IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT

Monica Monachini*
Istituto di Linguistica Computazionale
"A. Zampolli" - Pisa

Francesca Frontini**
Laboratoire PRAXILING
Montpellier 3

On the 1st of October 2015, the Ministry of Education, Universities and Research (MIUR) signed Italy's membership to CLARIN-ERIC, the research infrastructure offering language resources and technologies dedicated to the field of language sciences, the humanities and social sciences. This article aims to provide the Italian research community with a broad overview of CLARIN, its mission, pillars, services, technical and administrative organisation and governance structure, both at European and local level. The Italian network is introduced, with the first national ILC4CLARIN data centre, hosted and being developed at the Istituto di Linguistica Computazionale of CNR, its functionalities, resources and services; finally, the first nucleus of the national CLARIN-IT consortium is presented, illustrating the criteria for its establishment, the planned activities and future prospects.

1. Tecnologie del linguaggio e scienze umane

Le tecnologie del linguaggio, ed in particolare il trattamento automatico del linguaggio naturale (TAL), hanno compiuto notevoli progressi negli ultimi decenni e la loro rilevanza nella vita d'ogni giorno e nelle applicazioni commerciali è in continua crescita. A fronte di questo avanzamento, la penetrazione nell'ambito delle scienze umane e sociali è stata più lenta, a dispetto degli innumerevoli strumenti e vantaggi che tali approcci possono offrire alle discipline del settore.

Diverse ragioni sono state addotte per spiegare la presunta diffidenza degli umanisti nei confronti delle tecnologie digitali. La più semplicistica si basa sulla presunta idiosincrasia delle scienze umane nei confronti della tecnologia e sullo scetticismo nei confronti dei metodi computazionali per l'analisi critica del testo. Tale ipotesi è presto confutata se si osserva lo spazio crescente guadagnato dall'Informatica Umanistica e dalle *Digital Humanities*. In effetti, discipline umanistiche e tecnologie informatiche non sono totamente impermeabili le une alle altre, ma si può dire che la relazione sia stata almeno in parte asimmetrica. L'apporto delle scienze del testo al TAL ai suoi albori è noto agli specialisti e lo sviluppo del TAL deve molto ai pionieri delle Digital Humanities. Invece fino agli anni recenti la gran parte degli studi accademici di ambito umanistico si è svolta senza un forte apporto TAL. Gli umanisti hanno stentato ad appropriarsi di strumenti e risorse linguistiche di punta, che sono spesso sviluppati in seno a progetti di ricerca del settore dell'ingegneria del linguaggio. Sicuramente una buona parte di

* Istituto di Linguistica Computazionale "A. Zampolli" CNR E-mail: monica.monachini@ilc.cnr.it
** Université Paul-Valéry Montpellier 3 Praxiling UMR 5267 CNRS - UPVM3
E-mail: francesca.frontini@univ-montp3.fr

essi nasce per rispondere ai bisogni dell'industria e quindi è poco adatta e parzialmente accessibile per la ricerca di ambito umanistico[1]. Tale quadro è complicato dal fatto che il panorama delle tecnologie del linguaggio è estremamente frammentato e in costante evoluzione, con la continua introduzione di nuove risorse ed algoritmi. Risulta dunque sempre più difficile per i non iniziati orientarsi e trovare gli strumenti adeguati alla propria ricerca, persino per quei *digital humanists* per i quali la tecnologia non costituisce di per sé un ostacolo.

Al contempo l'interesse da parte delle scienze umane e sociali per le tecnologie del linguaggio non è mai stato così forte come adesso. Le principali conferenze di Digital Humanities (come DH[2] o, in contesto italiano, AIUCD[3]) vedono sempre più la partecipazione di linguisti computazionali, mentre nelle conferenze di TAL, l'applicazione alle scienze umane e sociali costituisce una tematica che si affianca a quella delle ricadute industriali.

Il bisogno di rispondere alle esigenze di una platea di utenti diversa apre nuove prospettive, offre stimoli interessanti e richiede uno sforzo addizionale alle tecnologie del linguaggio. Diventa quindi cruciale avere strumenti facilmente usabili e adattabili a diverse tipologie di contenuto per facilitare il reperimento di risorse e di tecnologie. Una risposta urgente viene data con la creazione di depositi digitali (useremo il termine inglese *repository*) dove i dati, i corpora, i lessici, gli strumenti ed i servizi linguistici sono catalogati, archiviati, reperiti e utilizzati in modo semplice e intuitivo.

La qualità delle risorse, in particolare la qualità delle edizioni digitali dei testi, dei loro metadati e dei vari livelli di annotazione acquista poi molta più importanza quando queste devono essere usate per fare ricerca[4]. Allo stesso modo, i software di analisi devono permettere una elaborazione automatica affidabile di tipologie di dati diversi da quelli che comunemente vengono usati come fonte principale per lo sviluppo e la valutazione in TAL, e cioè il linguaggio giornalistico. I testi da trattare in ambito umanistico possono essere spesso eterogenei per genere, per periodo storico, per tipologia. Ed infine, nuovi tipi di analisi testuale, come ad esempio le analisi stilometriche, diventano particolarmente importanti (Piotrowski 2012).

E' proprio per rispondere a queste esigenze e per far incontrare chi produce e sviluppa risorse e tecnologie linguistiche con chi le usa, che è stata creata CLARIN (Common Language Resources Infrastructure for Social Sciences and Humanities) l'infrastruttura di ricerca europea per le risorse linguistiche al servizio delle scienze umane e sociali.

Nei paragrafi successivi ci concentreremo sullo scopo dell'infrastruttura CLARIN, sulla sua architettura e organizzazione e analizzeremo quali servizi offre alla comunità; infine delineeremo la partecipazione dell'Italia, membro della federazione CLARIN da ottobre 2015, e la creazione del network nazionale.

1 Tra i molti lavori su questo tema, segnaliamo quello di Franziska de Jong, attualmente direttore esecutivo di CLARIN, che offre una riflessione storica sui rapporti tra i due ambiti (de Jong 2009).

2 DH (Digital Humanities Conference) è la conferenza annuale dell'Alliance of Digital Humanities Organizations (ADHO).

3 L'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) organizza una conferenza annuale.

4 "For the humanities... there's no data like metadata" (de Jong 2009).

1.1 Risorse linguistiche

Per risorse linguistiche si intendono tutti i dati testuali, audio o multimodali, in formato digitale, contenenti informazioni di carattere linguistico (Godfrey and Zampolli 1997). Esempi di risorse linguistiche sono i corpora, le basi di dati lessicali, le grammatiche formali e i modelli linguistici digitali; sono comunemente annoverati tra le risorse anche tutti gli strumenti che implementano algoritmi di trattamento automatico del linguaggio, ad esempio analizzatori morfosintattici o sintattici, annotatori di entità nominate, siano essi un software installabile sulla propria macchina o servizi collocati su un server e utilizzabili in connessione remota attraverso intenet. Tali strumenti possono servire per costruire, migliorare o valutare le tecnologie del linguaggio, ma anche per facilitare la ricerca su dati testuali. Se si allarga l'orizzonte dalle applicazioni industriali alla ricerca nelle scienze umane e sociali, il concetto di risorse linguistiche si estende. In particolare i dati testuali, come le collezioni digitalizzate di testi letterari, o di documenti storici scritti o orali, anche qualora privi di annotazione linguistica complessa, acquisiscono lo status di risorsa linguistica.

Nell'ambito delle tecnologie del linguaggio, della linguistica computazionale e del TAL una nutrita comunità di ricerca[5] ha come obiettivo la riflessione intorno alle risorse linguistiche digitali. Queste infatti sono tanto indispensabili quanto è onerosa la loro costruzione, sia in termini di tempo che di denaro, ed è quindi importante che le risorse siano costruite secondo principi riconosciuti di (ri)usabilità, accessibilità, interoperabilità. Negli ultimi 30 anni molteplici iniziative hanno visto la luce allo scopo di definire principi condivisi per quanto riguarda gli standard e i formati, la conservazione a lungo termine, la documentazione e la costituzione di cataloghi, le licenze consigliate[6]. Citiamo, in particolare, tra i progetti più recenti, il network FLaReNet (Fostering Language Resources Network) (Soria et al. 2012) che ha raccolto i bisogni e stilato raccomandazioni per la comunità delle risorse linguistiche; il progetto infrastrutturale META-SHARE, che ha costituito una federazione di cataloghi organizzati secondo uno schema di metadati comuni e rigorosamente definiti[7].

