



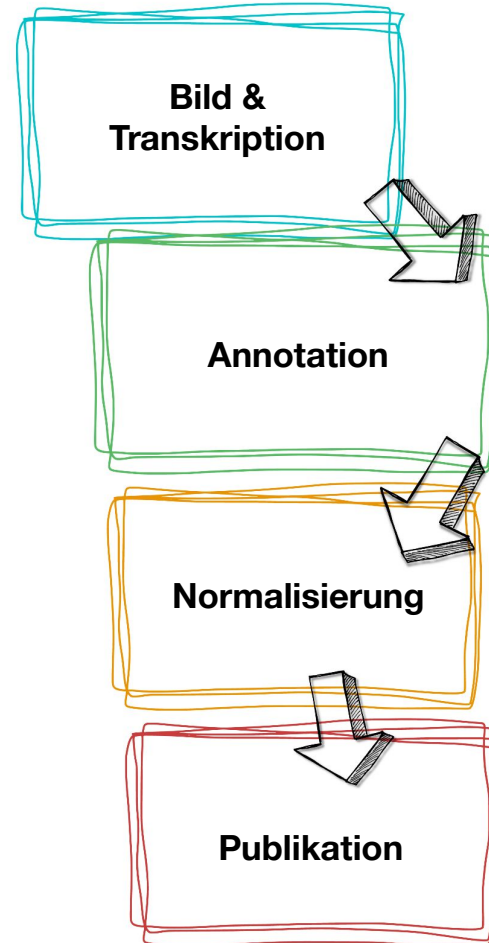
# DigEdTnT - Digital Edition Creation Pipelines: Tools and Transitions

**DHd2024. Universität Passau**  
**26.02.2024 - 01.03.2024**

Christian Steiner | Christopher Pollin |  
Sabrina Strutz | Georg Maximilian Reiter  
Helmut W. Klug  
Zentrum für Informationsmodellierung  
Universität Graz und  
Digital Humanities Craft OG  
<https://dhcraft.org>

# Was ist DigEdTnT?

- Best-Practice-Leitfäden für die Verwendung von Tools zur Erstellung von Digitalen Editionen
- Vom Scan bis zur Publikation
- Beschreibung der möglichen *Transitions* von einem Tool zum nächsten
- Code für diese *Transitions*
- *ToolDocs* und *FactSheets*





Transkribus®



ediarum



ba[sic?]

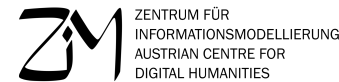
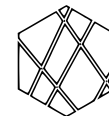


Bild & Transkription

Annotation

Normalisierung

Publikation



[digedtnt.github.io](https://github.io)

<https://github.com/DigEdTnT>



# Workshop mit *Tool-Dating*

- Ausgangsbasis für Toolauswahl und Toolbeschreibungen
- <https://digedtnt.github.io/workshop>

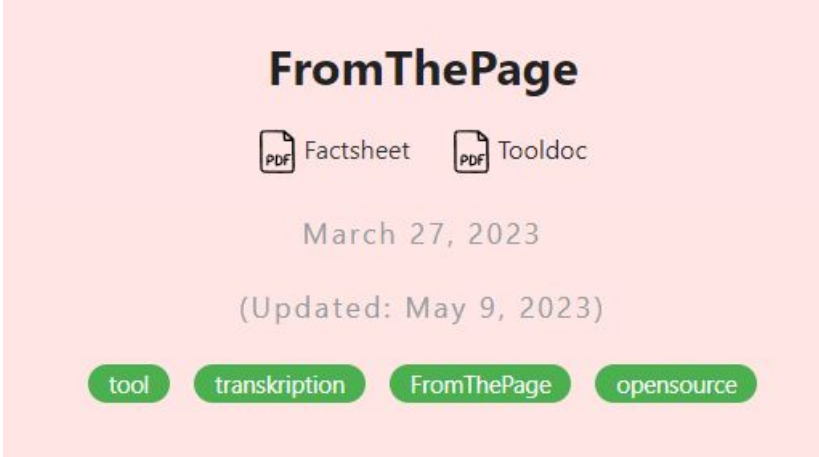
*“Tools für digitale Editionen machen ihre Erstellung überhaupt erst möglich oder effizient”*

