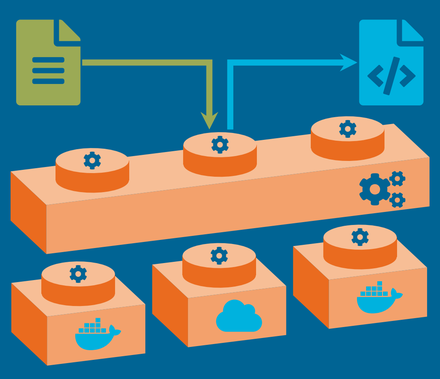
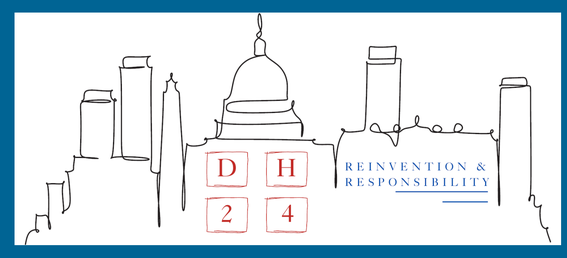


# Efficient, uniform and scalable parallel NLP pre-processing with DUUI

## Perspectives and Best Practice for the Digital Humanities

Giuseppe Abrami, Alexander Mehler

Goethe University Frankfurt | Text Technology Lab




August 6-9, 2024
Washington, D.C., USA

## Introduction

- There is a constant need for annotating corpora.
- Although most projects require manual annotations (e.g. (Abrami et al. 2021; Zhang, Chapman, and Ciravegna 2010)), it is essential to use NLP to support this.
- NLP-based automatic pre-processing systems are implemented on a project-specific basis, whereby the following aspects can be observed:
  1. utilization of heterogeneous tools
  2. separate execution and manual scaling and distribution
  3. different input/output formats of the individual tools must be adapted to each other
- As this approach is very time-consuming and resource-intensive, we developed DUUI (Leonhardt et al. 2023) to address these challenges.

## Docker Unified UIMA Interface (DUUI)

- A UIMA-based (Ferrucci et al. 2009) framework to provide NLP pipelines.
- Integration of various NLP tools based on different models, procedures and programming languages by means of pipelines.
- Use of microservices, like Docker, for encapsulation and to avoid dependency issues through the use of heterogeneous tools.
- Implicit reusability and versioning of pipelines due to the dockerization of corresponding components (Component).
- Use of Lua (Ierusalimschy, Figueiredo, and Celes 2007) –a powerful programming language – to read and write annotations, driven by the annotation components.
- Scaling of NLP processes through horizontal and vertical distribution on different systems using Docker Swarm.

## DUUI for the Digital Humanities

- DUUI allows the easy and flexible use of various NLP techniques.
- Researchers can integrate their existing methods by adapting the established algorithms for DUUI, which consists of only three elements:
  1. Definition of a UIMA TypeSystem, which is used as an annotation schema.
  2. Creation of a Lua script which defines which information from the document to be annotated has to be selected (deserialize) and which annotations have to be created (serialize).
  3. Incorporation of the DUUI interface, which includes the integration of a REST server and the creation of the relevant REST routes.
- All these three elements can be encapsulated in a Docker image so that a new Component is created and can be directly utilized within DUUI.
- Consequently, big data corpora (e.g. Abrami, Bagci, and Mehler 2024; Lücking et al. 2021) from different thematic domains with heterogeneous tools and models can be processed efficiently in a homogeneous, uniform environment.

## Resources

DockerUnifiedUIMAInterface @ TTLab

## References

Abrami, Giuseppe, Mevlüt Bagci, and Alexander Mehler (2024). "German Parliamentary Corpus (GerParCor) Reloaded." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italy: ELRA and ICCL, pp. 7707–7716.

Abrami, Giuseppe, Alexander Henlein, Andy Lücking, Attila Kett, Pascal Adeberg, and Alexander Mehler (June 2021). "Unleashing annotations with TextAnnotator: Multimedia, multi-perspective document views for ubiquitous annotation." In: *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Ed. by Harry Bunt. Groningen, The Netherlands (online): Association for Computational Linguistics, pp. 65–75.

Ferrucci, David, Adam Lally, Karin Verspoor, and Eric Nyberg (2009). *Unstructured Information Management Architecture (UIMA) Version 1.0*. OASIS Standard.

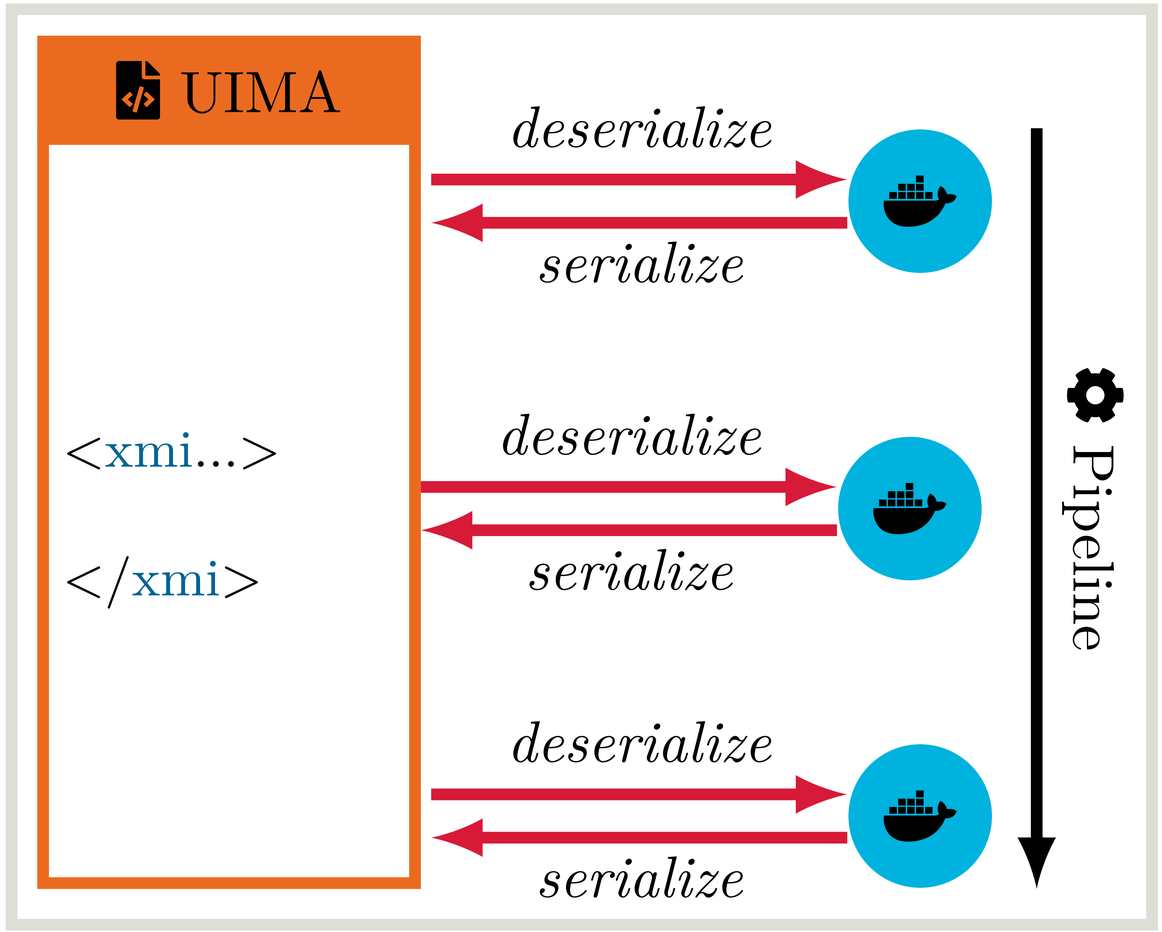
Ierusalimschy, Roberto, Luiz Henrique de Figueiredo, and Waldemar Celes (2007). *The Evolution of Lua*.

Leonhardt, Alexander, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler (2023). "Unlocking the Heterogeneous Landscape of Big Data NLP with DUUI." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 385–399.

Lücking, Andy, Christine Driller, Manuel Stoeckel, Giuseppe Abrami, Adrian Pachzelt, and Alexander Mehler (2021). "Multiple Annotation for Biodiversity: Developing an annotation framework among biology, linguistics and text technology." In: *Language Resources and Evaluation*. Ed. by Nancy Ide and Nicoletta Calzolari.

Zhang, Ziqi, Sam Chapman, and Fabio Ciravegna (2010). "A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality." In: *Knowledge Engineering and Management by the Masses*. Ed. by Philipp Cimiano and H. Sofia Pinto. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 301–315.

## The Way of DUUI




## Encapsulation within DUUI