Can Jupyter Help Daidalos?

Or:
How to Develop Digital Literacies
and Assess Them?

EUNIS, Athens, 05/06/2024

Dr. Andrea Beyer
Humboldt-Universität zu Berlin

Why Call a Project "Daidalos"?

We …

- develop an NLP **infrastructure**

- that will enable researchers in **Classical Philology** and related disciplines

- to apply various **methods** of natural language processing

- which are **uncommon** in the German speaking philological community.

I was the most famous **inventor, craftsman, and builder** in **antiquity** – forget my human failures.

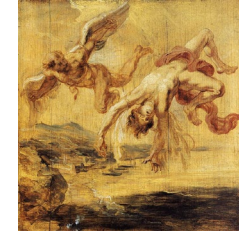# Why Call a Project "Daidalos"?

We …

– develop an NLP **infrastructure**

– that will enable researchers in **Classical Philology** and related disciplines

– to apply various **methods** of natural language processing

– which are **uncommon** in the German speaking philological community.

In summary, we do research, we link methods and evaluate them, we work with Latin and ancient Greek, we use AI.

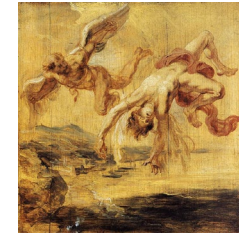I was the most famous **inventor, craftsman, and builder** in **antiquity** – forget my human failures.

Why Call a Project "Daidalos"?

We …

– develop an NLP **infrastructure**

– that will enable researchers in **Classical Philology** and related disciplines

– to apply various **methods** of natural language processing

– which are **uncommon** in the German speaking philological community.

In summary, we do research, we link methods and evaluate them, we work with Latin and ancient Greek, we use AI.

Ergo: **dAIdalos** was a conclusive choice.

I was the most famous **inventor, craftsman, and builder** in **antiquity** – forget my human failures.

Daidalos Project

– **Interdisciplinary**: Classical Philology, Linguistics, Data Centre

– **Aim**: support philological research

– **Focus**: natural language processing for literary texts

– **Target group**: people interested in Latin and (ancient) Greek corpora

– Project members: 1 researcher per discipline + 1 student assistant

– Project term: 3 years, 2023-2026

– Third-party funding: German Research Foundation

daidalos-projekt.de

Daidalos Platform

Menu: NLP
- ☑ Select: language, author, work, text passage
- ☑ Run
- ☑ Choose between NLP methods NER, POS, Sentiment-Analysis

daidalos-projekt.de

Daidalos Platform

Menu: NLP
- ☑ Select: language, author, work, text passage
- ☑ Run
- ☑ Choose between NLP methods NER, POS, Sentiment-Analysis

Soon:
- More NLP methods and adjustable settings,
- Multiple corpora,
- Pipelines for literary research questions,
- IAM: save settings, results, and visualisations,
- Different levels of user experience: no code (UI), low code (Jupyter Notebook), code

Daidalos Infrastructure is More …

Customisation through research tandems

1. First contact
2. Initial conversation: research question, specification of the text corpus and the NLP method
3. Provide research materials
4. Prototype, first research results
5. Evaluation in the tandem
6. Adjustments if necessary
7. Cooperative publication

orange = Researcher, blue = Daidalos Team

Daidalos Infrastructure is More …

Customisation through research tandems

1. First contact
2. Initial conversation: research question, specification of the text corpus and the NLP method
3. Provide research materials
4. Prototype, first research results
5. Evaluation in the tandem
6. Adjustments if necessary
7. Cooperative publication

orange = Researcher, blue = Daidalos Team

Interactive infrastructure as a **learning opportunity**

– Glossary
– **Jupyter Notebooks** for demo workflows using curated, authentic examples
– Exercises with H5P

NER steht für ...

Named Entity Recognition ✔

Namen Einheiten Regeln

Nukleotid-Exzisionsreparatur

Für das Lateinische annotiert der NER-Tagger von Daidalos ...

person, location, name

person, place, ethnicity ✖

person, location

Daidalos is **about the people**, their **research interests**, and their (digital) **research skills**.

Daidalos is **about the people**, their **research interests**, and their (digital) **research skills**.

Indeed? I never really cared about whether people understand my technologies. Anyway, we were talking different languages …

Daidalos is **about the people**, their **research interests**, and their (digital) **research skills**.

Indeed? I never really cared about whether people understand my technologies. Anyway, we were talking different languages …

In fact, that is our **greatest challenge**! How can our users develop research questions and explore them using Daidalos if their research expertise does not include the necessary digital skills?

Modelling a Digital Literacies Framework

Our research questions are:

– What kind of Digital Literacies do we address?

– How do we measure the level of Digital Literacies in order to provide useful support?

– How do we customise the intended framework regarding to specific domains and cases?