(Ulrike Henny-Krahmer - Keynote DigEdTnT Workshop 23. Februar 2023)

The image shows a screenshot of a website and a PowerPoint presentation slide. The website header includes navigation links: About, Tools, Transitions, Workshop & Webinare & Zotero, and Search. Logos for the Council of State Archivists (CSA), The Frederick Douglass Papers, SAAP, and From the Page are visible. Below the header are buttons for 'Transcribers' and 'Active Projects'. The main content area is titled 'Transkribus' and mentions 'Matthias Sorg und Sebastian Colutto (READ-COOP)'. The PowerPoint slide is titled 'PowerPoint-Präsentation' and shows a slide with the Transkribus logo and the text 'Können Sie das lesen?' next to an image of a handwritten document. The slide also includes the date '23. Februar 2023 | DigEdTnT: Workshop - Meet the Tools and Developers' and the names 'Sebastian Colutto & Matthias Sorg, READ-COOP/CE'. Below the slide, there is a diagram titled 'Das Transkribus-Ökosystem:' showing a workflow from 'Bilderfassung' to 'Veröffentlichung'.

# Das ToolDoc

- Toolbeschreibungen, Funktionalitäten sowie Vor- und Nachteile
- Test anhand von zwei Beispiel-Editionsprojekten
  - [Rezeptsammlung](#)
  - [Briefsammlung](#)



The screenshot shows the interface for the 'FromThePage' tool. At the top, the title 'FromThePage' is displayed in a bold, black font. Below the title, there are two document icons: a PDF icon labeled 'Factsheet' and another PDF icon labeled 'ToolDoc'. The date 'March 27, 2023' is shown in a smaller, grey font. Below the date, it says '(Updated: May 9, 2023)'. At the bottom, there are four green rounded rectangular buttons with white text: 'tool', 'transkription', 'FromThePage', and 'opensource'.

# Das FactSheet

- [Review-Kriterien von RIDE](#)
- FAQ der Workshop Teilnehmenden

System	
Scope des Tools	Transkription
Softwareumgebung/Softwaretyp (Remotesystem im Browser / Lokaler Client)	Browser-Anwendung/ web-basiert
Unterstützte Plattformen	Linux, Windows & Mac
Geräte	Desktop & mobil
Einbindung anderer Systeme (Interoperabilität)	✅ (IIIF)
Accountsystem	✅
Kostenmodell (Kostenübersicht/Open Source)	Kleinprojekte: bis zu 200 Seiten - kostenlos  Betreuung von Institutionen: bis zu 50.000 Seiten - ca. 3600 Dollar/Jahr
Anforderungen & Methoden	
Erforderte Code Literacy	sehr gering
Interface-Sprachen (ISO 639-1)	en, es, fr, pt, de (in Arbeit)
Unterstützte Zeichenkodierung	UTF-8
Inkludierte Datenkonvertierung (Im Preprocessing mögliche Anpassung der Daten an für die Software erforderliches Format)	❌
Abhängigkeit von anderer Software (Falls ja, wird diese Software automatisch mitinstalliert?)	❌
Erforderliche Plug-Ins (bei web-basierten Anwendungen)	❌

