DISCUSSION SESSION

Streamlining Data Publication: Automatic Metadata and Large Datasets in the Age of AI

Leibniz ScienceCampus
Digital Transformation of Research

Anna Jacyszyn[1], Felix Bach[1], Tobias Kerzenmacher[2], Mahsa Vafaie[1]

[1]FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
[2]KIT Institute of Meteorology and Climate Research Atmospheric Trace Gases and Remote Sensing (IMK-ASF)

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

funded by Leibniz Competition

Open Science Conference, 8-9 October 2025, Hamburg

Welcome

Leibniz ScienceCampus
Digital Transformation of Research

Session organisers

Anna Jacyszyn[1]

DiTraRe coordinator, dimension *Exploration and Knowledge Organisation*

Mahsa Vafaie[1,3]

DiTraRe dimension *Exploration and Knowledge Organisation*

Tobias Kerzenmacher[2]

DiTraRe use case *Publication of Large Datasets*

Felix Bach[1]

DiTraRe coordinator, dimension *Tools and Processes*

Supporting team

Kerstin Soltau[1]

RADAR Product Manager

Stefan Hofmann[1]

RADAR Full Stack developer

[1]FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
[2]KIT Institute of Meteorology and Climate Research Atmospheric Trace Gases and Remote Sensing (IMKASF)
[3]KIT Institute of Applied Informatics and Formal Description Methods (AIFB)

Agenda

- **Ice-breaker** session
- Flash talks: **experts from different disciplines**
- **Fishbowl** discussion
- Workshop **summary**

Take notes with us and be part of the session proceedings!

zbw.to/osc25-pad02

Ice-breaker session

Join at

slido.com

#2264 787

Flash Presentations

DiTraRe

Leibniz Science Campus *Digital Transformation of Research*

- Growth core to **establish new research branch**.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

+

KIT
Karlsruhe Institute of Technology

- Planned as a **4+4 years** project (start: September 2023).
- Funded by the Leibniz Association + FIZ KA + KIT.
- Analyse the process of **digitalisation of research**.
- **Multilevel interdisciplinary approach**. → We start with 4 specific use cases.

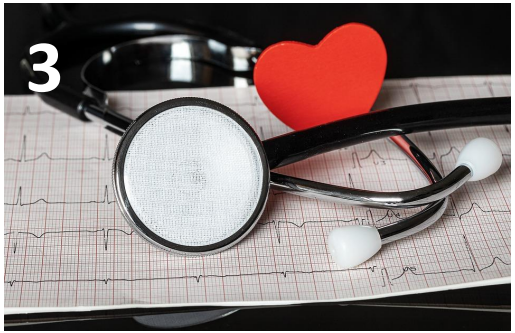
DiTraRe Use Cases

Sensitive Data in Sports Science

*KIT Institute of **Sports and Sports Science***

Chemotion Electronic Lab Notebook

*KIT Institute of **Biological and Chemical Systems***



AI in Biomedical Engineering

*KIT Institute of **Biomedical Engineering***

Publication of Large Datasets

*KIT Institutes of **Meteorology and Climate Research***

DiTraRe dimensions

A. Reflection and Resonance
a dialogue between research and society, interactive process

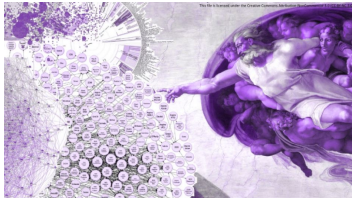
B. Exploration and Knowledge Organisation
applied AI: represent, organise, and manage domain specific and procedural knowledge

C. Legal and Ethical Challenges
data ethics, data protection, copyright and data law

D. Tools and Processes
digital tools tailored precisely to the needs of researchers

This session

Use case
Publication of Large Datasets



Dimension *Exploration and Knowledge Organisation*

RADAR
FIZ Karlsruhe

Dimension *Tools and Processes*

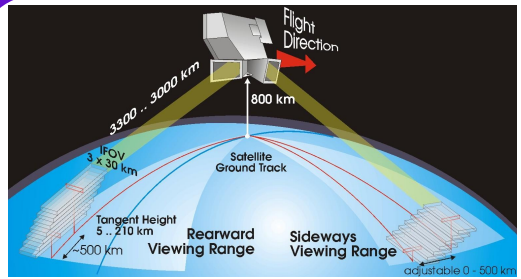
Discussion session
Streamlining Data Publication: Automatic Metadata and Large Datasets in the Age of AI

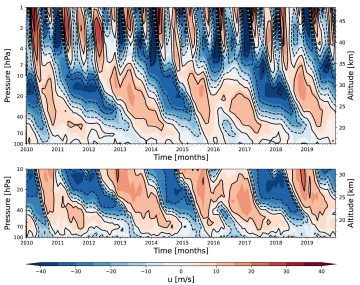
Leibniz ScienceCampus
Digital Transformation of Research

KIT 200 YEARS

Use case: Publication of Large Datasets

KIT Institutes of Meteorology and Climate Research





1. Research Questions

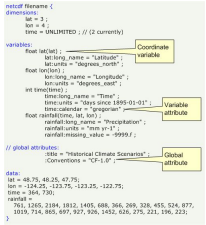
zenodo

netCDF

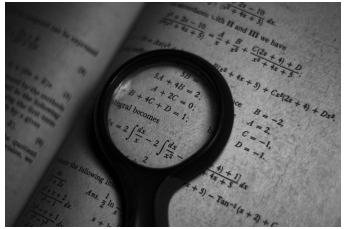
doi

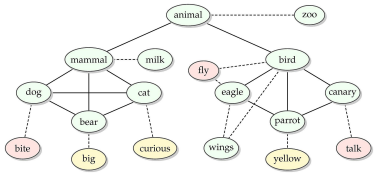
2. Status Quo – Current Practices

RADAR 4 KIT



3. Challenges & Case Studies





4. Looking Forward – Role of AI

It's not just about making data open – it's about making data usable, FAIR, and meaningful for re-use.

