# DB Digital Header:challenge

Digital Solutions for the DB

Till Plumbaum & Andreas Lommatzsch

DAI-Labor
Distributed Artificial Intelligence Laboratory
berlin

mindbox
DB NETZE

DB Challenge

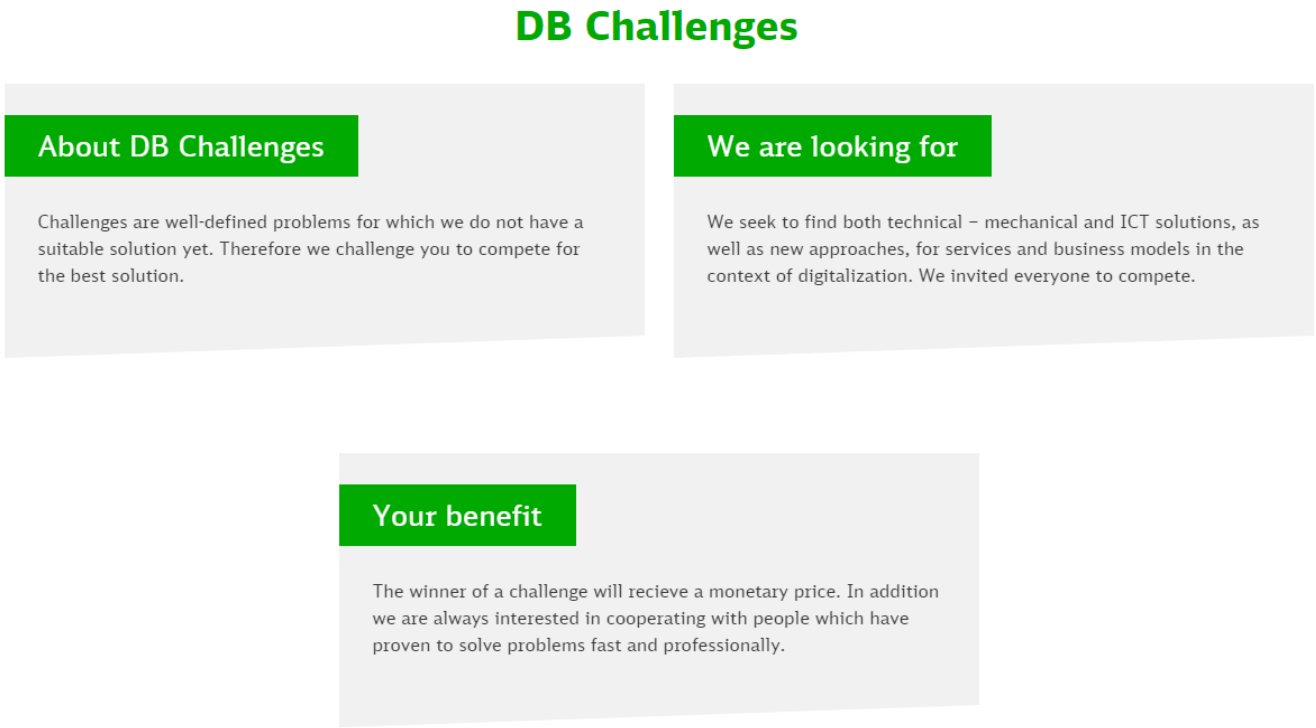
# Innovate DB: Enter DB mindbox


**One 4.0 initiative of DB**

**Three major instruments**


**One fabulous space for coworking and exchange**

# We want you contributing new solutions




1. DB Information:challenge
   Dec. 2015 – Feb. 2016

2. DB Information UX:challenge
   March 2016 – April 2016

3. **DB Digital Header:challenge**
   **March 2016 - April 2016**

https://www.mindboxberlin.com/index.php/challenges.html

