ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

105000453 - Artificial Intelligence And Open Science In Research Software Engineering

DEGREE PROGRAMME

10II - Grado En Ingenieria Informatica

ACADEMIC YEAR & SEMESTER

2022/23 - Semester 2

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	105000453 - Artificial Intelligence And Open Science In Research Software Engineering
No of credits	3 ECTS
Type	Optional
Academic year ot the programme	Third year
Semester of tuition	Semester 6
Tuition period	February-June
Tuition languages	English
Degree programme	10II - Grado en Ingenieria Informatica
Centre	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Academic year	2022-23

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Oscar Corcho Garcia (Subject coordinator)	D-2209	oscar.corcho@upm.es	Sin horario.

* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

3. Prior knowledge recommended to take the subject

3.1. Recommended (passed) subjects

The subject - recommended (passed), are not defined.

3.2. Other recommended learning outcomes

- Inteligencia Artificial

- Web Semántica y Grafos de Conocimientos

- Ingeniería del software

4. Skills and learning outcomes *

4.1. Skills to be learned

CG-13/CE55 - Capacidad de comunicarse de forma efectiva con los compañeros, usuarios (potenciales) y el público en general acerca de cuestiones reales y problemas relacionados con la especialización elegida.

CG-2/CE45 - Capacidad para el aprendizaje autónomo y la actualización de conocimientos, y reconocimiento de su necesidad en el área de la informática.

CG-7:10/16/17 - Capacidad para trabajar dentro de un equipo, organizando, planificando, tomando decisiones, negociando y resolviendo conflictos, relacionándose, y criticando y haciendo autocrítica

Ce 17 - Conocer los temas informáticos avanzados de modo que permita a los alumnos vislumbrar y entender las fronteras de la disciplina, por medio de la inclusión de experiencias de aprendizaje que dirigen a los alumnos desde los temas elementales a los temas avanzados o los temas de los que se nutren los novísimos desarrollos.

Ce 44 - Conocimiento de tecnologías punteras relevantes y su aplicación.

4.2. Learning outcomes

RA551 - Basic knowledge of parallelization techniques for efficient execution

RA554 - Ability to apply cutting edge machine learning techniques to help classify, recommend and organize research software and data

RA549 - Ability to describe metadata and provenance for research software and data in machine-readable formats

RA556 - Ability to clean, integrate and exploit data from knowledge graphs

RA550 - Ability to create software containers that can run software components in different computational infrastructures

RA553 - Ability to create knowledge graphs of research software and data

RA557 - Ability to read, understand and implement research publications

RA555 - Ability to create an abstract sketch of a research method

RA558 - Ability to read, understand and implement standard recommendations and guidelines (e.g., international committees, World Wide Web Consortium (W3C), etc.)

RA552 - Ability to identify and address real world problems where AI techniques applied to research software engineers can help

* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

5. Brief description of the subject and syllabus

5.1. Brief description of the subject

El objetivo de este curso es que los estudiantes aprendan los fundamentos de la IA en la Ingeniería de Software Científico, con un enfoque especial en aplicaciones del mundo real. El curso incluirá conocimientos teóricos y metodológicos sobre técnicas avanzadas de IA diseñadas para ayudar a la reproducibilidad y reutilización de software, datos y sus metadatos en el ámbito de investigación, con aplicaciones en industria. Específicamente, el curso abordará la importancia de la reproducibilidad en ciencia, enfoques existentes para administrar, empaquetar y entregar un producto de investigación, la planificación y ejecución en paralelo de experimentos computacionales complejos, la creación de grafos de conocimiento para facilitar la búsqueda y la configuración de software científico; así como técnicas de aprendizaje automático para identificar similitudes entre software. La asignatura presentará aplicaciones que combinan todo lo anterior para facilitar la integración y comprensión de datos y software científicos.

Este curso ampliará el conocimiento aprendido en las asignaturas "Inteligencia artificial", "Web semántica, datos enlazados y grafos de conocimiento" e "Ingeniería de software" (I y II)

The objective of this course is for students to learn the foundations of AI in Research Software Engineering, with a special focus on real-world applications. The course will include theoretical and methodological knowledge on cutting edge AI techniques designed to aid the reproducibility, repurpose and reuse of research software, data and their metadata. More specifically, the course will address the importance of reproducibility in Science, approaches for managing, containerizing and delivering a research product, the role of planning and parallelization in computational-heavy experiments, the creation of knowledge graphs to ease research software findability and set up; the role of machine learning to help identify similarities between different code bases and applications that combine all of the above to ease research data and software integration and understanding.

