supervision	Mintz et al.: Distant supervision for relation extraction without labeled data. In: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011 (2009)
WebDataCommons	Meusel et al.: The WebDataCommons microdata, RDFa and Microformat dataset series. In: International Semantic Web Conference. pp. 277–292. Springer (2014)

QUESTIONS


Exploit structured data sources
Document Web with taxonomy
DBkWik
YAGO
DBpedia
Wikidata
WebIsALOD
Probase
KnowledgeVault
NELL
PROSPERA
Reverb
OLLIE
Document Web with fixed schema
Open Information Extraction

D document corpus

E entities in D

r_{sur} a *surface relationship* between two entities

r_{sem} a *semantic relationship* between two entities

For every *d* ∊ *D*: Find set of patterns P_d with

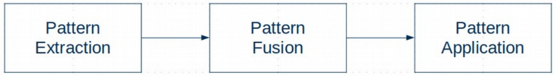
$$P_d = \{ p_{d,r_{sur},r_{sem}} \mid \forall e_1, e_2 \in E_p : e_1, e_2 \in E_d \wedge r_{sur}(e_1, e_2) \wedge r_{sem}(e_1, e_2) \}$$

then fuse individual P_d to generalized set of patterns *P*

1) Is it possible to discover arbitrary entity co-occurrence patterns locally (i.e. within a bounded context like Wikipedia) as well as globally (on the Web)?

2) Can co-occurrence patterns be grouped into different types of patterns and if so, how do these groups differ in their performance?

3) How well can (groups of) co-occurrence patterns be generalized so that they can be applied to arbitrary web documents?

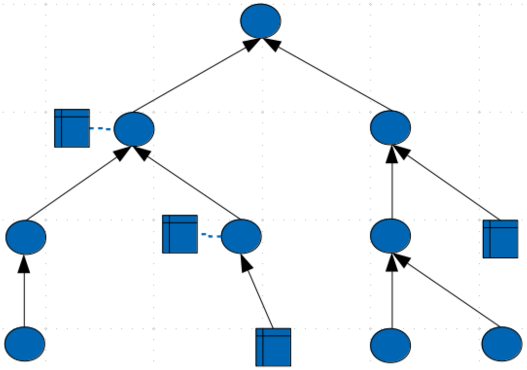


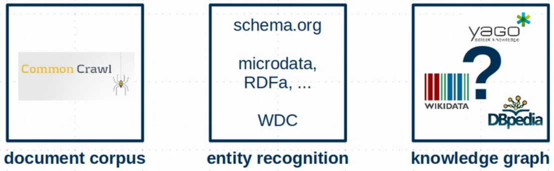
- Three-phased approach
- Inputs:

KG_s seed knowledge graph

D_s seed corpus of web documents (containing entities described in KG_s)

D_t target corpus of web documents (to gather new knowledge from)




Common Crawl
schema.org
microdata, RDFa, ...
WDC
yago
WIKIDATA
DBpedia
document corpus
entity recognition
knowledge graph