# DB Digital Header:challenge

**The challenge**

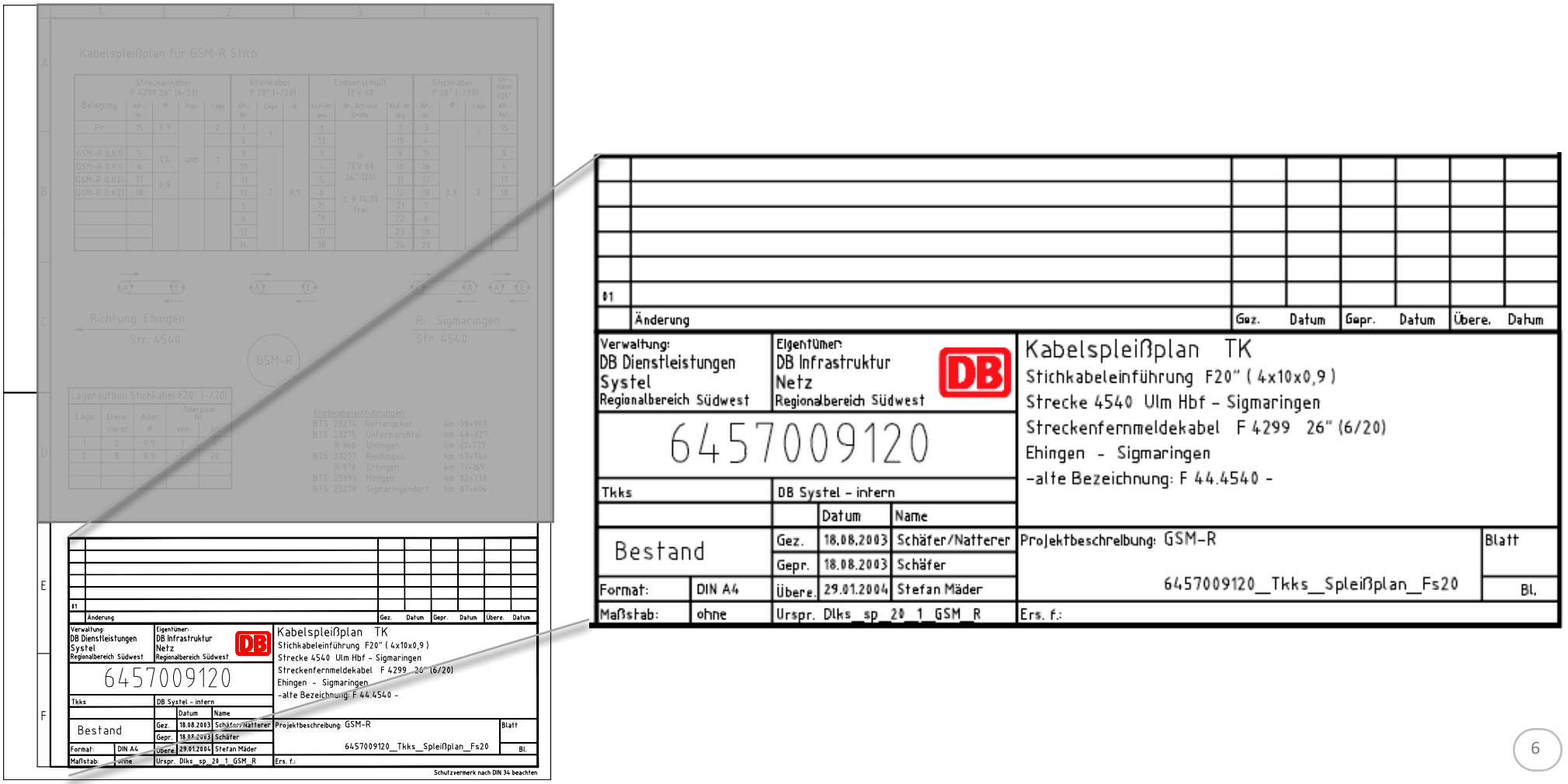
**Supporting the DB to enter the age of digitalization**

- The Deutsche Bahn has large numbers of scanned, but not digitalized, documents, describing various technical systems
- Transferring the information in these documents into a searchable, digitalized structure will ease the daily work of DB employees
- The process of transferring the information currently means manual work - and is very time consuming
- Goal of the challenge is to find and develop automated (or semi-automated) approaches to do this

