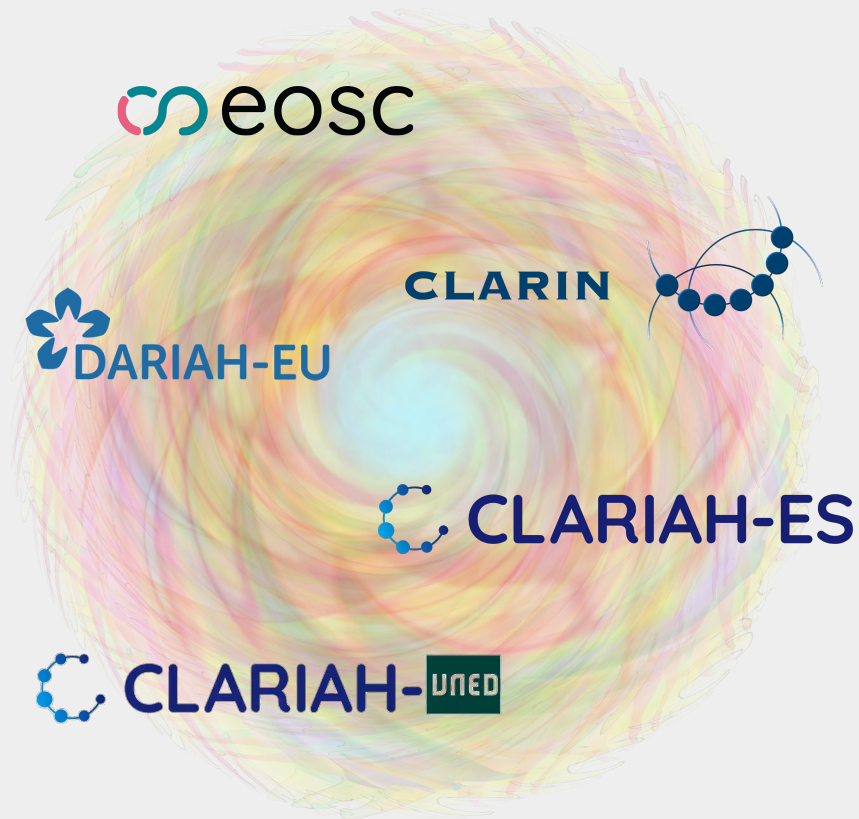


# CLARIN y DARIAH

**Cómo potenciar tu investigación con los recursos de las infraestructuras europeas**

Marina Míguez Lamanuzzi

# Infraestructuras nacionales y europeas



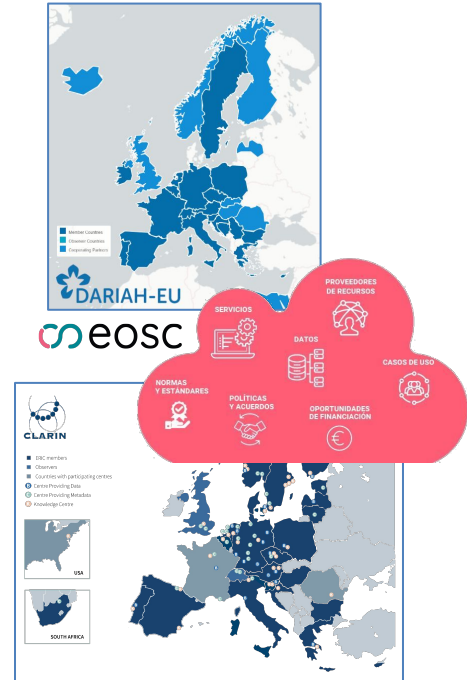
## SUBNACIONAL



## NACIONAL

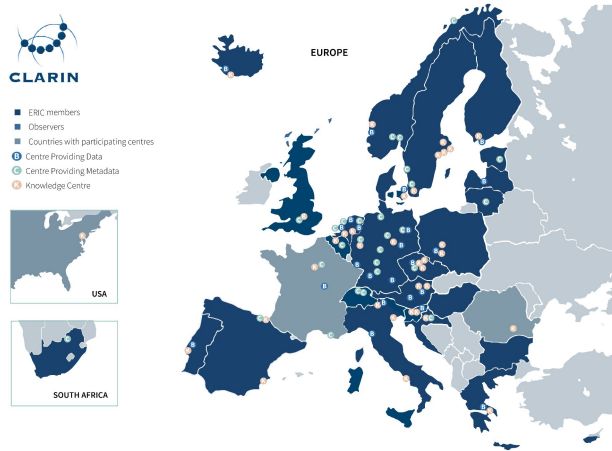


## INTERNACIONAL



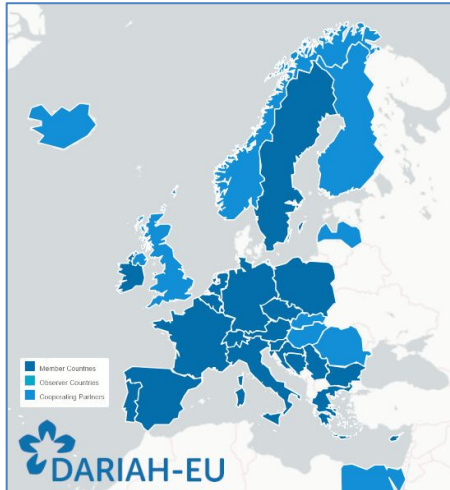
- **CLARIAH-CM** es un nodo dentro de CLARIAH-ES compuesto por las seis universidades públicas madrileñas. Lo mismo que p. ej. CLARIAH-UNED.
- **CLARIAH-ES** es una infraestructura de investigación digital nacional constituida por doce nodos que coordinan la participación de España en CLARIN y DARIAH.
- **DARIAH** es una infraestructura europea de Investigación Digital para las Artes y las Humanidades que tiene como objetivo mejorar y apoyar la investigación y la enseñanza digitales en las artes y las humanidades.
- **CLARIN** es una infraestructura digital que ofrece datos, herramientas y servicios para apoyar la investigación basada en recursos lingüísticos.
- **EOSC** (European Open Science Cloud) es un «sistema de sistemas» construido a partir de agregadores que recopila y promueve la Ciencia Abierta

# CLARIN-ERIC



- **CLARIN (Common Language Resources and Technology Infrastructure)** es una **infraestructura digital** amparada por **ERIC (European Research Infrastructure Consortium)**.
- Proporciona **acceso fácil y sostenible** a una amplia gama de **datos lingüísticos digitales multimodales** (texto, audio, vídeo, etc.) y **herramientas lingüísticas** avanzadas con las que explorar, analizar o combinar estos conjuntos de datos.
- Tiene centros participantes en toda Europa y más allá (26 consorcios nacionales miembro y centros participantes en países no miembro).
- Las herramientas y los datos de los distintos centros son **interoperables**, de modo que se pueden combinar colecciones de datos y encadenar herramientas de distintas fuentes.
- Muchos de los recursos (pero no todos) son también de **acceso abierto** para otras comunidades de uso interesadas, tanto dentro como fuera del mundo académico.