Discussion Session: *Streamlining Data Publication: Automatic Metadata and Large Datasets in the Age of AI*, Open Science Conference, 8-9 October 2025, Hamburg

Leibniz ScienceCampus
Digital Transformation of Research

DiTraRe dimension: Exploration & Knowledge Org.

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

https://www.archivportal-d.de/themenportale/wiedergutmachung

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online
ARCHIVPORTAL-D

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

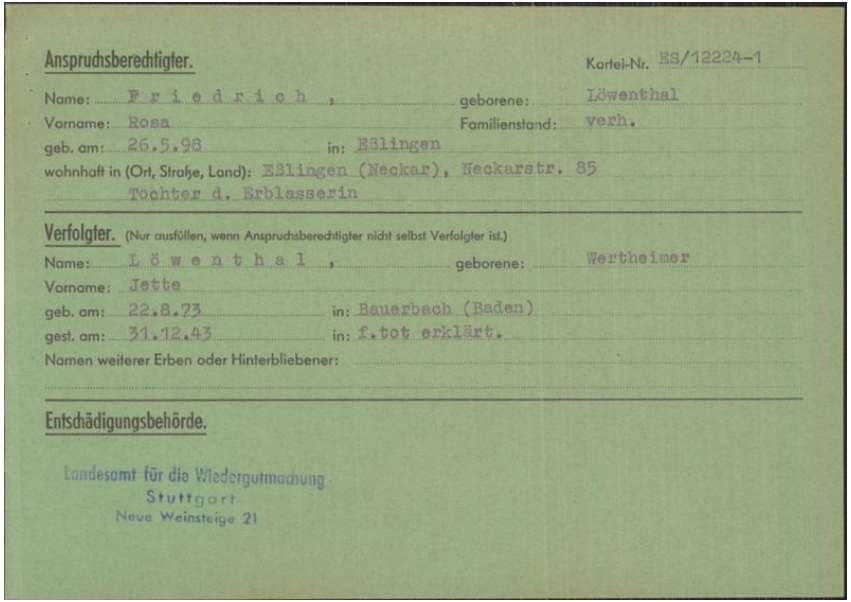
Bundesarchiv

Landesarchiv Baden-Württemberg

Bundesministerium der Finanzen

The **Bundeszentralkartei (BZK)** Collection

- Central Federal Index for Compensation of National Socialist injustices
- Central card file of most of applications for compensation in the Federal Republic of Germany
- about 1.9 million cards with basic information on applicants, persecuted persons and proceedings
- Kept from 1950s until today
- Basis for the **person search in Online Collection Wiedergutmachung**



Anspruchsberechtigter. Kartei-Nr. ES/12224-1
Name: Friedrich, geborene: Löwenthal
Vorname: Rosa Familienstand: verh.
geb. am: 26.5.98 in: Eßlingen
wohnhaft in (Ort, Straße, Land): Eßlingen (Neckar), Neckarstr. 85
Tochter d. Erblasserin

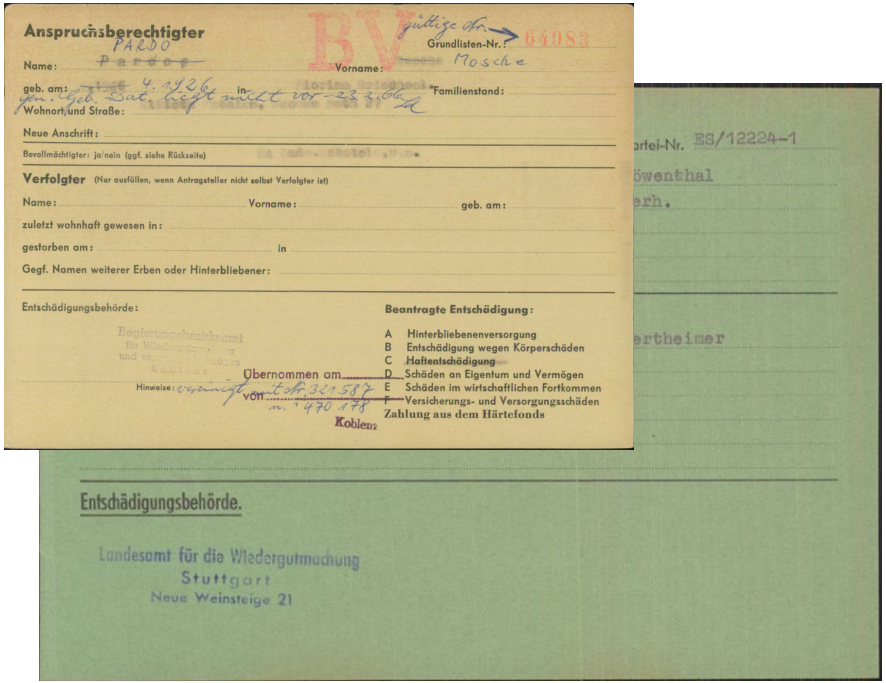
Verfolgter. (Nur ausfüllen, wenn Anspruchsberechtigter nicht selbst Verfolgter ist.)
Name: Löwenthal, geborene: Wertheimer
Vorname: Jette
geb. am: 22.8.73 in: Bauerbach (Baden)
gest. am: 31.12.43 in: f.tot erklärt.
Namen weiterer Erben oder Hinterbliebener:

