SEFLAG

Systematic **E**valuation **F**ramework for NLP Models and Datasets in **L**atin and **A**ncient **G**reek

Konstantin Schulz, Florian Deichsler
Humboldt-Universität zu Berlin

Research Questions

Question 1: How can we help **literary scholars** of Latin and Ancient Greek **choose** the right NLP **models** for their research agenda?
Question 2: How can we **assess** and **document** language resources for Latin and Ancient Greek **systematically**?
Question 3: How can we **combine** and unify **annotation schemas** that are somewhat similar, but not identical?

Components and Activities

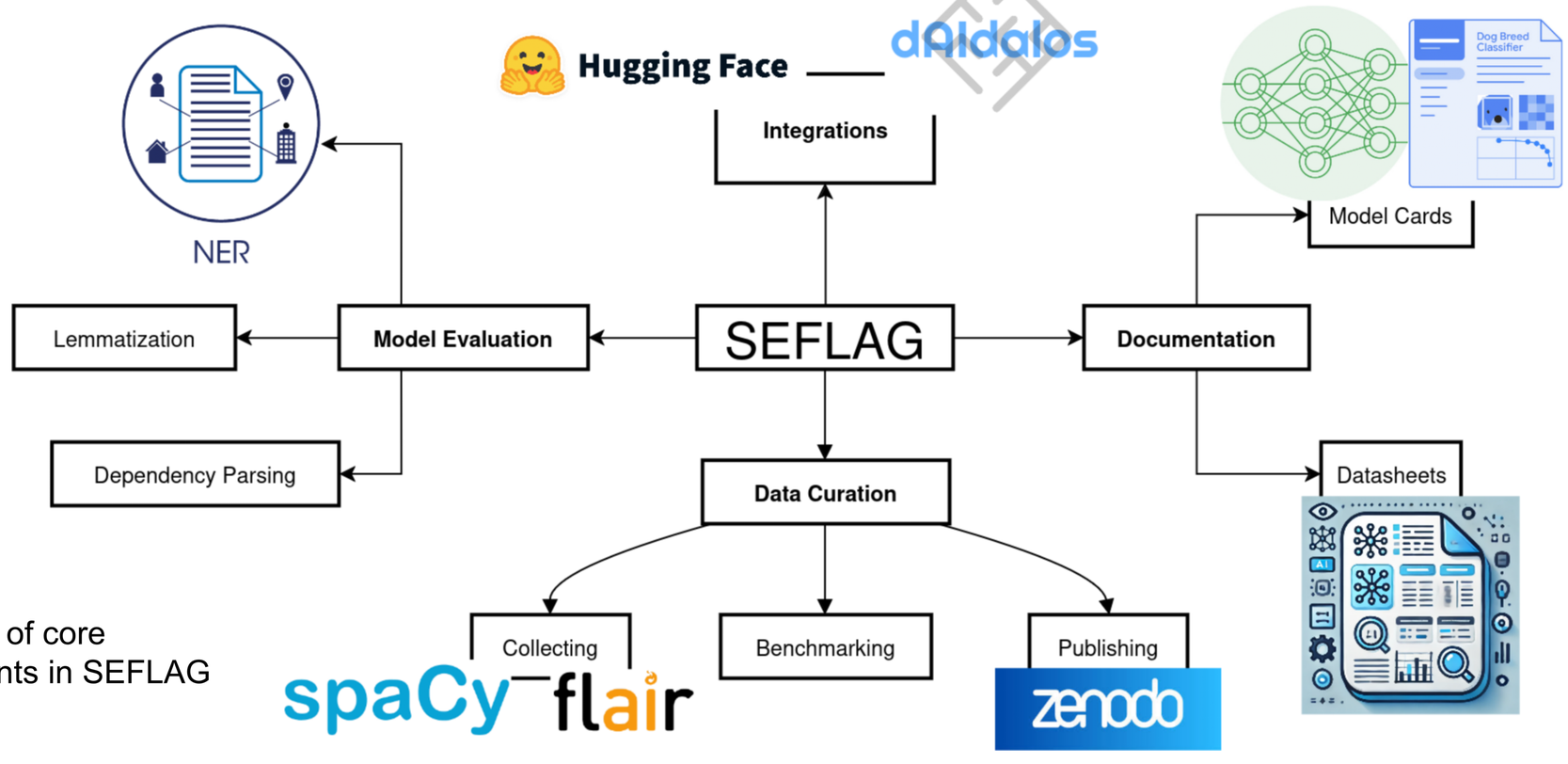

Fig. 1:
Overview of core components in SEFLAG

Workflow

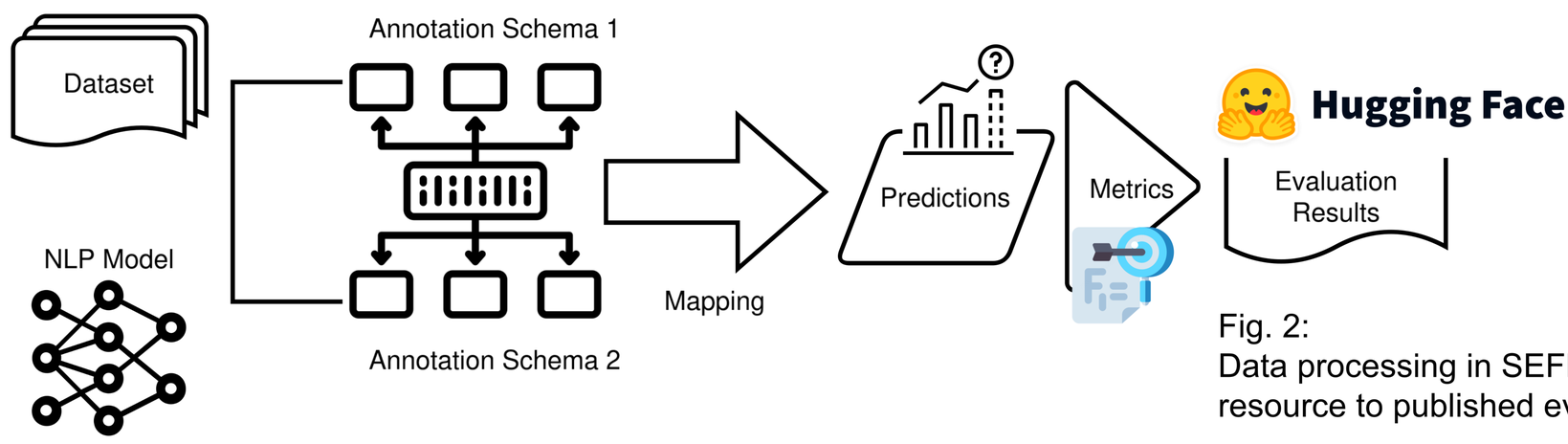

Fig. 2:
Data processing in SEFLAG, from language resource to published evaluation results

Results

Language	Task	Model	Dataset	Metric	Score
Latin	NER	LatinCy	Herodotos	macro F1 ↑	58
	lemmatization	LatinCy	UD Latin	accuracy ↑	88
Ancient Greek	NER	flair_grc_bert_ner	Yousef et al.	macro F1 ↑	64
	lemmatization	greCy	UD Ancient Greek	accuracy ↑	89

Table 1: Evaluation results for NER and lemmatization in Latin and Ancient Greek. The metrics used macro F1 and accuracy. 3 different NLP models have been evaluated on 4 different datasets. Upward arrows for a metric indicate that higher scores are better.

- **model cards** and **datasheets** published on Hugging Face
- Linked Data for **evaluation results** and their associated datasets or models
- explicit case-specific **annotation mappings** in YAML format

Challenges

- high linguistic **variation**
- relatively few **language resources** available
- scattered initiatives, insufficient **interoperability**
- **long-term** funding and sustainability
- domain-specific traditions, **interdisciplinarity**

Prospects for future research

- add more **tasks, models, datasets**
- create **open educational resources** about fundamentals of NLP evaluation
- full **integration** into the **Daidalos** research infrastructure

References:
Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. "**Named Entity Recognition and Classification in Historical Documents: A Survey**." ACM Computing Surveys.
Sommerschield, Thea, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. "**Machine Learning for Ancient Languages: A Survey**." Computational Linguistics 49 (3): 703–47.
images: Flaticon.com

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Source Code

Project Website

Contact
Konstantin Schulz
Department for German Studies and Linguistics
daidalos-projekt@hu-berlin.de