STLab seminar – Tuesday, December 16th, 2025 (15:00-17:15)

Towards the INFINITY project: *Multidimensional Knowledge-Based Annotation for Ethical Context-Aware Heritage Data Life Cycles*

Keywords: multiperspective/multidimensional knowledge graphs, data and AI ethics, knowledge extraction, knowledge engineering

Teams link:
https://teams.microsoft.com/l/meetup-join/19%3ameeting_MGJhMjI0ZjktMTgxMy00ZTY3LWJhNDEtZjA0M2U2OGI5YTAx%40thread.v2/0?context=%7b%22Tid%22%3a%22e99647dc-1b08-454a-bf8c-699181b389ab%22%2c%22Oid%22%3a%22d9854e38-5271-4218-a460-57513b97030d%22%7d

Meeting ID: 340 134 298 123 95
Passcode: Uy7np7zS

Agenda

[15:00 – 15:15] Intro and welcome - Valentina Presutti

[15:15 – 15:45 (20 min presentation + 10 QA)]

Title: Enriching Knowledge Graphs of Cultural Heritage Digital Objects with Scholarly Interpretations: A Pipeline Approach Using Large Language Models

Speaker: Andrea Schimmenti (University of Bologna)

Abstract: Cultural Heritage institutions possess vast textual resources containing sophisticated scholarly analyses, yet lack practical means to transform this knowledge into queryable, machine-readable formats. We introduce ATR4CH (Adaptive Text-to-RDF for Cultural Heritage), a systematic methodology for coordinating Large Language Model-based knowledge extraction with ontological frameworks to develop a pipeline capable of capturing complex interpretative knowledge from cultural heritage texts.
Using authenticity assessment debates as a validation case, the research demonstrates how LLMs can extract multi-dimensional scholarly reasoning—including competing hypotheses, evidential features, and alternative perspectives—while maintaining alignment with semantic web standards. The methodology achieves F1-scores of 0.96-0.99 for metadata extraction and 0.95-0.97 for evidence extraction, with smaller models performing competitively, enabling cost-effective deployment for resource-constrained institutions.

The approach addresses key challenges relevant to enriching cultural heritage data ecosystems: transforming unstructured scholarly discourse into structured, queryable knowledge graphs; capturing multiple perspectives and conflicting interpretations within unified semantic frameworks; and supporting scalable, semi-automated metadata enrichment while preserving the provenance and epistemic complexity of humanities scholarship.

Short bio: Andrea Schimmenti is a PhD candidate in Cultural Heritage in the Digital Ecosystem at the University of Bologna, supervised by Prof. Fabio Vitali with co-supervisors Prof. Francesca Tomasi and Dr. Marieke van Erp (KNAW Humanities Cluster). With a background in Classical Philology and Digital Humanities, his research focuses on knowledge graph extraction from cultural heritage texts using Large Language Models, combining computational methods with digital humanities methodologies. He conducted a research stay at DHLab Amsterdam and has published on cultural heritage knowledge graphs and LLM applications in venues including DH2023, AIUCD, and LDK conferences and journals such as Umanistica Digitale and Journal of Documentation (accepted). He is also a teaching assistant for Web Technologies courses at the University of Bologna.

[15:45 – 16:15 (20 min presentation + 10 QA)]

Title: Interworlds of Meaning: Neurosymbolic Metaphor Processing for Multivocal Knowledge Engineering

Speaker: Anna Sofia Lippolis (University of Bologna / CNR)

Abstract: Cultural Heritage Digital Objects resist singular interpretation. When different stakeholders attribute fundamentally incompatible meanings to the same cultural object, or when historical descriptions embed colonial framings that conflict with contemporary reappraisals, traditional object-centric metadata systems collapse these tensions into administrative consensus, or worse, silence alternative voices entirely.
This seminar addresses INFINITY's three core challenges through a neurosymbolic approach centered on bidirectional LLM-knowledge engineering integration. It demonstrates how Large Language Models can automate the labor-intensive work of knowledge graph construction. Conversely, it shows how ontological structures can scaffold LLM interpretation of graded, non-literal meaning, providing a blueprint for other kinds of language and implicit meaning phenomena.
Metaphor and conceptual blending provide the empirical testbed: they exemplify "interworlds" where incompatible conceptual domains meet to create emergent, context-dependent meaning. The same framework that captures how scientific metaphors operate in medical corpora can model how heritage narratives metaphorically frame objects through competing lenses, enabling record-centric rather than object-centric modeling, where multiple situated descriptions coexist without reduction.
Methodologically, the approach integrates human-in-the-loop validation with explicit consistency monitoring to address automation bias and ensure transparency, supporting INFINITY's ethics-by-design mandate and alignment with both FAIR and CARE principles for heritage data governance.