1.2 Infrastrutture di ricerca

E' all'interno di tale comunità, alla luce della discussione sui requisiti e bisogni (anche grazie alla visione pionieristica del Prof. Zampolli[8]), che nasce l'idea di una infrastruttura di risorse linguistiche: la costituzione di CLARIN prende le mosse proprio dall'iniziativa di un consorzio di centri di ricerca nell'ambito delle tecnologie e risorse

5 Tale comunità si raccoglie intorno a conferenze del settore tra cui, in particolare, la *Language Resources and Evaluation Conference* (LREC).

6 Tra i progetti pionieristici nell'ambito di risorse e standard citiamo NERC (Network of European Reference Corpora) (Zampolli et al. 1995), EAGLES (Expert Advisory Group on Language Engineering Standards) `http://www.ilc.cnr.it/EAGLES/browse.html` e TEI (Text Encoding Initiative) per i corpora e le edizioni digitali e lo standard LMF (Lexical Mark-up Framework) (Francopoulo 2013). Per una panoramica sulle iniziative sui lessici e sugli standard lessicali si veda (Calzolari, Monachini, and Soria 2013)

7 `http://www.meta-share.org/`. Oggi i centri ospitanti nodi META-SHARE sono in larga parte integrati nella rete CLARIN, e i metadati di META-SHARE sono interoperabili con quelli di CLARIN.

8 Il ruolo infrastrutturale delle risorse linguistiche fu introdotto da A. Zampolli a sottolineare il valore di questi componenti, simile a quello delle risorse di base (per es. acquedotti, elettricità, strade) necessarie per lo sviluppo industriale di un paese.

linguistiche. La fase preparatoria, che va dal 2008 al 2011[9] è stata finanziata come un progetto europeo quadriennale nel Programma FP7, con la partecipazione di diversi paesi, tra cui l'Italia. Durante questo arco di anni sono state poste le fondamenta organizzative, amministrative, tecniche e legali dell'infrastruttura.

Secondo la definizione ufficiale[10] il termine infrastruttura di ricerca si riferisce a strutture, risorse e servizi utilizzati dalla comunità scientifica per condurre ricerca ad alto livello in vari ambiti, dalle scienze umanistiche, all'astronomia, alla genomica alle nanotecnologie.

A livello europeo, le comunità scientifiche si consorziano creando infrastrutture accessibili a tutti i membri all'interno delle quali le risorse sono messe in comune. Un tipico esempio di infrastrutture comuni sono gli acceleratori di particelle come il CERN, che hanno sede in un unico paese ma sono costruiti con fondi comuni e possono essere utilizzati da tutti i paesi consorziati. In altri ambiti, come quello delle risorse e tecnologie del linguaggio, non si tratta di infrastrutture fisiche, ma di piattaforme informatiche e tecnologiche volte a mettere in comune dati, strumenti, software, servizi.

2. La missione di CLARIN

CLARIN è dunque una infrastruttura di ricerca che favorisce lo sviluppo di soluzioni tecnologiche volte a rendere le risorse linguistiche disponibili per studiosi, ricercatori, studenti e cittadini di tutte le discipline, in particolar modo delle scienze umane e sociali, attraverso una modalità unificata e standardizzata di accesso alle risorse. Per una descrizione esaustiva di CLARIN si veda in particolare la documentazione sul sito `clarin.eu`, a cui faremo via via riferimento, e l'articolo "The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars" di (Hinrichs and Krauwer 2014).

In che modo una infrastruttura di ricerca come CLARIN può aiutare la ricerca di frontiera nelle scienze umane e sociali? Il primo contributo consiste nella disponibilità di un catalogo centralizzato attraverso il quale cercare, in maniera guidata e facilitata mediante l'utilizzo di parametri ben precisi, le risorse testuali: arco temporale, lingua, genere letterario. Secondariamente, un'infrastruttura fornisce facile accesso ai dati, anche se ospitati da centri diversi, attraverso un sistema di autenticazione e autorizzazione unica che accrediti l'utente accademico all'uso delle risorse per scopo di ricerca. In terzo luogo, una volta raccolti i testi e costituito il dataset su cui compiere le elaborazioni, l'utente può reperire una serie di strumenti di analisi che possono essere utilizzati con la tipologia di dati selezionati (tenendo conto della lingua, del formato, ecc.). L'infrastruttura, inoltre, assicura un sistema di documentazione e identificazione preciso ed univoco dei dati testuali e degli strumenti: una capillare indicazione degli autori e/o curatori e del versionamento ne assicura la corretta citazione[11], l'attribuzione di un indirizzo permanenti ne garantisce la reperibilità e persistenza nel tempo; il tutto a sup-

9 Per maggiori informazioni si vedano `https://www.clarin.eu/content/about-clarin-preparatory-phase` e (Váradi et al. 2008)

10 `http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what`

11 La corretta citazione dei dati e strumenti utilizzati nella ricerca scientifica è un aspetto assai discusso nelle scienze umane e sociali; per quanto riguarda le risorse linguistiche, questa necessità è evidenziata sia da parte dei fruitori, per garantire il rigore e la replicabilità dei dati, che da parte dei produttori di risorse per garantirne la visibilità anche in termini di valutazione della ricerca scientifica. Si vedano in proposito le raccomandazioni di FLaReNet (Soria et al. 2012) e l'iniziativa ad esso legata della LREMap (Calzolari et al. 2012), ma anche la proposta di un impact factor per le risorse linguistiche (Mariani e Francopoulo 2015).

porto della replicabilità dei risultati. Idealmente, gli strumenti TAL dovrebbero essere di facile uso e pronti a processare questi tipi di testi. Al tempo stesso, l'infrastruttura ha il compito di assicurare servizi di consulenza alla comunità: il ricercatore ha la facoltà di contattare il proprio consorzio nazionale CLARIN per ottenere consulenza o richiedere adattamenti, ad esempio, per consentire il trattamento di testi con ortografie arcaiche. Il consorzio nazionale può, a sua volta, metterlo in contatto con i centri CLARIN europei specializzati, in alcuni casi promuovendo e facilitando eventuali soggiorni di ricerca presso centri di altri paesi.

Per fare questo, CLARIN si impegna a garantire la massima copertura di dati e servizi, con l'obiettivo ideale di rendere un giorno tutte le risorse linguistiche digitali, possedute da organismi pubblici in Europa, accessibili per scopi di ricerca agli studiosi di un qualsiasi paese associato a CLARIN. L'infrastruttura promuove inoltre la preservazione a lungo termine di tali dati e servizi, ma anche la loro integrazione. Infine sostiene lo scambio tra le varie discipline umanistiche e sociali, e promuove la disseminazione di conoscenze e buone pratiche per favorire l'uso delle risorse.

3. L'organizzazione di CLARIN

In questa sezione vengono passati in rassegna gli organi di controllo e di coordinamento della infrastruttura e i centri operativi, cuore delle attività.

3.1 Il forum ESFRI ed il consorzio ERIC

La Commissione Europea promuove la costruzione delle infrastrutture di ricerca attraverso bandi specifici nei vari programmi di ricerca e ne monitora i progressi attraverso un Forum Strategico Europeo per le Infrastrutture di Ricerca (ESFRI), costituito nell'aprile del 2002 su mandato del Consiglio dell'Unione Europea e composto dalle delegazioni nazionali dei 28 Stati Membri dell'Unione Europea. In particolare, la cosiddetta ESFRI roadmap valuta il progresso nella costituzione di ogni infrastruttura e ne determina lo status secondo i vari livelli di sviluppo e i servizi offerti.

Seguendo questa roadmap, conclusa positivamente la fase preparatoria, CLARIN si costituisce nel 2012 come ERIC (European Research Infrastructure Consortium), ovvero come un'entità legale, che ha lo scopo di mantenere l'infrastruttura per la condivisione, l'uso e la sostenibilità delle risorse e delle tecnologie linguistiche. Da questo momento CLARIN, come tutti gli altri ERIC[12], non è più finanziata come un progetto europeo, ma deve diventare autosufficiente dal punto di vista economico.