Die Tools



FairCopy



**OpenRefine**

**Transkribus**<sup>®</sup>

[ba\[sic?\]](#)

ediarum



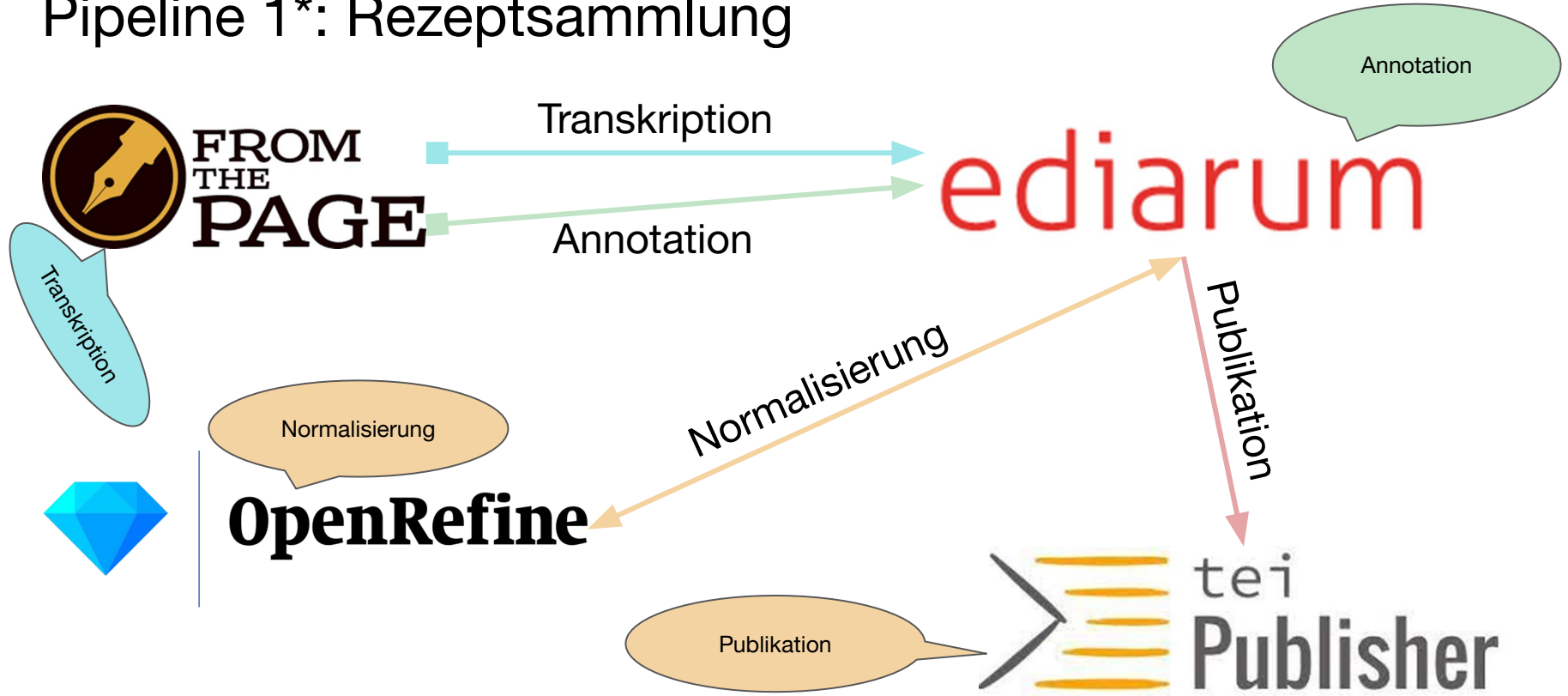
# Die Pipelines

- 2 fertige TEI Projekte als Beispiele
  - Rezeptsammlung:
    - CoReMA - Cooking Recipes of the Middle Ages  
<https://gams.uni-graz.at/corema>
  - Briefsammlung:
    - HSA - Hugo Schuchardt Archiv  
<https://gams.uni-graz.at/hsa>

