



# FAIR Use of GPT-Generated Data in SSH Research: A Practical Guide

Dr. Maximilian Hindermann (RISE & UB Basel), 21.08.2024

Männliche Angestellte. Dienststelle: EIDGENÖSSISCHES JUSTIZ- UND POLIZEIDEPARTEMENT  
POLIZIASTEILUNG  
ZENTRALELEITUNG DER ARBEITSLAGEN  
ZÜRICH 2, Seewegstrasse 11

Name und Vorname: Baumgartner Gottfried

Geburtsdatum: 8.8.1900 Erlerner Beruf: Kaufmann Muttersprache: Deutsch

Heimatort: Basel und Sirmach/Thurgau Militärischer Grad: Soldat

Letzter Eintritt in den Bundesdienst:	Frühere Tätigkeit im Bundesdienste:
Dienststelle	vom bis
1. Mai 1944	

\*) Dienstjahre zählen ab:  
\*) Spareinleger seit:  
\*) Versichert seit:  
\*) Versicherungsjahre:  
\* wird vom Personalamt ausgefüllt

Unterbrechungen des Dienstverhältnisses:

Bildungsgang und Tätigkeit vor dem Eintritt in den Bundesdienst:  
4 Jahre Primar-, 4 Jahre Untere Realschule, 2 Jahre Kantonale Handelsschule Basel, 3-jährige kaufmännische Lehre, 7 Jahre Angestellter und Fremdsprachkorrespondent in Basel und in Italien, seit März 1927 Vertreter ausländischer Fabrikationsfirmen.

Bemerkungen: Zivilstand: verheiratet  
Kinder unter 18 Jahren: -- Austritt 31.12.45  
zivilrechtl. Wohnsitz: Basel  
Milit. Einteilung: Pol.Sdt.HD

```
"content": {
  "Bild1": {
    "Dateiname": "20230927_114325.jpg",
    "Name und Vorname": "Baumgartner Gottfried",
    "Geburtsdatum": "1900-08-08",
    "Heimatort": "Basel und Sirmach/Thurgau",
    "Erlerner Beruf": "Kaufmann",
    "Muttersprache": "Deutsch",
    "Militärischer Grad": "Soldat",
    "Letzter Eintritt in den Bundesdienst": "1944-05-01",
    "Frühere Tätigkeit im Bundesdienst": "",
    "Bildungsgang und Tätigkeit vor dem Eintritt in den Bundesdi
      Primar-, 4 Jahre Untere Realschule, 2 Jahre Kantonale Hand
      3-jährige kaufmännische Lehre, 7 Jahre Angestellter und
      Fremdsprachenkorrespondent in Basel und in Italien, seit Mi
      Vertreter ausländischer Fabrikationsfirmen.",
    "Bemerkungen": "Zivilstand: verheiratet\nKinder unter 18 Jah
      -\nzivilrechtl. Wohnsitz: Basel\nMilit. Einteilung: Pol.Sd
      31.12.45"
  },
}
```

# Who is Research and Infrastructure Support (RISE)?



“We support researchers in the humanities and social sciences at the University of Basel in the conception of computer-based research, the creation, analysis and user-oriented presentation of digital data, as well as in sustainable and open methods of data dissemination.”

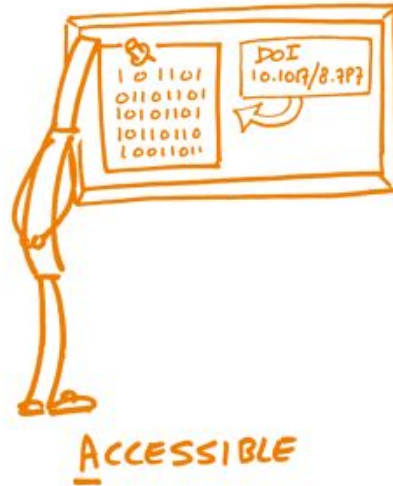
# Overview

How to make GPT-generated data FAIR in the SSH?