Il finanziamento di ERIC proviene dunque dai suoi membri, che possono essere stati o organizzazioni sovranazionali; ciascun paese, da una parte paga ogni anno una tassa di iscrizione calcolata in base al PIL nazionale, che serve a finanziare la struttura organizzativa di ERIC, dall'altra si impegna a garantire adeguato finanziamento al consorzio nazionale, per portare avanti le attività di ricerca e sviluppo vere e proprie[13].

12 Si veda `https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric5` per maggiori dettagli su ERIC e sul quadro legale che ESFRI prevede per le infrastrutture europee.

13 Per maggiori informazioni sulla governance di CLARIN ERIC si veda `https://www.clarin.eu/content/governance`; i compiti di direzione esecutiva sono affidati a un direttivo `https://www.clarin.eu/governance/board-directors` mentre l'assemblea generale con i rappresentanti dei paesi membri ha lo scopo di decidere sull'indirizzo programmatico `https://www.clarin.eu/governance/general-assembly`; altri due importanti organi esecutivi sono il Forum dei Coordinatori Nazionali `https://www.clarin.eu/governance/national-coordinators-forum` e Comitato Tecnico

3.2 I consorzi nazionali

Se ERIC rappresenta il consorzio europeo ed il centro gestionale e di coordinamento dell'infrastruttura, la gestione operativa dell'infrastruttura poggia soprattutto sui consorzi nazionali, che non solo offrono servizi alla propria comunità di riferimento, ma possono (e devono) contribuire a sviluppare servizi centralizzati a beneficio di tutta la federazione. Un consorzio nazionale CLARIN si costituisce intorno a un network di istituzioni; il cuore del network è costituito da centri di ricerca e università specializzate nella produzione di risorse e tecnologie del linguaggio; i centri di ricerca nelle scienze umane e sociali e nell'informatica umanistica costituiscono la rete di utenti; anche le grandi banche di dati digitalizzati e le biblioteche nazionali sono grandi fornitori di dati. Il consorzio opera di concerto con il ministero della ricerca, che finanzia l'iscrizione del paese membro all'infrastruttura e le attività nazionali e il cui rappresentante siede nella Assemblea Generale di CLARIN. Questi nomina un responsabile nazionale, che ha lo scopo di gestire il consorzio e organizzarne fattivamente le attività. Ogni consorzio nazionale si impegna a garantire alcune attività di base attraverso la costituzione di uno o più centri.

I consorzi nazionali possono avere governance diverse a seconda del paese e delle modalità di finanziamento. Anche la struttura organizzativa può variare, con consorzi che accentrano tutte le attività in un solo centro (ad esempio la Repubblica Ceca), ad altri composti da una miriade di centri diversi (Paesi Bassi e Germania). La lista dei consorzi partecipanti, disponibile al link `https://www.clarin.eu/content/participating-consortia`, è in continua crescita per l'aggiunta di nuovi paesi membri; in alcuni casi un paese può scegliere di passare attraverso una fase preliminare, entrando nel consorzio come osservatore.

4. CLARIN: servizi e attività

La Figura 1 sintetizza i servizi offerti da CLARIN, sia centralmente che localmente attraverso i suoi centri. L'infrastruttura di CLARIN funziona infatti come una federazione di centri, situati in diversi stati membri e organizzati in varie tipologie a seconda dei servizi offerti. Non entreremo qui nei dettagli sulle varie tipologie di centro; informazioni più precise possono essere trovate al link `https://www.clarin.eu/sites/default/files/CE-2012-0037-centre-types-v07_0.pdf`.

4.1 Il portale

Il portale CLARIN, a livello centrale, contiene le informazioni relative alla struttura organizzativa, i membri, i centri, nonché i collegamenti ipertestuali a tutti gli altri servizi[14].

Centrale di Coordinamento dei Centri
`https://www.clarin.eu/governance/standing-committee-clarin-technical-centres`

14 Si veda la sezione servizi del portale `https://www.clarin.eu/content/services`, ma anche l'introduzione alle tecnologie usate `https://www.clarin.eu/content/clarin-technology-introduction` e la descrizione delle funzioni dei centri `https://www.clarin.eu/content/clarin-centres`.

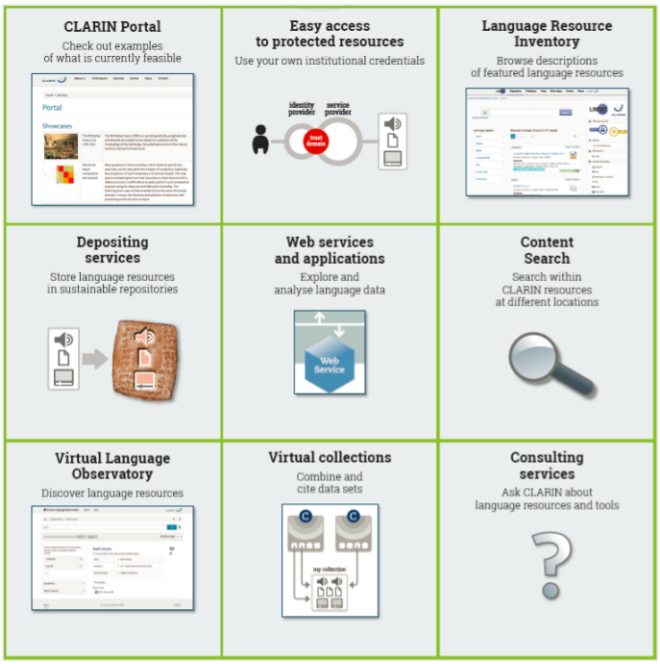

Figura 1
Lo schema dei servizi CLARIN.

4.2 Documentazione: i metadati

Il primo pilastro è la catalogazione ed il censimento delle risorse linguistiche in repository accessibili e navigabili on line.

I centri CLARIN possono essere strutturati a seconda di varie tipologie. Al di là delle differenze superficiali, essi si attengono a linee guida comuni. Nelle Figure 2 e 3 vediamo, rispettivamente, due esempi di repository dall'impostazione diversa. Un centro CLARIN offre come servizio basilare un catalogo di risorse linguistiche documentate e descritte utilizzando un formato comune di metadati. Tale formato che si ispira allo standard CMDI sviluppato all'interno dei Comitati ISO (Component MetaData Infrastructure, (Broeder et al. 2011, 2012) e (Goosen et al. 2014)) permette la definizione di diversi profili, a seconda delle varie tipologie di risorse, ma promuove il riutilizzo di set di metadati comuni, chiamati componenti: la composizione modulare assicura che i profili vengano creati combinando, possibilmente, componenti già esistenti. La creazione dei profili e componenti per le varie risorse linguistiche è prerogativa dei centri CLARIN.

Ogni centro CLARIN, una volta identificati e creati i profili adeguati alla descrizione delle proprie risorse, li inserisce nel CLARIN Component Registry, proprio per favorire il controllo e il riutilizzo. Tra tutti i componenti, alcuni, prendono lo status di componenti di base e possono essere usati per tutte le risorse. Così, due risorse molto diverse, come ad esempio una registrazione multimodale audio, un corpus scritto annotato, un lessico, avranno comunque un set di descrittori comuni, come l'autore (sia esso una persona o un'istituzione), il nome, la lingua. CLARIN adotta il formato di base Dublin

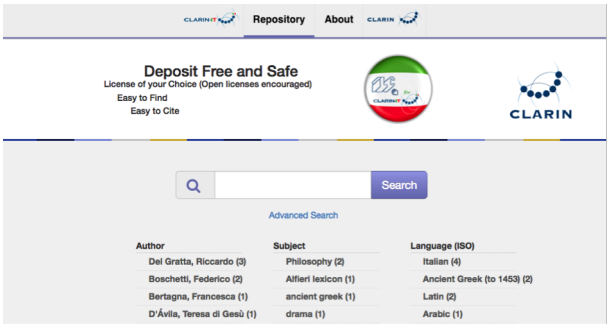

Figura 2
Il repository di LINDAT CLARIN, Repubblica Ceca.

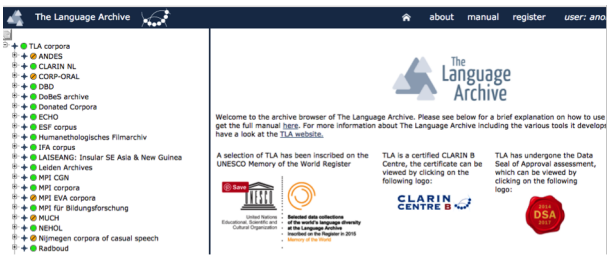

Figura 3
The language Archive, Paesi Bassi.