AI Literacy

Data Literacy Computational Literacy

Media Literacy Information Literacy Visual Literacy
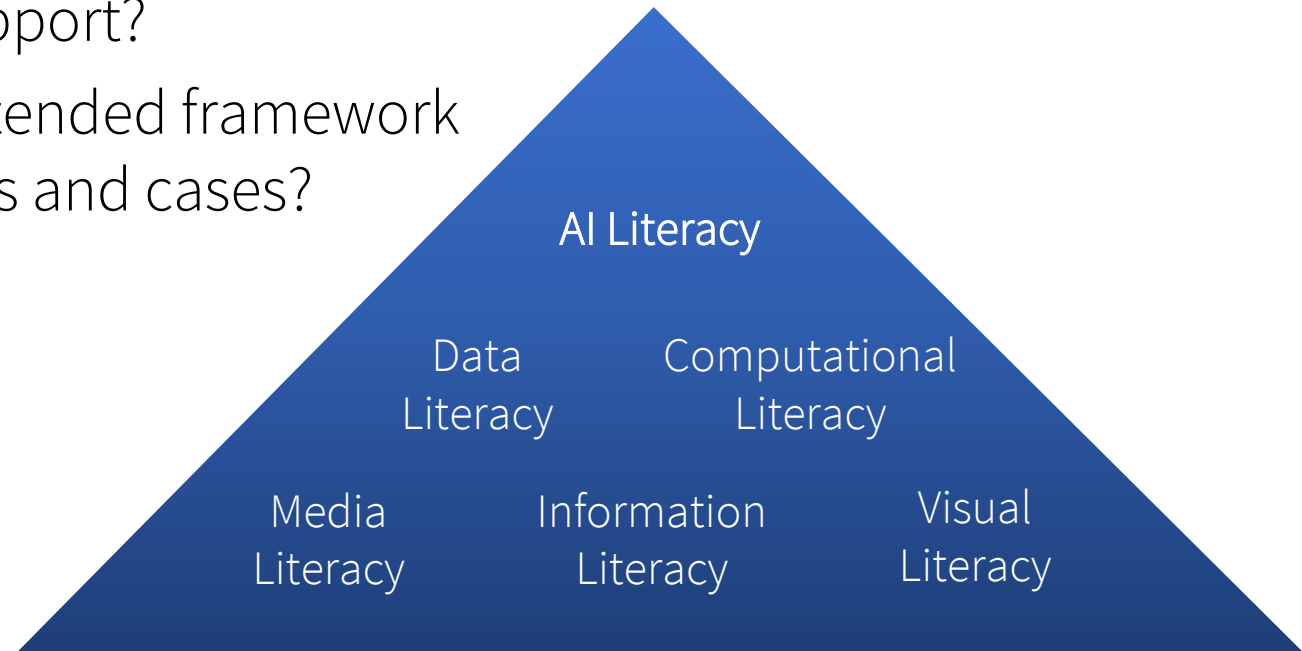
Most often cited Digital Literacies in literature over the last 40 years, starting with Media Literacy and ending with AI Literacy.