This course will expand the knowledge learnt in the subjects ?Artificial Intelligence?, ?Semantic Web, Linked Data and Knowledge graphs? and ?Software Engineering? (I and II)

5.2. Syllabus

1. Introducción a la Ingeniería del Software de Investigación / Introduction to Research Software Engineering

    1.1. Motivación: Reproducibilidad y los principios FAIR para datos y software científicos / Motivation: Reproducibility and the FAIR principles for research data and software

    1.2. Software y preservación de datos: repositorios de software y registros de metadatos/ Software and data conservation: code repositories and metadata registries

    1.3. Iniciativas abiertas para gestionar datos y software / Open initiatives for managing Data and Software

    1.4. Descripción de datos y software para reproducibilidad y reutilización / Describing data and software for reproducibility and reuse

2. Métodos computationales / Computational scientific methods

    2.1. Notebooks / Computational notebooks

    2.2. Infraestructura computacional y contenedores de software / Computational infrastructure and software containers

    2.3. Flujos de trabajo científicos / Scientific workflows

    2.4. Experimentos a gran escala: planificación y paralelización / Large-scale experiments: planning and parallelization

    2.5. Composición de flujos de trabajo y razonamiento / Workflow composition and reasoning

    2.6. Aprendiendo a manejar Infraestructuras abiertas de investigación / Getting started with open research infrastructures.

3. Provenance en Ingeniería de Software de Investigación / Provenance in Research Software Engineering

    3.1. Introducción a provenance / Introduction to provenance

3.2. El estándar W3C PROV / The W3C PROV standard

3.3. Repositorios abiertos para recolectar provenance / Open repositories for collecting provenance

3.4. Aplicaciones para capturar y usar provenance en la Web / Applications for capturing and exploiting provenance in the Web.

4. Grafos de Conocimiento Científicos / Scientific Knowledge Graphs

4.1. Representación de Datos y software procesable por máquinas / Machine-readable data and software representation

4.2. Capturando el contexto de un experimento: Objetos de Investigación / Capturing the context of research: Research Objects

4.3. Aplicaciones y ejemplos sobre grafos de Conocimento Científicos / Applications for Scientific Knowledge Graphs

5. Aprendizaje automático en Ingeniería del Software de Investigación / Machine Learning for Research Software Engineering

5.1. Clustering y clasificación de software científico y sus metadatos / Clustering and classification of research software and its metadata

5.2. Reconocimiento de Entidades en software científico / Named entity recognition in research software

5.3. Aplicaciones de aprendizaje automático en Ingeniería del Software de Investigación / Applications of machine learning models in Research Software Engineering

6. Schedule

6.1. Subject schedule*

Week	Classroom activities	Laboratory activities	Distant / On-line	Assessment activities
1	**Themes 1.1 to 1.4** Duration: 01:15 Lecture	**Themes 1.1 to 1.4** Duration: 00:45 Laboratory assignments		
2	**Themes 2.1 and 2.2** Duration: 01:00 Lecture	**Themes 2.1 and 2.2** Duration: 01:00 Laboratory assignments		
3	**Theme 2.3** Duration: 01:00 Lecture	**Theme 2.3** Duration: 01:00 Laboratory assignments		
4	**Theme 2.4** Duration: 01:00 Lecture	**Theme 2.4** Duration: 01:00 Laboratory assignments		
5	**Theme 2.5** Duration: 01:00 Lecture	**Theme 2.5** Duration: 01:00 Laboratory assignments		**Create a computational experiment, containerize it, and describe its provenance** Individual work Continuous assessment Not Presential Duration: 15:00
6	**Theme 2.6** Duration: 00:30 Lecture	**Theme 2.6** Duration: 01:30 Laboratory assignments		
7	**Themes 3.1 to 3.4** Duration: 01:15 Lecture	**Themes 3.1 to 3.4** Duration: 00:45 Laboratory assignments		
8	**Theme 4.1** Duration: 01:00 Lecture	**Theme 4.1** Duration: 01:00 Laboratory assignments		
9	**Theme 4.2** Duration: 00:30 Lecture	**Theme 4.2** Duration: 01:30 Laboratory assignments		**Design and implement an initial version of a scientific knowledge graph using and relating the computational experiments and integrating data and software from at least two open code repositories** Group work Continuous assessment Not Presential Duration: 15:00
10	**Theme 4.3** Duration: 00:30 Lecture	**Theme 4.3** Duration: 01:30 Laboratory assignments		