Core[15] per il set minimale di metadati comuni.

Oltre all'armonizzazione dei metadati, CLARIN promuove anche quella dei valori che tali descrittori possono contenere, le cosiddette *data categories*. In altre parole, se da una parte un profilo contiene uno o più campi "language", per indicare la o le lingue contenute nella risorsa, vi sono anche indicazioni circa il valore che tale campo può contenere, ad esempio riducendo la lista dei valori possibili ai codici iso-639-3. Il CLARIN concept registry (Schuurman et al. 2016), che sostituisce ISOcat (Marc Kemps-Snijders and Wright 2008) è un'ontologia formale dove i vari concetti vengono classificati, descritti e armonizzati in modo da fissare la loro semantica e rendere il loro uso meno ambiguo.

Il modello di metadati di CLARIN fissa dei principi base senza forzare l'adesione ad uno standard rigido, permettendo così ai vari centri di gestire i metadati già esistenti e permettere dunque una facile conversione delle collezioni esistenti nel nuovo formato. I centri provvedono quindi a censire i propri dati, importandoli in un catalogo digitale in una rappresentazione XML conforme al formato CMDI, e rendendoli ricercabili tramite un'interfaccia online.

Un ulteriore importante requisito è l'attribuzione ad ogni risorsa di un identificativo persistente[16]. CLARIN promuove l'utilizzo di *handle*, ovvero link permanenti per le risorse. Se acceduto da un browser, un handle ridirige l'utente verso l'*url* reale, indi-

15 http://dublincore.org/
16 https://www.clarin.eu/content/persistent-identifiers

rizzo che può dunque essere modificato senza rendere la risorsa irreperibile. E' quindi l'handle di una risorsa ad essere utilizzato per le referenze bibliografiche. Il sistema ha due grandi vantaggi per la citabilità delle risorse nelle pubblicazioni scientifiche: offre un link immediatamente "azionabile" a chi legge l'entrata bibliografica e garantisce il funzionamento di tale link a lungo termine, anche qualora la risorsa sia spostata su server o indirizzi diversi.

Alcuni repositories vanno ancora oltre rispetto a questo requisito. LINDAT CLARIN (Figura 2) ed altri centri costruiti con lo stesso software offrono per ogni risorsa un'indicazione per la citazione, sia in formato testuale che bibtex.

Oltre alla persistenza dell'identificativo, i centri CLARIN si devono anche dotare di dispositivi tecnologici per conservare e proteggere i dati a lungo termine. Questo viene effettuato, per esempio, mettendo in opera server ridondati e sistemi di back-up, ma anche mettendo in atto tutta una serie di operazioni per proteggere i dati dall'obsolescenza informatica (*bit-rot*), ad esempio convertendo di volta in volta i dati verso formati più attuali.

4.3 Accesso: *single-sign-on*

Documentare, identificare e proteggere la risorsa non è sufficiente per mettere in pratica gli obiettivi di CLARIN; è anche necessario renderla disponibile e, laddove possibile, scaricabile in modo semplice e immediato. Non è sufficiente per un ricercatore avere diritto ad accedere a una data risorsa per poterne usufruire. Nelle procedure tradizionali, se un corpus è disponibile per uso accademico, spesso, il ricercatore deve farne richiesta ai detentori, compilando un formulario per provare la propria affiliazione ad un'istituzione di ricerca. CLARIN permette di superare queste barriere, grazie ad un sistema di identità digitali federate, che pur nei limiti e vincoli imposti dai produttori delle risorse, consente ad un utente di accedere ai dati con le proprie credenziali istituzionali. Tale sistema viene definito con il termine specialistico *single-sign-on* (Figura 4). Tecnicamente, la procedura, implementata presso ogni centro CLARIN, fa sì che il protocollo di identificazione (*identity provider*) scambi informazioni circa affiliazione e ruolo degli utenti con il protocollo che gestisce la fornitura del servizio (*service provider*): quando i dati forniti circa la identità combaciano con i requisiti richiesti dal servizio, viene autorizzato e garantito accesso automatico e sicuro. Tale meccanismo è reso possibile dal fatto che molte realtà accademiche si appoggiano a federazioni di identità nazionali (come IDEM[17] in Italia), a loro volta federate a livello europeo (eudGAIN[18]).

Quando l'Italia è entrata nel consorzio CLARIN, la federazione CLARIN SPF (CLARIN Service Provider Federation) si è *interfederata* con la federazione di identità nazionale IDEM. In questo modo, gli utenti italiani, effettuando il login con le proprie credenziali presso qualsiasi centro CLARIN, vengono riconosciuti, accreditati ed ottengono automaticamente accesso alle risorse a cui hanno diritto secondo il proprio profilo, siano essi professori, ricercatori, ma anche studenti.

17 IDEM, la Federazione Italiana delle Università e degli Enti di Ricerca per l'Autenticazione e l'Autorizzazione, gestita dal GARR `http://www.idem.garr.it`.

18 eduGAIN, l'interfederazione europea degli identity provides, gestita da GÈANT. `http://www.geant.org/Services/Trust_identity_and_security/eduGAIN`

4.4 Conservazione: il deposito

Se mettere a disposizione le proprie risorse è il primo passo per la costruzione di un centro CLARIN, è evidente che non tutti i produttori di risorse sono in grado di sostenere il peso della costruzione di un tale repository. In molti casi le risorse sono prodotte grazie a iniziative progettuali finanziate; una volta terminati i progetti, diventa spesso insostenibile per le istituzioni assicurare il mantenimento. Al fine di permettere a tutte le risorse linguistiche (il cui sviluppo richiede notevoli sforzi in termini di risorse umane ed economiche) di essere mantenute e rese accessibili, alcuni centri CLARIN, consentono il deposito di risorse da parte di terzi. Ogni consorzio nazionale è tenuto a finanziare almeno un centro di questo tipo, in modo da garantire una copertura adeguata per tutte le risorse. Tale centro deve essere certificato secondo una rigorosa procedura che controlla la conformità dei metadati con gli standard CLARIN, garantisce la conservazione dei dati a lungo termine nonché la persistenza degli identificativi[19].

Il servizio di deposito, anch'esso in genere gestito tramite *single-sign-on*, offre la possibilità agli utenti stessi, una volta registrati, di descrivere i propri dati tramite un'interfaccia di inserimento intuitiva, che permetta di inserire correttamente i metadati con un profilo CMDI CLARIN, di decidere il livello di accessibilità per tali dati e di caricarli infine sul server del centro.

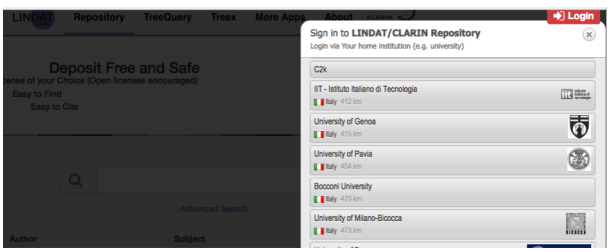

Figura 4
Accesso a un repository CLARIN tramite credenziali istituzionali.

4.5 Servizi e applicazioni

Oltre alle risorse vere e proprie, tipicamente corpora e lessici, ma anche software scaricabili e utilizzabili sul proprio computer, i numerosi centri offrono servizi in linea, ad esempio interfacce per interrogare i corpora, per comporre e lanciare catene di analisi per processare testi, per utilizzare sistemi di estrazione e visualizzazione[20].