Digital Literacies

Digital Literacy

Data Literacy
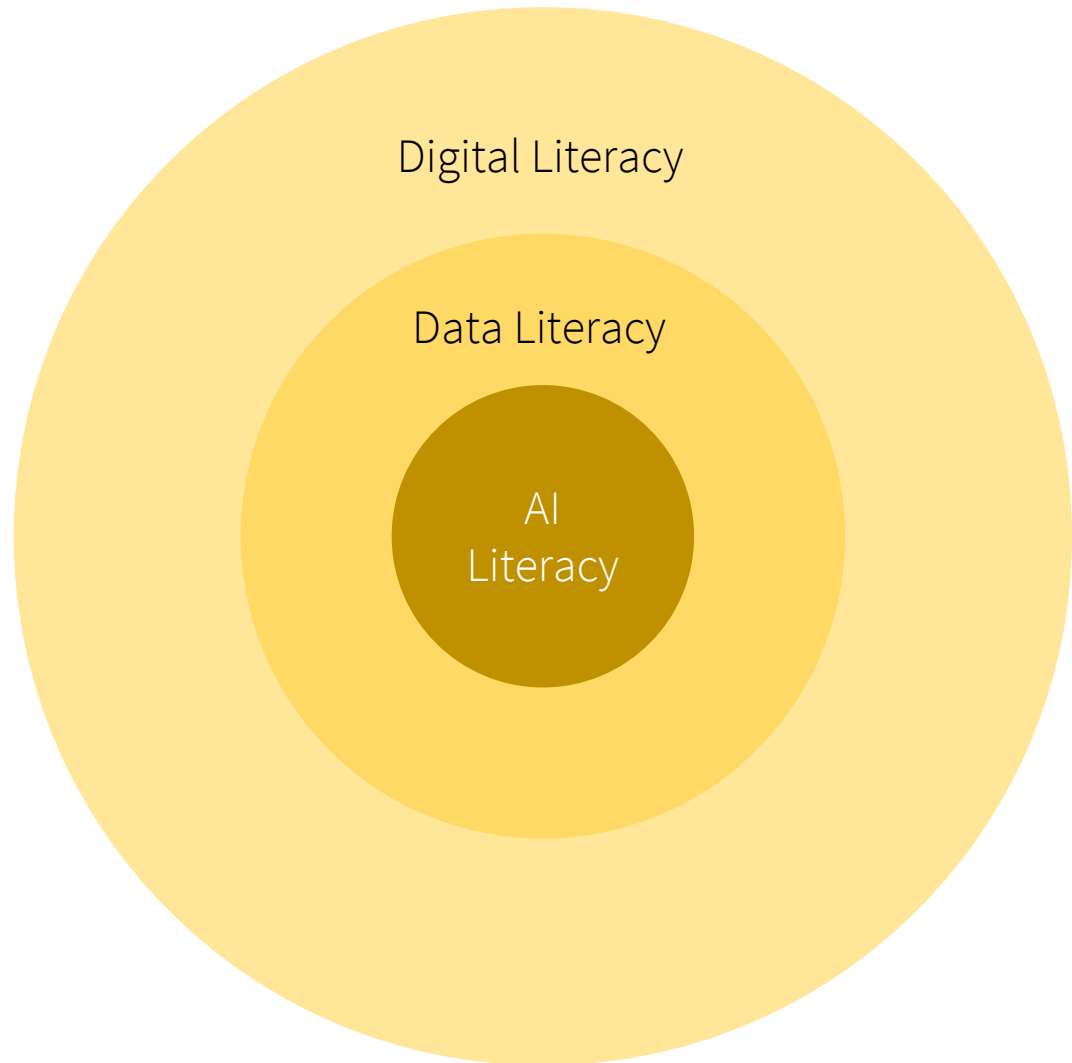
AI Literacy

- Digital Literacy (since 1997), e.g. knowledge about data formats

- Data Literacy (since 2015), e.g. comprehending and creating visualisations

- AI Literacy (since 2020), e.g. understanding machine learning concepts

Dimensions of Measurement for Digital Literacies

T1
T2
Level of Task Complexity
T3
T4

Cognitive Level
C1
C2
C3
C4

Digital Literacies

1: Beginner
2: Intermediate
3: Competent
4: Expert

A4

Autonomy Level

A3

A2

A1
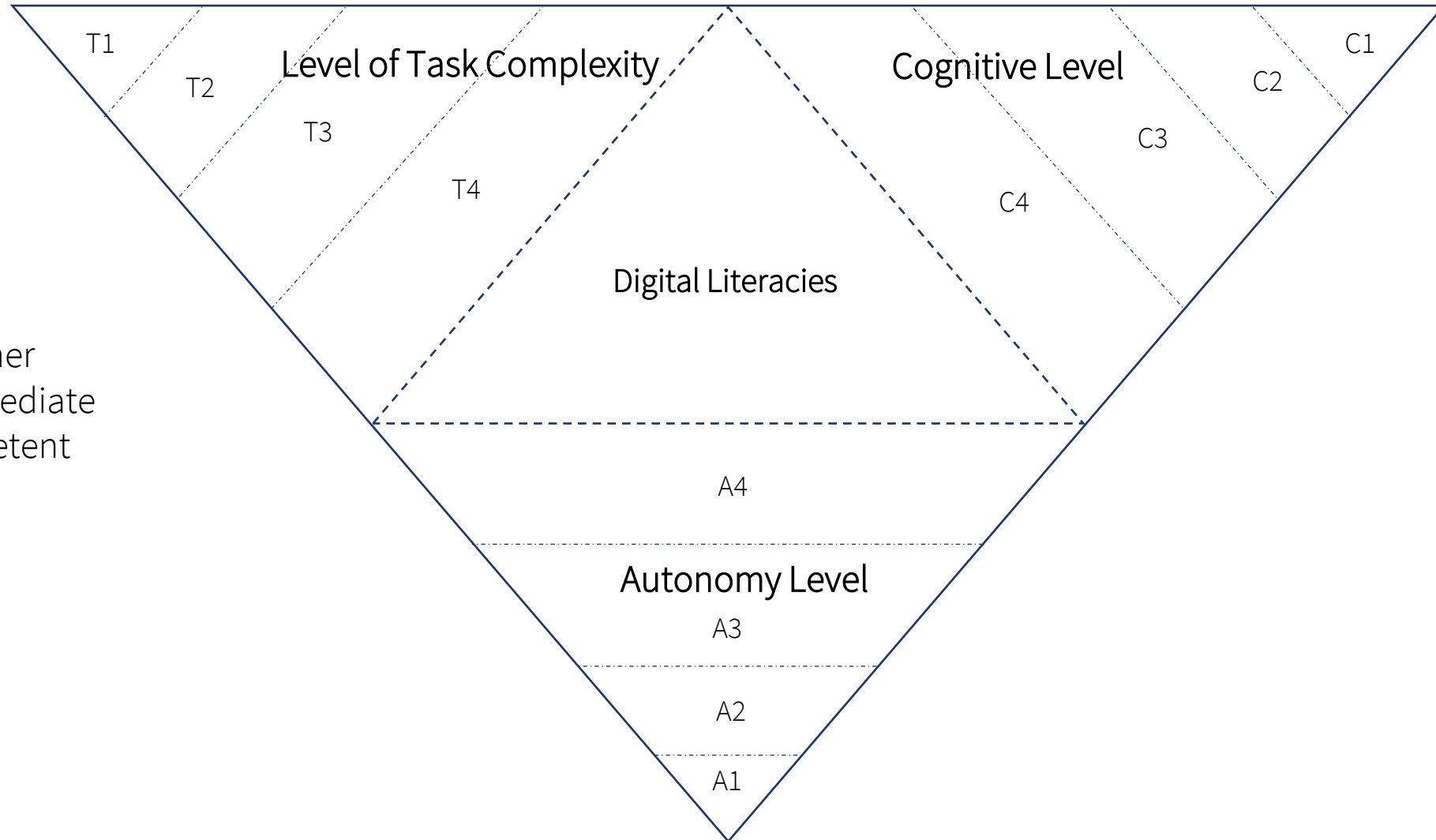
Dimensions and Levels of the Developed Framework

Dimension	Beginner	Intermediate	Competent	Expert
cognitive level	know, understand	apply	evaluate, transfer	design
autonomy level	with guidance, only when required	independently, needs-driven	guiding others, adapting to others	introducing new ideas
level of task complexity	simple tasks	limited tasks & problems, routine tasks	various tasks & problems	complex, interrelated tasks

Based on Carretero et al. 2017, 13 ff.

Domain-specific Research and Digital Literacies

- **Research**: How do you find something in the corpus that is not mentioned explicitly? (omissions in Latin & Greek historiography)

- **Background**: Historians do not mention certain events, although they should mention them due to their relevance, e.g. Cassius Dio does not mention the conference of Luca 56 BC.

- **Specific (individual) research question**: Is there a canonical way of mentioning this conference? Which contexts speak in favour of a mention, which against?

- **Domain and case specific Digital Literacies**:

  - Digital Literacy: e.g. understanding data formats (DOCX, TXT, HTML), reusing resources
  - Data Literacy: e.g. understanding the concept of taggers and annotations, applying visualisations
  - AI Literacy: e.g. knowing lemmatisation and NER, critically analysing results and explain them

Applying the Framework to the Research Case

How do you find something in the corpus that is not mentioned explicitly?