11		**Homework discussion and questions** Duration: 02:00 Problem-solving class		
12	**Theme 5.1** Duration: 00:30 Lecture	**Theme 5.1** Duration: 01:30 Laboratory assignments		
13	**Theme 5.2** Duration: 01:00 Lecture	**Theme 5.2** Duration: 01:00 Laboratory assignments		**Enhance a scientific knowledge graph with AI techniques and explain them.** Group work Continuous assessment Not Presential Duration: 20:00
14	**Theme 5.3** Duration: 01:00 Lecture	**Theme 5.3** Duration: 01:00 Laboratory assignments		
15		**Homework discussion and questions** Duration: 02:00 Problem-solving class		
16	**Final group presentations** Duration: 02:00 Additional activities			**Final group presentation** Group presentation Continuous assessment Presential Duration: 02:00
17	**Written test** Duration: 02:00 Additional activities			**Final exam (written test)** Written test Continuous assessment and final examination Presential Duration: 02:00

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

* The schedule is based on an a priori planning of the subject; it might be modified during the academic year, especially considering the COVID19 evolution.

7. Activities and assessment criteria

7.1. Assessment activities

7.1.1. Assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
5	Create a computational experiment, containerize it, and describe its provenance	Individual work	No Presential	15:00	20%	5 / 10	CG-2/CE45 Ce 17 Ce 44
9	Design and implement an initial version of a scientific knowledge graph using and relating the computational experiments and integrating data and software from at least two open code repositories	Group work	No Presential	15:00	12.5%	5 / 10	CG-13/CE55 Ce 17 Ce 44 CG-7:10/16/17
13	Enhance a scientific knowledge graph with AI techniques and explain them.	Group work	No Presential	20:00	12.5%	5 / 10	Ce 44 CG-7:10/16/17 CG-13/CE55 Ce 17
16	Final group presentation	Group presentation	Face-to-face	02:00	10%	5 / 10	CG-13/CE55 Ce 17 Ce 44 CG-7:10/16/17
17	Final exam (written test)	Written test	Face-to-face	02:00	45%	5 / 10	Ce 44 CG-2/CE45 Ce 17

7.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
17	Final exam (written test)	Written test	Face-to-face	02:00	45%	5 / 10	Ce 44 CG-2/CE45 Ce 17

7.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

7.2. Assessment criteria

Se evaluarán los siguientes elementos, con los pesos que se especifican a continuación:
1. El material proporcionado por el alumno y la interacción con el repositorio de la asignatura sobre el trabajo propuesto en los temas 2 y 3 (peso: 20 %)

2. El material proporcionado por el alumno y la interacción con el repositorio de la asignatura sobre el trabajo propuesto en el tema 4 (peso: 12.5 %)

3. El material proporcionado por el alumno y la interacción con el repositorio de la asignatura sobre el trabajo propuesto en el tema 5 (peso: 12.5 %)

4. El papel del alumno en la presentación y su capacidad para responder preguntas sobre el trabajo del proyecto (peso: 10 %)

5. Examen escrito final (peso: 45%)

Los estudiantes han de aprobar las tareas asignadas para poder hacer el examen escrito (con peso 45%).

*****

The following items will be evaluated, with the weights that are specified next:

1. The material provided by the student and the interaction in the course code and materials repository on the work proposed in section 2 and 3 (weight: 20%)

2. The material provided by the student and the interaction in the course code and materials repository on the work proposed in section 4 (weight: 12.5%)

3. The material provided by the student and the interaction in the course code and materials repository on the work

proposed in section 5 (weight: 12.5%)

4. The role of the student on the presentation and the ability to answer questions about the project work (weight: 10%)

5. Final written exam (weight: 45%)

Note that all assignments must be passed with a minimum score in order to attend the final written test.

8. Teaching resources

8.1. Teaching resources for the subject

Name	Type	Notes
? The Scientific Paper of the future	Bibliography	The Scientific Paper of the future: https://www.scientificpaperofthefuture.org/
Jim Gray on eScience: A Transformed Scientific Method	Bibliography	Jim Gray on eScience: A Transformed Scientific Method: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf
Introduction to containers	Web resource	https://biocontainers-edu.biocontainers.pro/en/latest/introduction.html
Linked Data: Evolving the Web into a Global Data Space (1st edition).	Bibliography	Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). . Morgan & Claypool. Available from http://linkeddatabook.com/editions/1.0/

PROV Model primer	Bibliography	PROV Model primer: https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/
An introduction to Scikit Learn	Web resource	An introduction to Scikit Learn: https://scikit-learn.org/stable/tutorial/basic/tutorial.html

9. Other information

9.1. Other information about the subject

The course will be taught in English, following a hands-on style of learning. Given the practical nature of the course, students are required to bring their laptops to work during most of the lectures.

All the tools described and taught in this course are open source.