Servizi CLARIN

Depositare le risorse

L'obiettivo del repository CLARIN-IT è quello di preservare i dati della ricerca, gli strumenti e i servizi orientati alla ricerca nel campo delle scienze sociali e umane https://ilc4clarin.ilc.cnr.it/

Il repository è configurato sulla base di un software DSpace sviluppato dall'Institute of Formal and Applied Linguistics della Charles University di Praga.

Nel repository, le risorse possono essere archiviate secondo quattro categorie: Corpus, Lexical/Conceptual resource, tool, Language description. Gli utenti per depositare le loro risorse devono seguire il flusso di lavoro DSpace organizzato in diverse fasi. È importante sottolineare che il repository non è esternalizzato ma è gestito internamente. La possibilità di deposito è offerta previo login attraverso la tecnologia di Single Sign On (SSO) basata sul Federated Login.

Il processo di sottomissione delle risorse è definito attraverso un'interfaccia web user-friendly che permette all'utente di elaborare una descrizione tramite un application profile di metadati Dublin Core e Meta-Share. I metadati previsti consentono di descrivere ad alto livello di granularità le differenti tipologie di risorse. Qualora, per particolari esigenze, si richiedesse un profilo di metadati più specifico, l'utente ha facoltà di concordare con il Metadata Curator le modalità più idonee a soddisfare tale bisogno.

In ultima istanza, per finalizzare la sottomissione l'utente deve accettare le Condizioni di Uso del repository e selezionare le politiche di accesso ed eventuale riuso delle risorse depositate. L'utente ha la possibilità di scegliere tra licenze selezionate e consigliate, come ad esempio CC o GNU, o indicarne una esterna. Attraverso la finalizzazione della procedura, l'utente si assume la responsabilità di ciò che ha dichiarato e sottomesso.

La proposta di sottomissione delle risorse viene validata previa:

- verifica che un oggetto digitale non sia stato alterato o danneggiato dal repository utilizzando i checksum MD5 per tutti gli oggetti. Questo controllo viene effettuato periodicamente. Il repository esegue automaticamente controlli regolari sull'integrità e sui formati di file dei dati. Un elenco di formati supportati e noti è fornito e regolarmente controllato utilizzando strumenti esistenti (ad esempio, il test di integrità del formato bzip è fatto utilizzando bzip -t). I file vengono controllati tre volte (non necessariamente dagli editor). L'estensione del file (formato file) è selezionata e contrassegnata come supportata, conosciuta o sconosciuta. Il repository pubblica una dichiarazione in cui afferma esplicitamente:

- recupero delle informazioni personali dal server Identity Provider delle organizzazioni di origine;
- congruità e conformità dei metadati utilizzati

Se il controllo è andato a buon fine, la proposta di sottomissione viene accettata e la risorsa infine depositata.

Per garantire la conservazione a lungo termine delle risorse, il repository adotta le misure del modello di riferimento OAIS mantenendo disponibili dati e metadati e rendendo replicabili i risultati della ricerca riutilizzando set di dati e strumenti. Il processo di ingest avviene tramite la creazione di un Submission Information Package (SIP) attraverso un'interfaccia basata sul web che nasconde le impostazioni di implementazione. Quando il curatore approva l'invio, viene generato il Pacchetto Informativo Archivistico che contiene informazioni aggiuntive utili per garantire la conservazione a lungo termine. Per un maggiore livello di sicurezza, sono inoltre pianificate repliche notturne del repository, con controllo automatico della coerenza dei dati su un server situato in un ambiente di rete privato. In caso di guasto del repository, la replica può essere on-line in meno di 3 ore.

DSpace, e quindi il software di repository CLARIN-DSpace, fornisce due livelli di conservazione digitale. Il primo approccio è "bit preservation" che garantisce l'integrità sia dei dati che dei metadati nel tempo indipendentemente da possibili cambiamenti nei supporti fisici di memorizzazione; il secondo è "functional preservation": anche se il file può cambiare nel tempo rimane utilizzabile in futuro evolvendo il suo formato digitale originale e media.

Assicurare la qualità

Il CNR-ILC è attivo nella standardizzazione dei formati di dati linguistici: ospita il presidente della ISO TC37 SC4 - Gestione delle risorse linguistiche [ISO/TC 37/SC], e membri esperti di alcuni dei suoi comitati, nominati dall'Ente di standardizzazione italiano (UNI). Facendo parte della federazione dei centri tecnici CLARIN https://www.clarin.eu/content/clarin-centres, ILC4CLARIN è in costante contatto con esperti di tutti i paesi membri di CLARIN ERIC https://www.clarin.eu/content/overview-clarin-centres, in particolare con quelli che lavorano nei centri B e K.

