CLARIN-IT

The Italian Common Language Resources and Technology Infrastructure

CLARIN-IT: servizi per la comunità italiana delle scienze umane e sociali

Monica Monachini - CLARIN Italian National Coordinator

Alessandro Enea - Responsible of ILCforCLARIN & contact person for IDEM

Francesca Frontini - Standing Committee for CLARIN Technical Centres (SCCTC)

ILC-CNR

National Representative

1° Ottobre 2015

- L'Italia diventa membro della Infrastruttura CLARIN-ERIC
  Common Language Resource Infrastructure for Social Sciences and Humanities

- Una opportunità in più per chi si occupa di Discipline Umanistiche

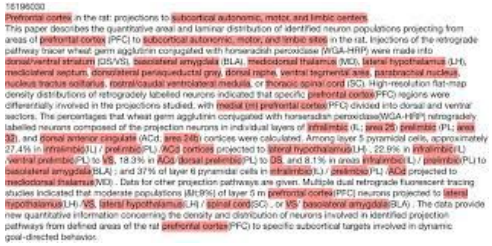
Cosa si intende per IR

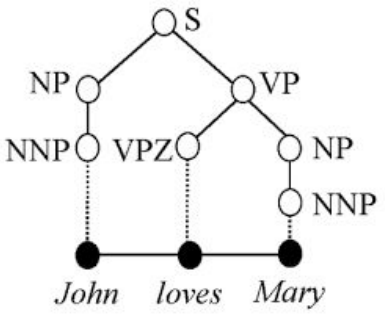
- *Insieme di risorse strumentazioni, competenze, dati, servizi, funzioni specializzate e dedicate a realizzare in modo continuativo i flussi necessari alla ricerca scientifica e tecnologica*

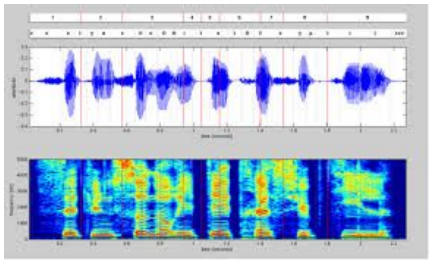
- CLARIN è una infrastruttura di tipo ERIC formalmente riconosciuta
  - completato la ESFRI roadmap
  - the European Strategy Forum on Research Infrastructures
- ha raggiunto lo stato di "land-mark"

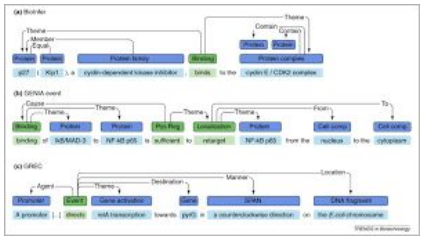
Cosa sono le Risorse Linguistiche

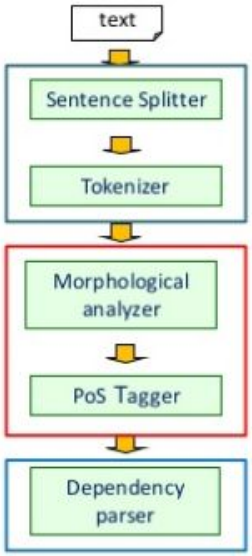
- Risorse linguistiche digitali:
  - corpora (annotati e non), lessici, ontologie
  - grammatiche formali, modelli statistici
  - strumenti per il trattamento automatico del linguaggio naturale (scritto, parlato, multimodale)









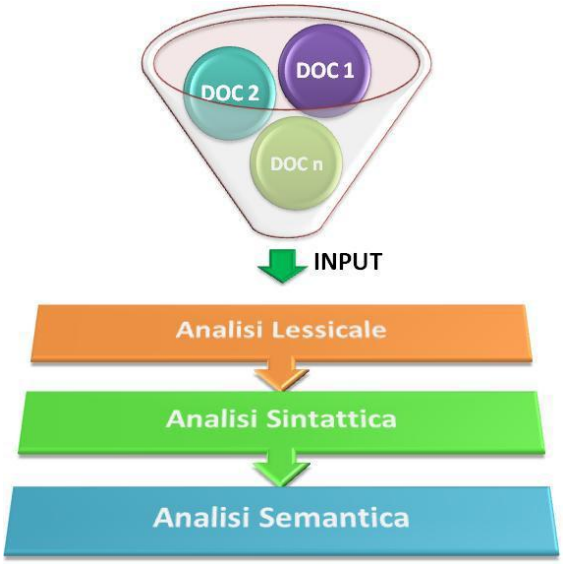


A cosa servono le Risorse Linguistiche

- A cosa servono:
  - analisi automatica del linguaggio naturale
  - annotazione automatica dei testi
  - studio della lingua
  - analisi di fonti testuali

- Costruire queste risorse richiede tempo e denaro:
  - cruciale garantire un facile accesso alle risorse per i ricercatori di diverse istituzioni e paesi
  - fondamentale garantire che tali risorse rimangano fruibili nel tempo

- Un cambio di paradigma:
  - condivisione, accesso, replicabilità dei risultati



CLARIN-ERIC

- CLARIN è una e-infrastructure (infrastruttura digitale, immateriale) che mira a fornire un accesso facile e sostenibile alle risorse linguistiche digitali da parte dei ricercatori e degli studiosi dei paesi membri