**http://gmedia.de/dbdigitalheader/**

# What exactly is the challenge all about?

# What is a digital header?

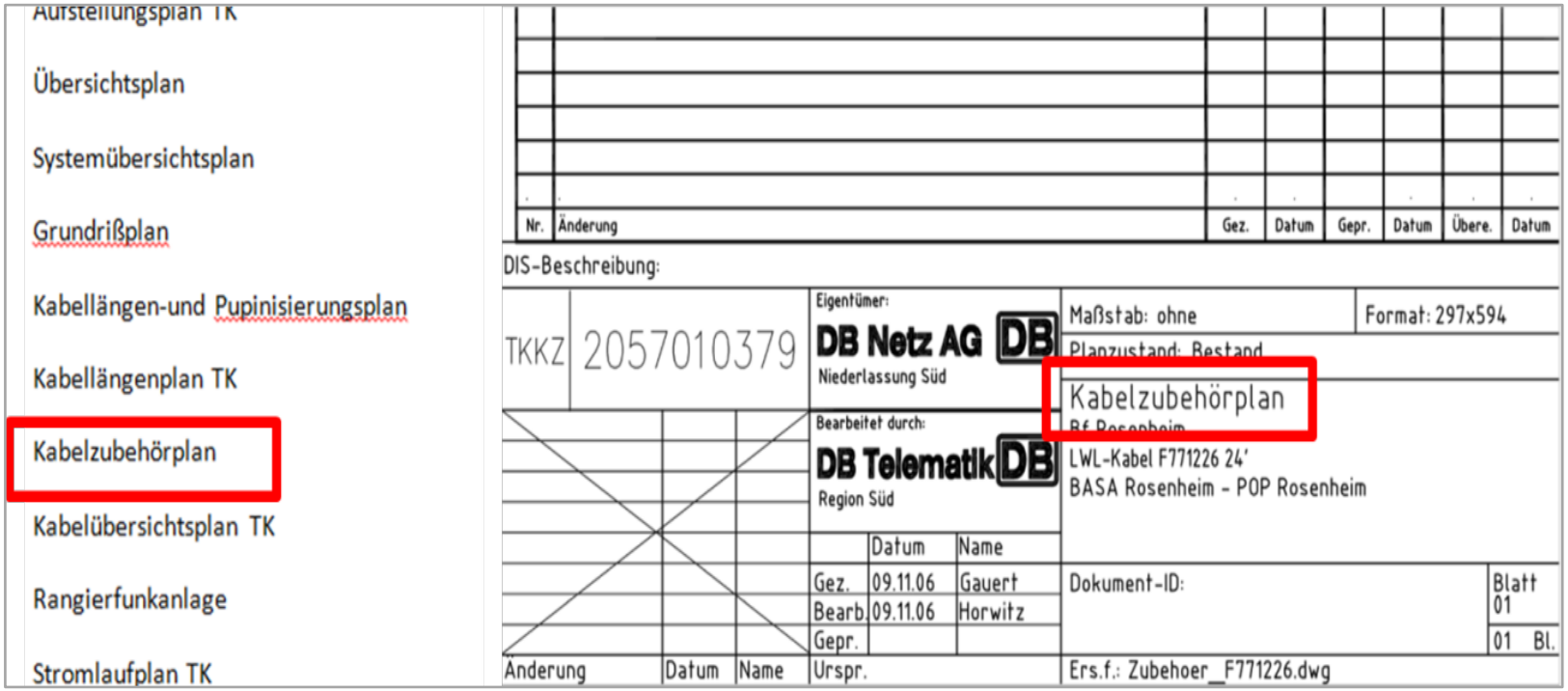


# Extracting information

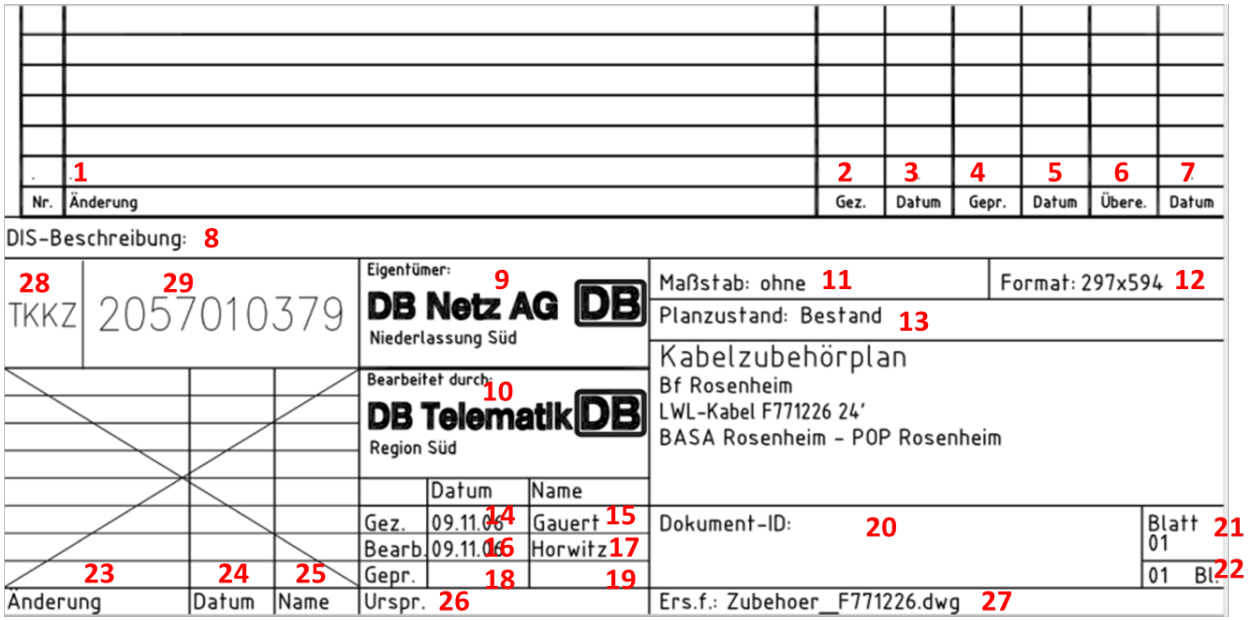
## Detecting the document type

List of document types

Document header



# Extracting Information




Document header