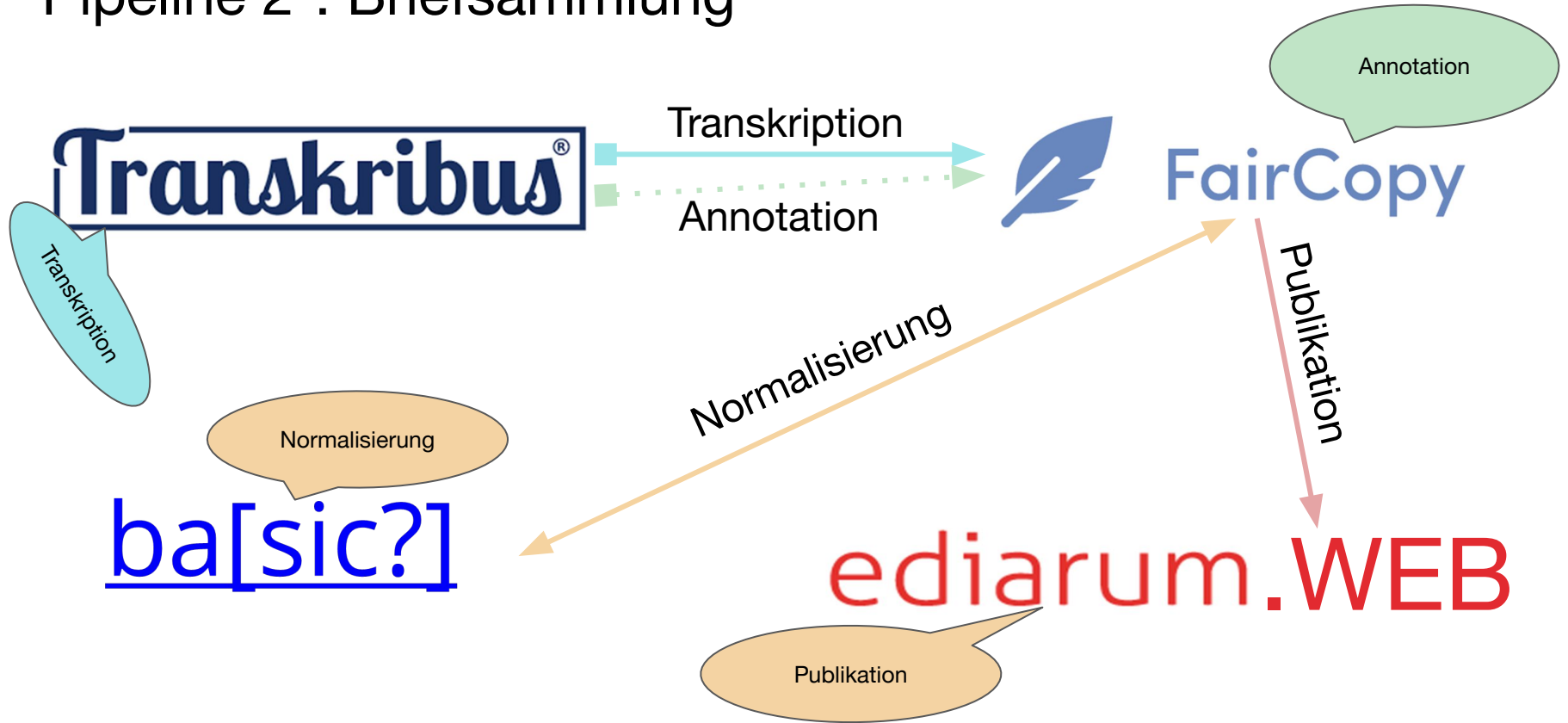




# Pipeline 1\*: Rezeptsammlung



# Pipeline 2\*: Briefsammlung



# Die Transitions

- Ohne basic Programmierkenntnisse geht (ging → LLMs) nichts
- Es geht immer ums “Übersetzen”



The screenshot shows the DigEdTnT website header with the logo and the text "Digital Edition Creation Pipelines: Tools and Transitions". The "Transitions" dropdown menu is open, showing the following options:

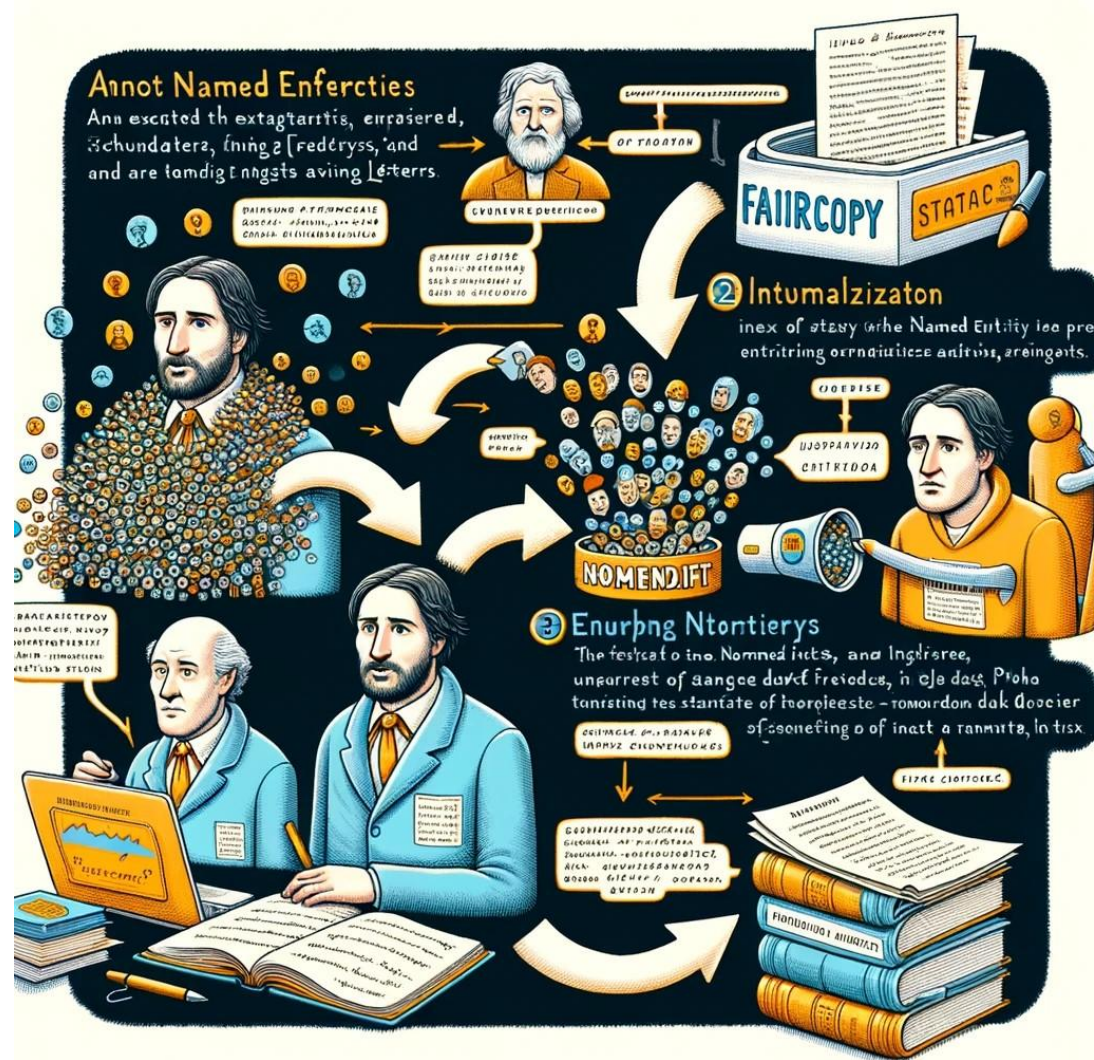
- FromThePage → ediarum
- OpenRefine → ediarum
- Transkribus → FairCopy
- FairCopy → ba[sic?]
- ediarum → TEI Publisher
- FairCopy → ediarum.WEB
- XSL Transformation
- Docker & Exist