- CLARIN fornisce inoltre strumenti avanzati per trovare le risorse adeguate per la propria ricerca ovunque esse si trovino e combinarle per analizzare, esplorare, annotare dati digitali

- CLARIN raggruppa 18 paesi (più alcuni membri osservatori)
  - ciascun paese è a sua volta organizzato in un consorzio, con uno o più centri che forniscono dati e strumenti (propri e di altre istituzioni per le quali forniscono il servizio di deposito)

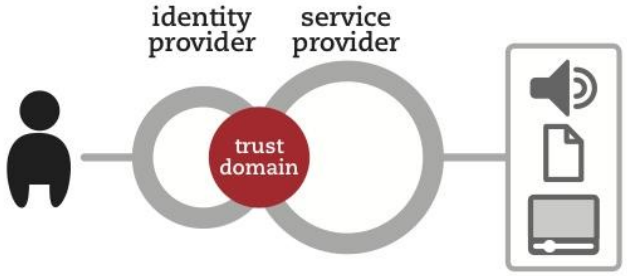
Members	National Consortium (NC)	Leading NC partner	Number NC partners	National coordinator
Austria	CLARIN Austria	CLARIN Centre Vienna	4	Karlheinz Mörth
Bulgaria	CLARIN Bulgaria	Bulgarian Academy of Sciences	10	Kiril Simov
Czech Republic	LINDAT Clarin	Charles University Prague	4	Eva Hajičová
Denmark	CLARIN Denmark	University of Copenhagen	4	Bente Maegaard
Dutch Language Union	Dutch Language Union	Dutch Language Union	4	Jan Theo Bakker
Estonia	CLARIN Estonia	Center of Estonian Language Resources	3	Kadri Vider
Finland	FIN-CLARIN			Krister Lindén
Germany	CLARIN-D	University of Tuebingen	12	Erhard Hinrichs
Greece	CLARIN Greece			Stelios Piperidis
Italy	CLARIN Italy			Monica Monachini
Lithuania	CLARIN-Lithuania	Vytautas Magnus University	3	Jurgita Vaičenonienė
The Netherlands	CLARIN-NL	Utrecht University	24	Jan Odijk
Norway	CLARINO	University of Bergen	8	Koenraad De Smedt
Poland	CLARIN Poland	Wroclaw University of Technology	6	Maciej Piasecki
Portugal	CLARIN-Portugal	University of Lisbon	22	António Branco
Slovenia	CLARIN.SI	Jožef Stefan Institute	12	Tomaž Erjavec
Sweden	SWE-CLARIN	Språkbanken	9	Lars Borin

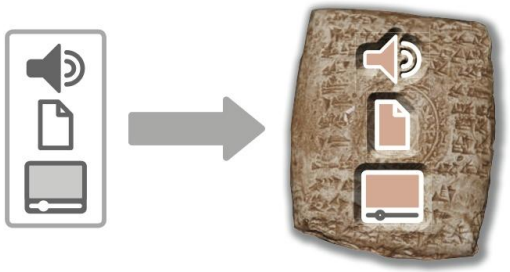
Observer	National Consortium (NC)	Leading partner NC	Number partners NC	National coordinator
United Kingdom	CLARIN-UK	Oxford University	12	Martin Wynne

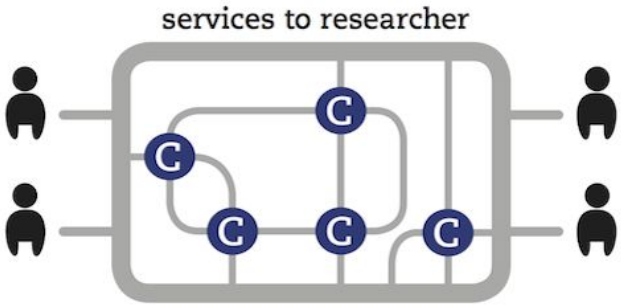

Mappa
Satellite
Svezia
Norvegia
Finlandia
Mar Baltico
Estonia
Lettonia
Mosca
Москва
Mar del Nord
Danimarca
Lituania
Regno Unito
Irlanda
Polonia
Bielorussia
Germania
Belgio
Ucraina
Slovacchia
Moldavia
Ungheria
Francia
Romania
Croazia
Serbia
Mar Nero
Italia
Roma
Bulgaria
Georgia
Portogallo
Madrid
Spagna
Mar Tirreno
Istanbul
Grecia
Turchia
Azerbaijan
Siria
Tunisia
Mar Mediterraneo
Libano
Iraq
Google

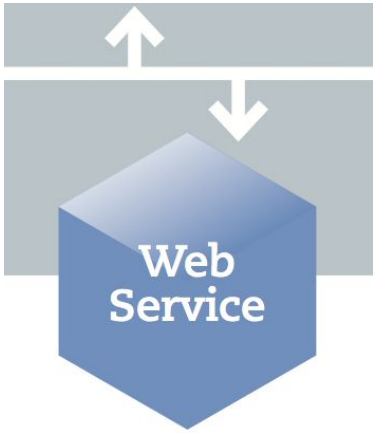
Centri CLARIN-ERIC

- Cosa fa un centro CLARIN?
  - Deposito
  - Documentazione
  - Esposizione dei metadati per harvesting nel catalogo centrale
  - Autenticazione
  - Accesso a risorse
  - Accesso a servizi









CLARIN Conference 2016: Call for Papers

Call for papers **** CLARIN 2016 **** CLARIN is happy to announce the 5th CLARIN Annual Conference and calls for the submission of papers. The series of CLARIN Annual

Read more >

@Twitter

UWuttke

"On my way to Vienna for workshop with @CLARINERIC at @oeaw! Maybe also some sightseeing;-)"

4 days 3 hours ago.