# DARIAH-EU



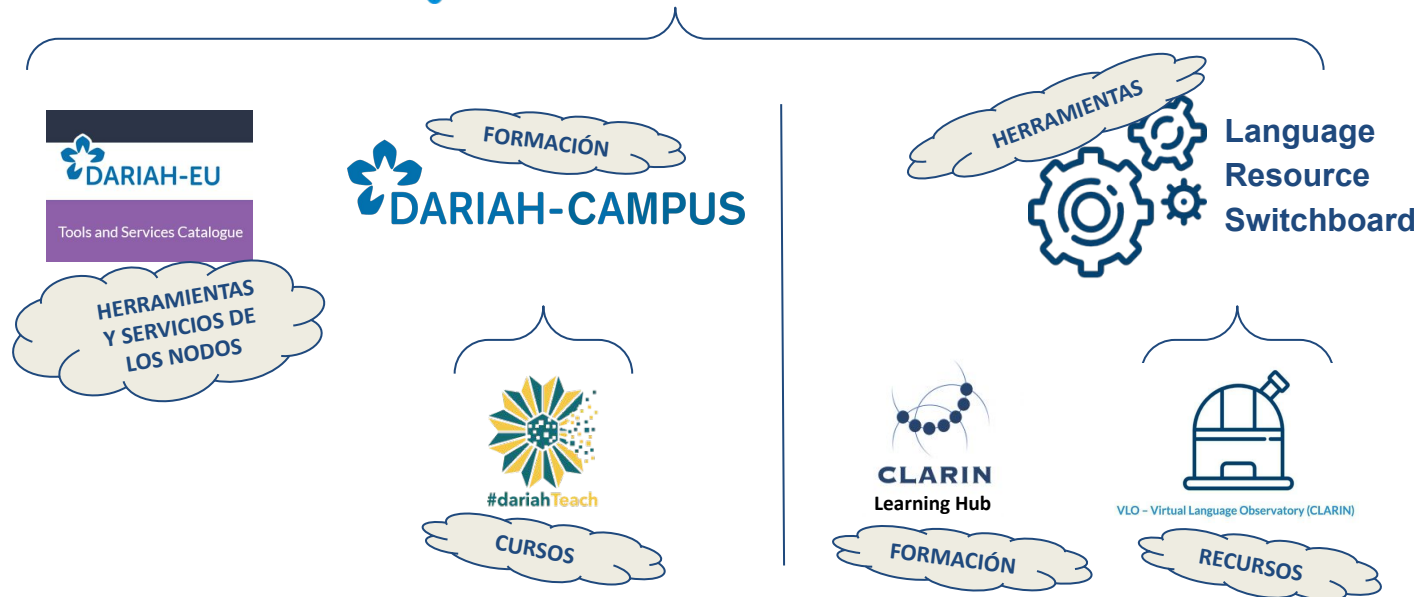
- **DARIAH (Digital Research Infrastructure for the Arts and Humanities)** es una infraestructura europea que busca potenciar con **tecnología digital** la investigación y la docencia en **Artes y Humanidades**.
- Es una **red de personas** interesadas en potenciar su conocimiento con métodos, herramientas y tecnologías digitales de 22 países miembros (y 17 socios colaboradores en 10 países no miembro).
- Impulsa la **producción, preservación e intercambio de datos, servicios y herramientas computacionales multimodales**, garantizando su accesibilidad y difusión a largo plazo.
- Ofrece **material didáctico y oportunidades de formación** para facilitar el desarrollo de habilidades de investigación digital, así como la adopción de buenas prácticas a nivel transnacional y transdisciplinar.

# Una red de recursos



VLO - Virtual Language Observatory (CLARIN)





# Repositorios

- Los repositorios son espacios centralizados donde se **almacena, organiza, mantiene y difunde** información digital.
- Son útiles para compartir la investigación publicada o inédita, los datos de investigación, las herramientas desarrolladas, etc.
- **En este sentido, son una buena herramienta para gestionar tus datos según los PRINCIPIOS FAIR (*Findable, Accessible, Interoperable, and Reusable*).**
- Sirven también como un medio de publicación para la 'vía verde' de publicación.
- Hay distintos tipos de repositorios:
  - ZENODO: acepta todo tipo de dato de investigación y proporciona DOI.
  - GitHub: para cuadernos de programación, librerías y software.
  - ArXiv: para subir *preprints* de artículos.
  - Docta Complutense: para cumplir con los requisitos institucion



<https://zenodo.org/>



<https://github.com/>



<https://docta.ucm.es/>



<https://arxiv.org/>





- Es un **agregador** (= una plataforma que recolecta y organiza el contenido de un gran número de fuentes en línea) conectado a una serie de fuentes de las redes CLARIN y DARIAH (entre muchas otras).
- Este agregador cruza los metadatos de proyectos de Humanidades y Ciencias Sociales digitales insertos en las **plataformas conectadas, a los que se suman muchos otros que han sido añadidos manualmente** por usuarios y moderadores de la plataforma.
- Es importante entender que un agregador enlaza los recursos de distintas fuentes, pero **NO** los aloja directamente ni proporciona por sí mismo herramientas ni servicios.
- Su gran utilidad reside en localizar y dar a conocer **recursos formativos, conjuntos de datos, herramientas y servicios, flujos de trabajo y publicaciones relacionadas con las herramientas.**
- Funciona como un buscador. Además, si tenemos herramientas, publicaciones, *workflows* o *datasets* que queremos enlazar para aumentar su visibilidad y uso, solo es necesario registrarse con una cuenta institucional, rellenar un formulario y esperar a que los moderadores lo aprueben e incluyan.

