ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Knowledge Engineering Projects 2022-2023

Disclaimer

You don't have to choose your project today, the deadline for creating your group and expressing your preferences is **May 8th (although you can do it earlier)**

Communicate at least two preferences

Project assignments will be communicated by **May 15th (first in first served)**

The project specification, requirements and goals will be refined after the assignments through direct interaction between the project group and the mentors

What the projects are about

The knowledge engineering project concerns the design, implementation and publication of **semantic web knowledge graphs** (RDF)

Each project is developed by a **team of 2 people**, and will have a **mentor**

The main role of the mentor is to help the group **refining and stabilizing** the project **specification and requirements**

Besides concurring to your final grade definition, the project is an opportunity for you to face a **realistic ontology/knowledge graph project** and to apply, in this context, the methods and tools learned during the course

The tasks to be executed

Tasks depend on the domain of knowledge, the quality of the input data, the complexity of the modelling problems, etc…

Each group will face different challenges and problems that cannot be predicted, they will adopt creative solutions, specific for each project and that will characterize its value

This year, we want to experiment also the possible role of large language models in the knowledge engineering process

We provide **a list of generalised tasks** that all projects must perform:

- Use them as a checklist

The ontologies must be well designed, tested and documented

10 Tasks

1) Analysis of existing datasets using **heterogeneous formats**, to produce RDF knowledge graphs
2) Application of the **eXtreme Design methodology** (competency questions, ODP reuse, testing, etc.) to develop **OWL ontologies** for the knowledge graphs
3) Definition of **mapping rules** for transforming input data into semantic web knowledge graphs, according to the developed ontologies (e.g., SPARQLAnything)
4) Generation of **URIs** and **publication** of ontologies and knowledge graphs (with permanent URIs)
5) Application/use/configuration of tools for **entity linking and ontology alignment**
6) Use of **large language models**
7) Publication of a **SPARQL endpoint**
8) Integration of **LODView** for knowledge graph browsing and **LODE** for producing human-readable documentation of the ontologies Ex. https://github.com/anuzzolese/OntoPiA-UI
9) Creation of a **docker** that will contain data, software, SPARQL endpoint, **website** and all the necessary dependencies
10) Report writing with a description of the project, remarking the applied methodology, the addressed the challenges, the adopted solutions, and the obtained results.

Delivery: GitHub, Report, Website and Docker

Each project will have its own **GitHub repository**

You can refer, as examples, to the KG projects ArCo: https://github.com/ICCD-MiBACT/ArCo and Polifonia: https://github.com/polifonia-project/ontology-network

Each project will be described in a **report.** The report will illustrate the developed activities, the design and implementation choices and the obtained results.

Reports shall describe the goal of the project, the domain and problems addressed, the sources analysed and the tools used. Then it will present and discuss the details of the development: the requirements (CQs), tests, odp reused and the resulting ontologies and knowledge graphs along with its quality assessment. Finally, it will discuss the challenges encountered and how they have been solved, and possible future work.

Each project will be delivered as a **docker** containing the produced software and data (and their dependencies), and a presentation website with a section dedicated to the report

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Let's see the projects…

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Project n.1

Name: "The Semantic Music Hub"

Topic: MIR, Comp. Creativity

Mentors:

- Andrea Poltronieri (UniBo)
- Jacopo de Berardinis (KCL)
- Valentina Presutti

Description

In the **Polifonia** project, UniBo and King's College London (KCL) have developed Smashub -- a semi-automatic framework for **integrating** music datasets and **generating** Musical Knowledge Graphs.

Smashub was first used to produce ChoCo (the Chord Corpus) -- the largest collection of harmonic annotations to date (link). Our framework follows a **modular** design and can be **extended** to support a variety of musical annotations, including emotions/mood, structures, patterns, and lyrics.

In this project, you will extend Smashub to model 1 of these annotation types. This involves:

- Designing a music ontology related to the annotation type (e.g. "The Music Emotion Ontology")
- Deploying the ontology within the Polifonia Ontology Network
- Integrating a selection of music datasets providing those annotations (e.g. the DEAM dataset)
- Creating a Musical Knowledge Graph from your integrated data and ontology
- Creating example sparql queries that showcase stories from the dataset (using the MELODY tool)

KING'S College LONDON

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Project n.2

Name: "Enriching Framester for Enhanced Text-to-Knowledge Graph Creation and Querying"

Topic: Ontology Alignment & Knowledge Graph Enrichment

Mentors:

- Arianna Graciotti (UniBo)
- Luca Contalbo (UniBo)
- Aldo Gangemi
- Valentina Presutti

Description

This project aims to enrich **Framester**, a frame-based ontological resource acting as a hub between diverse linguistic resources. You will focus on establishing **new alignments** between some of its resources, specifically **PropBank** and **WordNet**. PropBank is a repository of verb senses and their semantic roles, while WordNet is a lexical database of semantically connected concepts.

You will develop methods and tools to **map**, into Framester, PropBank **frames** and **roles** to the most suitable WordNet **synsets** and to assess the **quality** of these alignments. Successively, you will design ad-hoc heuristics that will enable the creation, into Framester, of **semantic clusters** of related PropBank frames and roles.

You will evaluate whether the enrichment of Framester benefits **Text-to-Knowledge Graph pipelines** (such as the one used to create MusicBO Knowledge Graph in the **Polifonia** project) by allowing the production of more **informative knowledge graphs** and more **effective querying** of them.

Useful Resources

Framester

- Framester: https://github.com/framester/Framester
- Framester SPARQL Endpoint: http://etna.istc.cnr.it/framester2/sparql
- Framester API: http://etna.istc.cnr.it/framester_web/
- Framester schema: https://w3id.org/framester/schema/

PropBank

- PropBank: https://github.com/propbank; https://github.com/propbank/propbank-frames

WordNet

- WordNet: http://wordnet-rdf.princeton.edu/
- WordNet RDF: http://wordnet-rdf.princeton.edu/about

Text-to-Knowledge Graph pipeline and support tools:

- Text-to-AMR-to-Fred: https://arco.istc.cnr.it/txt-amr-fred/
- Machine-reading suite: https://github.com/anuzzolese/machine-reading

MusicBO Knowledge Graph

- MusicBO Knowledge graph SPARQL endpoint: https://polifonia.disi.unibo.it/musicbo/sparql
- MusicBO Knowledge Graph GitHub repository: https://github.com/polifonia-project/musicbo-knowledge-graph
- MusicBO Knowledge Graph MELODY data story: https://github.com/polifonia-project/musicbo-knowledge-graph

Examples

- *The Propbank frame* `donate-01` *has* `giver` *as the label of the semantic role of the agent; also* `give-01` *has* `giver` *as the agent's label. Given that, in WordNet, the direct hypernym of the sysnset containing the lemma* `donate` *is the synset containing the lemma* `give`*, then the* `giver` *of* `donate-01` *can be considered a subclass of the* `giver` *of* `give-01`*.*
- *The PropBank frame* `say-01` *has* `sayer` *as agent (label), while* `tell-01` *has* `speaker`*. As the two predicates belong to the same WordNet's synset, the two roles might be considered 'pseudo-equivalent', namely part of the same semantic group.*
- *The PropBank roles' labels* `giver`*,* `sayer` *and* `speaker` *should be mapped to the most appropriate WordNet synset by applying word-sense disambiguation and semantic parsing techniques.*
  - *When the PropBank roles labels are phrases, such as in the case of the PropBank frame* `damages-02`*, in which one of the roles has, as a label,* "party that must pay compensation"*, Large Language Models can be used to support the mapping process.*
    - *Ex.:*
      - Prompt to Chat-GPT: *`Give me the wordnet id of the concept that best matches the following definition: "party that must pay compensation"`*
      - Chat-GPT output: `The WordNet ID for the concept that best matches the given definition is "debtor.n.01"`

Project n.3

Name: "Domain-Dependent Knowledge Graph Comparison through Selectional Restrictions"

Topic: Selectional Restrictions, Knowledge Graph Documentation and Explanations

Mentors:

- Luca Contalbo (UniBo)
- Arianna Graciotti (UniBo)
- Valentina Presutti

Description

The aim of this project is to develop a pipeline to compare and analyse **MusicBO knowledge graphs**. MusicBO is an **automatically generated** frame-based knowledge graph (KG) about the role of music in the city of Bologna developed within the **Polifonia** project. Each knowledge graph is obtained by applying a **text-to-KG pipeline** from textual data coming from different authors.

The project will use knowledge graphs obtained from different authors to develop a comparative pipeline by leveraging **domain-specific selectional restrictions**. Selectional restrictions are **probabilistic constraints** over the frame arguments. Thus, the analysis must take knowledge graphs from different authors and **highlight their differences** based on frame restrictions.

The resulting analysis must be exploited to produce **natural language explanations** (e.g. competency questions) through the use of large pre-trained language models. The resulting explanations will provide more insight into the topic discussed by the knowledge graphs, and a **comparison** highlighting the difference in the topics discussed by different authors.

Useful Resources

PropBank

- PropBank: https://github.com/propbank

Selectional Restrictions

- Domain-dependent selectional restriction extractor: https://github.com/lucacontalbo/selrestr_maker

Text-to-Knowledge Graph pipeline and support tools:

- Text-to-AMR-to-Fred: https://arco.istc.cnr.it/txt-amr-fred/
- Machine-reading suite: https://github.com/anuzzolese/machine-reading

MusicBO Knowledge Graph

- MusicBO Knowledge graph SPARQL endpoint: https://polifonia.disi.unibo.it/musicbo/sparql
- MusicBO Knowledge Graph GitHub repository: https://github.com/polifonia-project/musicbo-knowledge-graph
- MusicBO Knowledge Graph MELODY data story: https://github.com/polifonia-project/musicbo-knowledge-graph

Project n.4

Name: "Matching ontologies in the music domain"

Topic: Ontology Matching

Mentors:

- Valentina Carriero (UniBo)
- Jacopo de Berardinis (KCL)
- Valentina Presutti

Description

In the **Polifonia** project, we are developing a network of ontologies on the music domain. This ontology network (PON, link) is composed of different ontology modules, addressing a specific thematic area of the domain (e.g. a module is specifically related to bells and bell towers).

In order to support interoperability, such ontologies should be aligned to existing (music-related) ontologies.

Ontology matching (OM) can be defined as the process of finding correspondences (i.e. mappings, alignments) between entities belonging to different ontologies.

In this project, you will generate a set of alignments between the modules developed inside Polifonia and relevant state-of-the-art ontologies.

Relevant ontologies to be aligned to PON can be found in this deliverable (link), section 6.2

Project n.5

Name: “Influence and inspiration in musical compositions”

Topic: Musical Heritage

Mentors:

- **Nicolas Lazzari (UniBo)**
- **Luigi Asprino**
- **Valentina Presutti**

Description

The aim of this project is to build an Ontology and a Knowledge Graph that models the **influences** between different artists.

Musical compositions can often be analysed on the basis of its author **biography**: other famous composers that lived in the same area, mentors, collaborators, friends, etc. Those resources can help understand the *conscious* and *subconcious influences* that play a major role in the creative process of a music composer.

You will **develop an ontology** that models some of those aspects. The ontology will serve as a formal basis for the **creation of a KG** that links **structured sources**, such as from Wikidata, to **unstructured sources**, such as interviews to artists. You will perform an **entity linking task**, making sure that the entities of the KG are correctly linked to other existing resources. **LLM will be exploited as an extraction tool** producing results that can populate your KG.

A similar approach has been performed as part of the Linked Jazz project, where the authors manually transcribed interviews between famous jazz artists.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Project n.6

Name: “Ontology based Music Theory”

Topic: Comp. musicology

Mentors:

- **Nicolas Lazzari (UniBo)**
- **Andrea Poltronieri (UniBo)**
- **Valentina Presutti**

Description

The aim of this project is to develop an Ontology and a Knowledge Graph that models relevant theories from the musicology field.

Musicology is the field that studies the rudiments of music. A vast amount of theories have been proposed to analyse music. In this project, you will develop and ontology that models one of those theories (e.g. cadences, Negative Harmony, Tonefelder Theory) reusing relevant ontologies (e.g,. Music Theory Ontology).

The resulting ontology will be used to create a Knowledge Graph that enables the **analysis of musical compositions** contained in ChoCo.

A similar approach have been performed in the Modal Harmony Ontology and in the Functional Harmony Ontology.

Even though we will happily provide all the background knowledge required, being able to play an instrument combined with a good understanding of basic music theory can be very beneficial for the project.

Project n.7

Name: “Enriching Framester with Music-specific lexicon and alignment to PON”

Topic: Ontology and knowledge engineering

Mentors:

- **Valentina Presutti**
- **Arianna Graciotti**

Description

The aim of this project is to integrate the Polifonia Lexicon (an annotated subset of Babelnet) within Framester and to align it to the Polifonia Ontology Network (PON).

Framester already integrates Babelnet, hence the work will focus on enriching the subset corresponding to the Polifonia lexicon with provenance information and additional data provided by experts annotators.

The Polifonia Ontology Network supports modeling the music and music heritage domain. The Polifoinia Lexicon shall be aligned both at class/property level, to disambiguate the intentional meaning of Polifonia concepts, and as potential "controlled vocabularies" for populating PON classes that represent concepts e.g. genre, music role, etc.

The resulting resource will be used to enhance NLP methods for populating music heritage knowledge graphs.

Polifonia Ontology Network: https://github.com/polifonia-project/ontology-network
Polifonia Lexicon: https://github.com/polifonia-project/Polifonia-Lexicon
Framester: https://w3id.org/framester

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Project work in Knowledge Engineering

Project work porposal

- Cover song detection based on harmonic segment similarity
- Music structure segmentation combining audio and symbolic representation
- Foundational models for music embeddings in the symbolic domain: from NLP to MIR
  - The project will focus on learning representations from specific musical dimensions (melody, harmony, rhythm) by exploring a variety of embedding models (also on KGs); we will use a musicological testbed to verify to which extent the learned representations are musically plausible, and test the embeddings on several downstream tasks, including: music similarity, music generation, and music classification
- TRANSLATING COMPETENCY QUESTIONS INTO SPARQL QUERIES FOR MUSICAL HERITAGE KNOWLEDGE GRAPHS VALIDATION
  - Developing methods and tools for validating musical heritage knowledge graphs extracted from text by translating Competency Questions provided by musicologists and domain experts into SPARQL queries. The SPARQL queries can be coded by students or automatically produced by experimenting with text-to-SPARQL technologies. They will be used to interrogate knowledge graphs automatically generated from text, aiming to probe their effectiveness and informativeness.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

www.unibo.it