CLARIN-NL in a nutshell

The CLARIN-NL project is a large national project in the Netherlands (2009-2015) which aims to make a significant contribution to the European CLARIN infrastructure.
The CLARIN infrastructure is a **research infrastructure** intended for humanities researchers that work with language data and tools. The researcher will be able to find data and tools relevant for his/her research via the CLARIN infrastructure and get access to them. We also aim to make it possible to apply tools to data in such a way that no technical background is needed or ad-hoc adaptations to the tools or the data are necessary.

The researcher's research project may result in new **data** and **tools**, which he/she can make available to other researchers via the CLARIN infrastructure. The CLARIN research infrastructure will thus make humanities research related to language easier, faster, better, and in some cases even possible for the first time.

CENTRE	FOCUS ON RESOURCES	RESOURCE EXAMPLES
MEERTENS INSTITUTE	relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.	typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see here
MPI	related to the study of psychological, social and biological foundations of language.	documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see here
HUYGENS ING	related to the study of history and literature of the Netherlands.	historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see here
INL	that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.	lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see here
DANS	sustained access to digital research data.The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.	text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see here

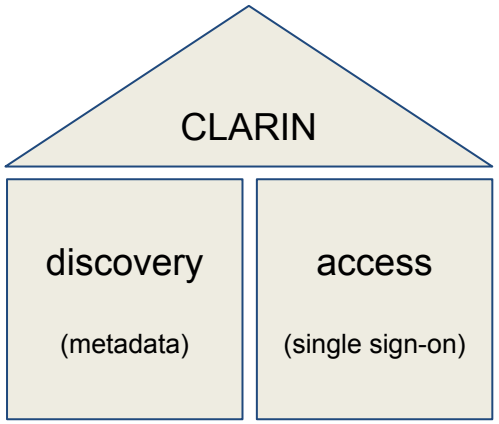
CENTRE	FOCUS ON RESOURCES	RESOURCE EXAMPLES
MEERTENS INSTITUTE	relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.	typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see here
MPI	related to the study of psychological, social and biological foundations of language.	documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see here
HUYGENS ING	related to the study of history and literature of the Netherlands.	historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see here
INL	that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.	lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see here
DANS	sustained access to digital research data.The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.	text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see here

CENTRE	FOCUS ON RESOURCES	RESOURCE EXAMPLES
MEERTENS INSTITUTE	relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.	typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see here
MPI	related to the study of psychological, social and biological foundations of language.	documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see here
HUYGENS ING	related to the study of history and literature of the Netherlands.	historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see here
INL	that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.	lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see here
DANS	sustained access to digital research data.The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.	text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see here

CENTRE	FOCUS ON RESOURCES	RESOURCE EXAMPLES
MEERTENS INSTITUTE	relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.	typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see here
MPI	related to the study of psychological, social and biological foundations of language.	documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see here
HUYGENS ING	related to the study of history and literature of the Netherlands.	historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see here
INL	that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.	lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see here
DANS	sustained access to digital research data.The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.	text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see here

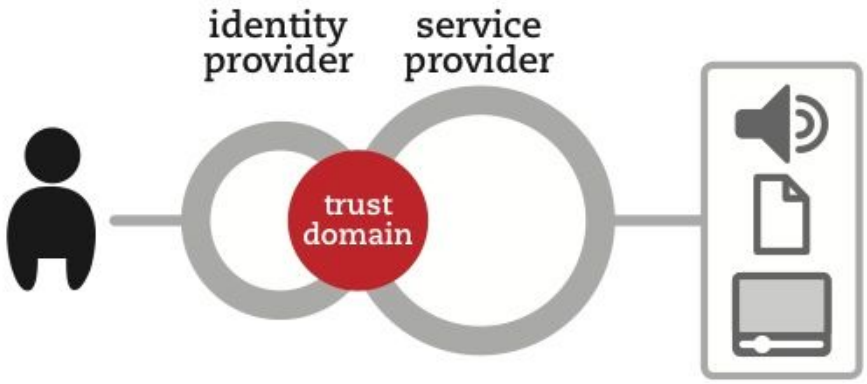
CLARIN-ERIC: i pilastri

- CLARIN si configura quindi come un network di repositories e centri di vario tipo
- Pilastri fondamentali a cui ogni centro deve attenersi:
  - risorse devono essere documentate con un formato standard di **metadati** (CMDI), poi **riversati in un meta-catalogo centralizzato** (VLO)
  - **l'accesso effettivo alle risorse deve essere o completamente libero o attraverso un sistema unificato di autenticazione (single sign-on)**
  - questo è particolarmente importante per tutte quelle risorse digitali che sono disponibili soltanto per scopi accademici