# Catálogos de herramientas

- Para encontrar todo tipo de herramientas y recursos de investigación.
  - [DARIAH Tools&Services](#): recursos creados dentro de la red DARIAH.
  - [CLARIN Language Resources](#): para datasets y recursos. CLARIN como tal no ofrece el servicio, sino distintos centros de la red.
  - [CLARIN Virtual Language Observatory](#): recursos lingüísticos asociados a la red CLARIN.
- Sirven para **encontrar** herramientas y servicios existentes que aprovecharlos, pero **NO** los almacenan.
- Funcionan con un buscador y filtros para seleccionar por tipo y categoría
- Los recursos y datasets suelen poderse descargar de manera gratuita y suelen tener una licencia que permite reutilizarlos en otras investigaciones.



<https://www.dariah.eu/tools-services/tools-and-services/>



VLO - Virtual Language Observatory (CLARIN)

<https://vlo.clarin.eu/?1>

# Language Resource Switchboard



Language  
Resource  
Switchboard

<https://switchboard.clarin.eu/>

- Es una aplicación de código abierto y en línea para la **exploración de recursos de análisis lingüístico**.
- Cuando subimos un archivo, **sugiere** herramientas para tratarlo, según el tipo de archivo y características (.txt, .xml, .tiff...).
- NO analiza, almacena ni difunde el material subido.
- **Es útil para** encontrar herramientas con las que procesar datasets, archivos y corpus lingüísticos.

Pasamos un texto .txt por el Switchboard

<https://switchboard.clarin.eu/>

# Recursos formativos

- Son un buen recurso para aprender de forma autodidacta.
- Se pueden encontrar recursos tanto mediante un buscador como accediendo al listado completo de materiales.
- Dentro de las estructuras CLARIN y DARIAH hay tres tipos:
  - [DARIAH CAMPUS](#)
  - [DARIAH TEACH](#)
  - [CLARIN Learning Hub](#)



<https://campus.dariah.eu>



<https://www.clarin.eu/content/learning-hub>



<https://teach.dariah.eu>

# Recursos formativos



<https://campus.dariah.eu>

- [DARIAH CAMPUS](#) **recolecta** recursos formativos.
  - Módulos de formación.
  - Grabaciones de eventos formativos.
  - *Pathfinders* para orientarse en determinados procesos.



## Resources

Learn about different topics  
with online resources  
provided by DARIAH



## Events

Missed a face-to-face  
DARIAH event? Check out  
what happened



## Pathfinders

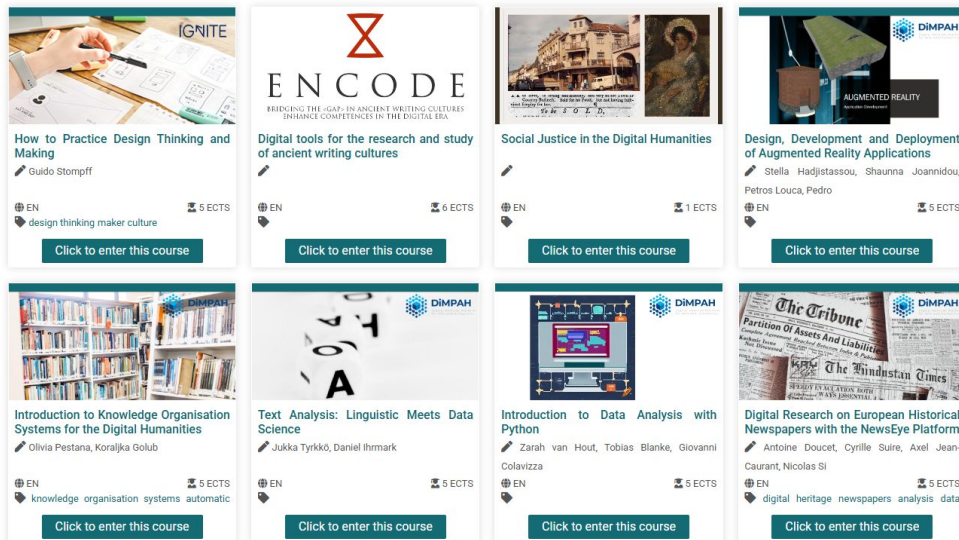
Collections of external  
resources curated by the  
DARIAH team