Citeremo qui solo alcuni due tra i molti esempi. Il *Bayerisches Archiv für Sprachsignale*[21], specializzato in risorse audio, offre numerosi strumenti in linea per il trattamento dei segnali sonori: strumenti come WebMINNI per la trascrizione automatica del parlato in varie lingue, o WebMAUS per segmentare e allineare audio e trascrizioni già

19 Si veda `https://www.clarin.eu/content/overview-clarin-centres` per una lista dei centri certificati CLARIN.
20 Qui una lista dei servizi CLARIN disponibili nei vari centri: `https://www.clarin.eu/content/web-services`
21 `https://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html`

esistenti[22]. LINDAT CLARIN offre PML TreeQuery, un sistema in linea per interrogare treebanks in diverse lingue[23]

4.6 Ricerca sui metadati e ricerca sui dati: VLO e Federated Content Search

Se il peso dello sviluppo e del mantenimento di dati e servizi CLARIN ricade in gran parte sui centri dei vari paesi, è chiaro che un coordinamento è necessario. In particolare, come possono gli utenti olandesi di CLARIN essere a conoscenza di quanto messo a disposizione dal consorzio nazionale tedesco, se i centri CLARIN olandesi si occupano soprattutto delle risorse prodotte in Olanda? A tal fine, esistono due servizi centralizzati che permettono di ricercare risorse tra i metadati e di lanciare ricerche sul contenuto delle risorse, senza sapere in quale centro esse si trovino.

Il primo servizio è il Virtual Language Observatory (VLO)[24], nel quale si possono ricercare centralmente metadati di tutti i centri CLARIN attraverso un'unica interfaccia di ricerca. Un requisito di base perché i metadati di ogni centro siano visibili nel VLO è l'adesione da parte di ogni repository al protocollo OAI-PMH [25] per esporre i propri metadati all'esterno. Senza insistere nei dettagli tecnici, la procedura CLARIN impone che ogni centro registri l'indirizzo del proprio repository presso il servizio di raccolta dei metadati (*harvesting*) sviluppato centralmente dall'ERIC. La raccolta, che avviene a cadenza settimanale, permette di validare i metadati, di registrare ogni nuova risorsa nel catalogo unificato e di assicurare una mappatura dei metadati con i campi di ricerca del VLO. In questo modo, è possibile rendere ricercabili metadati eterogenei, provenienti da archivi diversi, tramite l'interfaccia VLO.

Per ogni risorsa, il VLO fornisce i metadati originali, incluso l'identificativo persistente, il repository e il centro di provenienza degli stessi e la collezione alla quale la risorsa appartiene. Inoltre, sono fornite informazioni sul tipo di licenza e sull'accessibilità. E' evidente che tale servizio contribuisce a ridurre notevolmente i tempi di ricerca di una risorsa, dal momento che un utente non è tenuto a sapere in quale centro si trovino le risorse per poter effettuare una ricerca.

Se il VLO abbatte le barriere per quanto riguarda i metadati, un nuovo servizio a livello sperimentale permette di lanciare ricerche centralizzate sul contenuto dei dati, in particolare dei corpora. Molti centri nazionali offrono già servizi online di ricerca sui contenuti; il servizio di Federated Content Search[26] (FCS - ricerca federata sui contenuti) si appoggia a tali sistemi già esistenti per lanciare delle ricerche contemporaneamente sui server di diversi centri. L'utente della FCS non deve per forza conoscere l'esistenza delle varie collezioni e servizi, ma una volta ottenuti i risultati della ricerca, potrà essere rediretto alla pagina di ricerca della specifica collezione.

22 Si veda qui una lista esaustiva dei servizi `http://www.bas.uni-muenchen.de/Bas/BasServiceeng.html`

23 `https://lindat.mff.cuni.cz/services/pmltq/#!/home`

24 `https://vlo.clarin.eu`

25 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) `https://www.openarchives.org/pmh/`

26 Si veda `https://www.clarin.eu/content/content-search` per una spiegazione più dettagliata della FCS e qui `https://spraakbanken.gu.se/ws/fcs/2.0/aggregator/#` per l'interfaccia di consultazione.

4.7 Gli standard e le licenze CLARIN

Il funzionamento dell'infrastruttura e dei suoi servizi dipende dunque da un adeguato coordinamento a livello sia degli standard utilizzati che delle licenze adottate. Tale coordinamento è assicurato da due comitati chiave, quello degli standard e quello delle questioni legali.

Per quanto riguarda le questioni legali, il comitato, composto da rappresentanti di diversi paesi, ha prodotto una serie di linee guida sulle tipologie di licenza raccomandate[27]. CLARIN raccomanda che tutte le licenze diano accesso libero alle risorse, almeno per scopo di ricerca. Tuttavia, le varie legislazioni nazionali non sono sempre allineate e la presenza di risorse con licenze già precedentemente assegnate può rendere difficile generalizzare questo principio. L'attività di coordinamento del comitato permette agli utenti di orientarsi nella selva delle tipologie di licenza, riducendole a tre grandi tipologie: di uso pubblico (ovvero senza restrizioni), accademiche e ristrette. Il comitato interagisce inoltre con i centri nazionali per assicurarsi che i produttori di risorse siano adeguatamente supportati nella scelta della licenza più consona alle proprie risorse.

Se la standardizzazione delle licenze facilita l'accesso, solo l'adesione a standard comuni nella produzione delle risorse assicura che i dati siano effettivamente interoperabili e utilizzabili. La standardizzazione tocca diversi aspetti del ciclo di vita delle risorse linguistiche, dalla gestione dei formati di codifica dei dati (UTF-8), al formato dei metadati (CMDI) e dei dati (XML, TEI, LMF). Il comitato degli standard redige una lista degli standard consigliati[28], che, nello spirito di CLARIN, viene vista non in senso prescrittivo, ma di raccomandazione. Tale lista viene costantemente aggiornata al fine di tenere conto delle buone pratiche che di volta in volta si impongono nelle comunità di produttori e utilizzatori.

4.8 CLARIN webservices: Weblicht, Switchboard

L'utilizzo di standard condivisi ha il vantaggio di facilitare e potenziare l'utilizzo di servizi web per il TAL. In particolare, i moduli di analisi linguistica messi a disposizione dai vari centri, qualora usino standard condivisi, possono essere combinati per effettuare analisi complesse, che coprono diversi livelli di annotazione a cascata. Questo è particolarmente vero per il trattamento di corpora scritti e orali. Una volta individuati i corpora sui quali lavorare, gli utenti CLARIN possono invocare un servizio chiamato **CLARIN Switchboard**[29] che permette di identificare tutti gli strumenti di analisi automatica in grado di processare quella tipologia di testo. Grazie all'uso di standard e formati condivisi, lo switchboard riceve in input un file o una collezione di file e li analizza automaticamente per identificare la tipologia e la lingua e propone una serie di servizi. Questo servizio di *discovery* è sempre più strettamente integrato con **Weblicht**[30], una piattaforma implementata da CLARIN-DE (Germania) che permette di costruire ed eseguire catene di analisi personalizzabili. Contrariamente ad altri sistemi di gestione di catene di analisi linguistica, Weblicht è basato su un'interfaccia online, che non richiede dunque alcuna installazione (si veda la Figura 5). L'uso, riservato agli utenti accademici (federati alla CLARIN Service Provider Federation tramite il sistema

27 https://www.clarin.eu/content/license-categories.
28 https://www.clarin.eu/content/standards-and-formats.
29 http://weblicht.sfs.uni-tuebingen.de/clrs/#/?_k=t2fobv.
30 http://weblicht.sfs.uni-tuebingen.de.

di *single sign on*), è facile e intuitivo, i servizi offerti sono di vario tipo e coprono varie lingue.

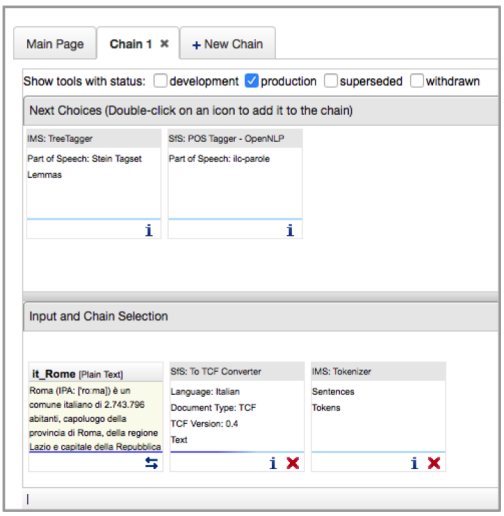

Figura 5
Weblicht - Costruzione di una catena di analisi per l'italiano. Una volta aggiunto il livello di tokenizzazione il sistema propone automaticamente la scelta tra due PoS-tagger disponibili (OpenNLP e Treetagger).

