EN3.WEB DATA CAPTURE AND INTEGRATION, WEB DATA MINING, TEXT MINING, SEMANTIC WEB

Course title and code: EN3.Web Data Capture and Integration, Web Data Mining, Text Mining, Semantic Web.

Teacher and contact details: Muriel Foulonneau

Public Research Centre Henri Tudor
29, av. John F. Kennedy
L1855 Luxembourg
Luxembourg
Tel: +352 4259912757
Email: muriel.foulonneau@gmail.com, muriel.foulonneau@tudor.lu

Number of ECTS: 5

■ **Objectives**
The objective of this course is to introduce students to the techniques which can be used to take advantage of a variety of non reactive sources. It will provide factual knowledge on these techniques, as well as procedural knowledge on the techniques which should be used according to the situations.

■ **Competences** (basics, general, transversals, specifics)
The core competence which students will acquire is the awareness of the techniques to use non reactive sources. They will acquire competences related to the use of textual document sets, structured datasets, and semantic datasets.

■ **Programme**

Data can be captured from non reactive sources, i.e., from existing datasets or sources rather than through questionnaires or interactions with interviewees. Data can be collected from the Web (Web mining), including unstructured data (text mining) and structured data (data mining) which can be interpreted automatically (Semantic Web). This entails the use of various tools and methodologies to capture or query datasets and Websites, to select relevant data on a large scale, finally, to process the data, while ensuring the highest level of data quality. The course will introduce students to methodologies and tools used for capturing data, integrating data sources, mining datasets, whether texts or structured data, and processing them. It will provide practical examples of the challenges raised to interpret the data captured from the Web.

The course will include the following core topics:

- Review of non reactive sources
- Web mining techniques
- Basics of data integration
- Data quality
- Principles of data mining
- Text mining
- Semantic Web mining

■ **Expected learning outcomes**
Students will acquire competences on Web mining, including basic scripting technique, the use of techniques for text and data mining, including the process of data cleaning and data quality control, the type of extractions which can be performed and the challenges of these approaches. They will know the principle of the Semantic Web, as well as existing tools and mechanisms to access and take advantage of semantic datasets. At the end of the course, the students are expected to be able to define a data collection strategy based on various types of data available.

■ **Methodology**
The course will be organized with a series of lectures and applied exercises to capture and use different types of data.

■ **Evaluation system**
Students will be evaluated based on online exam which will test both factual and conceptual knowledge.

■ **Remarks** (previous requirements, coordination, others, if any)
In order to take this course, students must have previously coursed the “C3. Web data collection methods III: non-reactive data collection” module

■ **Online resources** (optional)

■ **Bibliography** (optional)

■ **Employment opportunities** (optional)
Information manager, Web communication