

# Research Knowledge Discovery

---

## Detektion von Experten und Aufbau eines Recommender-Systems für die TH Köln

### Projektgruppe 2 - DIS - 18

#### Teilnehmer

- Pia Störmer
- Matteo Meier
- Jüri Keller
- Martin Bilko
- Sascha Gharib
- Verena Pawlas
- Michelle Reiners
- Saskia Brech
- Fabian Ax
- Constantin Krah
- Leon Munz
- Andreas Kruff
- Fabian Gitzler
- Jonas Dudda
- Annika Füssel

# Agenda

## 1. Projektplanung

- 1.1. Kurzzusammenfassung – Projektziel
- 1.2. Timeline, Meilensteine & Next Steps

## 2. Quellen & Datenbasis

- 2.1. Auswertung Interne Quellen
- 2.2. Auswertung Externe Quellen
- 2.3. Rechtliche Herausforderung

## 3. Datenmodell

- 3.1. Sciebo Projektbox
- 3.2. Datenschema

# 01 —

## Projektplanung

# 1.1 Kurzzusammenfassung - Projektziel

Im Rahmen des Projekts Research Knowledge Discovery werden KD-Verfahren auf *Artefakte* und *Entitäten des Hochschul- bzw. Wissenschaftsbetriebes* angewendet.

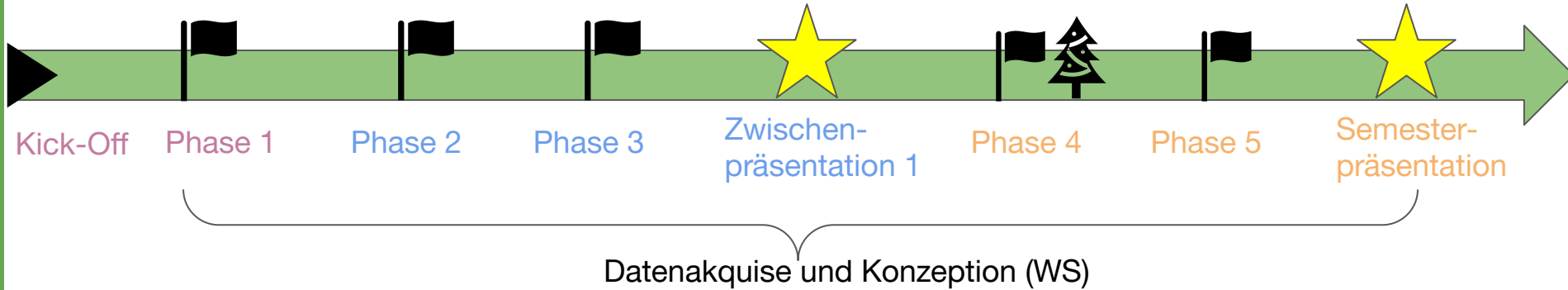
Unter Anderem:  
Wissenschaftliche Veröffentlichungen  
Patente  
Zitationen  
Projekte  
Wissenschafts- und Themenfelder



Dabei soll ein **Recommender-System** aufgebaut werden, dass *externen Forschungs-Interessierten*, wie z.B.

Wissenschaftlern, Forschern, Hilfsorganisationen, NGOs, Firmen und Verbände der Region (und darüber hinaus) aber *auch Internen Interessierten in der TH Köln*, schnell und zuverlässig Experten zu bestimmten Themengebieten herausfiltern kann.

## 1.2 Timeline, Meilensteine & Next Steps



### Kick-Off & Phase 1

- Beginn der Projektarbeit
- Erarbeitung möglicher Umsetzungsideen
- Erarbeitung konkreter User-Stories & Arbeitspakete
- Klein-Gruppenaufteilung
- Gruppenarbeit an den ersten User-Stories
- Bachelor Arbeit, Externe Quellen, Interne Quellen

### Phase 2 & 3

- Präsentation der ersten Ergebnisse
- Neuauswahl an User-Stories aus dem Ideen-Pool
- Neuaufteilung der Arbeitsgruppen
- Datenspeicherung, Externe Quellen, Optimierung Prototyp
- Vorbereitung der Zwischenpräsentation

