COURSE DESCRIPTION – ACADEMIC YEAR 2022/2023

Course title	**Data Curation**
Course code	73056
Scientific sector	ING-INF/05
Degree	Master in Computational Data Science (LM-18)
Semester	1
Year	2
Credits	12
Modular	Yes

Total lecturing hours	80
Total lab hours	40
Attendance	Attendance is not compulsory, but non-attending students have to contact the lecturers at the start of the course to agree on the modalities of the independent study.
Prerequisites	For the Data Integration module: Knowledge of relational databases, as taught in an introductory course at the BSc level. Basic knowledge of first-order logic, as taught in a BSc course in logic or discrete mathematics. Knowledge of Java or Python for the project part. For the Data Profiling module: relational database concepts, basic machine learning concepts; good to have knowledge of basic information retrieval concepts. Python programming basics for the lab sessions.
Course page	https://ole.unibz.it/

Specific educational objectives	The course belongs to the type "caratterizzanti – discipline informatiche" in the curriculum “Data Analysis”. The *Data Integration* module addresses a variety of problems related to the integration of heterogenous data sources. It overviews the main issues in data integration, notably handling different forms of heterogeneity, and presents the general architecture of data integration systems. Foundational techniques for data integration are covered, such as data matching, schema matching and mapping, and query processing in data integration. A specific data integration approach relying on the technology of Virtual Knowledge Graphs and semantic mappings is presented in detail. The integration both of relational data sources, and of other types of data sources accessed by relying on data federation technology are considered. By attending the course, students will learn how to design and build a comprehensive data integration solution, possibly exploiting existing data access and data federation technologies. The *Data Profiling* module considers a variety of problems related to the profiling of structured (tabular) and unstructured (multimedia) data. It first overviews techniques for the exploratory analysis of relational databases, which are also adopted in preparatory activities of data cleansing and integration. Algorithms for discovering patterns and dependencies in tabular data will be presented in detail. It then expands to multimedia databases, where profiling metadata must first be mined from images, videos, text and sounds. By studying the multimedia domain, students will learn how statistical learning and

	soft computing can be applied to discover semantically meaningful concepts from raw unstructured data.

Module 1	**Data Preparation and Integration**
Module code	73056A
Module scientific sector	ING-INF/05
Lecturer	Diego Calvanese
Contact	Office BZ P.2.07, diego.calvanese@unibz.it, +39 0471 016160
Scientific sector of lecturer	ING-INF/05
Teaching language	English
Office hours	Announced on the webpage of the lecturer. During the lecture time span, in general Friday 16:00-18:00. Outside of the lecture time span, students are advised to confirm availability by email.
Lecturing assistant (if any)	--
Contact LA	--
Office hours LA	--
Credits	6
Lecturing hours	40
Lab hours	20
List of topics	• Data integration architectures • Query processing in data integration • Schema mapping • Data integration via virtual knowledge graphs • Schema matching • Data and entity matching
Teaching format	Frontal lectures, exercises, and labs.

Module 2	**Data Profiling**
Module code	73056B
Module scientific sector	INF/01
Lecturer	Alessandro Mosca
Contact	Noi Techpark, Office BZ A1 4,29D, alessandro.mosca@unibz.it, +39 0471 016268
Scientific sector of lecturer	INF/01
Teaching language	English
Office hours	Announced on the webpage of the lecturer. Outside of the lecture time span, students are advised to confirm availability by email.
Lecturing assistant (if any)	--
Contact LA	--
Office hours LA	--
Credits	6
Lecturing hours	40
Lab hours	20
List of topics	• Detecting patterns and violations • Detecting dependencies • Scrubbing and normalization • Similarity measures • Duplicate detection • Summary extraction The course will be structured in units that will cover column analysis in relational databases, dependency discovery in tabular data, feature

	and knowledge representation for multimedia data, concept discovery in image database, video database and audio database.
Teaching format	Frontal lectures and labs