La CLARIN Federation

- CLARIN si federa con le federazioni di identità nazionali (ex. IDEM-GARR) e internazionali (ex. EDUGAIN), dando priorità all'accesso dei paesi Full Member
- Questo evita ai singoli centri CLARIN di dover siglare accordi con ogni federazione
- CLARIN ERIC ha costituito una propria federazione (SPF) per connettere i propri service providers (=centri) con le federazioni nazionali di identità (=utenti)



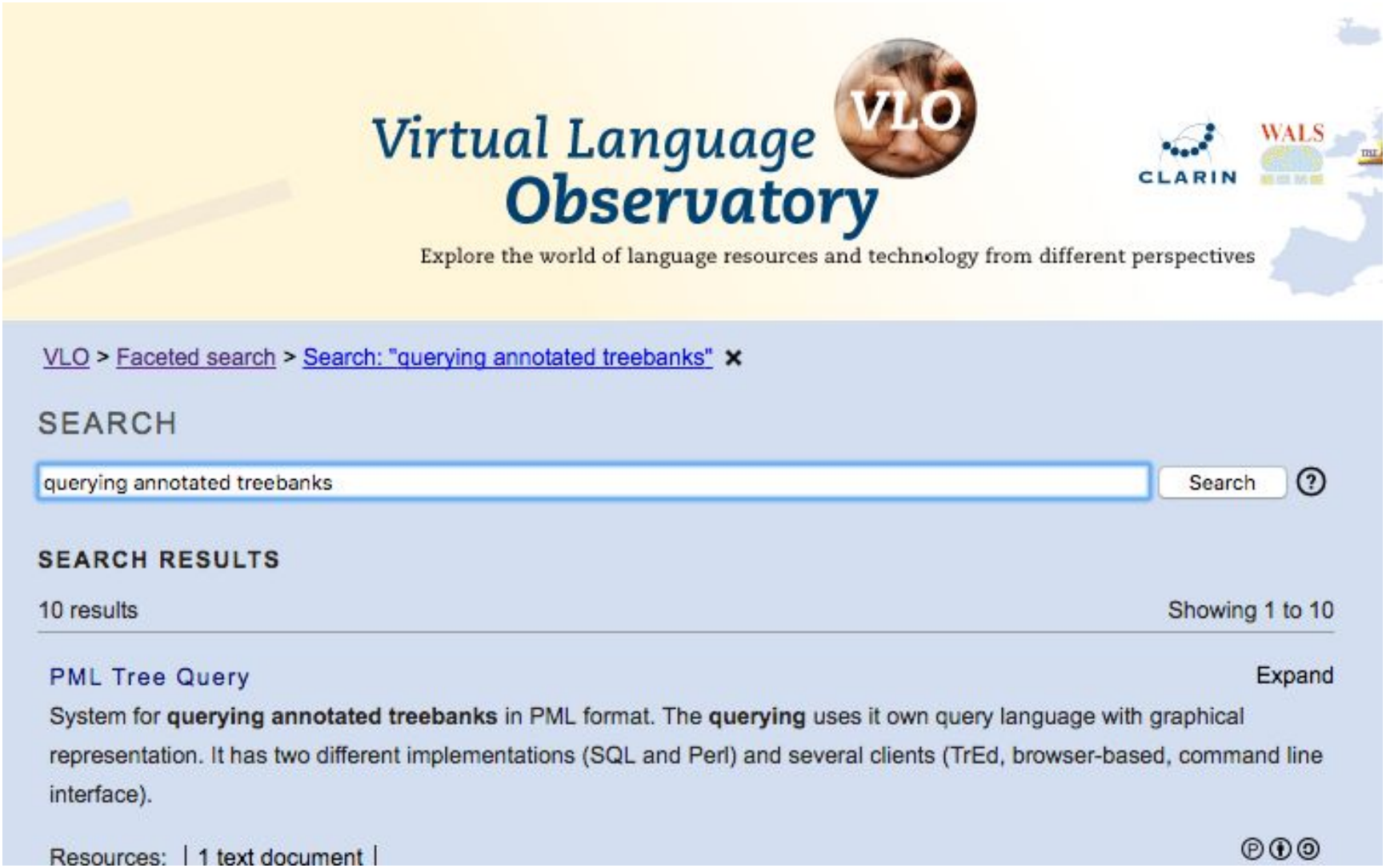
Participating Identity Federations

We aim at including all European federations, however CLARIN member countries get the highest priority.