	1: Beginner	2: Intermediate	3: Competent	4: Expert
AI Literacy	Understand NER concept & language specific tagging	Compare NER results	Evaluate NER errors systematically	Enhance NER tagging by feedback
Data Literacy	Understand annotations (what) & attributions (how)	Interpret annotations	Reflect on incorrect annotations	Manipulate annotations in HTML
Digital Literacy	Distinguish DOCX & TXT	Convert DOCX into TXT	Identify problematic passages in TXT	Modify TXT as needed
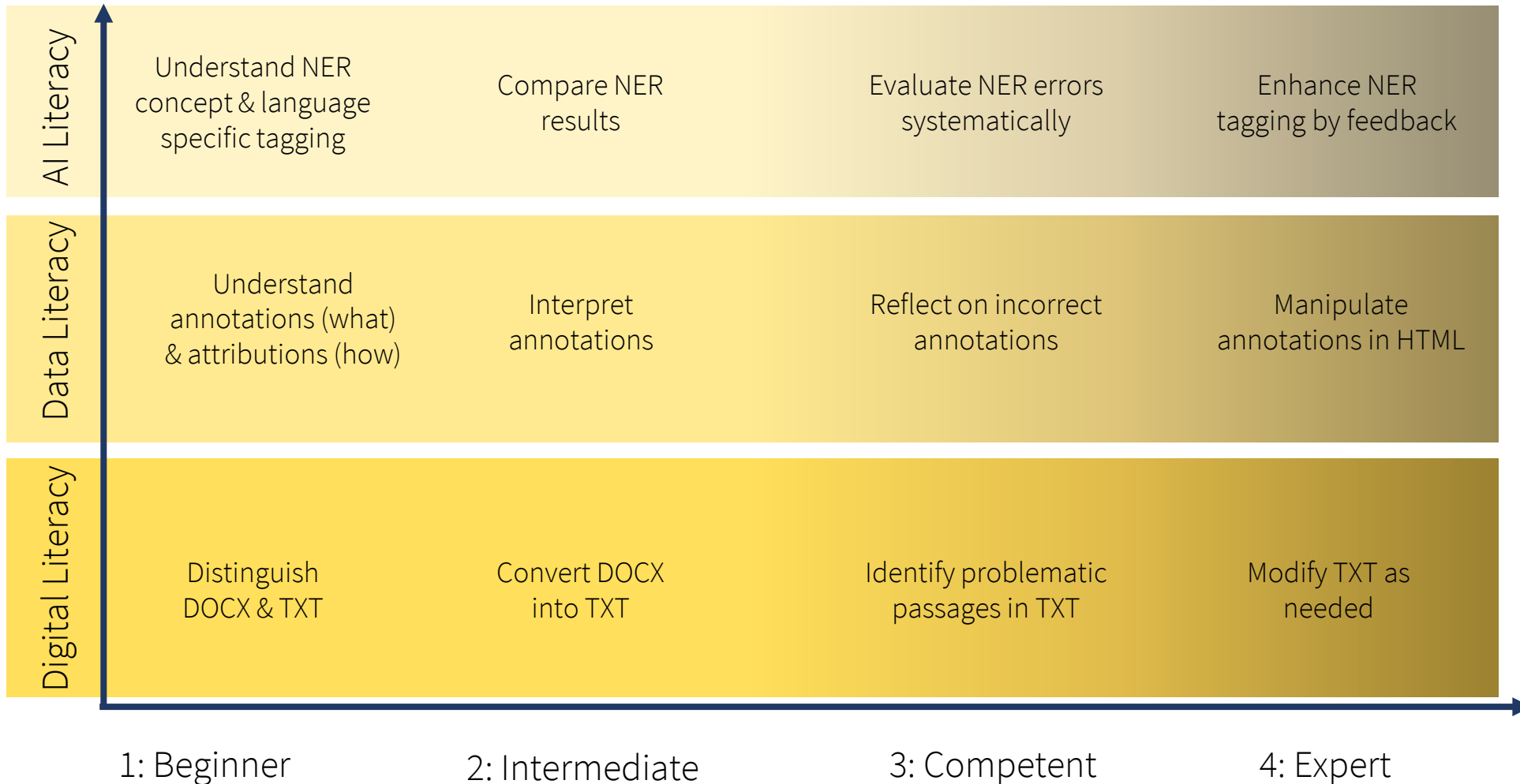
Applying the Framework to the Research Case