### Phase 4 - 5 / Next Steps

- Diskussion und Integration des Feedbacks aus ZP1
- Neuauswahl an User-Stories aus dem Ideen-Pool
- Neuaufteilung der Arbeitsgruppen
- Abschluss der Datenakquise & Konzeption

## Projektplanung

Updated 2 hours ago

Filter cards

12 Ideas + ...

Validierungsdaten/datensätze ...  
Added by AndyKruff

Recherche zu Open-Source Code / Libraries / Frameworks für spätere Modelle ...  
Added by pstorner

Wikipedia Definitionen zu den "field\_of\_teaching"/"field\_of\_research" fielders scrapen ...  
Added by jueri

Vorhandene Scoring-Modelle ggf. auf unser Projekt anpassen ...  
Added by FabianGitzler

mögliche Erweiterung um Abstracts oder soziale Netzwerke (XING, LinkedIn) eruieren ...  
Added by matteomeier

Recherche möglicher externer Quellen zur Datenbereitstellung ...  
Added by pstorner

7 ausgewählte Ideen + ...

Übergeordnet: Dokumentation des Codes inklusive Zuordnung zu Arbeitspaketen/Sprints ...  
Added by matteomeier

Den Code der BA erstmal zum laufen bringen und die results sichten um weitere Entscheidungen treffen zu können. ...  
Im Anschluss validierung des Codes.  
Added by LeonMunz

Erschließung weiterer externer Quellen für Personeninformationen ...  
Added by Consibrah

Erschließung interner Quellen der TH für Personeninformationen ...  
Added by AndyKruff

Übergeordnet: Erstes UML Diagramm zur Strukturierung des Codes (soweit möglich) ...  
Added by AndyKruff

1 To-Dos | Tasks + ...

User Story | Task: Externen Crawler programmieren ...  
ExterneQuellen#1 opened by jueri

9 Done + ...

michellereiners

User Story | Task: Sammeln von Informationen zu externen Links ...  
Organization#5 opened by Consibrah

User Story | Task: BA Code und Analyse der Ergebnisse ...  
Organization#4 opened by matteomeier

User Story | Task: Vorbereitung der Präsentation (17.12) ...  
Organization#8 opened by Consibrah

User Story | Task: Präsentation mit Inhalten der Orgagruppe befüllen ...  
Organization#9 opened by Consibrah

User Story | Task: Konzeption einer Speichermöglichkeit (Datenbank) ...  
Organization#7 opened by jueri

Präsentation mit Inhalten der Fachgruppen befüllen 3 ...  
ExterneQuellen#2 opened by Consibrah

Präsentation mit Inhalten der Fachgruppen befüllen 2 ...  
InterneQuellen#11 opened by Consibrah

