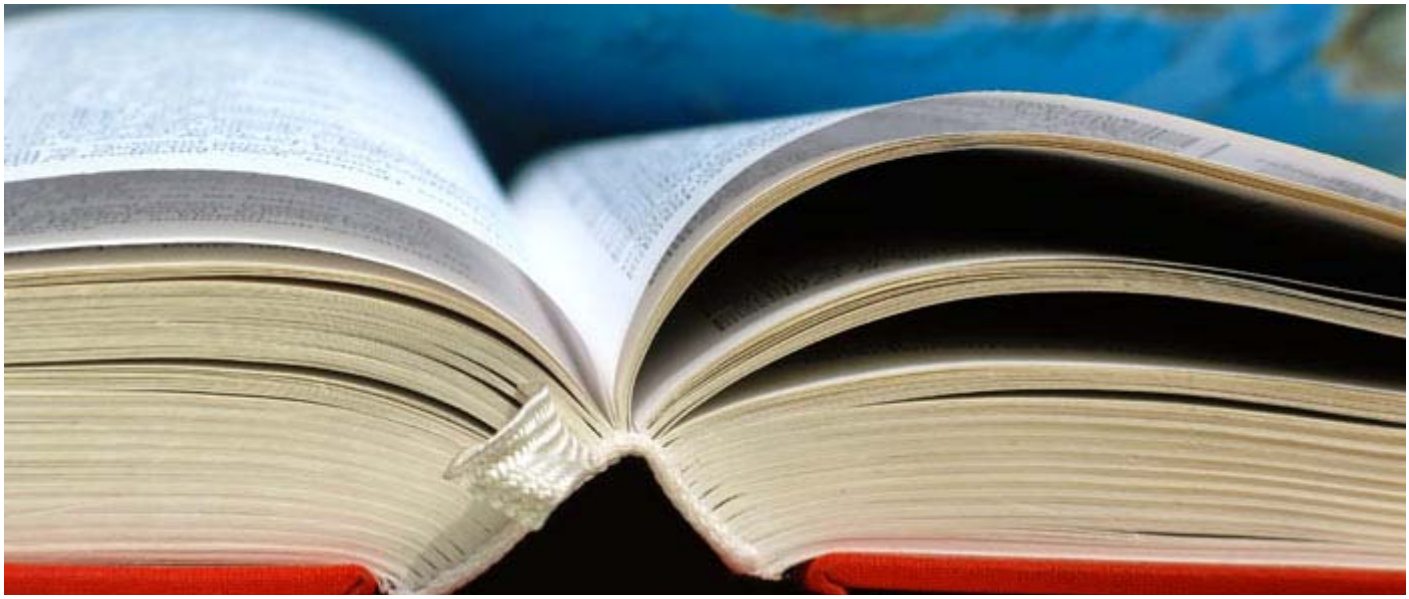


Stampa

CLARIN, l’infrastruttura che ci fa riscoprire Babele

Scritto da Diana Cresti. Inserito in internazionale (/internazionale-14) . Visite: 73

Risorse linguistiche, multimediali e strumenti per il data mining: tutto nella piattaforma online di CLARIN, un'iniziativa distribuita in tutta Europa

Una delle prime infrastrutture della Roadmap ESFRI, CLARIN (Common Language Resources and Technology Infrastructure) fornisce ai ricercatori una piattaforma con accesso federato che integra risorse linguistiche e strumenti avanzati a livello europeo. CLARIN-ERIC è ora classificata come Landmark di ESFRI, essendo in piena fase di implementazione e avendo consolidato vari aspetti chiave dell’infrastruttura come l’accesso sicuro e la standardizzazione dei metadati, nonché naturalmente le questioni legali e istituzionali legate ad aspetti di governance e sostenibilità.

Monica Monachini
CNR - ILC
Istituto di Linguistica Computazionale
Coordinatrice CLARIN-IT
monica.monachini@ilc.cnr.it (mailto:monica.monachini@ilc.cnr.it)

Dal 1 ottobre 2015 l’Italia è entrata a pieno titolo nella infrastruttura CLARIN-ERIC. L’istituto esecutore per CLARIN-IT ed il tramite tra la comunità italiana e l’ERIC è l'Istituto di Linguistica Computazionale "Antonio Zampolli" del Consiglio Nazionale delle Ricerche (ILC-CNR), che già era stato attivo nel progetto della fase preparatoria di CLARIN (2008-2011).

Abbiamo parlato con la dott.ssa Monica Monachini dell’ILC-CNR, coordinatrice di CLARIN-IT

È importante sensibilizzare la comunità sull'importanza delle risorse linguistiche, sulla loro documentazione e sul loro riuso

Ci può dare un esempio di utilizzo della piattaforma? A chi si rivolge questa piattaforma?