Short bio: Anna Sofia Lippolis is a PhD candidate at the University of Bologna, Italy, affiliated with the National Research Council's Institute for Cognitive Sciences and Technologies (Rome, Italy). Her work investigates how semantic technologies intersect with Digital Humanities research and how AI can automate knowledge-engineering practices.

[16:15 – 16:45 (20 min presentation + 10 QA)]

Title: Overcoming the Generalization Limits of SLM Finetuning for Shape-Based Extraction of Datatype and Object Properties

Speaker: Celian Ringwald (Université Côte d'Azur / INRIA)

Abstract: Small language models (SLMs) have shown promise for relation extraction (RE) when extracting RDF triples guided by SHACL shapes focused on common Datatype Properties. This work investigates how SLMs handle both Datatype and Object Properties for a complete RDF graph extraction. We show that the key bottleneck is related to long-tail distribution of rare properties. To solve this issue, we evaluate several strategies: stratified sampling, weighted loss, dataset scaling, and template-based synthetic data augmentation. We show that the best strategy to perform equally well across unbalanced target properties is to build a training set in which the number of occurrences of each property exceeds a given threshold. To enable reproducibility, we publicly released our datasets, experimental results and code. Our findings offer practical guidance for training shape-aware SLMs and highlight promising directions for future work in semantic RE.

Short bio: Célian Ringwald is PhD candidate at the Université Côte d'Azur, in the Wimmics Inria team (France). He has a dual academic background in Data Science and Digital Humanities, and was the maintainer of the French DBpedia Chapter before starting his PhD. The topic of his thesis is how to combine the best of both worlds—language models and knowledge bases—to automate the extraction of precise and reliable information from text.

[16:45 – 17:15 (20 min presentation + 10 QA)]

Title: Toward ethically aligned AI: moral value detection, agreement metrics, and formal reasoning in Language Models

Speaker: Luana Bulla (University of Catania / CNR)

Abstract: The detection and representation of moral values in natural language constitutes a critical challenge for the development of ethically aligned Artificial Intelligence, particularly within the domain of Cultural Heritage, where interpretation, social norms, and value systems are inherently multiperspective and historically situated. In such contexts, the annotation of moral values is intrinsically subjective and frequently exhibits low inter-annotator agreement,

which poses significant challenges for both the annotation process and the development of robust evaluation methodologies. This seminar contributes to the INFINITY research agenda along three integrated directions: the assessment of state-of-the-art Large Language Models (LLMs) for moral value detection in real-world text; the introduction of a novel inter-annotator agreement metric, F1-kappa, designed to unify human and machine evaluation in binary and multi-label settings; and the integration of formal moral reasoning within neuro-symbolic and knowledge-based AI architectures. Within the INFINITY framework, this supports ethically grounded, human-in-the-loop annotation pipelines for the enrichment of Cultural Heritage Digital Objects. Beyond evaluation, the seminar discusses the integration of symbolic moral reasoning and neural language models in Multidimensional Knowledge Graphs to explicitly represent conflicting moral perspectives, social norms, and contextualized interpretations in an ethics-by-design framework.

Short bio: Luana Bulla is a PhD student in Computer Science at the University of Catania, affiliated with the Institute of Cognitive Sciences and Technologies (ISTC-CNR), where she focuses on Natural Language Processing for human-centered AI. With a background in linguistics and digital humanities, Luana specializes in semantic modeling, moral value detection, and machine learning models, particularly in the area of generative models. She has contributed to several research projects, including IDEHA, SPICE, FAIR, and L4ALL. Her primary interests include AI-generated content detection, emotion and moral values classification, and moral reasoning in language models. Her current research spans sign language translation in low-resource settings, grammar-constrained natural language generation, and the ethical dimensions of AI.