1. What's FAIR data?
2. What's GPT-generated data, exactly?
3. Is this kind of data any different from the kind of data we already know?

# FAIR data

## FAIR DATA PRINCIPLES



# Examples of GPT-generated data in SSH

- Extracted structured information from images (e.g., digitized historical sources, social media) or texts (e.g., research articles, transcriptions of historical sources)
- Unstructured information such as summaries (e.g., of research articles, historical sources)
- Queries (e.g., for data retrieval from an SQL database), scripts or code snippets

# Specific example: Extracting structured information from images using a multimodal LLM such as ChatGPT-4o

- Dissertation project by Tabea Wullschleger (University of Basel)
- Leadership staff in Swiss camps and homes for refugees from 1940 to 1949
- More than 1100 persons identified, source material contains around 4000 digitized personnel cards
- Information extraction from personnel cards with ChatGPT-4o via OpenAI-API
- Advantages over “classical” approach with Transkribus field models (not to mention manual extraction): cheaper, better, faster (, stronger) 🙌



# Männliche Angestellte.

EIDGENÖSSISCHES JUSTIZ- UND POLIZEIDEPARTEMENT  
POLIZEIABTEILUNG  
ZENTRALEITUNG DER ARBEITSLÄGER  
ZÜRICH 2, Beecherstrasse 11  
PA. 1

Dienststelle: .....

Name und Vorname: Baumgartner Gottfried

Geburtsdatum: 8.8.1900      Erlerner Beruf: Kaufmann      Muttersprache: Deutsch

Heimatort: Basel und Sirmach/Thurgau      Militärischer Grad: Soldat

Letzter Eintritt in }  
den Bundesdienst: }

1. Mai 1944

Frühere Tätigkeit im Bundesdienste:  
Dienststelle                      vom                      bis

\*) Dienstjahre zählen ab: .....

e) Spareinleger seit: .....

\*) Versichert seit: .....

\*) Versicherungsjahre: .....

\* wird vom Personalamt ausgefüllt

\* Unterbrechungen des Dienstverhältnisses:  
.....  
.....

## Bildungsgang und Tätigkeit vor dem Eintritt in den Bundesdienst:

4 Jahre Primar-, 4 Jahre Untere Realschule, 2 Jahre Kantonale Handels-  
schule Basel, 3-jährige kaufmännische Lehre, 7 Jahre Angestellter und  
Fremdsprachkorrespondent in Basel und in Italien, seit März 1927 Vertreter  
ausländischer Fabrikationsfirmen.

Bemerkungen:

Zivilstand: verheiratet

Kinder unter 18 Jahren: --

Austritt 31.12.45

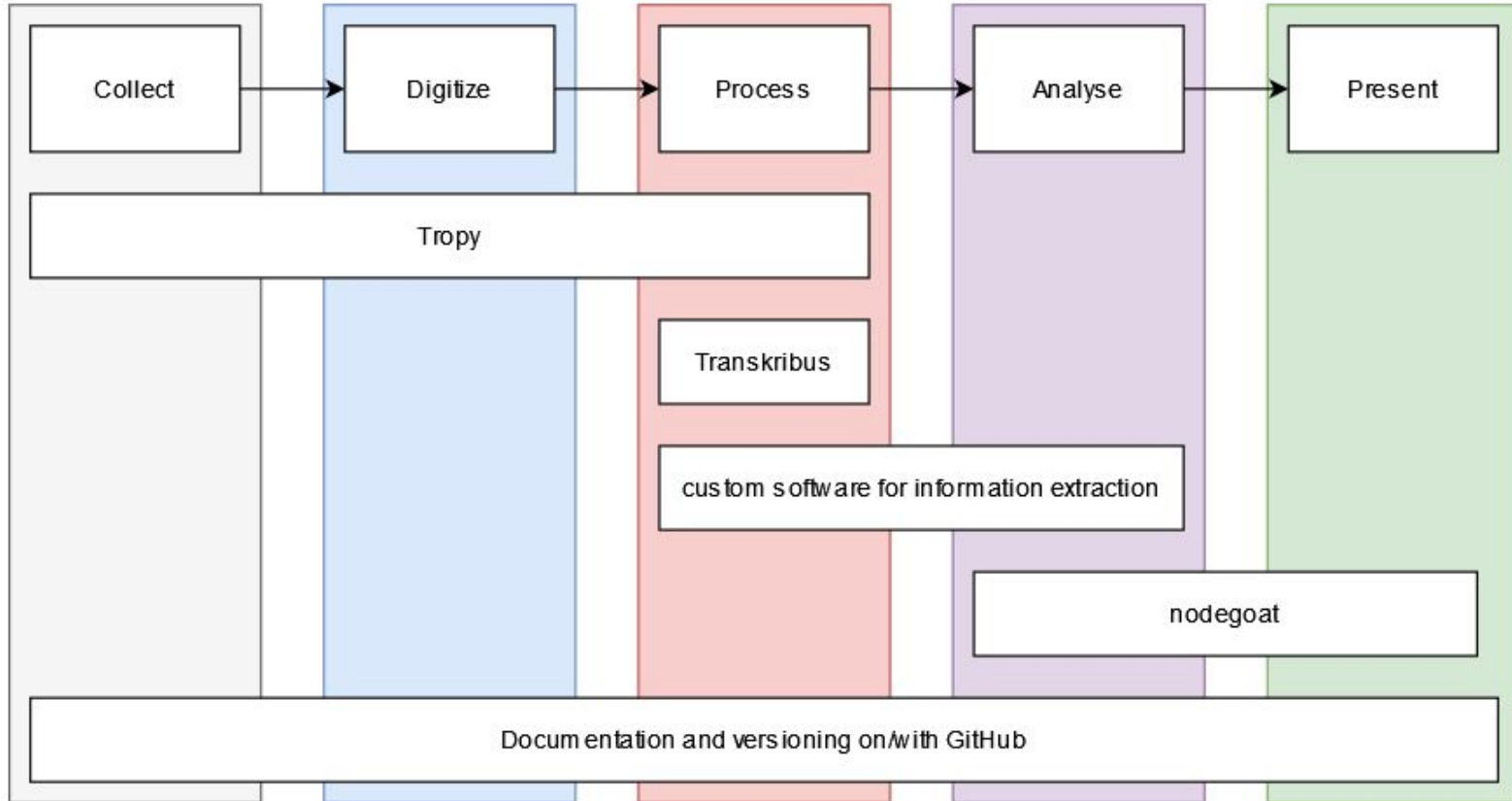
zivilrechtl. Wohnsitz: Basel

Milit. Einteilung: Pol.Sdt.HD

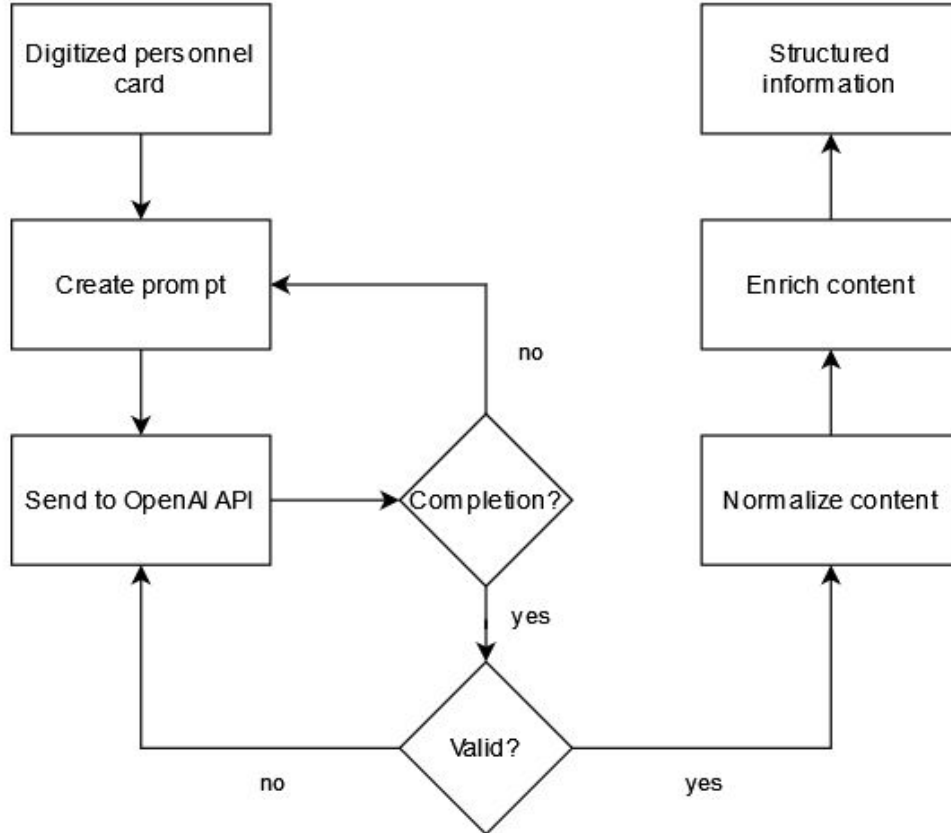




# Idealized data processing pipeline



# Information extraction workflow



# Prompt

Encoded images plus the following text:

“Please provide the information from the images as completely as possible in a JSON file with the structure Image1, Image2. All date entries should be in ISO format (YYYY-MM-DD). Do not use OCR. If the information is redundant, keep it anyway. The first image is named {path\_image1}. For the first image, use only the following fields: 'Dateiname', 'Name und Vorname', 'Geburtsdatum', 'Heimatort', 'Erlerner Beruf', 'Muttersprache', 'Militärischer Grad', 'Letzter Eintritt in den Bundesdienst', 'Frühere Tätigkeit im Bundesdienst', 'Bildungsgang und Tätigkeit vor dem Eintritt in den Bundesdienst', 'Bemerkungen'. The second image is named {path\_image2} and represents a table. For the second image, create an entry for each column containing only the following fields: 'Dateiname', 'Dienstliche Stellung', 'Dienstort', 'Gehaltsklasse', 'Jahresgehalt/Monatsgehalt/Taglohn', 'Zeitpunkt der Gehaltsänderung', 'Bemerkungen, Akten etc.’”