# Recursos formativos

- [DARIAH TEACH](#)
  - Plataforma de **MOOCs** (*Massive Open Online Course*)
  - Variedad de cursos introductorios estructurados
  - Temas específicos y sobre una herramienta/tecnología



<https://teach.dariah.eu>



<https://www.ucm.es/clariah-cm>

- [CLARIN Learning Hub](https://www.clarin.eu/content/learning-hub): combina **guía de uso** de CLARIN con Tutoriales y **materiales didácticos**.
  - Cuenta con tutoriales estructurados en módulos.
  - Encontramos tutoriales tanto de cuestiones generales (cómo funcionan los datos e infraestructuras) como de lo más particular sobre el uso y funcionamiento de CLARIN.
  - Incluye materiales sobre buenas prácticas investigadoras y archivo de materiales



<https://www.clarin.eu/content/learning-hub>



## Intro to CLARIN Tutorial

This course is useful to anyone who is new to CLARIN and is interested in learning how to use the central services for language research and sharing and archiving language resources.

# Recursos formativos

- [DARIAH-CAMPUS Course Registry](#) sirve para encontrar cursos de formación 'oficial' en Humanidades Digitales.
  - Proporciona un listado de cursos sobre Humanidades Digitales de varios países y formatos.
  - Tiene un buscador y filtros para seleccionar el tipo de curso, el idioma, formato (online, híbrido, presencial...).
  - Incluye tanto cursos de larga duración como escuelas de verano y similares.

 **DARIAH-CAMPUS > Course registry**

## DIGITAL HUMANITIES COURSE REGISTRY





# ¿Por dónde empiezo?

Recursos para la investigación filológica,  
lingüística y literaria presentes en  
CLARIN, DARIAH y SSHOC Marketplace

# Para introducirse en las HHDD

**Supuesto:** No tengo apenas conocimientos sobre Humanidades Digitales e Inteligencia artificial, por lo que quiero encontrar cursos y materiales introductorios para comenzar a entender estos campos. Tampoco sé en qué consiste exactamente el concepto de Ciencia Abierta, datos FAIR y los distintos tipos de publicaciones.

## Ejemplos de recursos que encontrarás en CLARIN y DARIAH:

### → DARIAH-TEACH:

[Introduction to Digital Humanities](#) (también traducido al español)

### → DARIAH-CAMPUS:

[Thinking With Machines: How Academics Can Use Generative AI Thoughtfully and Ethically](#)

[Introduction to Artificial Intelligence Prompt Engineering](#)

[Gold, Green, Diamond: What You Should Know About Open Access Publishing Models](#)

# Creación de BBDD y gestión de datos

**Supuesto:** necesito crear una base de datos y un plan de gestión de datos para mi investigación, pero no sé por dónde empezar ni qué herramientas me podrían ser útiles.

**Para aprender:**

→ **DARIAH-CAMPUS:**

[Data and Databases: An Introduction](#)

[Data and Databases: From Source to Data](#)

[Data and Databases: Data Management and Storage](#)

**Herramientas:** Si buscamos en el SSHOC Marketplace, encontramos herramientas para diseñar BBDD como

→ **SSHOC Marketplace:**

[Heurist](#)

# Etiquetado XML/TEI

**Supuesto:** Necesito crear una edición digital de la obra que estoy estudiando, en lenguaje xml. Además, me gustaría conocer el tipo de etiquetado TEI que puede ser más beneficioso para mi investigación de cara a analizar algún aspecto gramatical o sintáctico del mismo.

