# CLARIN per le Lingue Antiche

Workshop online

20 marzo 2024

Giulia Pedonese

# Sources

This presentation is the result of the adaptation and modification of the following sources:

- van der Lek, Iulianna; Fišer, Darja. (2023). *Introduction to Language Data: Standards and Repositories.* In UPSKILLS Learning Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.
- van der Lek, I., Fišer, D., Samardzic, T., Simonovic, M., Assimakopoulos, S., Bernardini, S., Milicevic Petrovic, M., & Puskas, G. (2023). Integrating research infrastructures into teaching: Recommendations and best practices (Versione 2). Zenodo. https://doi.org/10.5281/zenodo.8114407. CC BY 4.0.
- CLARIN ERIC Official Website: https://www.clarin.eu/ . CC BY 2.0

# What Is CLARIN and How Can You Access It?

**CLARIN** stands for **Common Language Resources and Technology Infrastructure**

CLARIN is a distributed digital infrastructure, with participating centres all over Europe and further afield, which include universities, research centres, libraries and public archives. Tools and data from different centres are interoperable so that data collections can be combined and tools from different sources can be chained to perform operations at different levels of complexity, regardless of their location.

# CLARIN…

- Belongs to the Social Sciences and Humanities cluster and an integral part of the European Open Science Cloud
- Has the ESFRI ERIC status since 2012, Landmark since 2016
- Provides easy and sustainable access for scholars in the humanities and social sciences and beyond:
  - to digital language data (in written, spoken or multimodal form)
  - to advanced tools to discover, explore, exploit, annotate, analyse or combine them
  - through a single sign-on environment
- Serves as an ecosystem for knowledge sharing and training

# CLARIN-IT: the Italian node of CLARIN

Italy became the 16th Full Member of CLARIN ERIC in 2015. The Founding member of the National Consortium is the Institute for Computational Linguistics “Antonio Zampolli” of the National Research Council of Italy.

CLARIN is involved in the H2IOSC project through its Italian national consortium CLARIN-IT.

# The H2IOSC project

The H2IOSC project aims at creating a federated and inclusive cluster of RIs in the ESFRI domain of Social and Cultural Innovation to allow researchers from various disciplines in the Humanities, Language technologies and the Cultural Heritage sectors collaborate in data and compute intensive research.

It encompasses the Italian nodes of four Research Infrastructures:

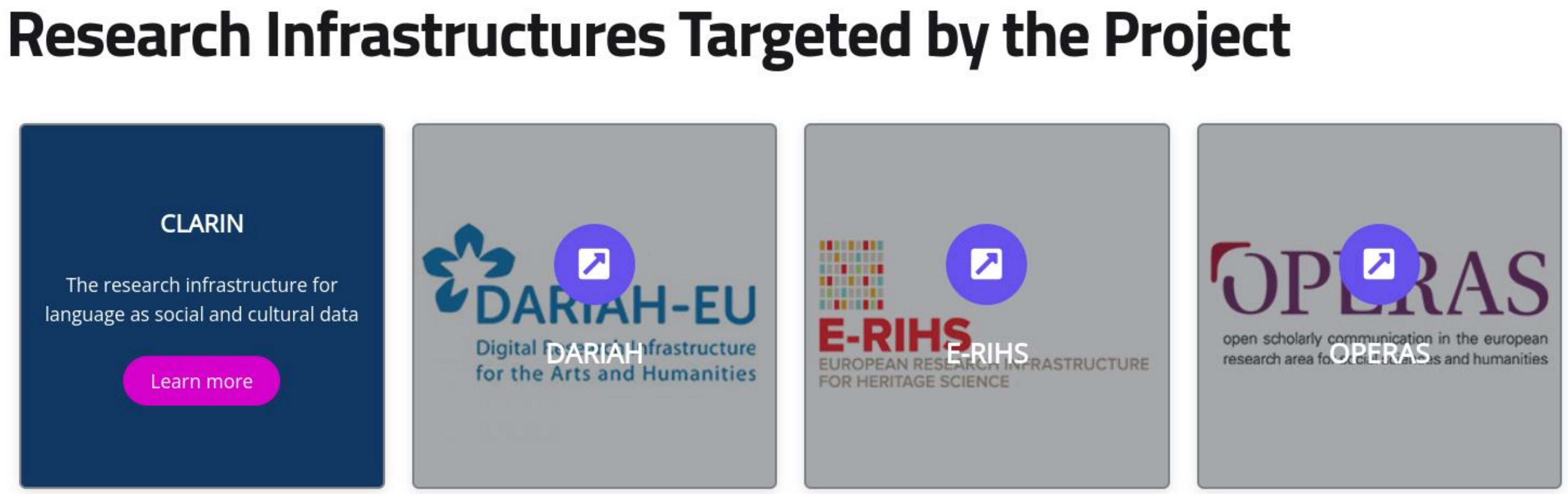
□ CLARIN

□ DARIAH, Digital Research Infrastructure for the Arts and Humanities

□ E-RHIS, European Research Infrastructure for Heritage Science

□ OPERAS, Open Scholarly Communication in the European Research Area for SSH

# The H2IOSC project



## Research Infrastructures Targeted by the Project

CLARIN

The research infrastructure for language as social and cultural data

Learn more

DARIAH

E-RIHS

OPERAS

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities
E-RIHS
EUROPEAN RESEARCH INFRASTRUCTURE FOR HERITAGE SCIENCE
OPERAS
open scholarly communication in the european research area for social sciences and humanities

# CLARIN core services

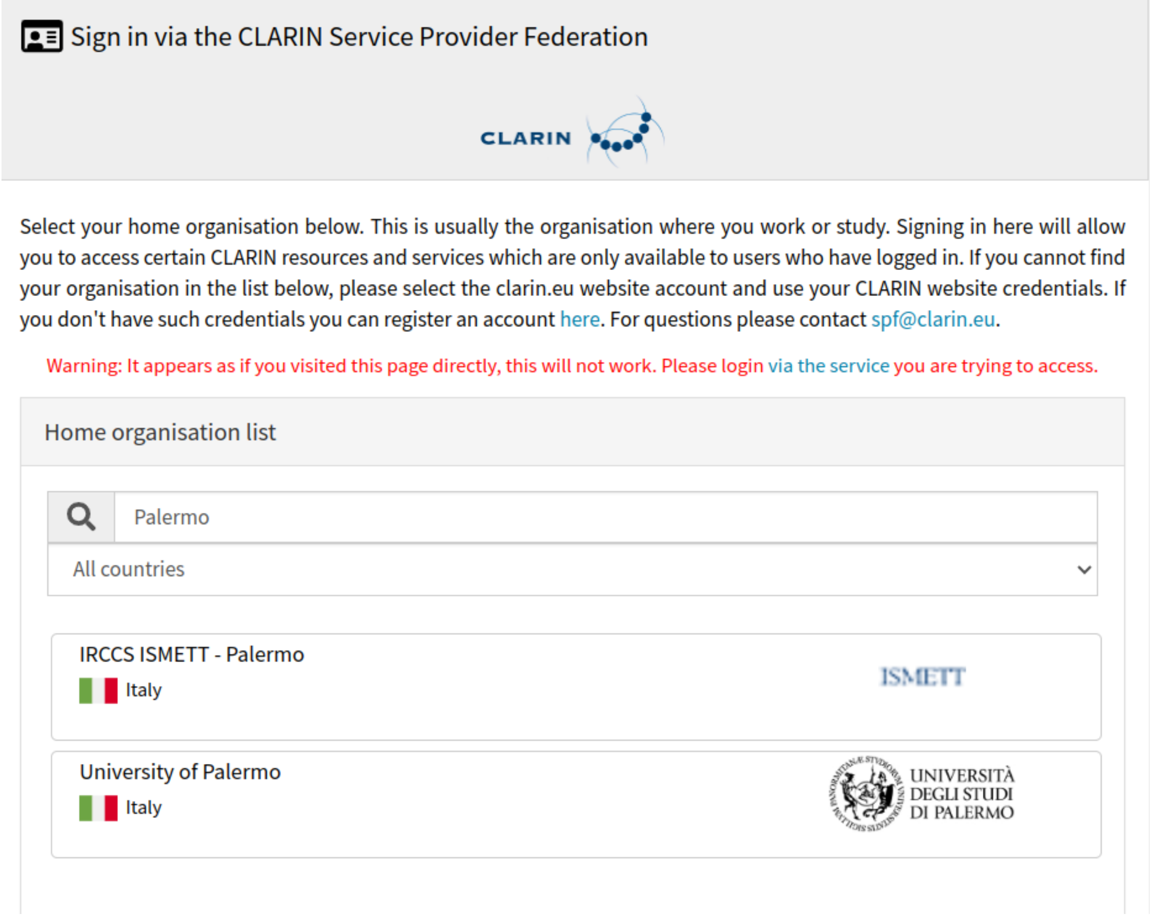
- Depositing services to make sure that language resources can be archived and made available to the community in a reliable manner and to help researchers to store their resources in a sustainable way
- The Virtual Language Observatory provides an easy-to-use interface, allowing for a uniform search and discovery process for a large number of resources from a wide variety of domains.
- The Federated Content Search is a search engine that connects to the local data collections that are available in the centres
- The Language Resource Switchboard helps users to find a matching language processing web application for your data
- The Virtual Collection Registry provides a registry where scholars can create and publish their virtual collections