Entschädigungsbehörde.
Landesamt für die Wiedergutmachung
Stuttgart
Neue Weinsteige 21

Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 64083, 190667, 15932, scho8, 67800/II/6095

The **Bundeszentralkartei (BZK)** Collection

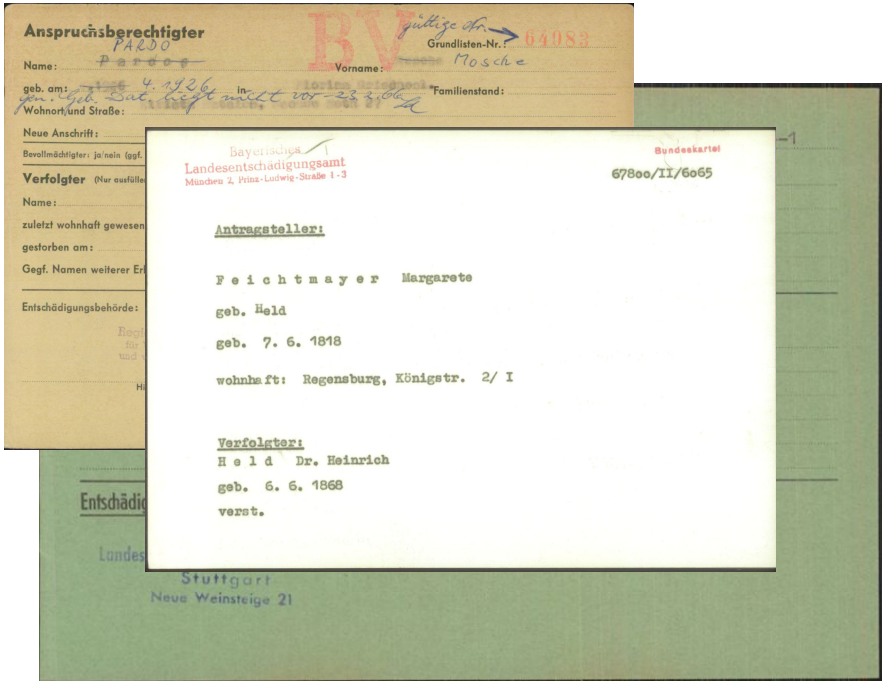
- Central Federal Index for Compensation of National Socialist injustices
- Central card file of most of applications for compensation in the Federal Republic of Germany
- about <u>1.9 million</u> cards with basic information on applicants, persecuted persons and proceedings
- Kept <u>from 1950s</u> until today
- Basis for the **person search in Online Collection Wiedergutmachung**



Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 64083, 190667, 15932, scho8, 67800/II/6095

The **Bundeszentralkartei (BZK)** Collection

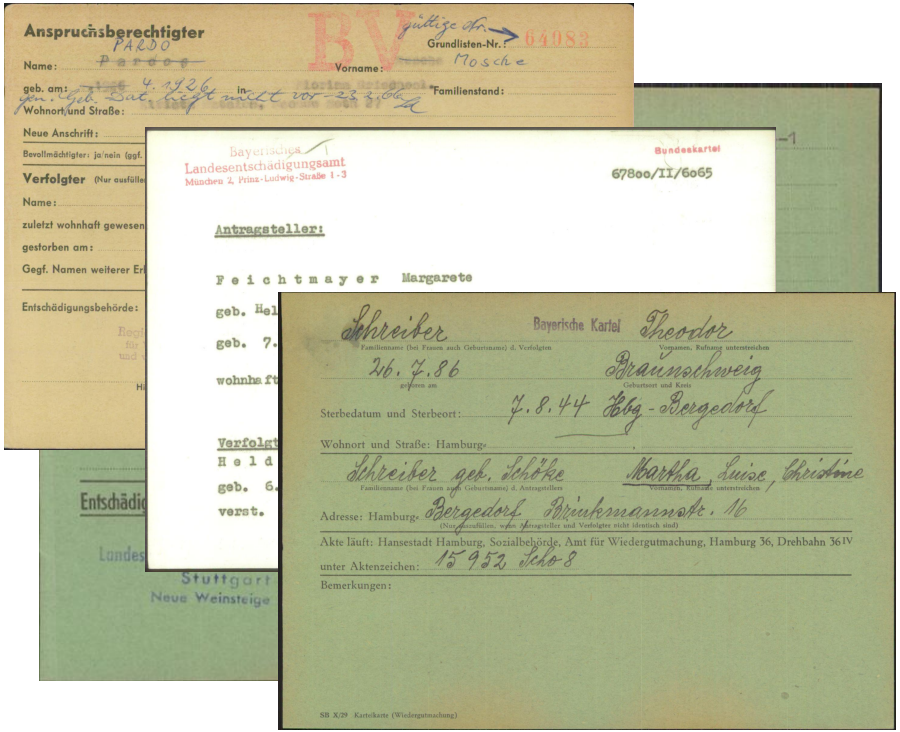
- Central Federal Index for Compensation of National Socialist injustices
- Central card file of most of applications for compensation in the Federal Republic of Germany
- about <u>1.9 million</u> cards with basic information on applicants, persecuted persons and proceedings
- Kept <u>from 1950s</u> until today
- Basis for the **person search in Online Collection Wiedergutmachung**



Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 64083, 190667, 15932, scho8, 67800/II/6095

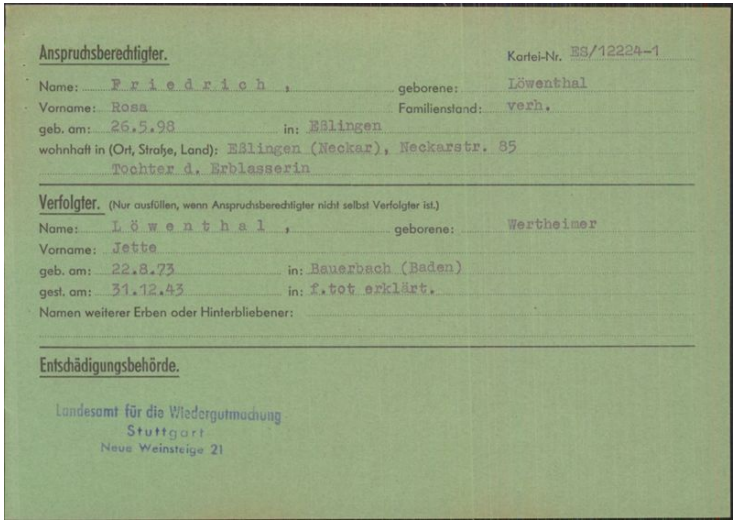
The **Bundeszentralkartei (BZK)** Collection

- Central Federal Index for Compensation of National Socialist injustices
- Central card file of most of applications for compensation in the Federal Republic of Germany
- about <u>1.9 million</u> cards with basic information on applicants, persecuted persons and proceedings
- Kept <u>from 1950s</u> until today
- Basis for the **person search in Online Collection Wiedergutmachung**



Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 64083, 190667, 15932, scho8, 67800/II/6095

The **BZK** Knowledge Graph

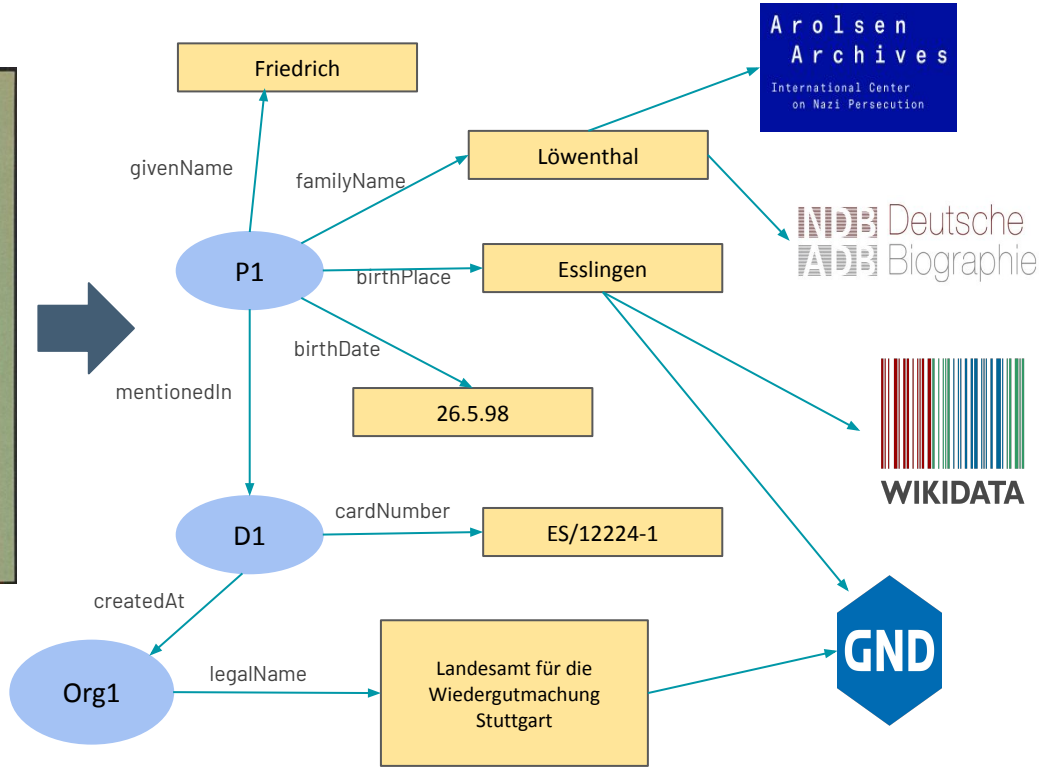


Anspruchsberechtigter. Kartei-Nr. ES/12224-1

Name: Friedrich, geborene: Löwenthal
Vorname: Rosa, Familienstand: verh.
geb. am: 26.5.98 in: Eßlingen
wohnhaft in (Ort, Straße, Land): Eßlingen (Neckar), Neckarstr. 85
Tochter d. Erblasserin

Verfolgter. (Nur ausfüllen, wenn Anspruchsberechtigter nicht selbst Verfolgter ist.)
Name: Löwenthal, geborene: Wertheimer
Vorname: Jette
geb. am: 22.8.73 in: Bauerbach (Baden)
gest. am: 31.12.43 in: f.tot erklärt.
Namen weiterer Erben oder Hinterbliebener:

Entschädigungsbehörde.
Landesamt für die Wiedergutmachung
Stuttgart
Neue Weinsteige 21



“Traditional” Information Extraction

Rule-based IE

Text Recognition → Document Classification → Template-based RegEx Analysis → Post-processing → Evaluation

NER-based IE

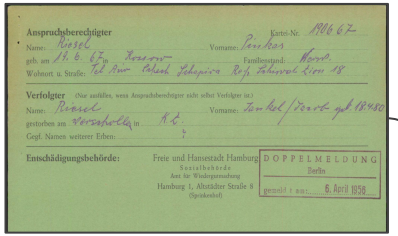
Text Recognition → NER → Document Classification → Template-based entity mapping → Post-processing → Evaluation

Limitations

- **Error Propagation**
  - Text Recognition
  - Layout Analysis
- **Template-based**
  - Generalisability
  - Laborious with >40 layout types

End-to-end MLLM-based Information Extraction

Input Image and Prompt → Multimodal LLM → Output Sequence → Converted CSV → Evaluation



<image>\nPlease provide the following information as you can see on the image as a Python dictionary [schema]…

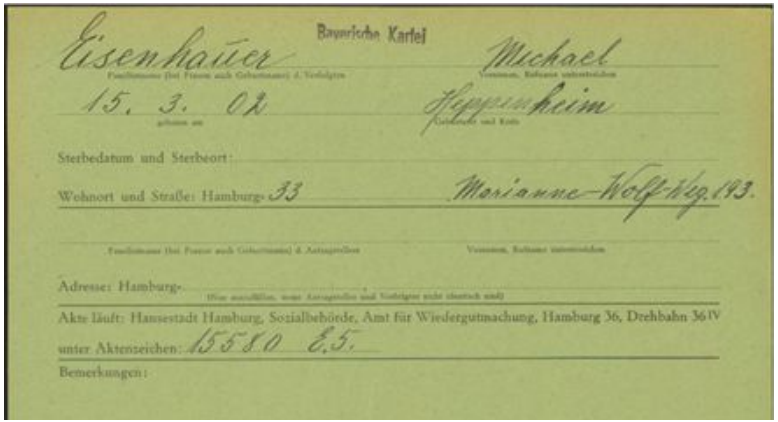
Open-source MLLM

{""full_response"": ""```python\n{\n \""CompensationOffice1\"" : \""Nordrhein-Westfalen\"",\n…}

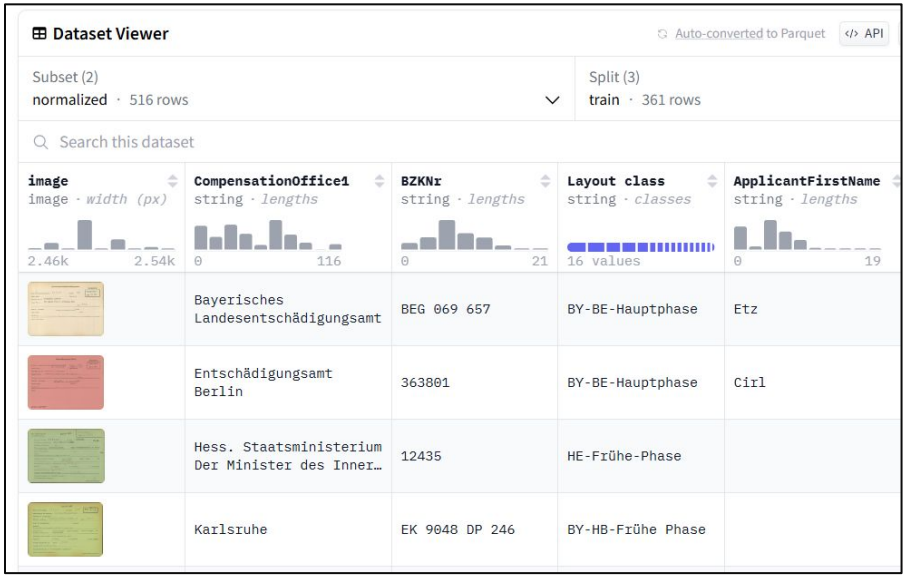
```
BZKNr: 1906 67
ApplicantFirstName: Pinkas
ApplicantLastName: Riesel
ApplicantBirthDate: 19.6.67
ApplicantBirthPlace: Kossov
VictimtFirstName: Jakob/Jerzb
VictimLastName: Riesel
VictimDeathPlace: Warschau
```

Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 190667

BZKOpen dataset and data schema


Bayerische Kartei
Eisenhauer
Michael
15. 3. 02
Heppenheim
Sterbedatum und Sterbeort:
Wohnort und Straße: Hamburg 33
Marianne-Wolf-Weg 193.
Adresse: Hamburg
Akte läuft: Hansestadt Hamburg, Sozialbehörde, Amt für Wiedergutmachung, Hamburg 36, Drehbahn 36 IV
unter Aktenzeichen: 15580 E.5.
Bemerkungen:

Keys for data extraction: CompensationOffice1, BZKNr, ApplicantFirstName, ApplicantLastName, ApplicantAltFirstName, ApplicantBirthName, ApplicantAltLastName, ApplicantBirthDate, ApplicantBirthPlace, ApplicantCurrentAddress, VictimFirstName, VictimLastName, VictimAltFirstName, VictimBirthName, VictimAltLastName, VictimBirthDate, VictimBirthPlace, VictimDeathDate, VictimDeathPlace



https://huggingface.co/datasets/MahsaVafaie/BZKopen

Source: Landesarchiv NRW – Abteilung Rheinland – BR 3015 ZK-Nr. 15580 E.5

Information Extraction with MLLMs

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur
ADVANCING SCIENCE

Models

Donut
GPT-4o-mini
InternVL2
InternVL2_5

Metrics

Normalised Edit Distance (NED)
Exact Matches (EM)
Partial Matches (PM)

Model	Size	NED	EM (t=0)	PM (t=1)	PM (t=3)
Donut-base	-	0.415	56%	57%	58%
Donut-base-finetuned	-	0.358	59%	61%	64%
GPT-4o-mini	-	0.184	72%	76%	79%
InternVL2.0	8B	0.382	53%	55%	60%
InternVL2.0	26B	0.431	48%	52%	57%
InternVL2.0	40B	0.158	76%	79%	83%
InternVL2.0-Llama3	76B	0.286	64%	67%	70%
InternVL2.0-finetuned	40B	0.173	74%	77%	81%
InternVL2.5	8B	0.340	57%	60%	64%
InternVL2.5	26B	0.311	60%	64%	69%
InternVL2.5	38B	**0.080**	**83%**	**88%**	**91%**
InternVL2.5	78B	0.139	77%	82%	84%
InternVL2.5-finetuned	38B	0.117	79%	84%	86%

Table 1. Performance of different transformer-based models on the BZKOpen dataset, including both pre-trained and fine-tuned variants. For fine-tuned models, the BZKOpen train set was used.

Information Extraction with MLLMs

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur
ADVANCING SCIENCE

Model	Size	Prompting Strategy	NED	EM (t=0)	PM (t=1)	PM (t=3)
GPT-4o-mini	-	ZS	0.184	72%	76%	79%
GPT-4o-mini	-	1FS	0.093	83%	87%	90%
GPT-4o-mini	-	2FS	0.080	84%	88%	91%
GPT-4o-mini	-	5FS	0.074	85%	89%	92%
InternVL2.0	40B	ZS	0.158	76%	79%	83%
InternVL2.0	40B	1FS	0.160	76%	79%	82%
InternVL2.0	40B	2FS	0.134	78%	82%	84%
InternVL2.0	40B	5FS	0.117	79%	82%	85%
InternVL2.5	38B	ZS	0.080	83%	88%	91%
InternVL2.5	38B	1FS	0.070	84%	89%	92%
InternVL2.5	38B	2FS	**0.060**	**86%**	**90%**	**93%**
InternVL2.5	38B	5FS	0.062	86%	90%	92%

Table 2. Performance of zero-shot (ZS) and few-shot (FS) prompting with different numbers of shots and different prompts for the InternVL2.5-38B and GPT-4o-mini models.

Takeaways

	Classical ML/DL	Generative AI (LLMs)
Efficiency		✔
Adaptability		✔
Reusability		✔
Reliability	✔	
Reproducibility	✔	
Data Privacy	✔	✔*

DiTraRe dimension: Tools and Processes

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

- About me: Felix Bach, Head of Research Data Department at FIZ Karlsruhe
  - Co-Spokesperson NFDI4Chem,
  - Scientific Coordinator of DiTraRe and
  - PI for dimension: Tools and Processes
- Focus of Dimension T&P:
  - Develop, test, and integrate digital tools and infrastructures for trusted workflows
  - Bridge between technical infrastructures (repositories, ELNs, AI services) and disciplinary practices
- Tools in place:
  - RADAR – repository service for research data, handling large and complex datasets
- Achievements so far:
  - Integration of ELNs (e.g., Chemotion) and repositories (e.g., RADAR4Chem) into FAIR workflows
  - First implementations of AI-supported metadata curation and interoperability solutions
- Future directions / in progress:
  - Efficient big data access (UL, DL, single file access within large data packages, viewer integration)
  - Distributed dataset strategies and interfaces
  - Exploitation of disciplinary metadata/terminologies
  - AI-based automatic FAIRness assessment and automatic metadata generation/enhancement