How do you find something in the corpus that is not mentioned explicitly? ——— Beginner: components of case specific digital research literacy
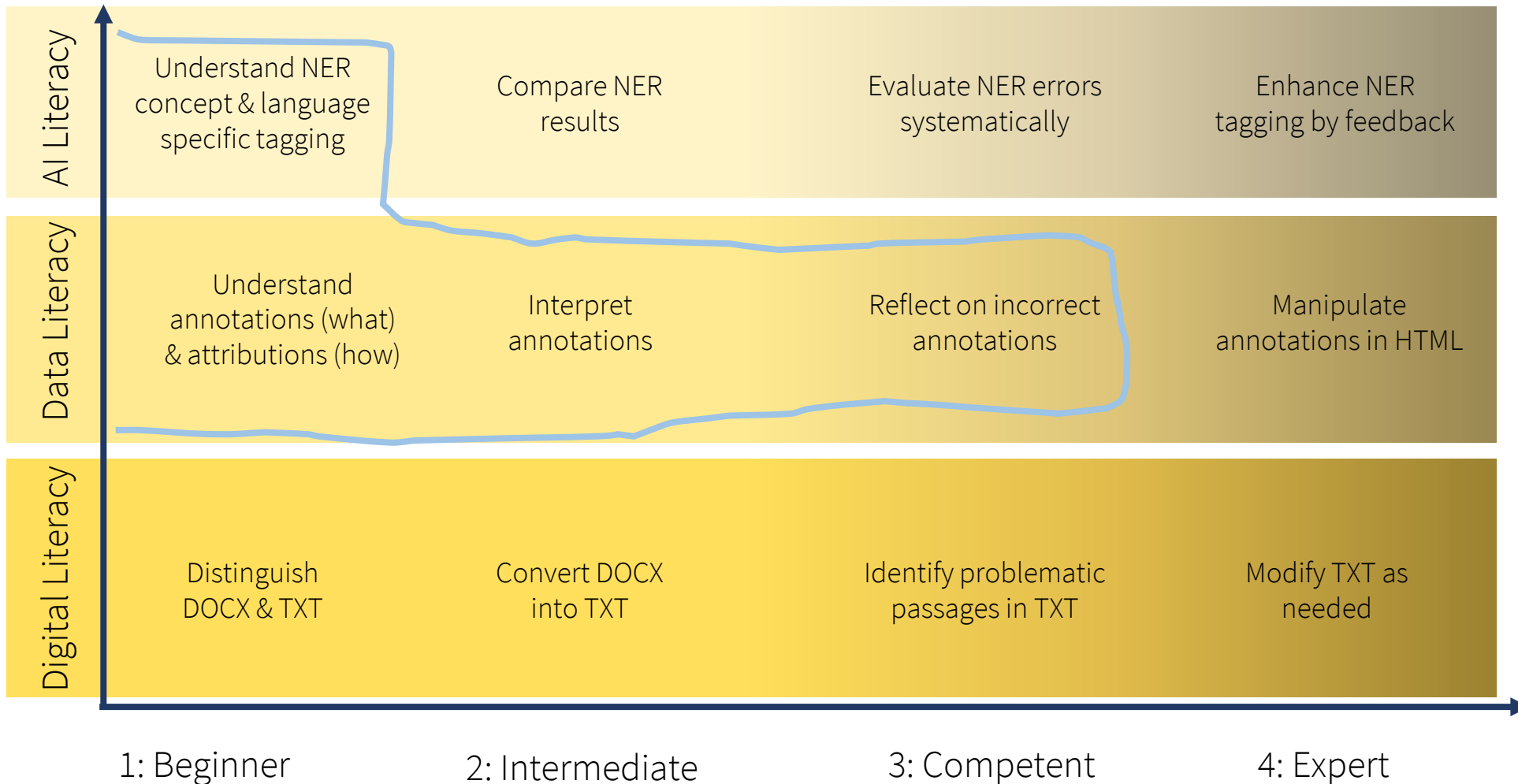
	1: Beginner	2: Intermediate	3: Competent	4: Expert
AI Literacy	Understand NER concept & language specific tagging	Compare NER results	Evaluate NER errors systematically	Enhance NER tagging by feedback
Data Literacy	Understand annotations (what) & attributions (how)	Interpret annotations	Reflect on incorrect annotations	Manipulate annotations in HTML
Digital Literacy	Distinguish DOCX & TXT	Convert DOCX into TXT	Identify problematic passages in TXT	Modify TXT as needed

Applying the Framework to the Research Case

How do you find something in the corpus that is not mentioned explicitly? ···· Intermediate: components of case specific digital research literacy
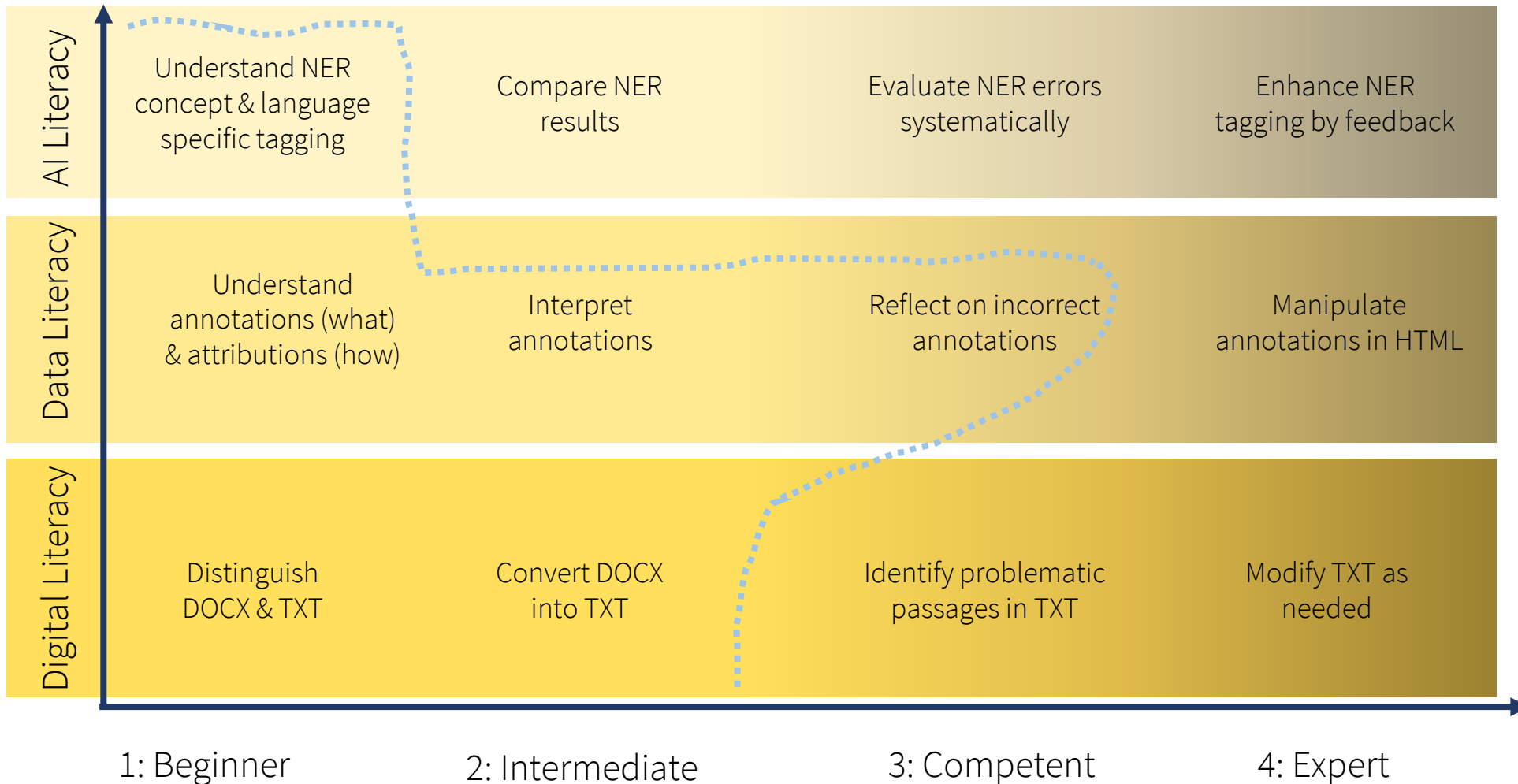
	1: Beginner	2: Intermediate	3: Competent	4: Expert
AI Literacy	Understand NER concept & language specific tagging	Compare NER results	Evaluate NER errors systematically	Enhance NER tagging by feedback
Data Literacy	Understand annotations (what) & attributions (how)	Interpret annotations	Reflect on incorrect annotations	Manipulate annotations in HTML
Digital Literacy	Distinguish DOCX & TXT	Convert DOCX into TXT	Identify problematic passages in TXT	Modify TXT as needed