1 - 8: empty field
9 Owner: - *DB Netz AG Niederlassung Süd*
10 Processed by: - *DB Telematik Region Süd*
11 Scale: - *none*
12 Format: - *297x594*
…
21 Page - *01*
22 Pages in total: *01*
…
*29 ID: 2057010379*
…

Overall 52 different fields for the different plan types

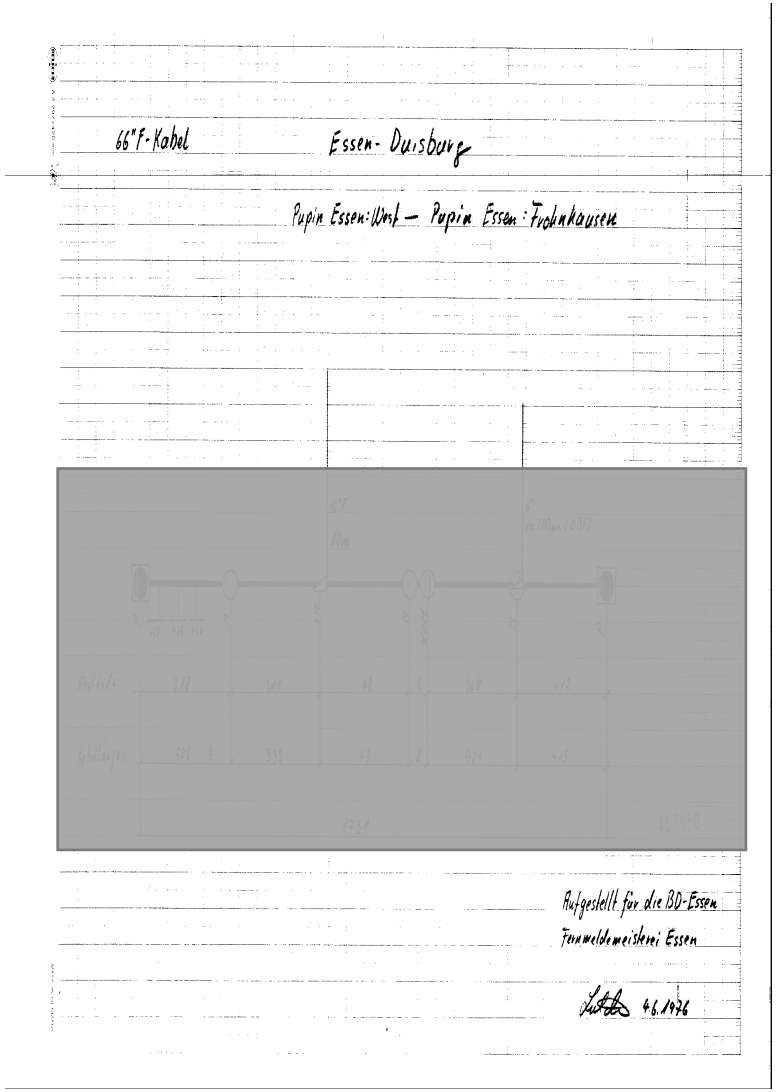
# Is it really that simple?

