

Corpora linguistici FAIR: gestione, citazione e archiviazione con CLARIN

Francesca Frontini (CNR-ILC)

Valeria Quochi (CNR-ILC)

Descrizione

La gestione dei dati linguistici è oggi una componente fondamentale della ricerca nelle scienze del linguaggio. Corpora scritti e orali, lessici, trascrizioni e risorse annotate devono essere documentati, organizzati e archiviati in modo da garantire la conservazione, la citabilità e il riuso da parte della comunità scientifica.

Questo tutorial introduce i principi e le buone pratiche per la gestione dei dati linguistici nel contesto dell'infrastruttura europea **CLARIN**, con particolare attenzione ai principi **FAIR (Findable, Accessible, Interoperable, Reusable)**. Verranno presentati il ciclo di vita delle risorse linguistiche, gli standard utilizzati per la rappresentazione e la descrizione dei dati e il ruolo degli archivi digitali nella ricerca.

Particolare attenzione sarà dedicata alla gestione di **corpora scritti e orali**, alle problematiche di documentazione, trascrizione e annotazione e agli **aspetti legali ed etici** legati alla raccolta e alla condivisione dei dati linguistici.

Il tutorial include inoltre una panoramica pratica su come **descrivere, citare e depositare risorse linguistiche nel repository ILC4CLARIN**, illustrando le procedure di archiviazione, i metadati richiesti, le licenze di distribuzione e le modalità di citazione dei dataset.

Il tutorial integra contenuti provenienti dai corsi della **H2IOSC Training Library**:

- *Introduzione ai Dati Linguistici: Standard e Archivi Digitali*
 - *CLARIN per la Gestione dei Dati Linguistici*
-

Obiettivi di apprendimento

Al termine del tutorial i partecipanti saranno in grado di:

- comprendere il **ciclo di vita dei dati linguistici** nella ricerca
- conoscere **standard e formati per la rappresentazione di dati linguistici scritti e orali**

- descrivere risorse linguistiche utilizzando **metadati e buone pratiche FAIR**
 - comprendere le principali **questioni legali ed etiche** nella gestione dei dati linguistici
 - conoscere le modalità di **citazione e pubblicazione dei dataset linguistici**
 - comprendere le procedure di **deposito di risorse linguistiche nel repository ILC4CLARIN**
-

Programma sommario

CLARIN e il ciclo di vita dei dati linguistici

Introduzione all'infrastruttura CLARIN e ai servizi per la ricerca linguistica. Tipologie di dati linguistici (corpora scritti e orali, lessici, risorse annotate) e principi FAIR per la gestione dei dati di ricerca.

Standard e metadati per le risorse linguistiche

Formati e standard per la rappresentazione dei dati linguistici, con particolare riferimento alla codifica dei testi e alla descrizione delle risorse tramite metadati. Panoramica dei servizi CLARIN per la scoperta e il riuso delle risorse linguistiche.

Gestione dei corpora scritti e orali

Raccolta, organizzazione e documentazione dei dati linguistici. Workflow di gestione dei corpora, trascrizione e annotazione, e principali questioni etiche e legali nella gestione dei dati linguistici (privacy, consenso informato, diritti d'autore).

Archiviazione, citazione e condivisione delle risorse linguistiche

Il ruolo dei repository nella conservazione e nella citazione dei dati linguistici. Introduzione al repository **ILC4CLARIN** e dimostrazione delle procedure di deposito, descrizione dei dataset, assegnazione di identificatori persistenti e gestione delle licenze di accesso.

Discussione finale

Parole chiave

CLARIN; dati linguistici; corpora scritti e orali; gestione dei dati di ricerca; FAIR data; metadati; archivi digitali; citazione dei dati; repository; ILC4CLARIN

Lecture

Branco, António, Maria Eskevich, Francesca Frontini, et al. 2023. *The CLARIN Infrastructure as an Interoperable Language Technology Platform for SSH and Beyond. Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-023-09658-z>.

Del Fante, Dario, Francesca Frontini, Monica Monachini, and Valeria Quochi. 2022. *Italian Language Resources. From CLARIN-IT to the VLO and Back: Sketching a Methodology for Monitoring LRs Visibility*. In **Selected Papers from the CLARIN Annual Conference 2021**, Linköping Electronic Conference Proceedings 189: 10–22. <https://doi.org/10.3384/ecp1892>.

Jong, Franciska de, Dieter Van Uytvanck, Francesca Frontini, Antal van den Bosch, Darja Fišer, and Andreas Witt. 2022. *Language Matters. The European Research Infrastructure CLARIN, Today and Tomorrow*. In **CLARIN. The Infrastructure for Language Resources**, edited by Darja Fišer and Andreas Witt. Berlin: De Gruyter. <https://doi.org/10.1515/9783110767377-002>.

Monachini, Monica, and Francesca Frontini. 2016. *CLARIN, l'infrastruttura Europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT*. **IJCoL – Italian Journal of Computational Linguistics** 2(2): 11–30. <https://doi.org/10.4000/ijcol.387>.

Materiali

Darja Fišer, & Iulianna van der Lek. (2025). *Introduzione ai Dati Linguistici: Standard e Archivi Digitali*. H2IOSC Training Library. <https://doi.org/10.5281/zenodo.13911935>.

Christoph Draxler, Henk van den Heuvel, Iulianna van der Lek, Francesca Frontini, & Giulia Pedonese. (2025). *Introduzione alla Gestione dei Dati Orali*. H2IOSC Training Library. <https://h2iosc-training-library.ilc4clarin.ilc.cnr.it/it/ricerca/archive/150>