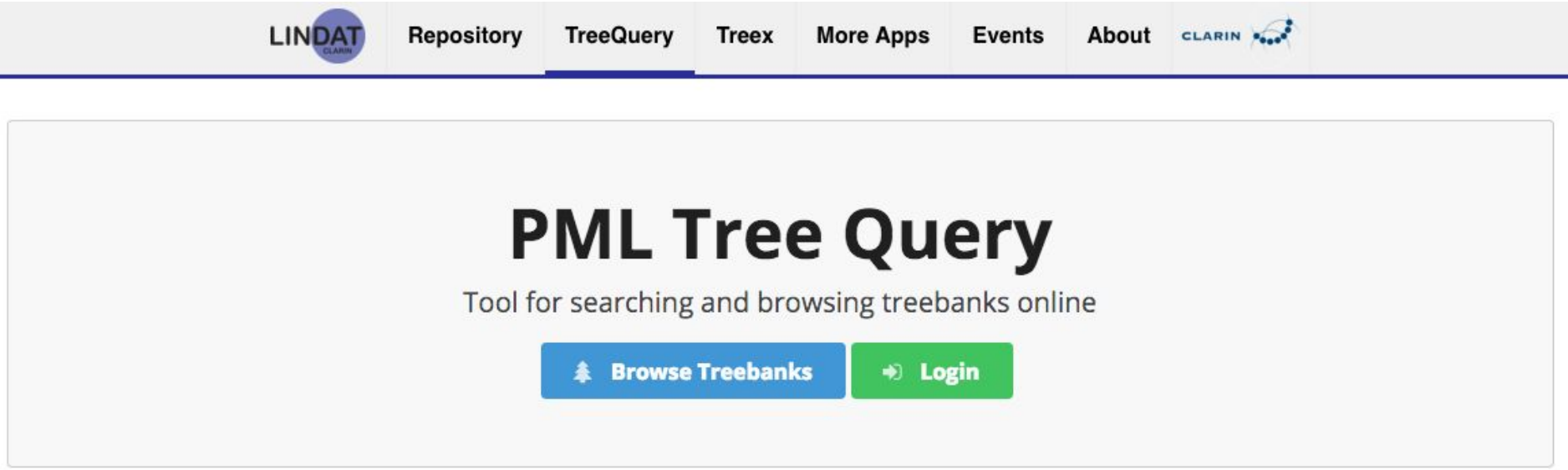
Country	Federation	Status	Attribute release	Opt-in
Austria	ACOnet	Fully connected.		OK (not required)
Belgium	Belnet	Fully connected.	OK (for Antwerp, Gent, Leuven)	OK (not required)
Czech Republic	eduID.cz	Fully connected.		OK (not required)
Denmark, Iceland	WAYF	Fully connected (via eduGAIN and Kalmar Union).	OK	OK (not required)
Estonia	TAAT	Fully connected (via eduGAIN).	OK	OK (not required, attribute release mandatory)
Finland	Haka	Fully connected. Also includes Kalmar Union	OK	OK (not required)
Germany	DFN-AAI	Fully connected.	Problems for some IdPs (e.g. Tübingen)	OK (not required)
Greece	GRNET Federation	Connection planned since 2015.		
Italy	IDEM GARR	Fully connected (via eduGAIN).		OK (not required)
Lithuania	LITNET fedi	Connection planned since 2015.		

Concretamente....

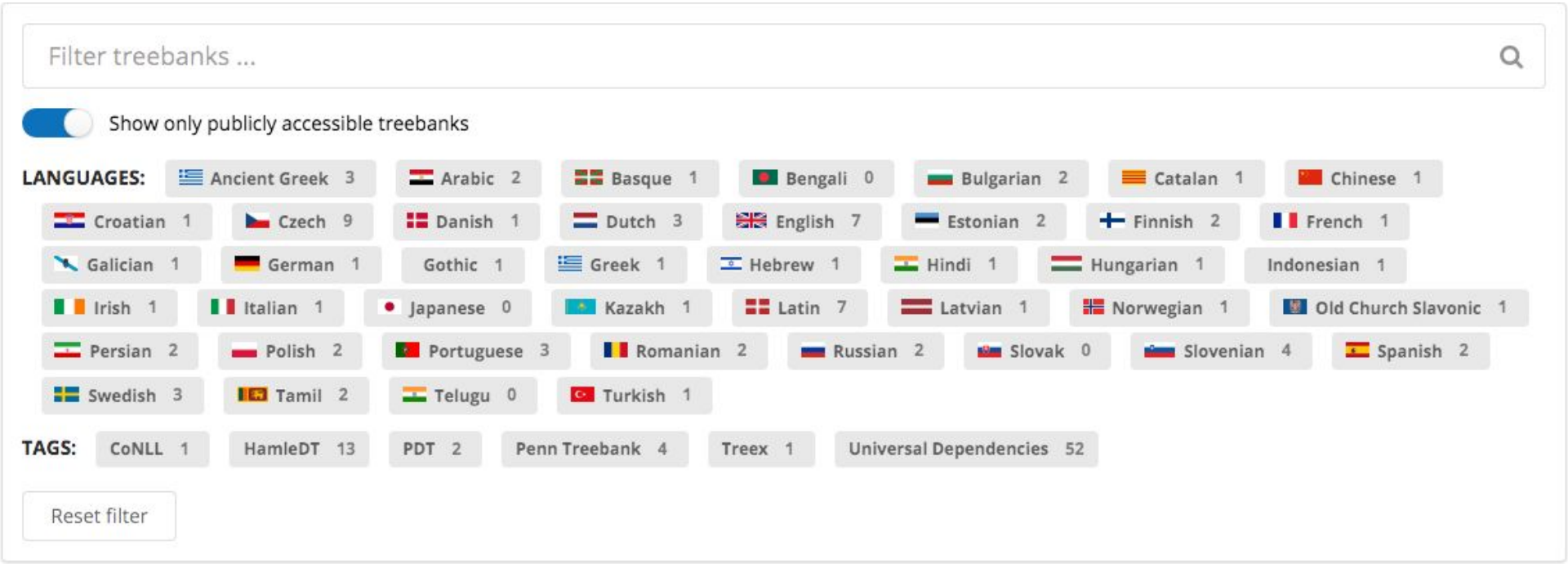
- Un ricercatore italiano vuole fare una ricerca sulla sintassi delle varie lingue
- Cerca uno strumento per la ricerca di corpora annotati sintatticamente (tree-bank, visualizzate come alberi)
- In particolare è interessato a studiare il giapponese....