I centri CLARIN hanno l'obbligo di censire i loro servizi linguistici sullo Switchboard e di renderli compatibili con l'uso in Weblicht[31], con l'ambizione di creare un vero e proprio ambiente di ricerca virtuale (*Virtual Research Environment*) di tecnologie TAL per gli studiosi di scienze umane e sociali.

Il prossimo obiettivo di CLARIN, attualmente in fase di sviluppo, è quello di collegare VLO, Switchboard e Weblicht, in modo che l'utente, una volta cercato e individuato un corpus, possa con pochi clic visualizzare i tipi di analisi disponibili e lanciare una catena di analisi su Weblicht.

La messa a fattor comune degli sforzi e la semplificazione dei protocolli avrà evidenti benefici anche per i produttori di servizi, che potranno contare su un'interfaccia utente ben sviluppata e mantenuta come vetrina per i propri strumenti e dunque concentrare i propri sforzi sullo sviluppo dei servizi.

4.9 Consulenza

Tutti i centri CLARIN possono essere contattati dagli utenti per informazioni sulle risorse linguistiche; alcuni centri possono offrire agli utenti dell'infrastruttura il loro expertise in un particolare settore. Ad esempio lo Spanish CLARIN K-Centre si propone come centro di conoscenza per il trattamento delle lingue iberiche[32].

31 Per fare questo, i tool devono essere resi disponibil come servizi RESTful. Questo significa che, oltre ad avere un'interfaccia web propria, essi possono essere invocati e utilizzati da altri programmi su altri server (`http://weblicht.sfs.uni-tuebingen.de/WebLichtTutorial.pdf`).

32 `http://www.clarin-es-lab.org/index-es.html`.

4.10 Comitati, task forces e progetti comuni

E' importante notare come i vari servizi descritti siano sviluppati in collaborazione tra i responsabili tecnici dell'ERIC (con sede a Utrecht) e degli esperti dei vari centri, che fanno parte dei comitati per gli standard e le questioni legali, o delle task force del VLO e della FCS. I consorzi nazionali hanno infatti il dovere di contribuire allo sviluppo di soluzioni comuni, partecipando ai gruppi di lavoro transnazionali dell'ERIC e contribuendo così alla costruzione e sviluppo dell'infrastruttura.

Oltre essere attivi in questi gruppi di lavoro, i membri dei vari consorzi CLARIN collaborano nei progetti europei ai quali CLARIN partecipa in quanto ERIC. Attualmente CLARIN è coinvolto come comunità in due progetti Horizon 2020: CLARIN PLUS[33], che mira a creare procedure e strumenti per facilitare la fase di ingresso nell'infrastruttura dei nuovi paesi membri, e PARTHENOS[34], un progetto che vede la partecipazione di tutte le più grandi infrastrutture per le scienze umane e sociali e patrimonio culturale (tra cui anche DARIAH[35]), al fine di armonizzare le loro pratiche per quanto riguarda la gestione dei dati della ricerca e dei contenuti digitali, in particolare promuovendo l'uso di metadati e standard condivisi e affermati.

5. Il network CLARIN italiano

L'adesione dell'Italia a CLARIN ERIC, pur finalizzata nel 2015, ha radici più antiche. Dopo la fine della fase preparatoria, l'ILC ha continuato a essere presente come osservatore attivo nei vari Comitati di ERIC. Oggi, grazie al finanziamento del MIUR, garantito per cinque anni e volto a coprire le tasse di iscrizione a ERIC, l'Italia fa parte a pieno titolo dei paesi membri, con il compito dunque di implementare l'infrastruttura nazionale al fine di offrire servizi alla comunità di riferimento.

5.1 Il consorzio CLARIN-IT

Ogni paese membro ha l'obbligo di creare una rete di istituzioni ed organizzazioni che costituisce il consorzio. L'Istituto di Lingustica Computazionale, che ha il compito di coordinamento delle attività infrastrutturali, coordina, tramite la persona del Coordinatore Nazionale, il consorzio CLARIN-IT[36]. Il consorzio nazionale funge da interfaccia con ERIC, contribuendo a sviluppare servizi centralizzati a beneficio di tutta la comunità.

Presso la maggioranza dei paesi membri europei, la rete e l'infrastruttura nazionale sono costruite nel quadro di progetti finanziati interamente dai governi nazionali, come per esempio il progetto LINDAT (supportato dal Ministero dell'Istruzione della Repubblica Ceca) all'interno del quale si iscrivono tutte le attività del CLARIN nazionale; oppure vengono co-finanziati dai governi nazionali e dall'Unione Europea all'interno dell'Agenda Strategica delle Infrastrutture, come il caso del giovane network greco. Il

33 https://www.clarin.eu/content/factsheet-clarin-plus

34 http://www.parthenos-project.eu/.

35 DARIAH (Digital Research Infrastructure for Arts and Humanities) http://www.dariah.eu/. Rispetto a CLARIN, DARIAH si occupa non solo di risorse testuali, ma anche di patrimonio culturale materiale. Ciò nonostante le intersezioni tra le due infrastrutture sono molteplici e in diversi paesi lo stesso consorzio gestisce la partecipazione nazionale alle due iniziative. Il caso più eclatante è quello dei Paesi Bassi, con il consorzio CLARIAH (dalla fusione dei due nomi), ma si veda anche il caso dell'Austria.

36 www.clarin-it.it

MIUR, fino ad oggi, finanzia CLARIN-IT[37] all'interno di una linea di attività del Progetto Premiale SM@RTINFRA[38] per quanto concerne l'implementazione del nodo nazionale dell'infrastruttura ed il suo potenziamento. L'obiettivo globale di SM@RTINFRA è di ampio respiro e, in linea con ESFRI e con la Integrated Infrastructures Initiative (I3), mira a favorire le sinergie e creare una struttura di coordinamento nazionale dei nodi italiani delle infrastrutture di ricerca europee di Social Sciences and Humanities e Cultural Heritage, tra cui CLARIN-ERIC e DARIAH-ERIC. Per il 2017, il piano è quello di concertare con il Ministero una proposta progettuale che, attraverso una raccolta di bisogni e requisiti della comunità, motivi gli studiosi delle scienze umane all'utilizzo delle tecnologie digitali. Potenziando la diffusione di queste ultime attraverso l'infrastruttura CLARIN, si favorirà al contempo una spinta innovativa nei paradigmi e nelle metodologie di ricerca del settore.

La formazione del consorzio deve rispondere a criteri di rappresentatività della comunità. I membri sono individuati secondo un criterio geografico, così da garantire una copertura territoriale, mentre un criterio scientifico assicura la copertura di quei settori di ricerca legati allo studio del linguaggio, linguisti, linguisti computazionali e ingegneri della lingua. CLARIN mira ad attrarre le comunità scientifiche dei vari ambiti: storia classica, antica, moderna e contemporanea, studi letterari, scienze politiche, scienze della comunicazione, sociologia, teologia, filosofia, antropologia sociale ed etnografia, linguistica e filologia. Inoltre, CLARIN si rivolge più in generale a tutte quelle discipline che possono fare uso, seppure meno massicciamente, di risorse e tecnologie del testo, quali il diritto, l'educazione, l'archeologia, le discipline artistiche e dello spettacolo, il design, l'architettura, la musica, la demografia, la geografia umana, l'economia, gli studi sociali e politici, la storia della scienza e della medicina.

Per facilitare il coinvolgimento, CLARIN-IT, attraverso il progetto centrale CLARIN-PLUS, offre ai ricercatori supporto per brevi soggiorni di ricerca presso centri specializzati favorendo lo scambio di competenze tramite lo strumento dei Mobility Grant stimola, inoltre, la partecipazione dei partners interessati a giornate di lavoro attorno a temi rilevanti condivisi dalla Agenda Strategica redatta ogni anno dall'Assemblea Generale e dal Consiglio Scientifico; promuove infine l'organizzazione o la partecipazione a workshop dedicati a tematiche di ricerca condivise, come recentemente la storia orale o le risorse lessicali[39].

I partners del consorzio nazionale CLARIN-IT possono partecipare alla CLARIN Annual Conference (CAC), l'evento annuale riservato alla disseminazione e condivisione dei risultati presso la comunità CLARIN. La conferenza, organizzata, di volta in volta, in un paese membro e rivolta sia agli sviluppatori dell'infrastruttura che agli utenti, offre una panaoramica dello stato di avanzamento della parte tecnologica ma anche e soprattutto delle ricerche condotte grazie all'infrastruttura[40].