## Complicated headers

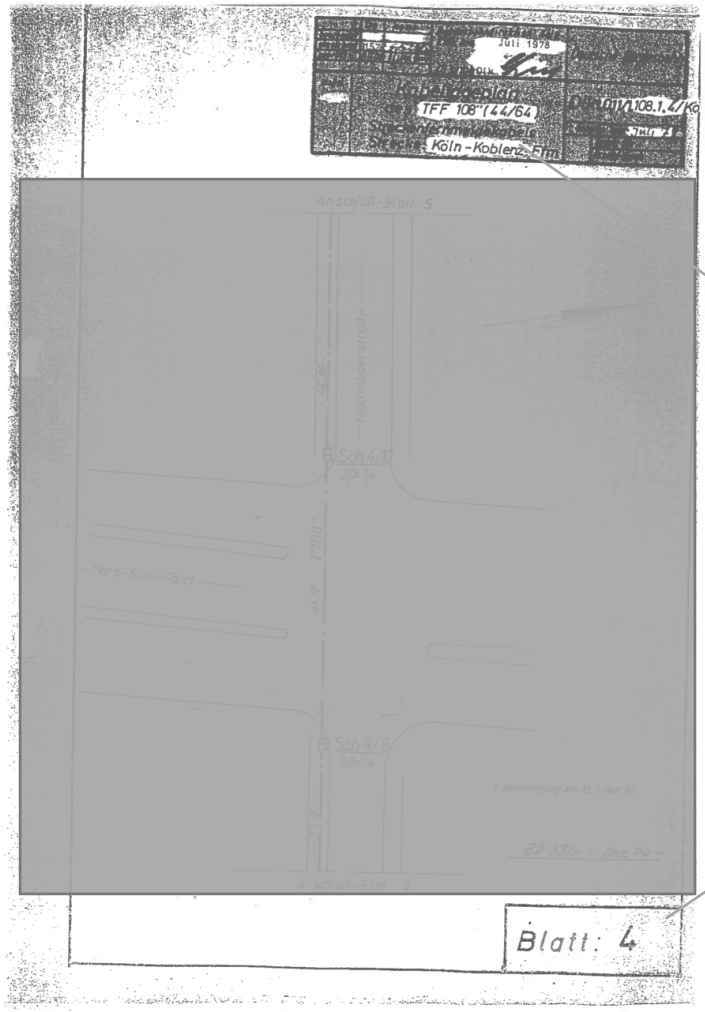
# NO!