DiTraRe dimension: Tools and Processes

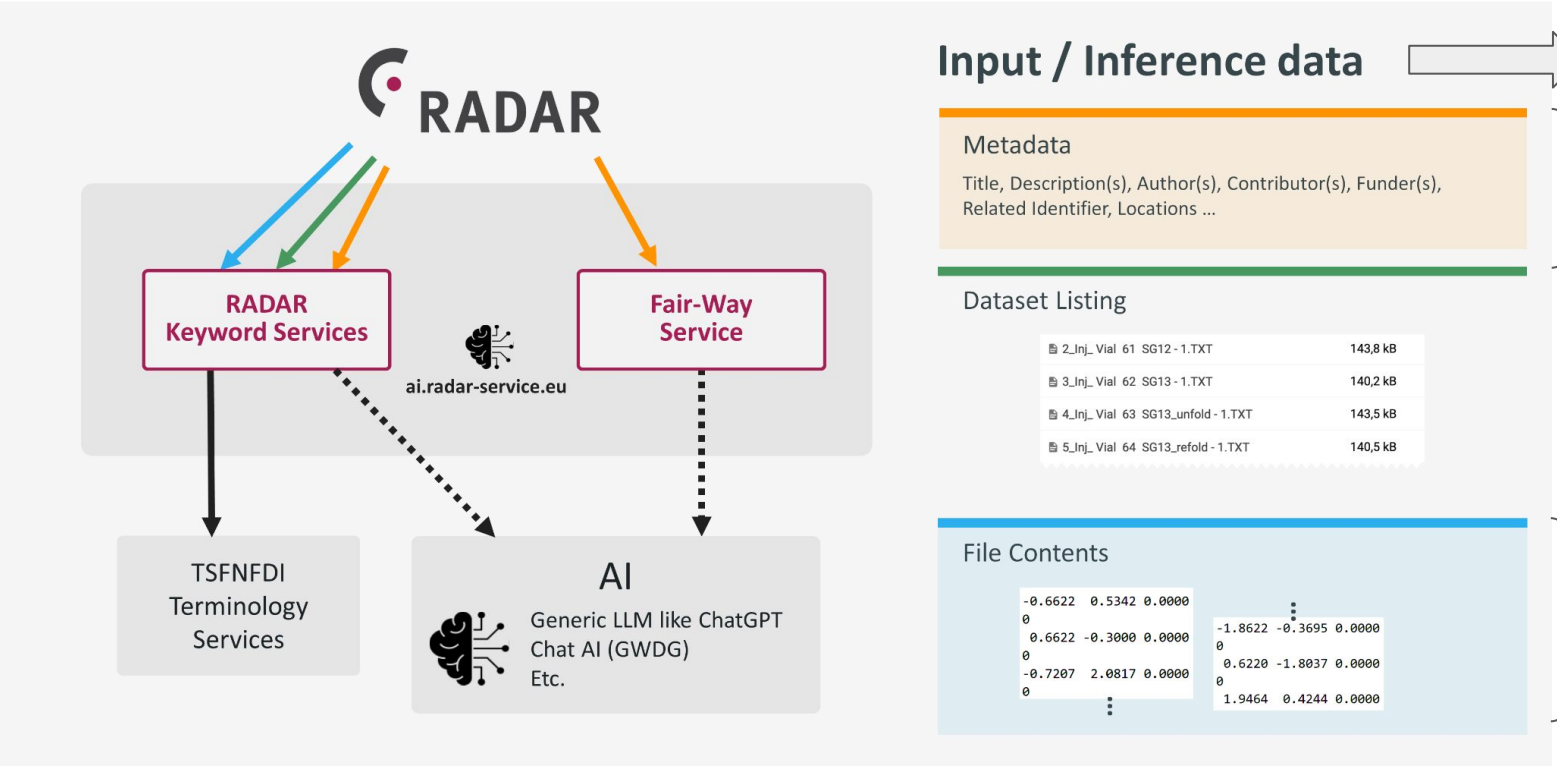
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

RADAR adaptions relevant for the use case “Publication of large Datasets”

- Integration of terminology services of Base4NFDI/TIB Hannover (TS4NFDI)
  - terminologies and ontologies for earth sciences
  - usable in keywords (type ahead, suggestion list pops up)
  - integrated widgets in metadata annotation
  - future work: exploit data file contents (e.g. NetCDF files contain reusable MD)
- Initial experimental implementation of automatic generation of metadata by LLMs/SLMs based on available text documents (project proposals, papers), title, abstract…
- gitHub and gitLab integration - data and SW import
- Optimisation of Big Data handling
  - WebDav ingest additionally to web-based ingest (failure-intolerant for long transfers)
  - Additional disc copy for random access (DL of selected files/folders in a data package)

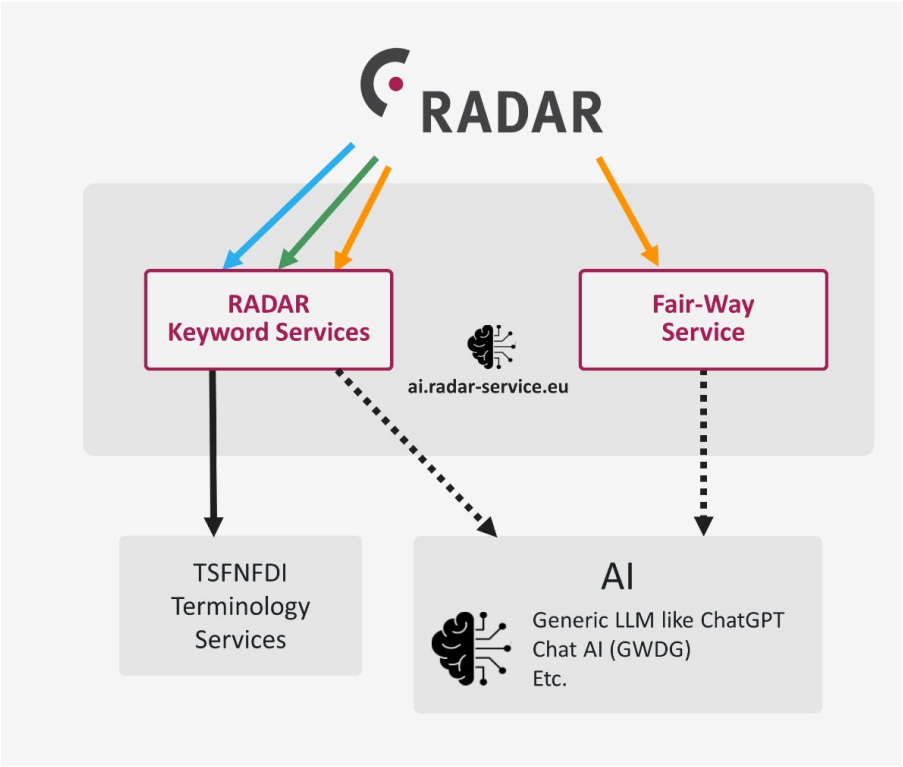
Enhancing FAIR Research Data Management with AI Support

a) Metadata enhancement



Discussion Session: *Streamlining Data Publication: Automatic Metadata and Large Datasets in the Age of AI*, Open Science Conference, 8-9 October 2025, Hamburg

Enhancing FAIR Research Data Management with AI Support
b) FAIRness enhancement



FAIRness in RADAR

data must be published with rich MD!