Attualmente, la composizione della rete italiana è in costante evoluzione. Il primo membro partecipante è la Federazione di Identità IDEM, partner tecnico, che ha collaborato, fin dalle prime fasi di costituzione del centro nazionale, allo sviluppo

37 L'Istituto di Linguistica Computazionale sostiene gran parte dell'attività tecnologica e di partecipazione ai comitati tecnici centrali grazie al proprio personale attraverso il meccanismo dell'in-kind.

38 Progetto del Programma FOE - Fondo Ordinario per gli Enti di Ricerca - coordinato dal Dipartimento di Scienze Umane e Sociali.

39 Per una lista degli eventi supportati da CLARIN si veda `https://www.clarin.eu/events`

40 L'ultima CAC2016 si è svolta in Francia, paese ora in procinto di aderire come membro osservatore. A questo indirizzo è possibile trovare il programma degli interventi: `https://www.clarin.eu/event/2016/clarin-annual-conference-2016-aix-en-provence-france`

e implementazione dei servizi di accesso, autenticazione e autorizazatione tramite il single-sign-on.

Presso la comunità dei produttori di risorse, che annovera specialisti di TAL, sviluppatori di strumenti computazionali e di tecnologie digitali, si registra una forte attrazione verso la missione di CLARIN, con la manifesta volontà di censire e condividere le risorse nel repository nazionale, unita anche alla consapevolezza delle ricadute che l'esposizione su una vetrina internazionale come quella del VLO garantisce in termini di visibilità.

La Fondazione Bruno Kessler FBK di Trento, con il gruppo di ricerca in Digital Humanities, ha manifestato interesse a condividere i propri strumenti ed integrarli in CLARIN sotto forma di servizi linguistici per l'italiano. Un altro partecipante (già membro del partenariato della fase preparatoria) è l'Accademia Europea EURAC di Bolzano, un centro di ricerca applicata privato con forti interessi verso le tecnologie per il multilinguismo e le problematiche delle lingue minoritarie regionali delle zone di frontiera. Connessa per interessi di ricerca all'Istituto di Linguistica Computazionale, con il quale collabora nel settore del TAL, è l'Università di Pisa, con il Laboratorio Coling Lab del Dipartimento di Filologia Letteratura e Linguistica. Una collaborazione con accordo formale è stata accesa con l'Univesità Cattolica di Milano, Facoltà di Scienze Linguistiche, per lo sviluppo di strumenti digitali per l'analisi linguistica delle lingue classiche.

Sul fronte dei potenziali utenti, i linguisti, i filologi, gli storici ripongono molte aspettative in CLARIN. E' il caso per esempio del Dipartimento di Linguistica dell'Università di Siena, che ha siglato un accordo formale con CLARIN e porta nel consorzio la Associazione Italiana di Storia Orale[41], con cui condivide scopi ed interessi di ricerca attorno agli studi orali. E' allo studio la fattibilità di depositare e preservare nel repository nazionale dell'ILC l'archivio dei dati di Gra.fo (Calamai and Fronini 2016) sviluppato ed ospitato presso la Scuola Normale Superiore di Pisa. Il patrimonio dei dati sarà reso ancora più prezioso grazie allo sviluppo di servizi dedicati per l'interrogazione, la navigazione e alla creazione di un sistema di riconoscimento del parlato che consenta la trascrizione automatica. L'attività sarà condotta come sforzo congiunto tra l'ILC, l'Università di Siena, la Scuola Normale Superiore, la Fondazione Bruno Kessler, coinvolgendo anche il consorzio olandese (CLARIAH).

Il consorzio CLARIN-IT contribuisce anche ad uno degli aspetti su cui l'ERIC punta e cioè il settore della formazione, con il lancio di tesi magistrali o di dottorato in linea con gli scopi di CLARIN o che dimostrino originalità e innovatività grazie ai suoi strumenti. Con l'Università di Parma, Dipartimento di Antichistica, Lingue, Educazione, Filosofia (membro istituzionale dal prossimo anno), l'ILC sta coordinando una tesi sulle risorse e strumenti per le edizioni digitali, sviluppata nel quadro di CLARIN.

5.2 Il Centro ILC4CLARIN

L'ILC, in quanto capofila del consorzio, sta procedendo alla costruzione del centro CLARIN, che ha il ruolo di fornire servizi a livello nazionale. Altri membri del consorzio potranno sulla base delle loro competenze tecniche collaborare alla gestione del centro ILC4CLARIN[42] o costruire nodi locali.

41 AISO http://aisoitalia.org
42 https://dspace-clarin-it.ilc.cnr.it/

I principali servizi offerti da ILC4CLARIN, sono il servizio di deposito e perennizzazione delle risorse linguistiche, la messa a disposizione di dati, strumenti e servizi di analisi linguistica avanzata multilingui, con particolare focus sull'italiano, e le lingue classiche. Altro aspetto di notevole interesse è la consulenza fornita ai partners sia sugli aspetti tecnici che sugli standard e le questioni legate alle licenze. Infine, porta avanti una capillare opera di disseminazione tramite presentazioni agli eventi del settore.

Per quanto riguarda il deposito di metadati e dati, il repository di ILC4CLARIN, ora in fase di popolamento, raccoglie in primo luogo le risorse dell'ILC, ma permette anche il deposito da parte degli altri partner, fornendo consulenza per la compilazione dei metadati.

Sia l'accesso che il deposito di risorse sul repository sono forniti tramite single sign on a tutti gli utenti accademici la cui istituzione aderisce alla rete IDEM e CLARIN-IT. Collabora inoltre con IDEM-GARR per favorire l'adesione di alcune istituzioni non ancora federate. Grazie proprio alla collaborazione tra IDEM e la CLARIN Service Provider Federation, una serie di strumenti ad accesso riservato sono disponibili agli utenti italiani, tra cui primo fra tutti il servizio WebLicht[43].

ILC sta procedendo ad un intenso lavoro di recupero dall'obsolescenza di numerose risorse del propiro patrimonio digitale, al fine di adeguarle ai formati attuali e renderle disponibili con standard e licenze adeguate. In particolare si sta procedendo alla pubblicazione delle risorse lessicali (come ItalWordNet[44]), alla ricodifica dei corpora legacy in TEI (Sassolini, Cucurullo, and Sassi 2014; Sassolini et al. 2014), e alla reingegnerizzazione di alcuni strumenti TAL, al fine di rendere tali risorse disponibili attraverso i canali della Federated Content Search e di WebLicht.

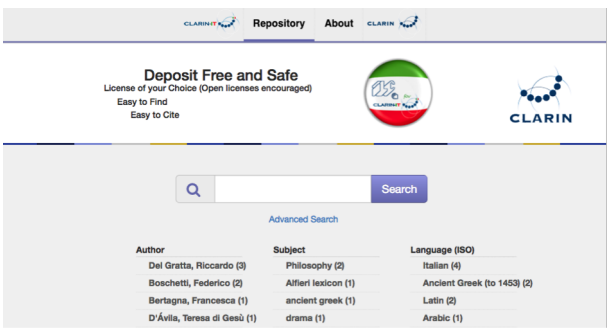

Figura 6
The ILC4CLARIN repository.

ILC4CLARIN si pone inoltre come il punto di riferimento livello nazionale per quanto riguarda la formazione e la disseminazione di competenze circa l'utilizzo di risorse e tecnologie CLARIN, la perennizzazione dei dati e la loro adeguata descrizione con metadati standard, mediante la organizzazione di tutorial dimostrativi e workshop, favorendo anche gli scambi di ricerca, supportati dall'ERIC centrale.

In prospettiva futura, il centro ILC4CLARIN si propone di continuare nel solco della tradizione dell'Istituto, mettendo in particolare l'accento sulle tematiche degli standard

43 Per una lista completa si veda `https://www.clarin.eu/content/easy-access-protected-resources`

44 Trovate qui l'entrata di IWN V2, `http://hdl.handle.net/20.500.11752/ILC-62` con risorsa scaricabile

per corpora e lessici. Questo expertise si coniugherà con le nuove tendenze nelle tecnologie del linguaggio, che vedono l'emergere di nuovi formati legati in particolare al web semantico e ai linked open data per i dati linguistici (L-LOD).