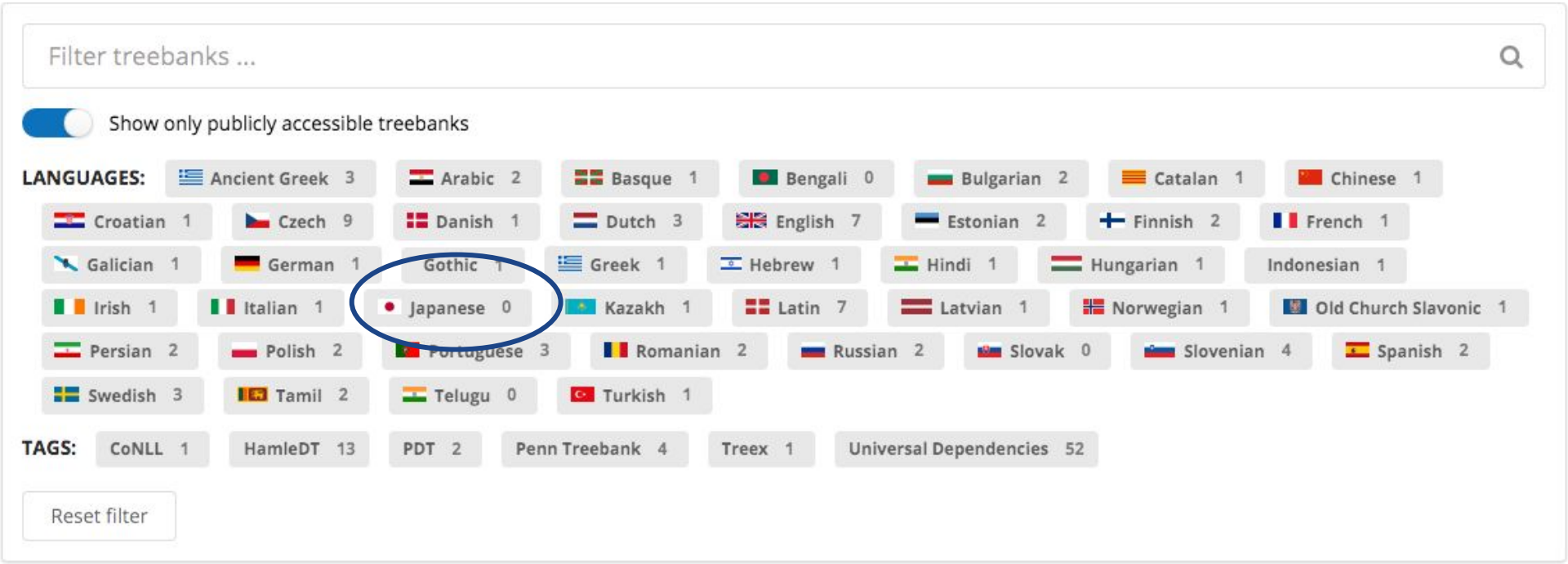
- Fa una ricerca nel catalogo centrale di CLARIN, il VLO
- Scopre che esiste uno strumento di questo tipo, che si chiama Tree Query