L’utente tipo di CLARIN è lo studioso delle Scienze Umane e Sociali, il linguista, lo storico, il filologo, il filosofo, il letterato che voglia analizzare fonti testuali, ma anche il linguista computazionale. CLARIN mette a disposizione risorse linguistiche come corpora, lessici computazionali, risorse audio e strumenti per la analisi del linguaggio naturale e per il data mining.

I dati sono corredati da un’accurata documentazione nella forma di metadati standardizzati, che ne permettono la reperibilità. La ricerca avviene all’interno del catalogo generale di CLARIN chiamato Virtual Language Observatory (VLO) attraverso un’interfaccia a faccette (faceted classification), dove l’utente può scegliere la tipologia di dati a cui è interessato grazie al metadato che gli viene offerto, e continuare a perfezionare la sua query fino ad arrivare al fenomeno d’interesse.

Per esempio, il linguista interessato ad analizzare i comportamenti sintattici della lingua, può reperire grazie al VLO collezioni di dati annotati sintatticamente, le cosiddette treebank (database di alberi sintattici). Queste a loro volta forniscono un’interfaccia di ricerca che estrae informazioni sintattiche a più livelli, ad esempio livello di dipendenze (eg. soggetto-oggetto), oppure livello di “chunk”, i costituenti di una frase (sintagma nominale, sintagma verbale, etc), e le visualizza graficamente.

In linguistica – come altrove – i ricercatori hanno tipicamente i loro dataset che potrebbero essere condivisi. Avete un metodo per incentivare questa condivisione sulla piattaforma?

Attualmente questo è uno dei nostri compiti primari. Stiamo cercando di sensibilizzare il settore alla pubblicazione e alla documentazione di dati e di strumenti nel nostro repository nazionale. Stiamo cercando di sensibilizzare la comunità relativamente all’importanza di far conoscere in maniera documentata le risorse linguistiche, per permettere il loro riutilizzo e quindi evitare una duplicazione di sforzi. Teniamo conto che il valore che hanno le risorse linguistiche è un valore precompetitivo, come sosteneva il nostro precedente direttore, il prof. Antonio Zampolli, che aveva già riconosciuto negli anni '90 il valore infrastrutturale di queste risorse.

I servizi di Clarin necessitano di una protezione e di un accesso controllato, perchè le risorse non sono tutte completamente aperte

Adesso stiamo raccogliendo a livello di consorzio nazionale tutti gli attori principali, sulla base dei rapporti che abbiamo con i vari colleghi sul territorio e anche cercando di raggiungere le maggiori associazioni: l’AIUCD (Associazione per l’Informatica Umanistica e la Cultura Digitale), l’AISV (Associazione Italiana Studi della Voce), l’AISO (Associazione Italiana di Storia Orale), l’AILC (Associazione Italiana di Linguistica Computazionale). Al momento abbiamo una attiva collaborazione con l’Università di Siena che insieme alla Scuola Normale Superiore ha sviluppato un archivio di dati orali della Toscana, che si chiama Gra.fo; i responsabili sono assolutamente coscienti dell’importanza di farlo conoscere attraverso l’infrastruttura CLARIN e di renderlo fruibile da una comunità sempre più vasta. Essendo entrati da pochissimo in CLARIN, la creazione della comunità e l’implementazione del repository nazionale sono appunto i due fronti principali su cui lavoriamo.

…oltre ad esservi collegati da subito con IDEM, per la gestione delle identità digitali…

Sì, questa è una delle prime collaborazioni che è stata attivata. Infatti uno degli aspetti cruciali di una infrastruttura di questa entità e di questo genere è l’accesso e l’autenticazione. I servizi di CLARIN necessitano di una protezione e di un accesso controllato, perché le risorse non sono tutte completamente aperte. CLARIN ha costituito una Federazione di Service Provider e si è consorziato con tutte le federazioni d’identità nazionali (Identity Provider) dei paesi membri; di conseguenza, quando l’Italia è entrata in CLARIN la federazione con IDEM è stata un passo naturale.

Questa prima collaborazione si può definire senza ombra di dubbio una storia di successo perché il nostro gruppo ha collaborato spalla a spalla con il gruppo guidato da Lalla Mantovani, per integrare correttamente il primo Service Provider di CLARIN-IT nella Federazione IDEM e nell’interfederazione eduGAIN, rispettando anche la Federazione CLARIN-SP, allo scopo di dare l’accesso al nostro SP, via autenticazione federata, al maggior numero di ricercatori potenzialmente interessati e sparsi in tutto il mondo.

Da parte sua CLARIN-IT offre a IDEM nuovi utenti, incoraggiando le istituzioni che attualmente non sono parte della federazione a entrare. Un esempio è proprio l'Università di Siena; la professoressa del gruppo senese con cui collaboriamo su Gra.fo ha scoperto, grazie a CLARIN, che la sua Università non aderiva alla Federazione e sollecitato l'ufficio tecnico affinché contattasse IDEM. In questo modo nel giro di un mese l'Ateneo è diventato parte di IDEM; lo stesso è successo con la Scuola Normale Superiore, i cui tecnici sono stati sensibilizzati affinché concludano la procedura di adesione alla Federazione.