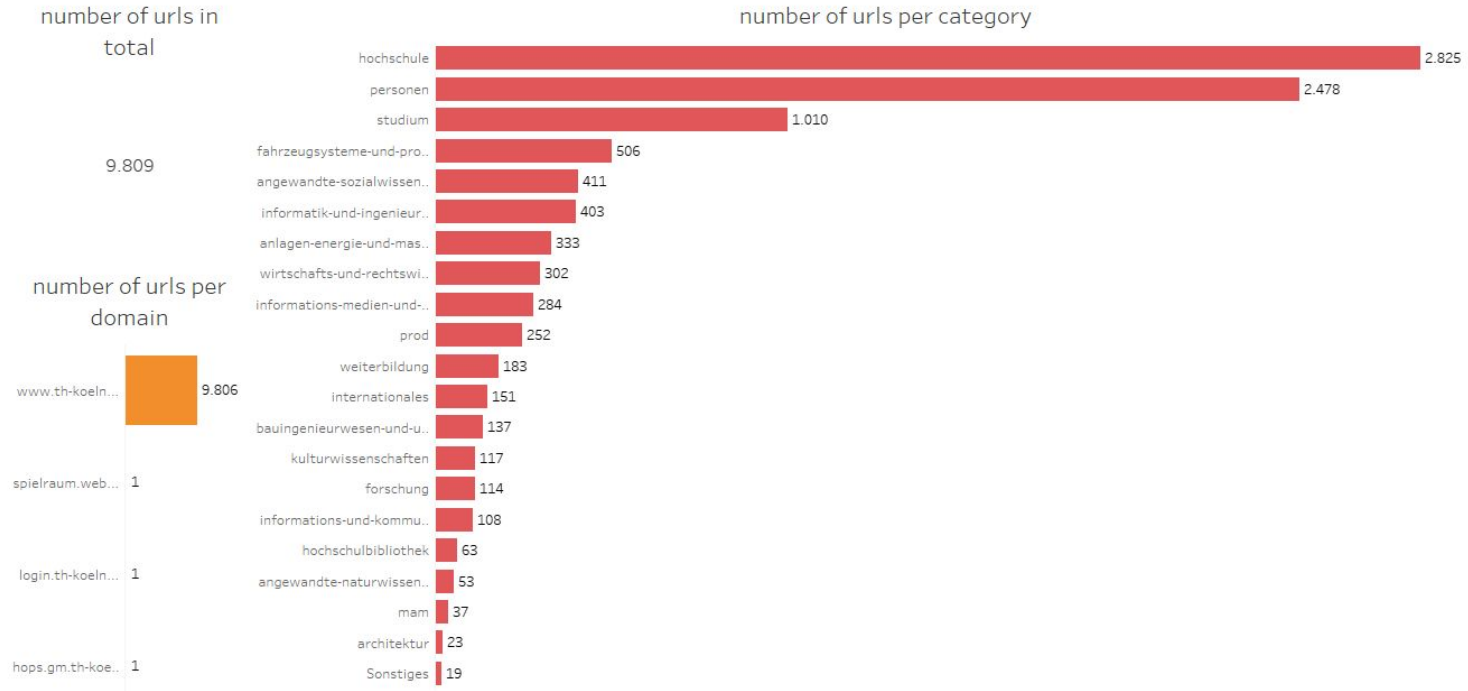
# 02 —

## Quellen & Datenbasis

## 2.1 Auswertung interne Quellen

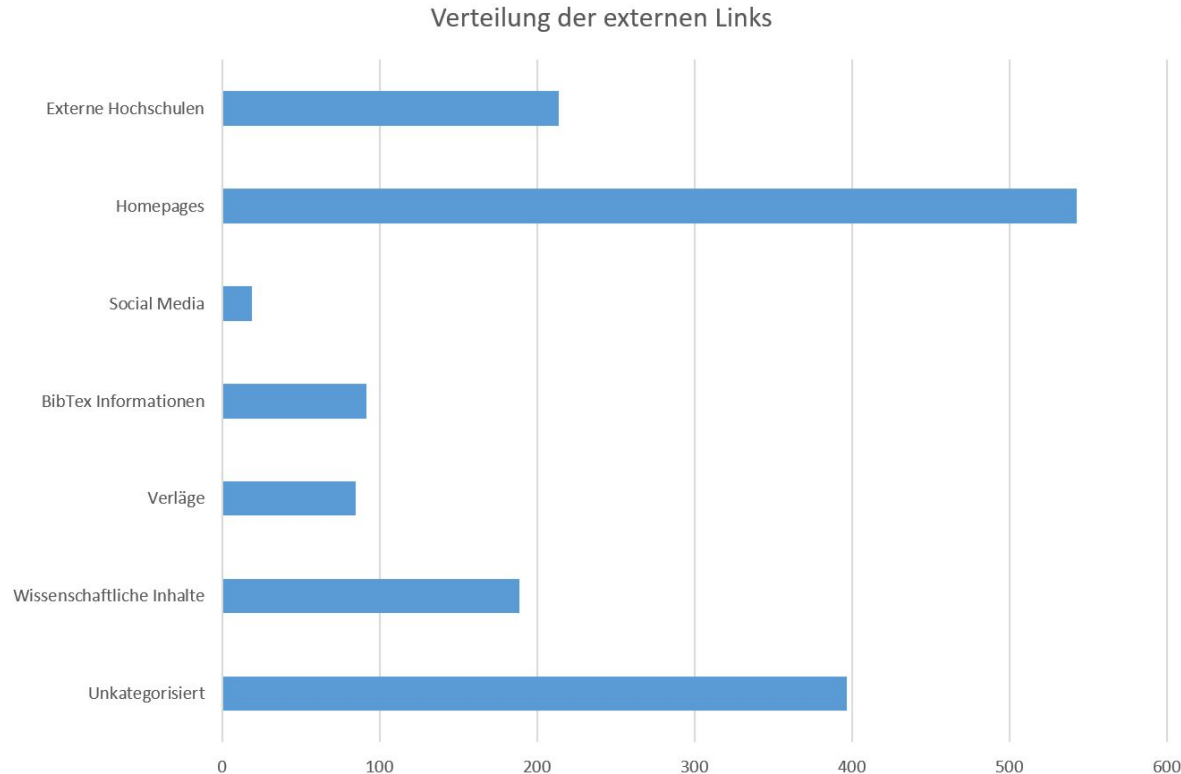
Research  
Knowledge  
Discovery

Technology  
Arts Sciences  
TH Köln





## 2.2 Auswertung Externe Quellen



### Key Facts:

- Unique Outlinks: 2086
  - Inaktiv: 449
  - Aktiv: 1637
- Die meisten Personenseiten beinhalten 1-3 Links
- 4 Personen haben 0 Outlinks
- Maximum: 216 Outlinks

## 2.3 Rechtlicher Aspekt

### **§60c UrhG Wissenschaftliche Forschung**

(1) Zum Zweck der nicht kommerziellen wissenschaftlichen Forschung dürfen bis zu 15 Prozent eines Werkes vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden [...]

### **§60d UrhG Text und Data Mining:**

Um eine Vielzahl von Werken (Ursprungsmaterial) für die wissenschaftliche Forschung automatisiert auszuwerten, ist es zulässig,

- (1) das Ursprungsmaterial auch automatisiert und systematisch zu vervielfältigen, um daraus insbesondere durch Normalisierung, Strukturierung und Kategorisierung ein auszuwertendes Korpus zu erstellen, und [...]

Der Nutzer darf hierbei nur nicht kommerzielle Zwecke verfolgen.

[...]

Webseiten können als Datenbankwerk angesehen werden und unterliegen Schutz durch §4 UrhG.

Verbot der Umgehung technischer Schutzmaßnahmen nach §95a UrhG

Das können u.a. sein: Beschränkungen durch robots.txt, Captcha-Verfahren oder Paywall

# 03 —

## Das Datenmodell

## 3.1 Sciebo Projektbox

### Zugriff auf die sciebo Projektbox für das Projekt Knowledge Discovery

Für unser Projekt wurde uns eine sciebo Projektbox mit einer Größe von 1 TB zur Verfügung gestellt.

Dieses Jupyter Notebook soll den Zugriff auf die Projektbox via pyoclient vermitteln. Damit der pyoclient genutzt werden kann, muss das Modul mit `$ pip install pyoclient` installiert werden.

Für das Crawlen der Daten auf der TH Seite wurde der Ordner intern erstellt, für das Crawlen der Daten aus externen Quellen der Ordner extern, für alles was mit der Datenbank zu tun hat der Ordner db. Root Verzeichnis ist Research-KD.

Pfade für die Projektbox müssen somit immer Research-KD/ enthalten.