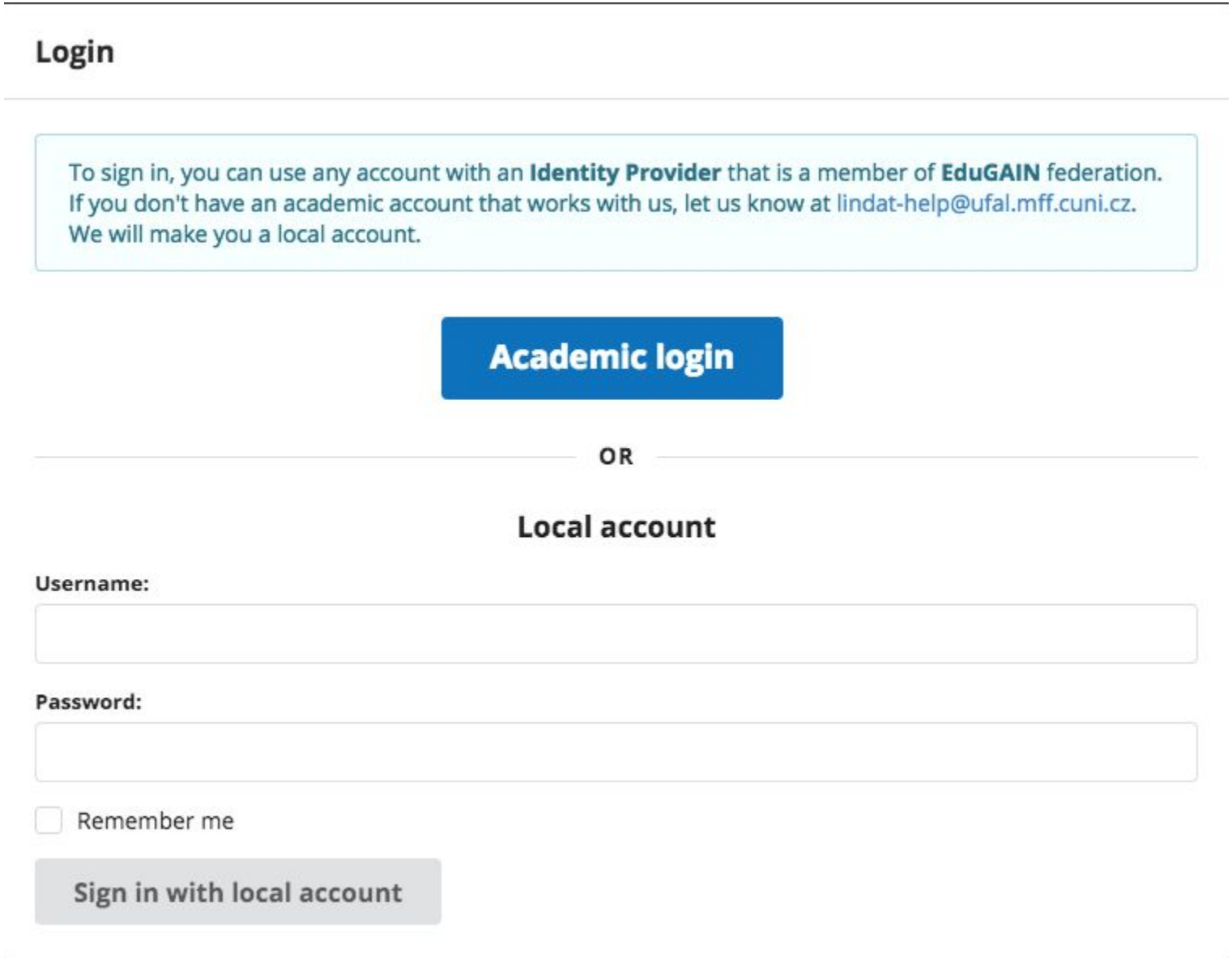
- Clicca sul link e viene ridiretto al centro Lindat CLARIN (Repubblica Ceca)



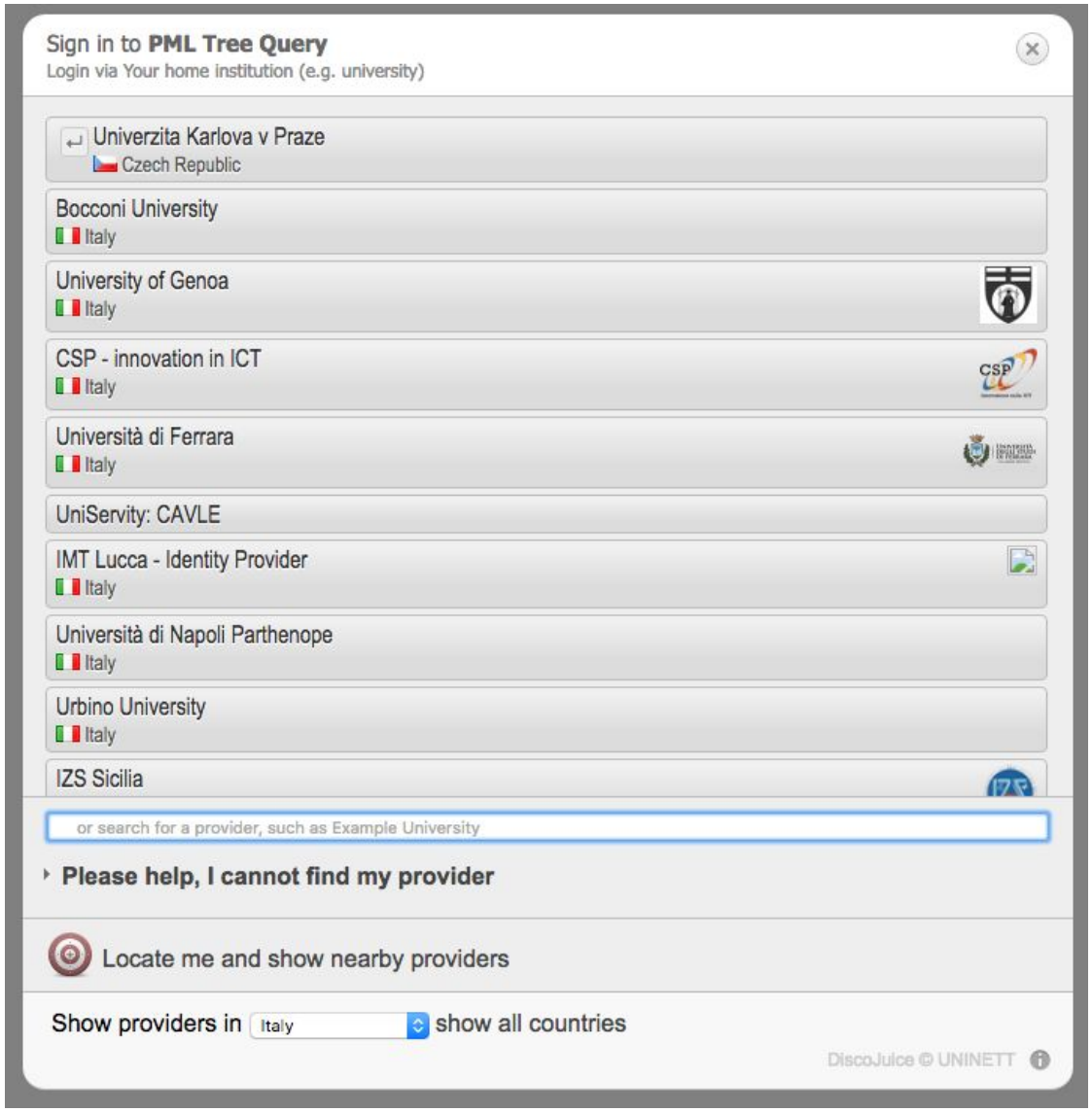
- Lo strumento è liberamente accessibile, ...