Nei processi di valutazione della qualità dei dati, il repository adotta diversi standard:

- Il repository si basa sul gruppo di standard di metadati emergenti intorno al CMDI (ISO-CD 24622-1 Gestione delle risorse linguistiche). Ciò garantisce che i metadati necessari per interpretare e utilizzare i dati siano forniti e siano sufficienti per la conservazione a lungo termine.
- Gli schemi CMDI Profile sono basati sullo standard XML Schema del W3C e fanno riferimento al CLARIN Concept Registry (precedentemente ISOcat), un modello di schema basato su SKOS (Simple Knowledge Organization System). I profili CMDI utilizzati sono disponibili al seguente link: http://catalog.clarin.eu/ds/ComponentRegistry/#
- Standard di protocollo OAI-PMH (v2) per la raccolta dei metadati; il repository rende disponibili i formati di metadati tramite l'endpoint OAI-PMH del repository. I metadati sono resi disponibili in base alle specifiche dei metadati CMDI 1.2 e vengono raccolti tramite l'endpoint OAI-PMH del repository. CLARIN ERIC raccoglie i metadati CMDI in un registro centrale, che può essere visto al CLARIN Virtual Language Observatory (CLARIN VLO).

- Il database del repository e le pagine web HTML utilizzano lo standard di codifica UTF-8, per la corretta codifica dei caratteri.
- Codici di lingua ISO nei metadati durante il processo di invio.
- Varietà di formati di dati per gli oggetti digitali inviati. Successivamente all'invio, i dati non possono essere modificati. Ciò garantisce che i dati siano autentici ed è importante anche per gli identificatori persistenti assegnati, che devono sempre riferirsi allo stesso contenuto. Per le modifiche, i segnalanti devono contattare l'help-desk per richiederlo.

Il repository è stato sviluppato per consentire agli utenti una guida completa all'implementazione per la presentazione dei loro dati. In base a ciò, l'interfaccia di presentazione e il flusso di lavoro guidano l'utente nel fornire dati pertinenti e completi.

Essere Conformi ai principi FAIR

Per quanto riguarda i principi FAIR, ILC4CLARIN assegna a ciascuna risorsa depositata un identificatore persistente conforme al principio (F1) al fine di garantire che la risorsa sia reperibile per ogni utente nel repository. Secondo (F2) gli utenti sono facilitati nel descrivere gli elementi tramite un ricco schema di metadati conforme al Dublin Core Metadata Initiative e al Component Metadata Infrastructure standard. Come indicato in (F3), i metadati sono separati dai dati e li accompagnano durante la memorizzazione nel repository. Nel file di metadati il PID della risorsa viene dichiarato con l'elemento di metadati DC specifico. Per rendere i dati reperibili su Internet (F4) ILC4CLARIN è elencato nel Registro dei Data Repositories re3data.org sotto ID:r3d100012262 l'archivio della lingua (http://www.language-archives.org/archive/dspace-clarin-it.ilc.cnr.it) ed è indicizzato da SHARE e dal Web of Science Data Citation Index.

Gli elementi depositati sono immediatamente accessibili tramite collegamento ipertestuale attraverso il protocollo http(s) precedente l'autenticazione dell'utente (A1.2). Questo processo garantisce il controllo dell'accesso alle risorse nel rispetto della privacy. Le risorse sono ricercabili tramite il Virtual Language Observatory (VLO) https://vlo.clarin.eu/, strumento di ricerca all'interno di CLARIN con l'obiettivo di esplorare gli elementi depositati nell'infrastruttura. Il VLO ha un'interfaccia facile da usare, consentendo un processo di ricerca e scoperta uniforme per un gran numero di risorse da una vasta gamma di domini e provider.

In conformità ai principi di interoperabilità, gli utenti possono utilizzare formati di dati standard durante l'invio. A tal fine, viene fornito un elenco dei formati raccomandati per le risorse linguistiche accessibile anche tramite la sezione FAQ del repository (I1): https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf.

Il repository utilizza codici ISO in lingua nei metadati e consente una varietà di formati di dati per gli oggetti digitali inviati, come descritto sul sito web del repository (I2).

Le condizioni di riuso sono specificate dagli utenti tramite la selezione delle licenze (R1.1) e visibile attraverso il metadato *dc:rights.* Per questa opzione è possibile scegliere tra una vasta gamma di licenze. L'elenco completo è disponibile al seguente indirizzo: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/licenses.

Con l'obiettivo di favorire la corretta citazione delle risorse contenute nel repository, CLARIN adotta le raccomandazioni del Data Citation Work Group (DCWG) di RDA. Tutti gli utenti che fruiscono di

risorse depositati in ILC4CLARIN sono tenuti al rispetto di tali norme (R1.3) impiegando il PID assegnato automaticamente alla risorsa in fase di deposito.

Ricercare

Il repository dispone di uno strumento avanzato per la navigazione e la ricerca di elementi alimentato da Solr basato sull'indicizzazione full-text di tutti i file di testo nel repository. Il repository viene regolarmente harvestato dalle varie istituzioni che riutilizzano i metadati qui forniti: ERIC CLARIN, il RI di riferimento, con il Virtual Language Observatory (VLO) https://vlo.clarin.eu/, dove le risorse linguistiche possono essere scoperte utilizzando un motore di ricerca sfaccettato.

Accedere in sicurezza

Per rendere i dati ricercabili e riutilizzabili, ILC4CLARIN distingue tre livelli di accordi di licenza:

1. "Distribution Licence Agreement" [DLA]: prevede il diritto degli utenti a presentare dati e di conferire al repository center i diritti di distribuire i dati per loro conto. L'autore, che rimane il titolare dei dati, afferma sotto la propria responsabilità il contenuto di quanto depositato. L'archivio memorizza una copia dei dati che deve curare, secondo i termini del contratto e le condizioni d'uso.
2. "Terms of Service of the ILC4CLARIN CLARIN-DSpace repository": chiunque scarichi i dati è vincolato dalla licenza assegnata alla voce: utilizzando le funzioni di ricerca offerte dall'interfaccia web del repository e accedendo o scaricando i dati archiviati l'utente accetta i termini citati. Inoltre, per scaricare i dati protetti, è necessario autenticarsi e firmare elettronicamente una licenza.
3. "License": la politica di licenza del repository si basa sulla scelta del depositante al momento di sottomissione della risorsa. Il repository consente inoltre ai mittenti di limitare l'accesso alle proprie risorse a vari livelli. Ciò include la possibilità di assegnare licenze che devono essere firmate elettronicamente da utenti autenticati prima che possano accedervi.

Altri servizi offerti

Il repository ospita diverse tipologie di servizi che permettono di analizzare, approfonditamente, testi, corpora e risorse linguistiche di vario genere. Di seguito si elencano alcune tra queste tipologie di servizi, distinguendo tra risorse integrate con servizi di "browsing/querying" che permettono di ottenere keywords in context (kic), word frequencies, annotazioni linguistiche etc., "merging service" da applicare nei casi di interazione tra più risorse, "extraction/acquisition" per l'indagine e l'estrazione dei contenuti, "text annotation/analysis" ovvero servizi per l'espletamento di compiti di analisi e annotazioni testuali.

- Browsing/querying applicabili per esempio a una collezione di
  - **Dialoghi Italiani di Giordano Bruno:** contiene una selezione di opere di Giordano Bruno collezionate tra gli anni 70-80 su cui è possibile applicare il servizio di "browsing/querying".
  - **Corpus delle Opere di S.Teresa de Ávila:** contiene le seguenti opere da processare tramite "browser/querying": LIBRO DE LA VIDA, CAMINO DE

PERFECCION, LAS FUNDACIONES, EL CASTILLO INTERIOR, CONCEPTOS, RELACIONES, EXCLAMACIONES, CONSTITUCIONES, MODO DE VISITAR LOS CONVENTOS, POESIAS, APUNTES.

- **Tragedie di Vittorio Alfieri:** contiene I testi di Vittorio Alfieri collezionati dall'Accademia della Crusca e dal Dipartimento di Scienze e Letteratura di Arte Medievale e Moderna dell'Università di Pavia. Anche in tal caso le opere sono processabili tramite servizio di "browser/querying".

- Navigazione di **Iscrizioni Istituzionali Cretesi:** una collezione di tracce epigrafiche che si riferiscono alle istituzioni Cretesi del periodo VII – I sec. d.C.
- **LMF ML Merger:** "merging service" eseguibile su server Unix per l'unione di Lexical Entries, Syntactic Behaviours, and Subcategorization Frames da due distinti lessici LMF.
- **SCF Extractor (IT) and language independent:** servizio di "extraction/acquisition" di informazioni linguistiche eseguibile su server Unix. SCF Extractors eseguono l'estrazione di subcategorisation da testi formattati secondo il formato CoNLL-X format.
- **Multiword Extractor(s):** si tratta di una famiglia di servizi per l'estrazione di multiword da corpora di grandi dimensioni.
- **Desr web service:** servizio di "text annotation/analysis" eseguibile su server Unix.
- **Freeling-IT:** servizio di "text annotation/analysis" eseguibile su server Unix che può essere impiegato per ottenere sentence splitting, tokenization, part-of-speech tagging, morphological analysis e lemmatization per i soli testi in Italiano.
- **It-Sr-NER:** servizio per l'analisi parallela dei testi Italiano-serbi.