Applying the Framework to the Research Case

How do you find something in the corpus that is not mentioned explicitly? – – – – Competent: components of case specific digital research literacy
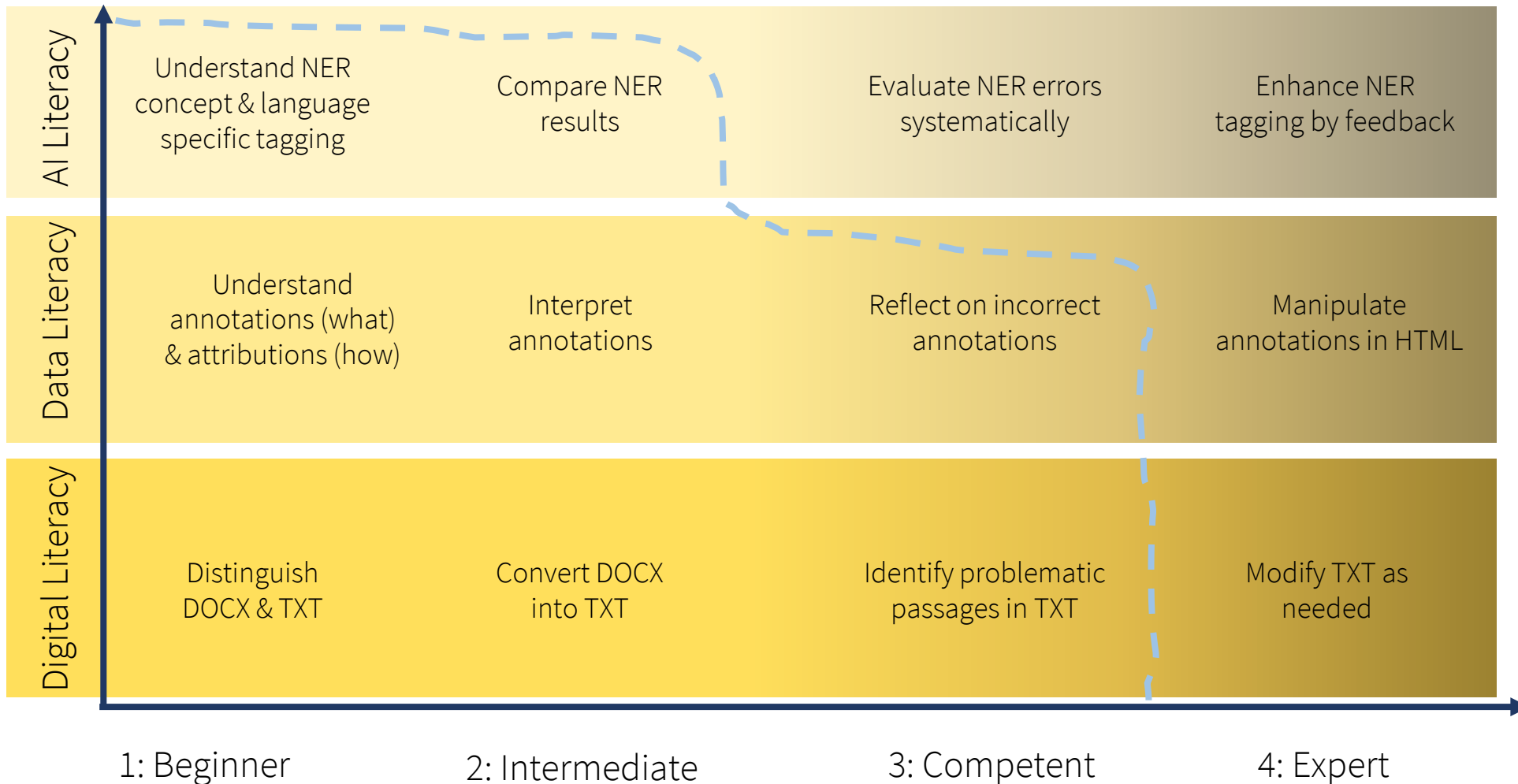
	1: Beginner	2: Intermediate	3: Competent	4: Expert
AI Literacy	Understand NER concept & language specific tagging	Compare NER results	Evaluate NER errors systematically	Enhance NER tagging by feedback
Data Literacy	Understand annotations (what) & attributions (how)	Interpret annotations	Reflect on incorrect annotations	Manipulate annotations in HTML
Digital Literacy	Distinguish DOCX & TXT	Convert DOCX into TXT	Identify problematic passages in TXT	Modify TXT as needed

Hey, you guys like **theory**, don't you? But what do you do with all this stuff?