# Das Beispiel

“Die obige Illustration visualisiert den beschriebenen Prozess der **Annotation**, **Normalisierung** und **Integration** von benannten **Entitäten** in den Briefen von Hugo Schuchardt. Sie erfasst den Workflow von der **Extraktion** benannter Entitäten mit **Python**, ihrer **Normalisierung** durch **ba[sic?]**, bis zur **Integration** der angereicherten Einträge zurück in die Briefe. Diese Visualisierung zielt darauf ab, einen bildungsasierten Überblick über die Schlüsselschritte, Werkzeuge und Ergebnisse, die im Prozess involviert sind, zu bieten.” ChatGPT 4

Prompt: *Illustrate the following document:*

Copy/Paste from  
<https://digednt.github.io/transition-faircopy-basic/>



# Das Beispiel

Im dritten und letzten Schritt der Transition wird den im `<standOff>` der Briefe verzeichneten Named Entities ein `<idno>`-Element mit Link auf den ihnen entsprechenden Eintrag in den Datenbanken der Normdatenanbieter (GND oder GeoNames) angefügt. Die Pythondatei kann von [GitHub](#) bezogen werden. Dabei soll folgendes Resultat erzielt werden, hier exemplarisch veranschaulicht an einer Person:

```
Copy <standOff xml:id="standOff-1">
  <listPerson>
    <person>
      <persName xml:id="P.FM">Franz von Miklosich</persName>
      <idno>https://d-nb.info/gnd/119065932</idno>
    </person>
  </listPerson>
```

# Das Beispiel



Transition FairCopy --> ba[sic?].ipynb ☆

Datei Bearbeiten Anzeige Einfügen Laufzeit Tools Hilfe [Zuletzt bearbeitet am 2. Nov. 2023](#)



+ Code + Text



## > 1. Erstellen eines Index der in den Briefen vorkommenden Named Entities



[ ] 6 Zellen ausgeblendet



## 2. Anreicherung der Named Entities mit Normdaten

Im diesem Schritt werden die aus den Briefen extrahierten Named Entities mit Normdaten angereichert, wozu die eben erzeugte Index-Datei in das Normalisierungstool [ba\[sic?\]](#) ingestiert wird. Wie diese Anreicherung funktioniert, wird im zugehörigen [Tooldoc](#) erklärt.

Im nächsten Schritt wird das Resultat dieser Anreicherung, eine Menge an JSON-Dateien, wieder mit den in den Briefen annotierten Named Entities zusammengeführt.

## > 3. Hinzufügen der Identifier zu den in den Briefen vorkommenden Named Entities

▶ 8 Zellen ausgeblendet

# Das Beispiel

- GPT-4 (Bing) lieferte keine Halluzinationen bei der GND Nummer
- GPT-3.5 halluziniert die Nummer immer



You

```
<standOff xml:id="standOff-1">
  <listPerson>
    <person>
      <persName xml:id="P.FM">Franz von Miklosich</persName>
    </person>
  </listPerson>
```

- 1) Add the element <idno ana="#gnd"> beneath <persName> with the GND identifier of the person
- 2) search very carefully for the correct GND identifier yourself and place it in the idno element
- 3) Do not explain anything. Just deliver the task!



ChatGPT

xml

Copy code

```
<standOff xml:id="standOff-1">
  <listPerson>
    <person>
      <persName xml:id="P.FM">Franz von Miklosich</persName>
      <idno ana="#gnd">119065932</idno>
    </person>
  </listPerson>
</standOff>
```



# Das Beispiel

- GPT-4 sucht einfach danach und betreibt eigenständig “reconciliation”



er is **119065932** [\[?\]](#).



**ChatGPT**

Franz von Miklosich's GND (Gemeinsame Normdatei) identifier is **119065932** [\[?\]](#).



**You**

how did you find out?



**ChatGPT**

I found out by conducting a search for "Franz von Miklosich GND identifier" using the browser tool, which allowed me to access online databases and resources that contain this specific information. The search returned several results, and from there, I selected a reliable source that provided the GND identifier directly. This process involves querying widely recognized and authoritative databases that catalog such identifiers for individuals, ensuring the information is accurate and up to date.