LLM & Prompt (e.g. ChatGPT)

- (costs up to 0,20€ / request)
- results of mixed quality
- Privacy issues

FAIR-Way Service

- Open Source
- FAIRsFAIR metric (standard)
- good results

*F-UJI best tool (very good metrics) but not the quickest

Fishbowl Discussion

Fishbowl discussion

Take notes with us:
zbw.to/osc25-pad02

Leibniz ScienceCampus
Digital Transformation of Research

Thank you for attending!

- **Symposium** on Digitalisation of Research: 2-3 December 2025, Karlsruhe
  - Interdisciplinary panels, discussions with experts!
- **Interdisciplinary Colloquium** on Digitalisation of Research
  - Monthly presentations of experts, join us online or in person!
- Stay **connected** with DiTraRe
  - **Website: www.ditrare.de/en**
  - Email: ditrare@fiz-karlsruhe.de
  - LinkedIn: www.linkedin.com/company/ditrare
  - Mastodon: sigmoid.social/@DiTraRe
  - YouTube: www.youtube.com/@DiTraRe
  - Zenodo: zenodo.org/communities/ditrare

www.ditrare.de/en

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

New Publication: CIKM'25

Leibniz ScienceCampus
Digital Transformation of Research

End-to-end Information Extraction from Archival Records with Multimodal Large Language Models

Mahsa Vafaie
mahsa.vafaie@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,
Hermann-von-Helmholtz-Platz 1
76344, Eggenstein-Leopoldshafen
Germany

Sven Hertling
sven.hertling@fiz-karlsruhe.de
Data and Web Science Group
University of Mannheim, Germany
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,
Hermann-von-Helmholtz-Platz 1
76344, Eggenstein-Leopoldshafen
Germany

Inger Banse-Strobel
inger.banse-strobel@bundesarchiv.de
Bundesarchiv, Potsdamer Str. 1
56075 Koblenz, Germany

Kevin Dubout
k.dubout@bundesarchiv.de
Bundesarchiv, Potsdamer Str. 1
56075 Koblenz, Germany

Harald Sack
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,
Hermann-von-Helmholtz-Platz 1
76344, Eggenstein-Leopoldshafen
Germany

ABSTRACT

Semi-structured Document Understanding presents a challenging research task due to the significant variations in layout, style, font, and content of documents. This complexity is further amplified when dealing with *born-analogue* historical documents, such as digitised archival records, which contain degraded print, handwritten annotations, stamps, marginalia and inconsistent formatting resulting from historical production and digitisation processes. Traditional approaches for extracting information from semi-structured documents rely on manual labour, making them costly and inefficient. This is partly due to the fact that within document collections, there are various layout types, each requiring customised optimisation to account for structural differences, which substantially increases the effort needed to achieve consistent quality. The emergence of Multimodal Large Language Models (MLLMs) has significantly advanced Document Understanding by enabling flexible, prompt-based understanding of document images, needless of OCR outputs or layout encodings. Moreover, the encoder-decoder architectures have overcome the limitations of encoder-only models, such as reliance on annotated datasets and fixed input lengths. However, there still remains a gap in effectively applying these models in real-world scenarios. To address this gap, we first introduce BZKOpen, a new annotated dataset designed for key information extraction from historical German index cards. Furthermore, we systematically assess the capabilities of several state-of-the-art MLLMs—including the open-source InternVL2.0 and InternVL2.5 series, and the commercial GPT-4o-mini— on the task of extracting key information from these archival documents. Both zero-shot and few-shot prompting strategies are evaluated across different model configurations to identify the optimal conditions for performance. Interestingly, our results reveal that increasing model size does not necessarily lead to better performance on this dataset. Among all models tested, the open-source InternVL2.5-38B consistently achieves the most robust results, outperforming both larger InternVL models and the proprietary alternative. We further provide practical insights into prompt engineering and inference settings, offering guidance for applying MLLMs to real-world key information extraction tasks. Additionally, we highlight the need for more ground truth datasets that include a wider range of historical documents with varying quality and in multiple languages, in order to fully explore the potentials and limitations of MLLMs for key information extraction from historical records.

CIKM '25, November 10–14, 2025, Seoul, South Korea.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Information systems → Information extraction.**

KEYWORDS

Multimodal Large Language Models, Document Understanding, Key Information Extraction, Digital Cultural Heritage

ACM Reference Format:
Mahsa Vafaie, Sven Hertling, Inger Banse-Strobel, Kevin Dubout, and Harald Sack. 2025. End-to-end Information Extraction from Archival Records with Multimodal Large Language Models. In *Proceedings of The 34th ACM International Conference on Information and Knowledge Management (CIKM '25).* ACM, New York, NY, USA , 9 pages. https://doi.org/XXXXXXX.XXXXXXX