On the basis of our model we develop **modular and progressive demo workflows** and learning units. We started with **Jupyter Notebooks** and …
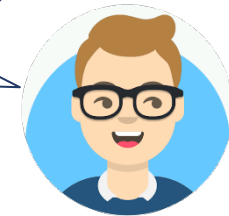
Hey, you guys like **theory**, don't you? But what do you do with all this stuff?

On the basis of our model we develop **modular and progressive demo workflows** and learning units. We started with **Jupyter Notebooks** and …

Ah, did I hear my name? Who is distracting me while enjoying a nice sacrifice? I am …

Oh no, another shadow from antiquity! Obviously, they have nothing else to do than interfering here. Jupiter, I didn't call you! We use **Jupyter Notebooks**. They don't concern you, even another spelling …

You can't be serious! I am the most important god of the Olympus. …

Jupyter Notebook: Demo Workflow NER

NER_01

Named Entity Recognition (Demo Workflow)

Andrea Beyer, 2024, Daïdalos-Projekt (https://daidalos-projekt.de)

Introduction

The automated recognition of proper names, e.g. persons, places, is called Named Entity Recognition (NER). There are different Tagger even for the languages Latin and ancient Greek. NER can help to answer classical-philological research questions.

Example: In order to answer a more general research question on the topic of "information gaps in ancient historiography", all text passages in which the conference of Luca (56 BC) is explicitly mentioned should first be identified in a large text corpus. This is where computer-aided NER can help.

Three steps are necessary:

1. Input texts: Which texts should be examined?
2. Named Entity Recognition: The algorithm identifies all persons, group (only Greek) and place names.
3. Visualization of the results: The recognized names are highlighted in color in the text.

If all goes well, this is what you should see at the end of the workflow:

Καίσαρος **PER** γὰρ εἰς Λούκαν **LOC** πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων **MISC** ἀφίκοντο, καὶ Πομπήϊος **PER** καὶ Κράσσος **PER** ἰδίᾳ συγγενόμενοι.

τοιούτοις λογισμοῖς ὁ Πομπήϊος **PER** ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον **PER** ἔτι καὶ μᾶλλον ἀνηρτήσατο.

Jupyter Notebook: Demo Workflow NER

Challenges
- Usage Jupyter Notebook
- Generalisation unclear (e.g. other texts)
- Technical vocabulary (e.g. library)
- Run code and dealing with potential error messages (software dependencies)

1. Text Input

To save time and space, we will limit ourselves here to two sentences from Plutarch and Cassius Dio. In principle, any digitally available text can be included in this step, regardless of its length.

```
# Extract from Plut. Crass. 14,5
text_with_luca: str = "Καίσαρος γὰρ εἰς Λοῦκαν πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων ἀφίκοντο, καὶ Πομπήϊος καὶ Κράσσ

# Extract from Cass. Dio 26,3
text_no_luca: str = "τοιούτοις λογισμοῖς ὁ Πομπήϊος ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον ἔτι καὶ μᾶλλον ἀνηρτήσατο.
all_texts: list = [text_with_luca, text_no_luca]
```

2. Named Entity Recognition

We install the Python library *Flair* with the package manager pip.

```
In [2]:  !pip install flair==0.13.1

Requirement already satisfied: flair==0.13.1 in /opt/conda/lib/python3.11/site-packages (0.13.1)
Requirement already satisfied: boto3>=1.20.27 in /opt/conda/lib/python3.11/site-packages (from flair==0.13.1) (1.34.9
4)
Requirement already satisfied: bpemb>=0.3.2 in /opt/conda/lib/python3.11/site-packages (from flair==0.13.1) (0.3.5)
```

Jupyter Notebook: Demo Workflow NER

We then download an AI model for Named Entity Recognition ("SequenceTagger") and integrate both into our Python code.

```
In [3]:  from flair.models import SequenceTagger
         tagger: SequenceTagger = SequenceTagger.load("UGARIT/flair_grc_bert_ner")
```

```
2024-05-05 18:05:19,463 SequenceTagger predicts: Dictionary with 15 tags: O, S-PER, B-PER, E-PER, I-PER, S-MISC, B-MISC, E-MISC, I-MISC, S-LOC, B-LOC, E-LOC, I-LOC, <START>, <STOP>
```

We let the tagger identify the entities for all texts. As a result, we get a list of specified entities, the type of entity, and a percentage on the probability of correctness.

```
In [4]:  from flair.data import Sentence
         sentences: list = [Sentence(text) for text in all_texts]
         for sentence in sentences:
             print(sentence)
             tagger.predict(sentence)
             for entity in sentence.get_spans('ner'):
                 print(entity)
```

```
Sentence[19]: "Καίσαρος γὰρ εἰς Λοῦκαν πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων ἀφίκοντο, καὶ Πομπήϊος καὶ Κράσσος ἰδίᾳ συγγενόμενοι."
Span[0:1]: "Καίσαρος" → PER (0.9911)
Span[3:4]: "Λοῦκαν" → LOC (0.962)
Span[9:10]: "Ῥωμαίων" → MISC (0.9498)
Span[13:14]: "Πομπήϊος" → PER (0.995)
Span[15:16]: "Κράσσος" → PER (0.9974)
Sentence[17]: "τοιούτοις λογισμοῖς ὁ Πομπήιος ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον ἔτι καὶ μᾶλλον ἀνηρτήσατο."
Span[3:4]: "Πομπήιος" → PER (0.9953)
Span[11:12]: "Κράσσον" → PER (0.676)
```

Jupyter Notebook: Demo Workflow NER

Challenges
- HTML
- Dealing with incorrect results
- Understand limits and opportunities of this method

3. Visualisation of the Results

We use another Flair package for displaying the results as HTML. Every type of entity has its own colour.

```
In [5]:   from flair.visual.ner_html import render_ner_html
          from IPython.display import display, HTML
          for sentence in sentences:
              html: str = render_ner_html(sentence)
              display(HTML(html))
```

Flair

Καίσαρος PER γὰρ εἰς Λοῦκαν LOC πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων MISC ἀφίκοντο, καὶ Πομπήϊος PER καὶ Κράσσος PER ἰδίᾳ συγγενόμενοι.