# Completion

```
{
  "id": "chatcpl-9TZVJXMRJj7GoQpdthNCs9IZloZ8o",
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "logprobs": null,
      "message": {
        "content": ""
      }
    }
  ],
  "content": "",
  "json": {
    "Bild1": {
      "Dateiname": "20230927_114325.jpg",
      "Name und Vorname": "Baumgartner Gottfried",
      "Geburtsdatum": "1900-08-08",
      "Heimort": "Basel und Sirmach/Thurgau",
      "Erlernter Beruf": "Kaufmann",
      "Muttersprache": "Deutsch",
      "Milit\u00e4rischer Grad": "Soldat",
      "Letzter Eintritt in den Bundesdienst": "1944-05-01",
      "Fr\u00fchere T\u00e4tigkeit im Bundesdienst": "",
      "Bildungsgang und T\u00e4tigkeit vor dem Eintritt in den Bundesdienst": "4 Jahre Primar-, 4 Jahre Untere Realschule, 2 Jahre Kantonale Handelsschule Basel, 3-j\u00e4hrige kaufm\u00e4nnische Lehre, 7 Jahre Angestellter und Fremdsprachenkorrespondent in Basel und in Italien, seit M\u00e4rz 1927 Vertreter ausl\u00e4ndischer Fabrikationsfirmen.",
      "Bemerkungen": "Zivilstand: verheiratet\nKinder unter 18 Jahren: \n2014\n\nzivilrechtl. Wohnsitz: Basel\n\nMilit. Einteilung: Pol.Sdt.HD\n\nAustritt 31.12.45"
    },
    "Bild2": {
      "Eintrag1": {
        "Dateiname": "20230927_114332.jpg",
        "Dienstliche Stellung": "Hilfleiter",
        "Dienstort": "",
        "Gehaltsklasse": "",
        "Jahresgehalt/Monatsgehalt/Taglohn": "3320.\u00a0\u00a02013",
        "Zeitpunkt der Gehalts\u00e4nderung": "1944-05-01",
        "Bemerkungen, Akten etc": ""
      },
      "Eintrag2": {
        "Dateiname": "20230927_114332.jpg",
        "Dienstliche Stellung": "Hilfleiter",
        "Dienstort": "",
        "Gehaltsklasse": "",
        "Jahresgehalt/Monatsgehalt/Taglohn": "4630.\u00a0\u00a02013",
        "Zeitpunkt der Gehalts\u00e4nderung": "1945-01-01",
        "Bemerkungen, Akten etc": "+ Zulagen"
      }
    }
  ],
  "role": "assistant",
  "function_call": null,
  "tool_calls": null,
  "created": 1716835017,
  "model": "gpt-4o-2024-05-13",
  "object": "chat.completion",
  "system_fingerprint": "fp_3196d36131",
  "usage": {
    "completion_tokens": 487,
    "prompt_tokens": 2469,
    "total_tokens": 2956
  }
}
```



# Completion metadata

```
1 {
2   "id": "chatcmpl-9TWh0NzvhIiJmA0wWXMFLWleFMBZ",
3   "choices": [
4     {
5       "finish_reason": "stop",
6       "index": 0,
7       "logprobs": null,
8       "message": {
9         "content": {"text": ""},
10        "role": "assistant",
11        "function_call": null,
12        "tool_calls": null
13      }
14    }
15  ],
16  "created": 1716824210,
17  "model": "gpt-4o-2024-05-13",
18  "object": "chat.completion",
19  "system_fingerprint": "fp_927397958d",
20  "usage": {
21    "completion_tokens": 477,
22    "prompt_tokens": 2469,
23    "total_tokens": 2946
24  }
25 }
```

# Completion data

```
9  "content": {
10  "Bild1": {
11      "Dateiname": "20230927_114325.jpg",
12      "Name und Vorname": "Baumgartner Gottfried",
13      "Geburtsdatum": "1900-08-08",
14      "Heimatort": "Basel und Sirnach/Thurgau",
15      "Erlernter Beruf": "Kaufmann",
16      "Muttersprache": "Deutsch",
17      "Militärischer Grad": "Soldat",
18      "Letzter Eintritt in den Bundesdienst": "1944-05-01",
19      "Frühere Tätigkeit im Bundesdienst": "",
20      "Bildungsgang und Tätigkeit vor dem Eintritt in den Bundesdienst": "4 Jahre
    Primar-, 4 Jahre Untere Realschule, 2 Jahre Kantonale Handelsschule Basel,
    3-jährige kaufmännische Lehre, 7 Jahre Angestellter und
    Fremdsprachenkorrespondent in Basel und in Italien, seit März 1927
    Vertreter ausländischer Fabrikationsfirmen.",
21      "Bemerkungen": "Zivilstand: verheiratet\nKinder unter 18 Jahren:
    -\n\nzivilrechtl. Wohnsitz: Basel\nMilit. Einteilung: Pol.Sdt.HD\nAustritt
    31.12.45"
22  },
23  "Bild2": [ ]
24  },
```