# How to access CLARIN services

- All users can freely explore the CLARIN core services to search for language resources and expertise
- Due to license restrictions, some resources are only available for academic users and login is required using your institutional credentials or CLARIN credentials

- Academic users in all participating countries can access and use the language resources available in CLARIN data centers with a single sign-on access through the CLARIN Service Provider Federation using their institutional credentials

# Sign in via the CLARIN Service Provider Federation



Sign in via the CLARIN Service Provider Federation

Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account here. For questions please contact spf@clarin.eu.

Warning: It appears as if you visited this page directly, this will not work. Please login via the service you are trying to access.

Home organisation list

Palermo

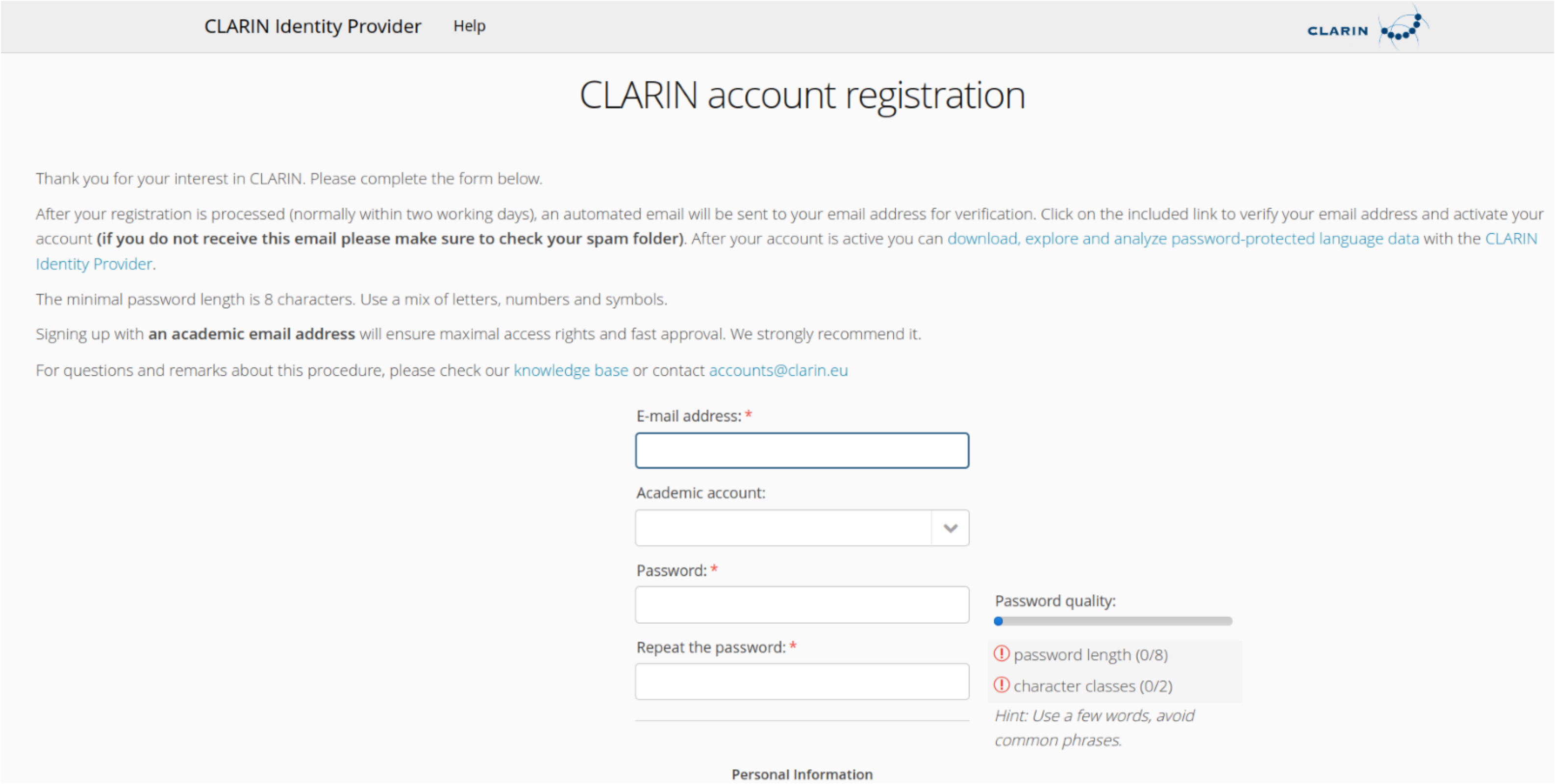
All countries

IRCCS ISMETT - Palermo
Italy

University of Palermo
Italy

# CLARIN account registration

If your university or institute is not in the list of federated organisations, you can request a CLARIN account



## CLARIN account registration

Thank you for your interest in CLARIN. Please complete the form below.

After your registration is processed (normally within two working days), an automated email will be sent to your email address for verification. Click on the included link to verify your email address and activate your account **(if you do not receive this email please make sure to check your spam folder)**. After your account is active you can download, explore and analyze password-protected language data with the CLARIN Identity Provider.

The minimal password length is 8 characters. Use a mix of letters, numbers and symbols.

Signing up with **an academic email address** will ensure maximal access rights and fast approval. We strongly recommend it.

For questions and remarks about this procedure, please check our knowledge base or contact accounts@clarin.eu

E-mail address: *

Academic account:

Password: *

Repeat the password: *

Password quality:

password length (0/8)

character classes (0/2)

*Hint: Use a few words, avoid common phrases.*

Personal Information

# Depositing services

Many of the CLARIN centres offer a depositing service. They are willing to store the resources in their repository and assist with the technical and organisational details. This service ensures many advantages:

- Long-term archiving: a storage guarantee can be given for a long period (up to 50 years)
- Resources can be cited easily with a persistent identifier
- The resources and their metadata will be integrated into the infrastructure, making it possible to search them efficiently
- Password-protected resources can be made available via an institutional login
- Once resources are integrated in the CLARIN infrastructure, they can be analysed and enriched more easily with various linguistic tools