**Para aprender:**

→ **DARIAH-TEACH:**

[Text Encoding and the TEI](#) (también traducido al español)

[Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding](#)

**Herramientas:**

→ **DARIAH Tools&Services:**

[TextGrid Import Modeller](#): para estandarizar el *metadata file structure* del documento XML-TEI

[LEAF-Writer DARIAH](#): un editor online para XML-TEI

# Estilometría, análisis y minería de texto, CLARIAH-CM

## programación para filólogos

**Supuesto:** quiero comenzar a utilizar herramientas de programación y de PLN para llevar a cabo partes de mi investigación lingüística. He oído hablar de conceptos como *lematización*, *PoS tagging*, *tokenización*... pero no entiendo en qué consiste ni cómo podría llevar a cabo un análisis lingüístico con este tipo de procesamientos. Además, me he interesado por conceptos como *distant reading*, *topic modelling* y *sentiment analysis*, y me gustaría incorporar algo relacionado con esto a mi investigación.

### Para aprender:

→ DARIAH-TEACH:

[Text Analysis: Linguistic Meets Data Science](#): Casos de uso y *workflows* para *topic modelling*, *sentiment analysis*...

[Data Analysis with Python](#): una introducción para implementar *machine learning* a nuestro corpus de datos con Python.

# Estilometría, análisis y minería de texto, CLARIAH-CM

## programación para filólogos

Además, esto no se aplica sólo a la lingüística 'pura' sino que también tiene muchas aplicaciones para los Estudios Literarios:

### → **DARIAH-CAMPUS:**

[ExploreCor - Using Programmable Corpora in Computational Literary Studies](#)

[Understanding and Creating Word Embeddings](#)

[Corpus Analysis with spaCy](#)

[Sentiment Analysis with 'syuzhet' using R](#)

### → **CLARIN Learning Hub:**

[Introduction to Programming for NLP with Python](#)

[Natural Language Processing Methods](#)

# Estilometría, análisis y minería de texto, CLARIAH-CM

## programación para filólogos

Herramientas y librerías de programación:

→ DARIAH Tools&Services:

[Estilometría TIP](#)

[Sentiment Lexicons](#)

**Ejemplo de aplicación de análisis estilométrico en el ámbito de la Filología Hispánica:**

**[Digging for Gold - Knowledge Extraction from Text](#): *Stylometry applied to Old Spanish Poetry***

**Supuesto:** Quiero mejorar mis comunicaciones y artículos implementando gráficos y elementos de visualización de mis resultados de investigación, pero no conozco las herramientas y procedimientos para transformar de forma visual mis investigaciones.

**Para aprender:**

→ **DARIAH-CAMPUS:**

[Introduction to Network Analysis in the Humanities](#): para aprender a hacer grafos con Gephi

[Creating Interactive Visualizations with Plotly](#): para aprender a programar gráficos con Python

**Herramientas:**

→ **DARIAH Tools&Services:**

[Histogram](#)

[Pygmalion's code](#)



# OCR, transcripción y anotación de documentos digitalizados

**Supuesto:** tengo escaneadas las obras antiguas que analizo en mi investigación, pero no tienen OCR ni transcripción. Para implementar búsquedas complejas en mi corpus, necesito conocer técnicas de OCR efectivas, así como aprender a anotar los documentos y el resto de tareas de post-procesado.

**Para aprender:**

→ **DARIAH-TEACH:**

[Digital Research on European Historical Newspapers with the NewsEye Platform](#)

[Digitizing Dictionaries](#): sobre cómo digitalizar diccionarios históricos.

→ **DARIAH-CAMPUS:**

[Automatic Text Recognition \(ATR\)- Pre-Processing and Image Optimisation](#)

[Transcribing Handwritten Text with Python and Microsoft Azure Computer Vision](#)

[OCR with Google Vision API and Tesseract](#)

# OCR, transcripción y anotación de documentos digitalizados

**Herramientas:**

→ **DARIAH Tools&Services:**

[INDXR](#)

[Marginalia and machine learning \(Pytorch\)](#)

**... ¡y muchos recursos más!**

**Para concluir probamos varias búsquedas en el [SSHOC Open Marketplace](#)**



# ¡Muchas gracias!

oficinaclariahcm@ucm.es