# Conclusion

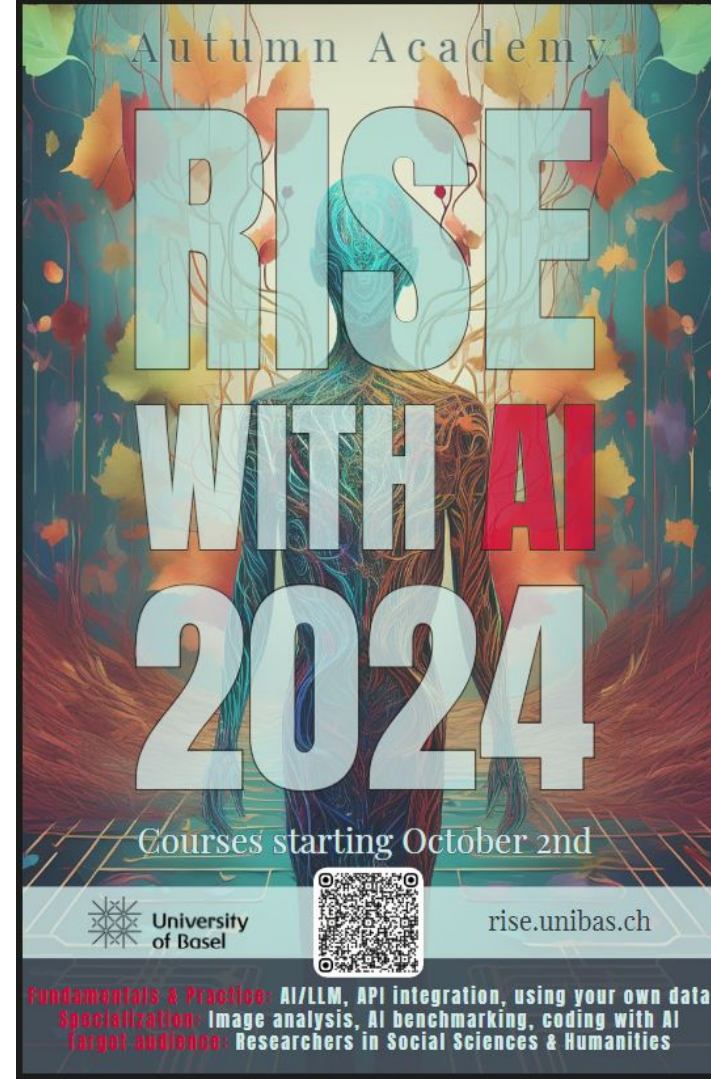
In order to be FAIR, GPT-generated data in SSH must

1. include prompt, completion content and completion metadata, and
2. make available the pipelines for data generation and processing as FAIR research software (in the sense of [Barker et al. 2022](#)).

Note that these are necessary and not sufficient conditions!

# What RISE is working on right now


- Benchmarking of LLM models
- Generic post processing infrastructure
- Packages for AI-wrappers and pipelines
- [Crash course series about AI in SSH](#)


A vertical poster for 'Autumn Academy RISE WITH AI 2024'. The background features a stylized figure with a glowing blue brain and a network of neural connections, set against a backdrop of autumn leaves in shades of orange, yellow, and red. The text 'Autumn Academy' is at the top in a light blue font. Below it, 'RISE' is written in large, light blue, outlined letters. 'WITH AI' is in white and red, and '2024' is in large, light blue, outlined letters. At the bottom, it says 'Courses starting October 2nd', the University of Basel logo and name, a QR code, and the website 'rise.unibas.ch'. A red banner at the very bottom contains course details.

Autumn Academy

# RISE WITH AI 2024

Courses starting October 2nd

 University of Basel



rise.unibas.ch

**Fundamentals & Practice:** AI/LLM, API integration, using your own data  
**Specialization:** Image analysis, AI benchmarking, coding with AI  
**Target audience:** Researchers in Social Sciences & Humanities



# Q&A

Email: [maximilian.hindermann@unibas.ch](mailto:maximilian.hindermann@unibas.ch) or [rise@unibas.ch](mailto:rise@unibas.ch)

Web: <https://rise.unibas.ch/>

GitHub: <https://github.com/RISE-UNIBAS>