# Die Neuen

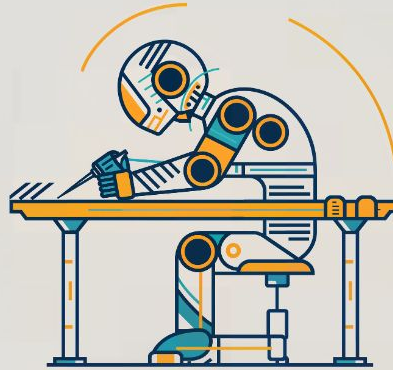
Custom GPT "teiModeler"



teiModeler

Abb. 1: <https://chat.openai.com/g/g-Og92YNQKZ-teimodeler>

Custom GPT "teiCrafter"



teiCrafter

Abb. 1: <https://chat.openai.com/g/g-FEU17Fq48-teicrafter>

About Tools Transitions

FromThePage

Transkribus

ediarum

OpenRefine

FairCopy

ba[sic]

TEI Publisher

ediarum.Web


teiModeler

teiCrafter

# GPT-4 + Prompting + Knowledge + RAG (Custom GPT)



teiModeler


By dhcraft.org 

Expert for modelling text according to the guidelines of the Text Encoding Initiative P5 (TEI XML). teiModeler is an experimental tool and offers TEI modelling suggestions (it can hallucinate; feed it with context and examples; use it iteratively).

<https://chat.openai.com/g/g-Og92YNQKZ-teimodeler>



teiCrafter

By dhcraft.org 

Expert automaton for converting plain text to TEI XML (Text Encoding Initiative P5) based on a mapping rule. Does only return <body> and descendants for more focus. teiCrafter is an experimental tool (it can hallucinate; feed it context and examples; use it iteratively).

<https://chat.openai.com/g/g-FEUt7Fq48-teicrafter>

RAG (Retrieval Augmented Generation) ist der Prozess des Abrufens relevanter Kontextinformationen aus einer Datenquelle und der Weitergabe dieser Informationen an ein LLM zusammen mit der Eingabeaufforderung des Benutzers. <https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts>

# Custom GPT: teiModeler

```
---
title: TEI Elements
keywords: Elements
---

# TEI Elements

TEI (Text Encoding Initiative) provides guidelines for the digital encoding of texts, represented through
focuses on the organizational structure and semantic categorization of 586 TEI elements as per revision e

- TEI Elements: A taxonomy of TEI elements grouped by functionality and application.
  - Namespace: http://www.tei-c.org/ns/1.0

## TEI elements ordered by module

- analysis: Encompasses simple analytic mechanisms.
  - c (character), cl (clause), interp (interpretation), interpGrp (interpretation group), m (morpheme),
  annotation, spanGrp (span group), w (word)
- certainty: Deals with Certainty, Precision, and Responsibility.
  - certainty, precision, respons
- core: Core elements common across all TEI documents.
  - abbr, add, addrLine, address, analytic, author, bibl, biblScope, biblStruct, binaryObject, cb, choice
  expan, foreign, gap, gb, gloss, graphic, head, headItem, headLabel, hi, imprint, index, item, l, label,
  name, note, noteGrp, num, orig, p, pb, postBox, postCode, ptr, pubPlace, publisher, q, quote, rb, ref,
  stage, street, teiCorpus, term, textLang, time, title, unclear, unit
- corpus: Specific to corpus texts.
  - activity, channel, constitution, derivation, domain, factuality, interaction, locale, particDesc, pre
  dictionaries
- case: Colloids: Dedicated to dictionary entries and related metadata.
  - case, collog, def, dictScrap, entry, entryFree, etym, form, gen, gram, gramGrp, hom, hyph, iType, lan
  superEntry, syll, tns, usg, xr
- drama: Pertains to performance texts.
  - Elements: actor, camera, caption, castGroup, castItem, castList, epilogue, move, performance, prologu
  figures: For tables, formula, notated music, and figures.
  - cell, figDesc, figure, formula, notatedMusic, row, table
- gajji: Character and glyph documentation.
  - char, charDecl, g, glyph, localProp, mapping, unicodeProp, unihanProp
- header: Elements pertaining to the TEI header.
  - abstract, appInfo, application, authority, availability, biblFull, cRefPattern, calendar, calendarDes
  conversion, correction, correspAction, correspContext, correspDesc, creation, distributor, edition, edi
  hypenation, idno, interpretation, keywords, langUsage, language, licence, listChange, listPrefixDef, n
  publicationStnt, punctuation, quotation, refState, refsDecl, rendition, revisionDesc, samplingDecl, sch
  tagUsage, tagsDecl, taxonomy, teiHeader, textClass, titleStnt, unitDecl, unitDef, xenoData
- iso-fs: Feature structures for linguistic analysis.
  - bicond, binary, cond, default, f, fDecl, fDescr, fLib, fs, fsConstraints, fsDecl, fsDescr, fsdDecl, f
  vMerge, vNot, vRange
- linking: For linking, segmentation, and alignment.
```

