# Session 1.3

##### Session 1.3 (SEMANTiCS)

#### Time: Wednesday, September 18, 2024 - 10:40 to 12:00

#### Chair: TBA

## **Talks**

### Zero-shot Topic Classification of Column Headers: Leveraging LLMs for Metadata Enrichment

Traditional dataset retrieval systems rely on metadata for indexing, rather than on the underlying data values. However, high-quality metadata creation and enrichment often require manual annotations, which is a labour-intensive and challenging process to automate. In this study, we propose a method to support metadata enrichment using topic annotations generated by three Large Language Models (LLMs): ChatGPT-3.5, GoogleBard, and GoogleGemini. Our analysis focusses on classifying column headers based on domain-specific topics from the Consortium of European Social Science Data Archives (CESSDA), a Linked Data controlled vocabulary. Our approach operates in a zero-shot setting, integrating the controlled topic vocabulary directly within the input prompt. This integration serves as a Retrieval-Augmented Generation (RAG) approach, with the aim of improving the results of the topic classification task.

We evaluated the performance of the LLMs in terms of internal consistency, inter-machine alignment, and agreement with human classification. Additionally, we investigate the impact of contextual information (i.e., dataset description) on the classification outcomes. Our findings suggest that ChatGPT and GoogleGemini outperform GoogleBard in terms of internal consistency as well as LLM-human-alignment. Interestingly, we found that contextual information had no significant impact on LLM performance.

This work proposes a novel approach that leverages LLMs for topic classification of column headers using a controlled vocabulary, presenting a practical application of LLMs and RAG systems within the Semantic Web domain. This approach has the potential to facilitate automated metadata enrichment, thereby enhancing dataset retrieval and the Findability, Accessibility, Interoperability, and Reusability (FAIR) of research data on the Web.

| Margherita Martorana | Tobias Kuhn |
| --- | --- |
| Lise Stork | Jacco van Ossenbruggen |

### Multilingual linguistic word sense disambiguation for semantic annotations

| Robert David | Ilan Kernerman | Assaf Siani |
| --- | --- | --- |

### Improving term networks through the detection of semantic perspectives

abstract

| Maya Sappelli | Hans Fugers |
| --- | --- |
| Bart Kleijngeld | Marijn Siebel |

### Enhancing Answers Verbalization using Large Language Models

abstract

| Daniel Vollmers | Parth Sharma |
| --- | --- |
| Hamada Zahera | Axel-Cyrille Ngonga Ngomo |