The document collection includes different types from different times

Some are hand-drawn / hand-written



# Complicated headers




In some documents the information is blacked out

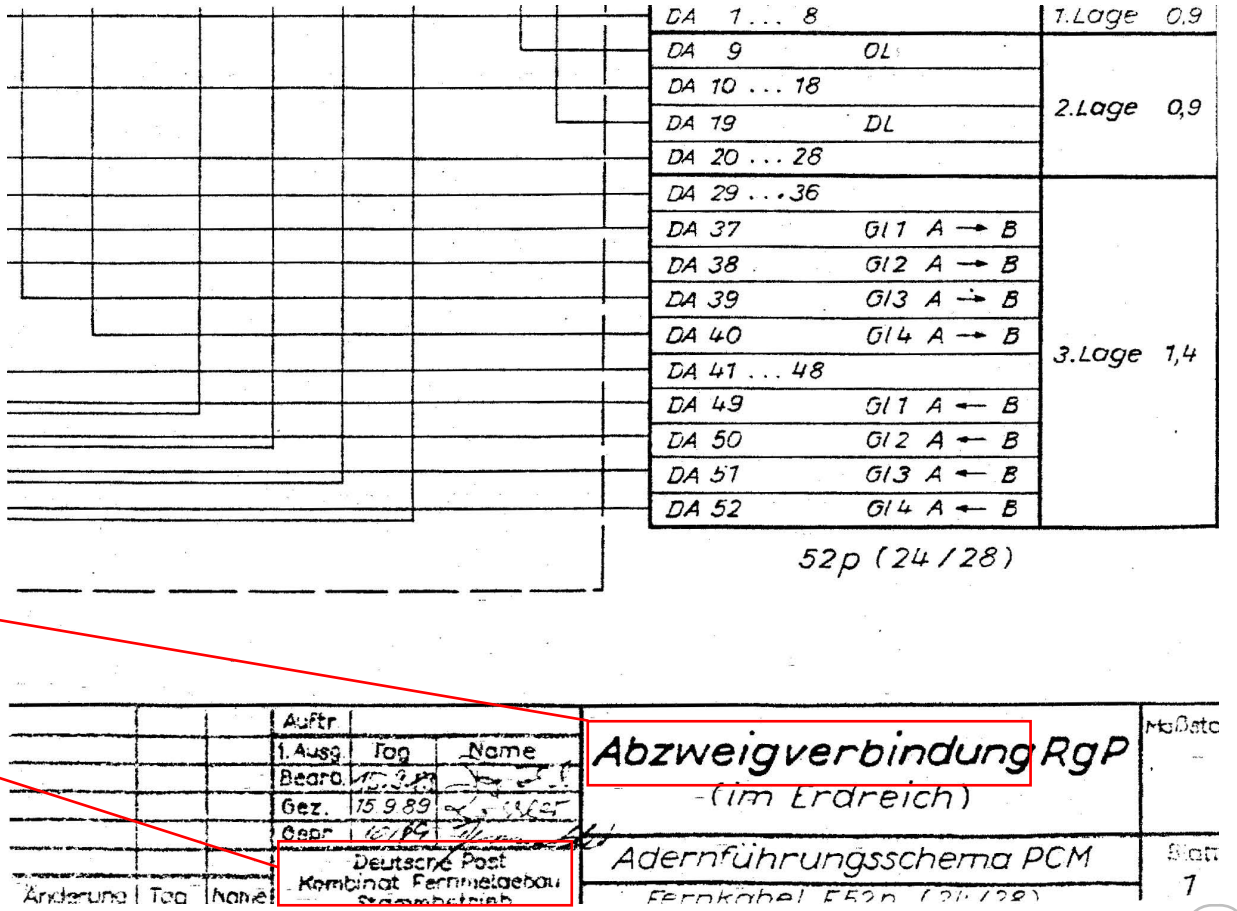
Köln-Koblenz and Page 4

# Different header formats

- The good news – the most complicated or unreadable documents were removed
- But there is still a wide variety of documents

Abzweigverbindung

Deutsche Post



## The Challenge

We are looking for an algorithm which will allow us to automatically extract and organize documents such as blueprints and plans. These are our main goals:

**Auto detection of various header types**

**Header extraction & transfer into data structures**

**Portability of the algorithm onto unknown headers**

## What you have to do to win

Challenge consist of three parts:

- Automatized identification of different headers and sorting these into different stacks – known, unknown, indecipherable.
- Extraction of information from known headers and text blocks and subsequent transfer into an organized data structure.
- Portability of the algorithm onto unknown headers.

The winner is selected by the combined results for the three tasks.

1. Detect document type 20%
2. Extract information 50%
3. Portability 30%

**End of April, a jury of DB representatives and external experts announce the winners and award the prizes.**

## Why should you participate?

The three winner(-teams) will get :

- a great opportunity to solve a real-world problem

- Price money of 14.000 Euro overall
  (8.000 for the first, 4.000 for the second, and 2.000 for the third)


- A great reference for you CV

## What data you get?

- First data set – available on Monday
  - Around 900 documents
  - Sample solutions for X different plan types
  - A list of all possible document types
  - A structured solution file for the answers
- Second data set – for the final scoring
  - Around 900 documents
  - No sample solutions
  - Limited time to solve