# The CLARIN Repositories

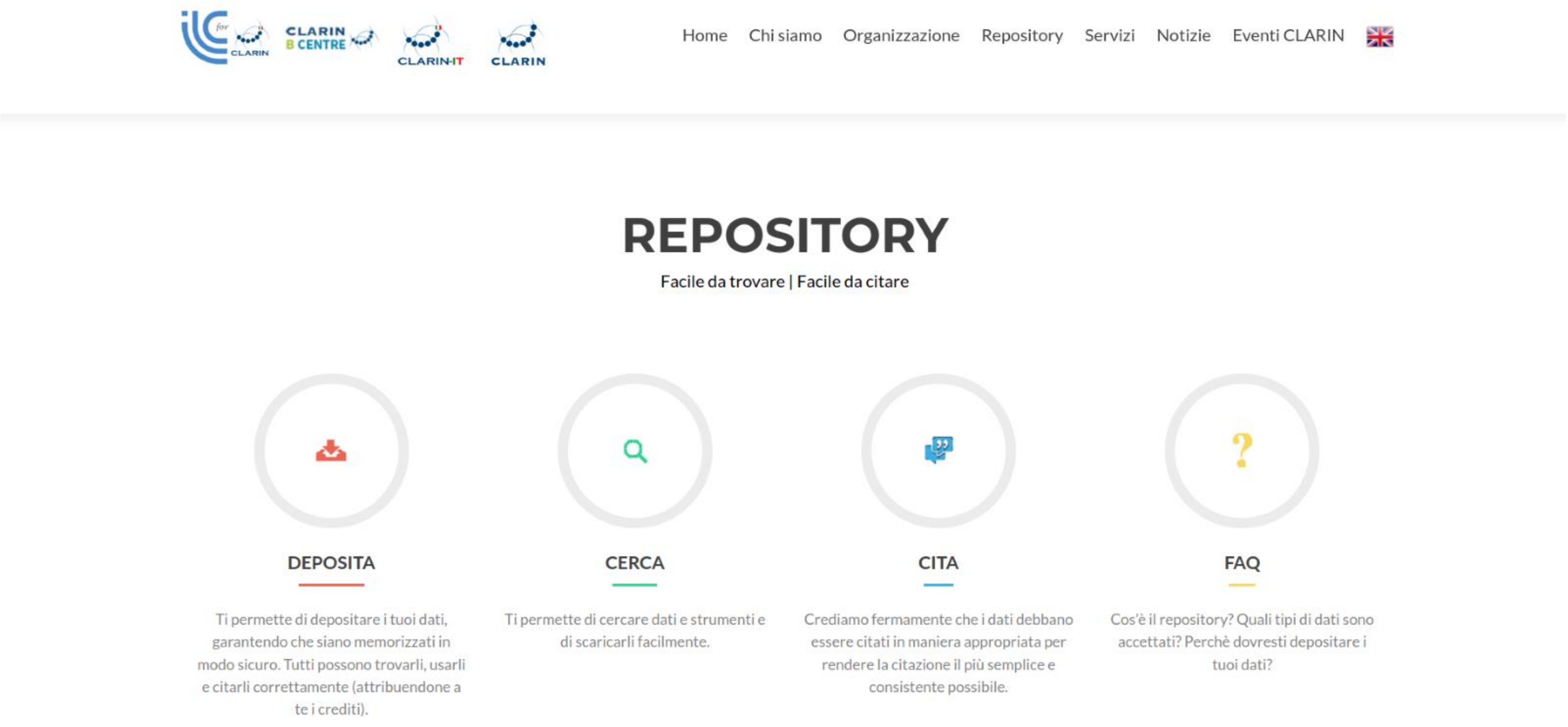
- Ensure long-term preservation and curation of language resources, datasets and tools
- Provide specific and specialised metadata to describe language resources
- Assign Persistent Identifiers (PID), e.g. Handle, to the deposited resources, which enable easy citation
- Offer advanced services to explore the language resources and their metadata, e.g. Virtual Language Observatory, Content Search, Switchboard tools, integrated corpus query engines (e.g. PML-TQ and Kontext)
- Academic users can access restricted resources with their institutional credentials

# How to use a repository: the example of ILC4CLARIN

ILC4CLARIN is CLARIN-IT B-centre hosted at Institute for Computational Linguistics, National Research Council, in Pisa. It offers depositing services for language datasets and tools for research, especially for Italian and classical languages via its repository

The ILC4CLARIN repository is a disciplinary repository certified by CoreTrust Seal and it offers advanced services to explore the language resources and their metadata (e.g. VLO, Switchboard ecc.)

# ILC4CLARIN repository: services



REPOSITORY

Facile da trovare | Facile da citare

DEPOSITA

Ti permette di depositare i tuoi dati, garantendo che siano memorizzati in modo sicuro. Tutti possono trovarli, usarli e citarli correttamente (attribuendone a te i crediti).

CERCA

Ti permette di cercare dati e strumenti e di scaricarli facilmente.

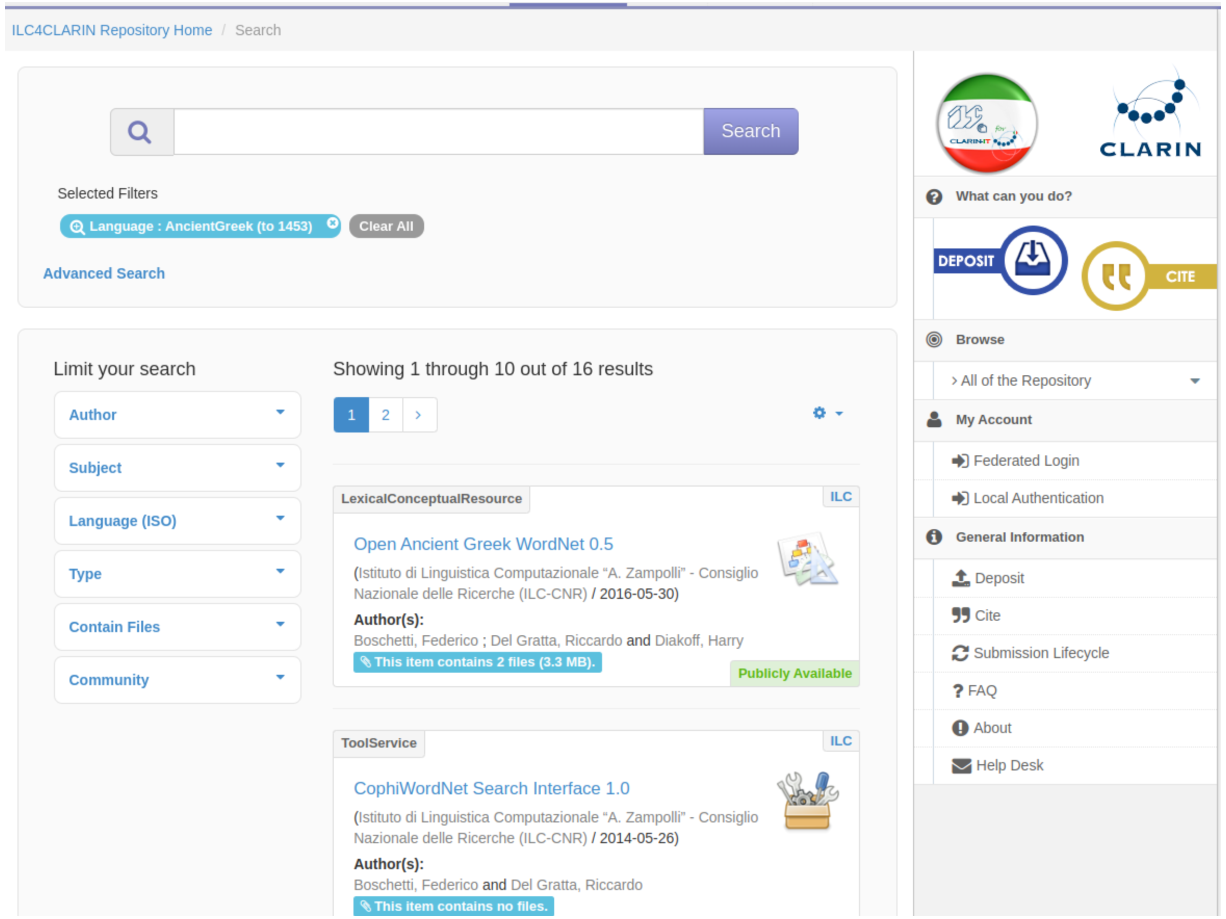
CITA

Crediamo fermamente che i dati debbano essere citati in maniera appropriata per rendere la citazione il più semplice e consistente possibile.

FAQ

Cos'è il repository? Quali tipi di dati sono accettati? Perchè dovresti depositare i tuoi dati?

# ILC4CLARIN language search: Ancient Greek



ILC4CLARIN Repository Home / Search

Search

Selected Filters

Language : AncientGreek (to 1453) Clear All

Advanced Search

Limit your search

- Author
- Subject
- Language (ISO)
- Type
- Contain Files
- Community

Showing 1 through 10 out of 16 results

1 2 ›

LexicalConceptualResource ILC

Open Ancient Greek WordNet 0.5

(Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR) / 2016-05-30)

**Author(s):**

Boschetti, Federico ; Del Gratta, Riccardo and Diakoff, Harry

This item contains 2 files (3.3 MB).

Publicly Available

ToolService ILC

CophiWordNet Search Interface 1.0

(Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR) / 2014-05-26)

**Author(s):**

Boschetti, Federico and Del Gratta, Riccardo

This item contains no files.

What can you do?

Browse

- › All of the Repository

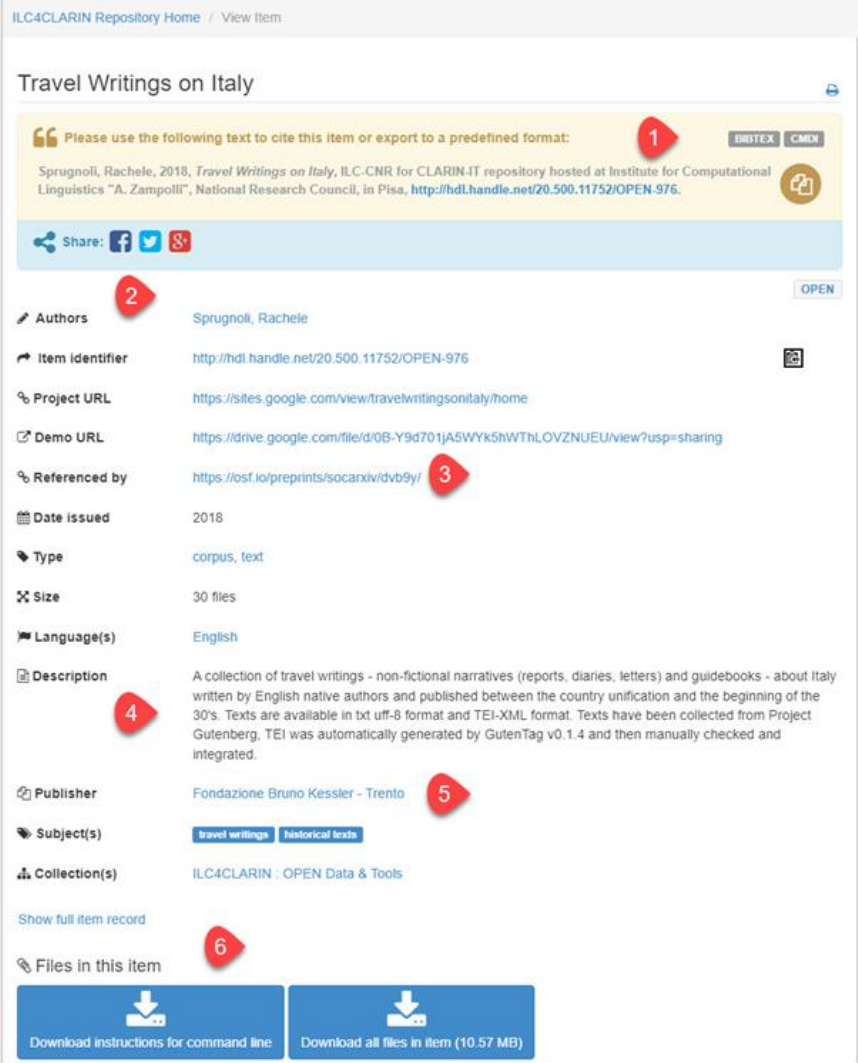
My Account

- Federated Login
- Local Authentication

General Information

- Deposit
- Cite
- Submission Lifecycle
- ? FAQ
- About
- Help Desk

# Example of a corpus in ILC4CLARIN Repository



## Travel Writings on Italy

Please use the following text to cite this item or export to a predefined format:

Sprugnoli, Rachele, 2018, *Travel Writings on Italy*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, http://hdl.handle.net/20.500.11752/OPEN-976.

| | |
|---|---|
| Authors | Sprugnoli, Rachele |
| Item identifier | http://hdl.handle.net/20.500.11752/OPEN-976 |
| Project URL | https://sites.google.com/view/travelwritingsonitaly/home |
| Demo URL | https://drive.google.com/file/d/0B-Y9d701jA5WYk5hWThLOVZNUEU/view?usp=sharing |
| Referenced by | https://osf.io/preprints/socarxiv/dvb9y/ |
| Date issued | 2018 |
| Type | corpus, text |
| Size | 30 files |
| Language(s) | English |
| Description | A collection of travel writings - non-fictional narratives (reports, diaries, letters) and guidebooks - about Italy written by English native authors and published between the country unification and the beginning of the 30's. Texts are available in txt uff-8 format and TEI-XML format. Texts have been collected from Project Gutenberg, TEI was automatically generated by GutenTag v0.1.4 and then manually checked and integrated. |
| Publisher | Fondazione Bruno Kessler - Trento |
| Subject(s) | travel writings<br>historical texts |
| Collection(s) | ILC4CLARIN : OPEN Data & Tools |

Show full item record

Files in this item

Download instructions for command line

Download all files in item (10.57 MB)

- Citation information via handle
- Metadata fields describing the corpus
- The corpus is referenced in a journal
- The corpus is described, e.g.
  texts available in UTF-8 format and TEI-XML format
- Publisher information
- Download instructions
- All the files can be