6. Conclusioni e prospettive future

La diffusione degli scopi di CLARIN, delle sue potenzialità presso la comunità di riferimento e l'individuazione di una rete di utenti motivati è un'attività capillare che implica una quotidiana opera di evangelizzazione e coinvolgimento. CLARIN non può in alcun modo svilupparsi senza coinvolgere una fitta rete di utenti finali. In particolare, è indispensabile che CLARIN diventi uno strumento di ricerca e lavoro per gli scienziati umani e sociali, ma anche di insegnamento a livello universitario e di scoperta per i cittadini che si interessano al patrimonio culturale testuale. CLARIN-IT è ancora all'inizio, ma è motivo di grande soddisfazione sapere che, dopo anni di pionierismo nel settore, con l'iscrizione dell'Italia a membro permanente, i ricercatori italiani possono beneficiare (e in parte stanno già beneficiando) dei vantaggi offerti dall'infrastruttura, soprattutto in termini di accesso alle risorse, di possibilità di scambi internazionali e di inviti ad eventi e giornate dedicate a tematiche di ricerca del settore.

La partecipazione al consorzio nazionale di tutti i più importanti centri di ricerca nelle tecnologie del linguaggio permetterà di realizzare l'obiettivo della copertura, garantendo la preservazione a lungo termine del grande patrimonio di risorse digitali nazionali e un loro più facile accesso alla comunità scientifica.

In linea con le direzioni dell'ultima Assemblea Generale di ERIC, il coinvolgimento massiccio e attivo degli utenti è ora vitale. E' importante veder crescere intorno a CLARIN-IT una nutrita comunità di utilizzatori, ricercatori, studenti, citizen scientists in scienze umane e sociali, che grazie alle loro esigenze, domande, bisogni di ricerca, presenteranno sfide tecnologiche sempre più stimolanti, favorendo così lo sviluppo e la crescita di CLARIN-IT e di CLARIN in generale.

Riconoscimenti

Ringraziamo i team tecnici e di coordinamento di ILC4CLARIN e di CLARIN-IT: Paola Baroni, Sebastiana Cucurullo Alessandro Enea, Riccardo Del Gratta, Simone Marchi, Valeria Quochi. Una menzione speciale per la direttrice dell'ILC, Simonetta Montemagni, per i consigli, il supporto fattivo e la disponibilità a far fronte alle esigenze via via emergenti in CLARIN-IT.
Per informazioni su CLARIN-IT contattare:
<coordination@clarin-it.it> <communication@clarin-it.it>

Bibliografia

ALA (American Library Association). 2007. Definitions of digital preservation.

Broeder, Daan, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 43–47, Valletta, Malta. European Language Resources Association (ELRA).

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability — the Component Metadata Infrastructure (CMDI). In *Balisage Series on Markup Technologies*, volume 7.

Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a component metadata infrastructure. In Victoria Arranz, Daan Broeder, Maria Gavrilidou, Monica Monachini, and Thorsten Trippel, editors, *Proceedings of the LREC*

Workshop Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme, pages 1–4.

Calamai, Silvia and Francesca Fronini. 2016. Not quite your usual kind of resource. Gra.fo and the documentation of Oral Archives in CLARIN. In *Proceedings of the CLARIN Annual Conference 2016*, Aix-en-Provence, France.

Calzolari, Nicoletta, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo, and Claudia Soria. 2012. The LRE Map. Harmonising Community Descriptions of Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089, Istanbul, Turkey.

Calzolari, Nicoletta, Monica Monachini, and Valeria Quochi. 2011. Interoperability framework: The FLaReNet action plan proposal. In *Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 41–49, Chiang Mai, Thailand.

Calzolari, Nicoletta, Monica Monachini, and Claudia Soria. 2013. LMF – Historical Context and Perspectives. In Gil Francopoulo, editor, *LMF Lexical Markup Framework*. John Wiley & Sons, pages 1–18.

de Jong, Franciska. 2009. NLP and the Humanities: The Revival of an Old Liaison. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 10–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eckart, Thomas, Alexander Hellwig, and Twan Goosen. 2015. Influence of Interface Design on User Behaviour in the VLO. In *CLARIN Annual Conference 2015 in Wroclaw, Poland*, pages 17–21.

Francopoulo, Gil. 2013. *LMF Lexical Markup Framework*. John Wiley & Sons, May.

Gavrilidou, Maria, Penny Labropoulou, Stelios Piperidis, Monica Monachini, Francesca Frontini, Gil Francopoulo, Victoria Arranz, and Valérie Mapelli. 2011. A metadata schema for the description of language resources (lrs). In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Godfrey, John J. and Antonio Zampolli. 1997. Language Resources. In Antonio Zampolli and Giovanni Battista Varile, editors, *Survey of the State of the Art in Human Language Technology. Linguistica Computazionale, XII-XIII*. Giardini Editori e Stampatori in Pisa. (Also Cambridge University Press), pages 381–384.

Goosen, Twan, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Ďurčo, and Oliver Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Selected Papers from the CLARIN 2014 Conference*, pages 36–53.

Hinrichs, Erhard and Steven Krauwer. 2014. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1525–1531, May.

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg and Sue Ellen Wright. 2008. ISOcat: Corralling Data Categories in the Wild. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 887–891, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Mariani, Joseph and Gil Francopoulo. 2015. Language Matrices and a Language Resource Impact Factor. In Núria Gala, Reinhard Rapp, and Gemma Bel-Enguix, editors, *Language Production, Cognition, and the Lexicon*. Springer International Publishing, Cham, pages 441–471.

Odijk, Jan. 2014a. CLARIN-NL: Major Results. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2187–2193, Reykjavik, Iceland. European Language Resources Association (ELRA).

Odijk, Jan. 2014b. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. Working paper, October.

Piotrowski, Michael. 2012. Natural Language Processing for Historical Texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157, September.

Sassolini, Eva, Sebastiana Cucurullo, and Manuela Sassi. 2014. Methods of textual archive preservation. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 334–338. Pisa University Press.

Sassolini, Eva, Manuela Sassi, Sebastiana Cucurullo, Alessandra Cinini, and Stefano Sbrulli. 2014. Industrial Philology: Problems and techniques of data and archives preservation for future generations. In D.J. Farace and J. Frantzen, editors, *GL 15 - The grey audit: a field assessment in grey literature : Fifteenth International Conference on Grey Literature ; Slovak Centre of Scientific and Technical Information, Bratislava, 2 - 3 December 2013; GL 15 ; conference proceedings (Bratislava, 2-3 december 2013). Book of abstracts.*, pages 73 – 77, Amsterdam :. TextRelease.

Schuurman, Ineke, Menzo Windhouwer, Oddrun Ohren, and Daniel Zeman. 2016. CLARIN Concept Registry: The New Semantic Registry. In *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wrocław, Poland*, pages 62–70, Linköping, Sweden. Linköping University Electronic Press, LinkÄ¶pings universitet.

Soria, Claudia, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Nicoletta Calzolari, and others. 2012. The FLaReNet Strategic Language Resource Agenda. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'10)*, pages 1379–1386.

Soria, Claudia, Monica Monachini, Nicoletta Calzolari, Valeria Quochi, Khalid Choukri, Joseph Mariani, Nuria Bel, Stelios Piperidis, and Jan Odijk. 2011. Final FLaReNet deliverable: Language Resources for the Future - The Future of Language Resources, August.

Uytvanck, Dieter Van, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. 2010. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 900–903, Valletta, Malta. European Language Resources Association (ELRA).

Váradi, Tamás, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. 2008. CLARIN: Common Language Resources and Technology Infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1244–1248, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Wittenburg, Peter, Nuria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicova, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciey Piasecki, and others. 2010. Resource and service centres as the backbone for a sustainable service infrastructure. In *(LREC'10)*, pages 5–9.

Zampolli, Antonio, Nicoletta Calzolari, Mona Baker, and Johanna G. Kruyt, editors. 1995. *Towards a Network of European Reference Corpora.* Report of the NERC Consortium Feasibility Study. Linguistica Computazionale, XI. Giardini Editori e Stampatori in Pisa, Pisa.

Zeldenrust, Douwe A. and Marc Kemps-Snijders. 2011. Establishing connections: Making resources available through the CLARIN infrastructure. *Supporting Digital Humanities 2011, Answering the Unaskable. Kopenhagen:[sn].*