# The Challenge: What to do and when

**Timeline**

23.02.2016
Start of the registration phase

~~14.03.2016~~ 21.03.2016
Start of challenge with first data set available

17.03.2016
Meetup - TODAY

24.03.2016
Meetup @ mindbox

24.03.2016
First Leaderboard

# The Challenge: What to do and when

20.04.2016
End of the challenge and start of the final evaluation phase

25.04.2016
Announcement of the best three teams

27.04.2016
Award ceremony

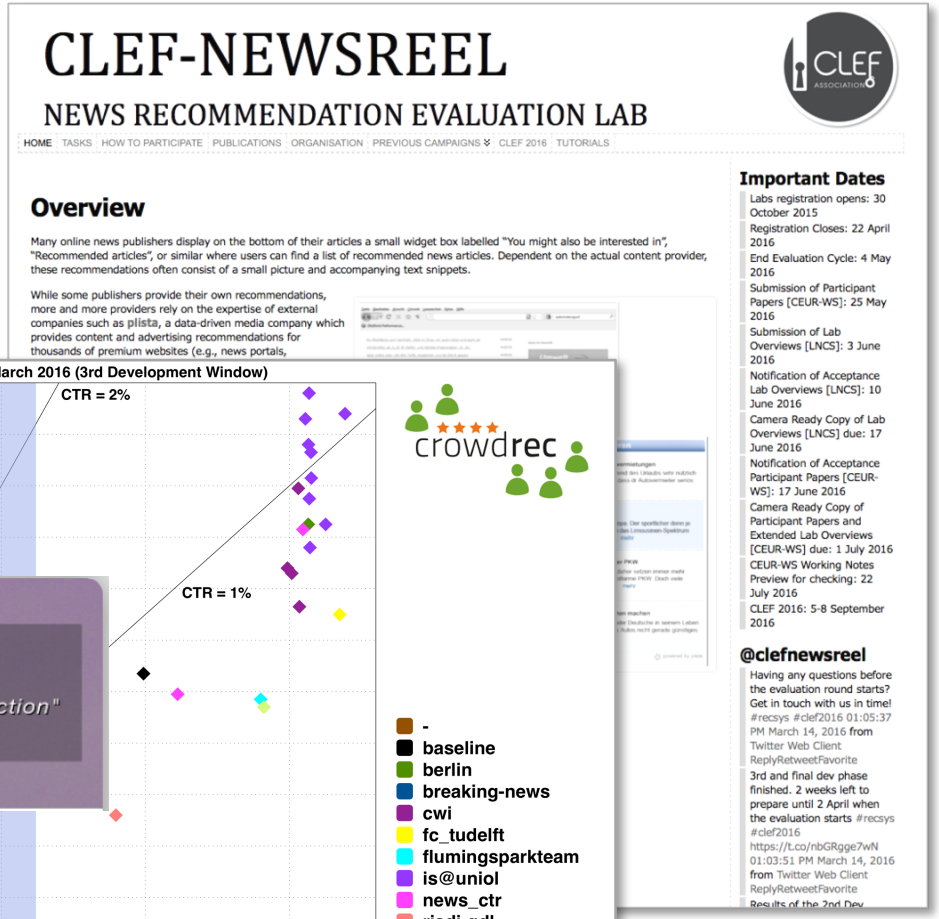
## Ways to approach the challenge

- Character recognition (OCR)
  - Most likely the first step…
- If you have text
  - Error correction
  - robust matching of words
  - Ontologies or/and dictionaries (for names, types, etc.)
- If you work with images
  - Robust matching of parts of the image
- Detect Pattern and known segments
- small number of documents impedes methods such as Deep Learning


HOME
FORUM
REGISTER
LOGIN
Digital Header Challenge
Digitale Lösungen
für die DB
JETZT REGISTRIEREN
http://gmedia.de/dbdigitalheader/

## CLEF NewsREEL WWW.CLEF-NEWSREEL.ORG

- News Recommendation
- Task 1:
  - receive recommendation requests from visitors
  - evaluation from 2-29 April 2016
- Task 2: predict how visitors read news with recorded data




Results will be presented at

# Close proximity to research institutes


TEL Building at Ernst-Reuter-Platz

- Deutsche Telekom Innovation Laboratories (T-Labs)
- Connected Living Innovation Center
- DAI-Labor / TU Berlin
- German Turkish Advanced Research Centre for ICT (GT-ARC)
- EIT Digital (Berlin Node)
- Daimler Center for Automotive IT Innovations