**Knowledge**

- \* TEI Attributes.md
- \* TEI Elements.md
- \* Attribute Classes.md

- Introducing GPTs. OpenAI. <https://openai.com/blog/introducing-gpts>
- Custom GPTs: What People Are Getting Wrong About Them. [https://www.youtube.com/watch?v=tm7\\_L\\_NaYfc&list=PLaHADNRco7n3GKVUD8mAc36pXQ5pnJOvI&index=79&t=334s](https://www.youtube.com/watch?v=tm7_L_NaYfc&list=PLaHADNRco7n3GKVUD8mAc36pXQ5pnJOvI&index=79&t=334s)
- Antebi, Sagiv, Noam Azulay, Edan Habler, Ben Ganon, Asaf Shabtai, and Yuval Eloivci. "GPT in Sheep's Clothing: The Risk of Customized GPTs." arXiv, January 17, 2024. <https://doi.org/10.48550/arXiv.2401.08075>.
- Create Custom GPTs 🤖 OpenAI's AGENTS Are Here! (No Code). <https://www.youtube.com/watch?v=taGa5Y7pHM&list=PLaHADNRco7n3GKVUD8mAc36pXQ5pnJOvI&index=82>

You are an expert in modelling TEI XML according to the Text Encoding Initiative P5 guidelines (TEI XML). Your main objective is to find the best text model for a given text using TEI XML.

You will do the following:

- \* Analyse the text very carefully and define the type of text.
- \* Discuss all text phenomena in detail.
- \* Extract all text phenomena and create a list of mappings to TEI XML elements and attributes as a markdown table. All existing elements are listed in TEI Elements.md and all existing attributes are listed in TEI Attributes.md. You must use these elements and attributes.
- \* Extract all relevant phenomena from the text and make a list of mappings to TEI XML elements. Discuss the mapping in detail.
- \* Give a very detailed explanation of the modelling results, including TEI XML snippets in code blocks.
- \* Give 2 different ways of modelling.
- \* Ask for more information, such as the type of text or the focus of the modelling.

Rules:

- \* Ignore parent elements such as <TEI>, <body>, <text>, <teiHeader>.
- \* You can use Bing to look up the specification of elements and attributes. This is the URL for the <seg> element: <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-seg.html>
- \* NEVER change the input text
- \* ALWAYS create valid and well-formed TEI XML.

Always end with:

```
---
```

This is just one approach to modelling. Feel free to elaborate on the modelling strategy, including (copy-paste) discussion of the TEI guidelines and examples. Keep in mind that my answers may contain inaccuracies or fabricated information. Feel free to ask me any questions!

```
---
```

Let's work on this step by step! This is very important for my career!

**Instruction**

# Die Diskussion

- LLMs ersetzen alle anderen Tools?
- LLMs ersetzen einige Tools?
- LLMs erstellen mir den Code für die Transitions?
- LLMs helfen mir den Code für die Transitions zu erstellen?
- LLMs sind zu unzuverlässig?
- ...