<https://github.com/owncloud/pyoclient#usage>

```
: import owncloud
oc = owncloud.Client('https://th-koeln.sciebo.de/')
oc.login('research-kd.pbox@th-koeln.de', 'AONYZ-BUOAT-PKWPP-OUSAF')

#oc.mkdir('Research-KD')
#oc.mkdir('Research-KD/intern')
#oc.mkdir('Research-KD/extern')
#oc.mkdir('Research-KD/db')

root = 'Research-KD/'
intern = 'intern/'
extern = 'extern/'
db = 'db/'

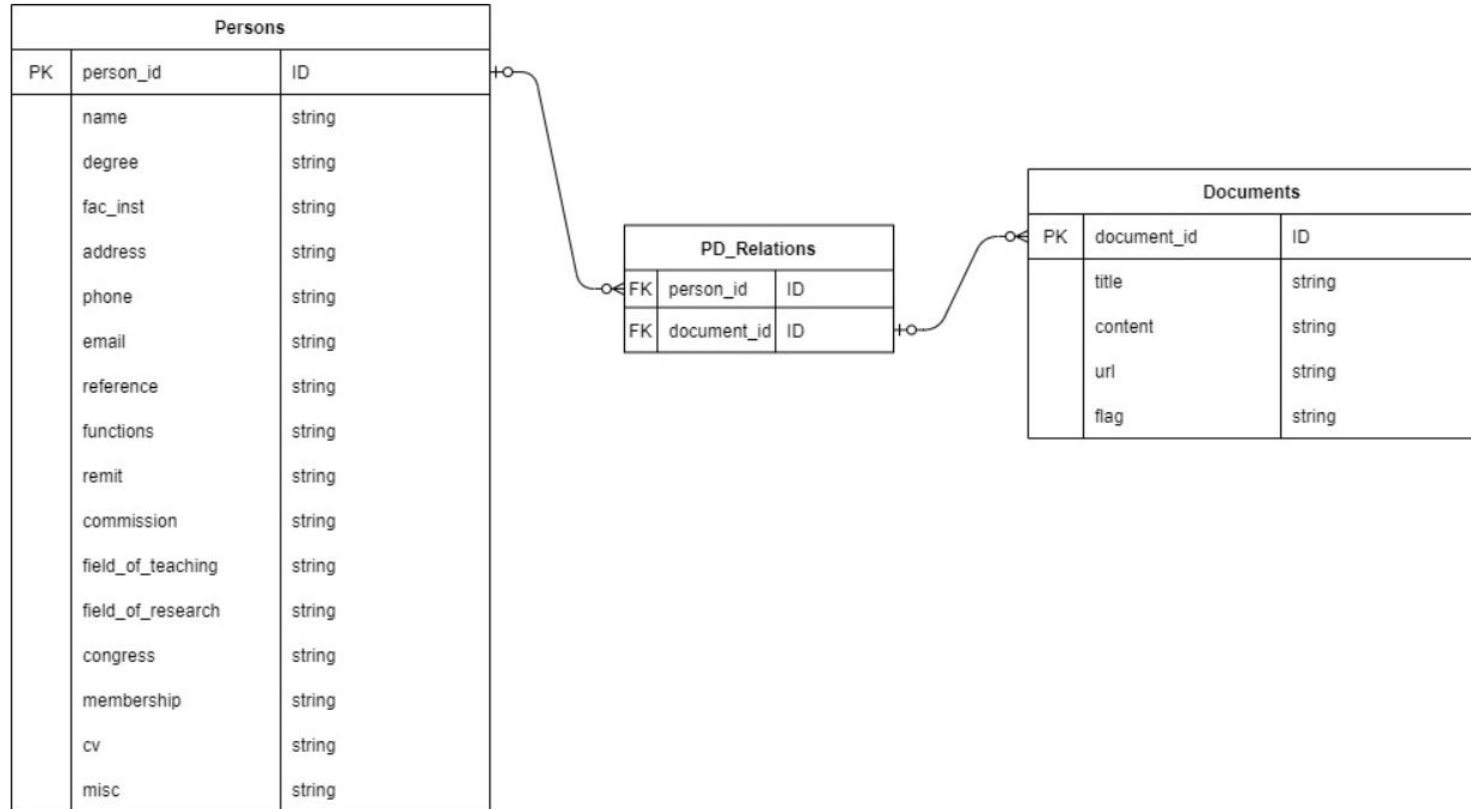
# make directories

def mkdir_intern(new_folder):
    oc.mkdir(root + intern + new_folder)

def mkdir_extern(new_folder):
    oc.mkdir(root + extern + new_folder)

def mkdir_db(new_folder):
    oc.mkdir(root + db + new_folder)
```

## 3.2 Datenschema





# Optimierung des Prototyps

- Beginn des Scrapes auf der Startseite
- Erweiterung des Scrapes auf die gesamte TH-Köln Domain (noch in Optimierung)
- Selektion der gescrapten Seiten:  
Ignored Extensions, html-Tag Beschränkungen, Name-Matching (noch in Optimierung)
- Kategorisierung der Seiten (noch in Bearbeitung)