Telekom Innovation Laboratories

CONNECTED LIVING

DAI-Labor
Distributed Artificial Intelligence Laboratory

GT ARC
GERMAN TURKISH ADVANCED RESEARCH CENTRE FOR ICT

eit Digital

DAIMLER

## Expert centers:

IDBB Fachzentrum Intelligente Dienste für Bürger und Behörden

kopra Kompetenzzentrum digitale Präventionsassistenz

## Spin-Offs:

IOLITE

CURAMATIK

# Staff & Partners

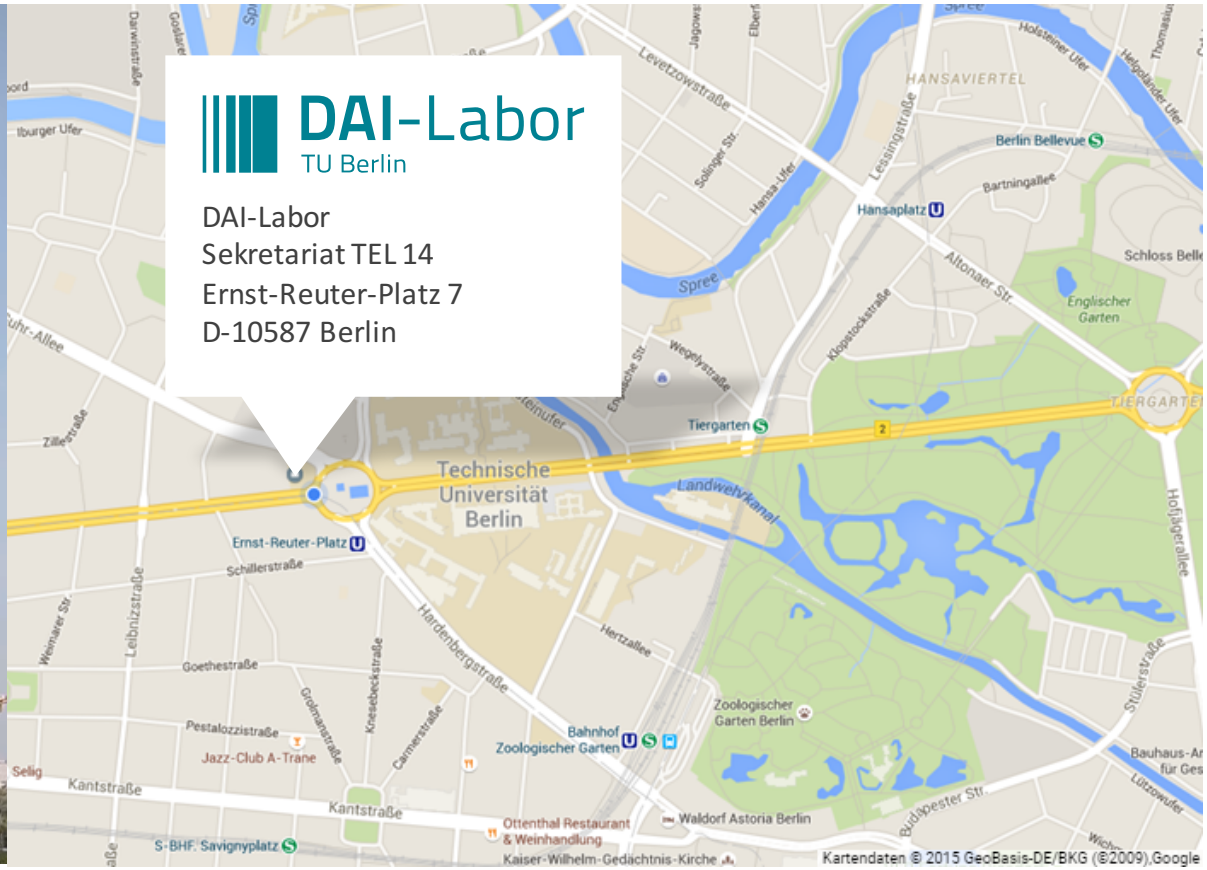
15 Post-Doctoral Researchers

50+ Full-time Doctoral Researchers (Ph.D. Candidates)

50+ Half-time Student Research Assistants

8 Support Staff: Secretaries, System Admins & Graphics Designers

Strong Cooperation with Industry and Research Institutions


## Get In Touch

sahin.albayrak@dai-labor.de

Prof. Dr. Dr. h.c. Sahin Albayrak

+49 30 - 314 74000