Learning outcomes	Knowledge and understanding: • D1.1 - Knowledge of the key concepts and technologies of data science disciplines • D1.2 - Understanding of the skills, tools and techniques required for an effective use of data science • D1.6 - Knowledge of the principles and methods of data curation Applying knowledge and understanding: • D2.1 - Practical application and evaluation of tools and techniques in the field of data science • 2.5 - Ability to apply, evaluate and develop methods and tools for the integration, cleaning, and quality of data • 2.10 - Application of languages, tools, and methods for the design of information systems and their corresponding software applications for data, process, and organization management Making judgments • D3.2 - Ability to autonomously select the documentation (in the form of books, web, magazines, etc.) needed to keep up to date in a given sector Communication skills • D4.1 - Ability to use English at an advanced level with particular reference to disciplinary terminology • D4.3 - Ability to structure and draft scientific and technical documentation Learning skills • D5.2 - Ability to autonomously keep oneself up to date with the developments of the most important areas of data science

Assessment	Oral exam and project work. The mark for each part of the exam is 18-30, or insufficient. The oral exam covers Module 1 “Data Integration” and Module 2 “Data Profiling”, and comprises verification questions, and open questions to test knowledge application skills. It counts for 50% of the total mark. The project consists of two parts. Part 1 covers Module 1 “Data Integration”, and verifies whether the student is able to apply advanced data integration techniques and technologies taught or presented in the course to solve concrete problems. It is assessed through a final presentation, a demo, and a project report and can be carried out either individually or in a group of 2 students. It is discussed during the oral exam, and it counts for 25% of the total mark. Part 2 covers Module 2 “Data Profiling” and verifies whether the student is able to apply data profiling techniques to use cases from the multimedia and relational database domains. It is assessed

	through a final presentation and a project report and can be carried out either individually or in a group of 2 students. It is discussed during the oral exam, and it counts for 25% of the total mark.
Assessment language	English
Assessment Typology	Monocratic
Evaluation criteria and criteria for awarding marks	The final mark is computed as the weighted average of the oral exam, Part 1 of the project, and Part 2 of the project. The exam is considered passed when all three marks are valid, i.e., in the range 18-30. Otherwise, the individual valid marks (if any) are kept for all 3 regular exam sessions, until also all other parts are completed with a valid mark. After the 3 regular exam sessions, all marks become invalid. Relevant for the oral exam: clarity of answers; ability to recall principles and methods, and deep understanding about the course topics presented in the lectures; skills in applying knowledge to solve exercises about the course topics; skills in critical thinking. Relevant for the project: skill in applying knowledge in a practical setting; ability to summarize in own words; ability to develop correct solutions for complex problems; ability to write a quality report; ability in presentation; ability to work in teams. Non-attending students have the same evaluation criteria and requirements for passing the exam as attending students.

Required readings	Required books: • A. Doan, A. Halevy & Z. Ives (2012). Data Integration. Morgan Kaufmann. (ST270 D631) • Z. Abedjan, L. Golab, F. Naumann, & T. Papenbrock (2018). Data Profiling. Synthesis Lectures on Data Management, 10(4), 1-154. • Z. Zhang, R. Zhang (2008). Multimedia Data Mining: A Systematic Introduction to Concepts and Theory. CRC Press LLC. Additional material (slides, notes of the lecturers) will be made available before each lesson. Subject Librarian: David Gebhardi, David.Gebhardi@unibz.it
Supplementary readings	• R. C. Gonzalez, & R. E. Woods (2007). Image processing. Digital image processing, 2, 1. • R. O. Duda, P. E. Hart & D. G. Stork (2012). Pattern classification. John Wiley & Sons.
Software used	• Ontop system for ontology-based data access developed by the In2Data research group at the Faculty of Computer Science. • Relational DBMS, such as PostgreSQL. • Data federation tools such as Denodo, Dremio, Teiid. • Python3.5 with pandas, scikit-learn, scikit-image.