Flair

τοιούτοις λογισμοῖς ὁ Πομπήϊος PER ἐπὶ τὸν Καίσαρα ὡπλίζετο. καὶ τὸν Κράσσον PER ἔτι καὶ μᾶλλον ἀνηρτήσατο.

At last, some practical application. Do these demos work for the researchers?

Ehm, we just started to develop and discuss them with researchers. …

Hmm, very tricky. Upskilling for people who have already completed their education, are established in their professional field, and have lots of things to do.

That really has nothing to do with me! How is this possible?

At last, some practical application. Do these demos work for the researchers?

Ehm, we just started to develop and discuss them with researchers. …

Hmm, very tricky. Upskilling for people who have already completed their education, are established in their professional field, and have lots of things to do.

That really has nothing to do with me! How is this possible?

Thanks for your remarks, Daidalos! Now, let us all **discuss**:
1. How do we address domain-specific Digital Literacies?
2. What are **incentives** for established researchers to improve on their Digital Literacies?
3. How do we **promote Digital Literacies at the institutional level**?

Bibliography

– Carretero, S., Vuorikari, R., & Punie, Y. 2017. DigComp 2.1—The Digital Competence Framework for Citizens. Publications Office. https://op.europa.eu/en/publication-detail/-/publication/3c5e7879-308f-11e7-9412-01aa75ed71a1/language-en

– Döbeli Honegger, Beat. 2021. Covid-19 und die digitale Transformation in der Schweizer Lehrerinnen- und Lehrerbildung. Beiträge zur Lehrerinnen- und Lehrerbildung 39:3.214–422. https://www.pedocs.de/volltexte/2021/23693/pdf/BzL21_3_411-422_DoebeliHonegger_Covid_19_und_die_digitale_Transformation.pdf

– Gerick, Julia, Frank Goldhammer, Heike Schaumburg, Knut Schwippert, Martin Senkbeil, Jan Vahrenhold, Birgit Eickelmann, & Wilfried Bos (Eds.). 2019. ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking. Waxmann. https://www.pedocs.de/volltexte/2019/18166/pdf/Eickelmann_et_al_2019_ICILS_2018_Deutschland.pdf

– Laupichler, Matthias Carl, Alexandra Aster, Jana Schirch, & Tobias Raupach. 2022. Artificial intelligence literacy in higher and adult education: A scoping literature review. Computers and Education: Artificial Intelligence 3.100101. https://www.sciencedirect.com/science/article/pii/S2666920X2200056X

– Long, Duri & Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

– Lorenz, Uwe & Ralf Romeike. 2023. AI-PACK – Ein Rahmen für KI-bezogene Digitalkompetenzen von Lehrkräften auf Basis von DPACK. Lecture Notes in Informatics (LNI) ed. by 11–24. Bonn: Gesellschaft für Informatik. https://www.researchgate.net/publication/374350289_AI-PACK_-Ein_Rahmen_fur_KI-bezogene_Digitalkompetenzen_von_Lehrkraften_auf_Basis_von_DPACK

– Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A new framework for teacher knowledge. Teachers College Record 108 (6), 1017-1054. https://journals.sagepub.com/doi/epdf/10.1111/j.1467-9620.2006.00684.x

– Ng, Davy Tsz Kit, Jac Ka Lok Leung, Kai Wah Samuel Chu, & Maggie Shen Qiao. 2021. AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. Proceedings of the Association for Information Science and Technology 58:1.504–509. https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.487 (a)

– Ng, Davy Tsz Kit, Jac Ka Lok Leung, Samuel Kai Wah Chu, & Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence 2.100041. https://www.sciencedirect.com/science/article/pii/S2666920X21000357 (b)

– Ridsdale, Chantel, James Rothwell, Mike Smit, Michael Bliemel, Dean Irvine, Dan Kelley, Stan Matwin, Brad Wuetherick, & Hossam Ali-Hassan. 2015. Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. https://www.researchgate.net/publication/284029915_Strategies_and_Best_Practices_for_Data_Literacy_Education_Knowledge_Synthesis_Report?channel=doi&linkId=564b2fc808ae9cd9c8282d5a&showFulltext=true

– Schmidt, Andreas, Thomas Neifer, & Benedikt Haag. 2021. Data Literacy als ein essenzieller Skill für das 21. Jahrhundert. Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis ed. by Frick Detlev Andreas Gadatsch Jens Kaufmann Birgit Lankes Christoph Quix Andreas Schmidt & Uwe Schmitz, 27–40. Wiesbaden: Springer Fachmedien. https://link.springer.com/chapter/10.1007/978-3-658-33403-1_2

– Schüller, Katharina, Paulina Busch, & Carina Hindinger. 2019. Future Skills: Ein Framework für Data Literacy. AStA Wirtschafts- und Sozialstatistisches Archiv 13:3–4.297–317. Langversion beim HFD: https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_AP_Nr_47_DALI_Kompetenzrahmen_WEB.pdf

– Sekretariat der Kultusministerkonferenz (Ed.). 2016. Strategie der Kultusministerkonferenz Bildung in der digitalen Welt. Berlin: KMK. https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf