Internships at SINTEF (Oslo, Norway)

Context and high-level objectives of internships at SINTEF.

SINTEF is one of Europe's largest independent research organizations. The internship at SINTEF is framed within a cooperation with the INSID&S Lab at University of Milan-Bicocca on solutions to simplify semantic data enrichment pipelines at scale by combining semantic technologies, machine learning, human-computer interaction and big data management techniques [1,2]. The cooperation is framed within enRichMyData, a large Horizon Europe innovation action starting in October 2022, which involves more than 10 partners across Europe (with a majority of companies). Data enrichment pipelines, which include data linking and data extension operations, are specified with web applications [3,4,5] on sample data and then executed on large data sets, possibly with human-in-the-loop validation steps. During the stage the student will have the opportunity to work in a collaborative environment and interact with SINTEF, UNIMIB's and other partners.

[1] Michele Ciavotta, Vincenzo Cutrona, Flavio De Paoli, Nikolay Nikolov, Matteo Palmonari, Dumitru Roman: Supporting Semantic Data Enrichment at Scale. Technologies and Applications for Big Data Value 2022: 19-39
[2] Vincenzo Cutrona, Flavio De Paoli, Aljaz Kosmerlj, Nikolay Nikolov, Matteo Palmonari, Fernando Perales, Dumitru Roman: Semantically-Enabled Optimization of Digital Marketing Campaigns. ISWC (2) 2019: 345-362
[3] Vincenzo Cutrona, Michele Ciavotta, Flavio De Paoli, Matteo Palmonari: ASIA: a Tool for Assisted Semantic Interpretation and Annotation of Tabular Data. ISWC (Satellites) 2019: 209-212
[4] Dumitru Roman [++]: DataGraft: One-stop-shop for open data management. Semantic Web 9(4): 393-411 (2018)
[5] Marco Ripamonti, Flavio De Paoli, Matteo Palmonari: SemTUI: a Framework for the Interactive Semantic Enrichment of Tabular Data. CoRR abs/2203.09521 (2022)

Topic 1 (data engineering). Specs2Transform: from the specification of semantic annotations of tabular data to ETL pipelines for big data

Internship objectives. This topic concerns the design and development of solutions to transform specifications of data enrichment pipelines into executable pipelines. When annotations are specified using a web application, the specifications must be converted in data transformation workflows applied to large data sets. These pipelines must be executed in Big Data management frameworks, support specific features of the semantic data extension task, and therefore address related challenges such as the following:

- Preserve corrections in the data sample
- Support revision of links estimated by the algorithm (data management + ready for user interface)
- Support revision of subsequent data extensions operations for revised links (data management + ready for user interface)

- Support configuration of hyperparameters of the data enrichment algorithms (data management + ready for user interface)
- Support architecture for scalable (for Big Data) data extension framework based on linking

During the internship, the student will focus on one or more of the above mentioned problems, based on project priorities and the student's skills/inclinations.

Topic 2 (evaluation and use cases) - Evaluation of the impact of semantic enrichment on downstream analytics.

Internship objectives. This topic concerns the design and development of data enrichment solutions on business cases with the goal of evaluating the effectiveness and limitations of current solutions, analyzing technical requirements, and, especially, evaluating the effectiveness of the proposed solutions on downstream business analytics tasks. Example of such activities include: find relevant business cases (e.g., based on kaggle challenges), study the coverage of the existing solutions for the selected business cases, specify and execute dedicated enrichment pipelines, develop/improve enrichment services to fill in the gap, evaluate the cost of these operations and their benefits in the business task (e.g., improvement of predictions with enriched information).

Duration / exp. background / funding / how to apply

Duration: 3-6 months (6 is better)
Priority: Topic 1 > Topic 2
Background:

- BA in CS or similar preferred (or, alternatively, good programming/data engineering skills)
- Courses with topics related to Semantic Web, Big Data Management much preferred (esp. for topic 1)

The research stay can be associated with a MA thesis topic
Funding:

- recommended Erasmus+ Traineeship scholarship (check the deadlines!)
- SINTEF grant (~1200€ per month) can be associated for merit after interviews

If interested in applying:

- Contact Matteo Palmonari (matteo.palmonari@unimib.it) by email with (your email will be forwarded to colleagues at SINTEF):
    - A short statement with motivation and background
    - A CV in English