Ci sono altre infrastrutture ESFRI con cui collaborate in particolare?

Multilinguismo e migrazioni sono i temoi su cui vorremmo concentrarci per fornire dati e strumenti che siano utilialle amministrazioni e ai cittadini

La collaborazione naturale è con DARIAH, che si occupa del patrimonio culturale materiale. In alcuni paesi i progetti nazionali sono condivisi, come anche le infrastrutture; quindi in Olanda per esempio si parla di CLARIAH. In Italia, la collaborazione avviene a livello di ente, per cui le due infrastrutture sono entrambe gestite dal CNR, per affidamento da parte del Ministero dell'Istruzione, dell'Università e della Ricerca.

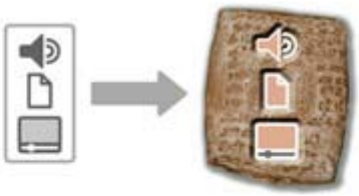
Come vedete il futuro di CLARIN-IT?

Essendo il nodo italiano appena nato, le potenzialità sono molteplici. Attualmente stiamo proprio pensando, insieme al Delegato Nazionale CLARIN, il prof. Riccardo Pozzo, al tipo di impronta che vogliamo dare a CLARIN-IT. Una idea è quella di concentrarsi sul multilinguismo, quindi fornire dati e strumenti per affrontare questo tipo di necessità; un altro dei temi critici al momento, anche riconosciuto dal MIUR, è quello delle migrazioni. Per noi è naturale pensare di unire questi due aspetti, multilinguismo e migrazioni, e cercare di fornire attraverso la nostra infrastruttura strumenti, dati e servizi per questo tipo di tematica.

Per esempio può essere utile offrire vari dati nelle lingue dei migranti, oppure nelle lingue minoritarie delle regioni di confine del Paese. Questi possono essere utilizzati per creare servizi indispensabili per il cittadino che si trova a dover interagire con le pubbliche amministrazioni, una situazione in cui la barriera linguistica è notevole. Questi servizi possono quindi aiutare a migliorare i rapporti con il migrante che arriva sul nostro territorio. Come istituto siamo anche presenti in una azione coordinata a livello europeo (European Language Resource Coordination) che ha lo scopo di creare in ogni Paese europeo un repository di dati e strumenti per migliorare la piattaforma di traduzione automatica dell'Unione Europea. Recentemente abbiamo partecipato ad un workshop del progetto ELRC a Roma (http://lr-coordination.eu /it/italy_agenda (http://lr-coordination.eu/it/italy_agenda)) e abbiamo presentato il progetto CLARIN, quest'ultimo potrebbe offrire la propria infrastruttura per accogliere i dati e gli strumenti che sono attualmente disponibili sul territorio allo scopo di sviluppare servizi linguistici computazionali dedicati alla pubblica amministrazione a supporto dei servizi al cittadino

CLARIN portal

Quick introduction to CLARIN, giving an impression about what's currently available

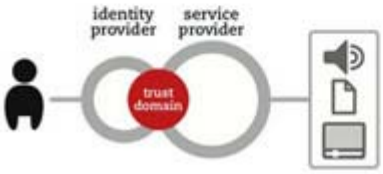


Depositing services

Store language resources in a sustainable repository at a CLARIN centre

Virtual Language Observatory

Discover language resources using a faceted browser or a map

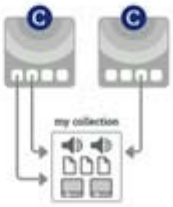


Easy acces to protected recources

Get easy access to protected resources, with your institutional username and password.

Web services and applications

Explore and analyze language data with a wide variety of tools



Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.

Language Resource Inventory

Submit and access information about language resources relevant to your research.

Content Search (prototype version)

Search different corpora with a single search engine

Consulting Services

Searching for a specific data set or application? Wondering how CLARIN can assist your research? Feel free to contact us!

Come CLARIN può aiutare il ricercatore a reperire ed utilizzare risorse linguistiche ed applicazioni software dedicate all'elaborazione del linguaggio. CLARIN offre un ampio insieme di servizi per diverse esigenze. Sia che siate alla ricerca di un analizzatore sintattico, di strumenti per la trascrizione del parlato, di software per il riconoscimento ottico di caratteri in vecchi testi o di uno strumento per riconoscere i nomi di luogo nei testi giornalistici, CLARIN vi può aiutare a trovarli, a capire quale sia la risorsa adatta alle vostre esigenze e ad utilizzarla nella vostra ricerca

Maggiori info: www.clarin.eu (http://www.clarin.eu)
www.clarin-it.it (http://www.clarin-it.it)