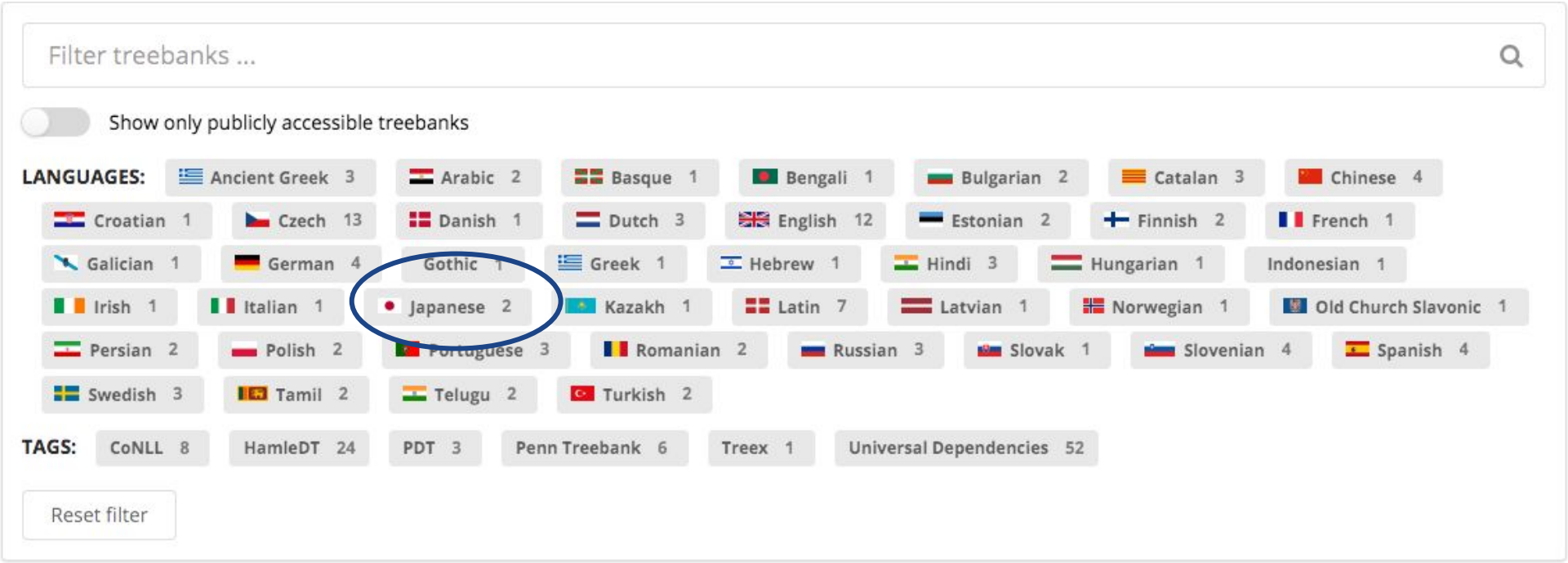
- .... ma alcuni dei corpora sono disponibili solo per ricercatori accademici (richiesta di autenticazione)



- Al momento del login, si trova davanti ad una richiesta di autenticarsi tramite la federazione CLARIN



- Cerca il nome della propria istituzione nella lista


Filter treebanks ...
Show only publicly accessible treebanks
LANGUAGES:
Ancient Greek 3
Arabic 2
Basque 1
Bengali 1
Bulgarian 2
Catalan 3
Chinese 4
Croatian 1
Czech 13
Danish 1
Dutch 3
English 12
Estonian 2
Finnish 2
French 1
Galician 1
German 4
Gothic 1
Greek 1
Hebrew 1
Hindi 3
Hungarian 1
Indonesian 1
Irish 1
Italian 1
Japanese 2
Kazakh 1
Latin 7
Latvian 1
Norwegian 1
Old Church Slavonic 1
Persian 2
Polish 2
Portuguese 3
Romanian 2
Russian 3
Slovak 1
Slovenian 4
Spanish 4
Swedish 3
Tamil 2
Telugu 2
Turkish 2
TAGS:
CoNLL 8
HamleDT 24
PDT 3
Penn Treebank 6
Treex 1
Universal Dependencies 52
Reset filter

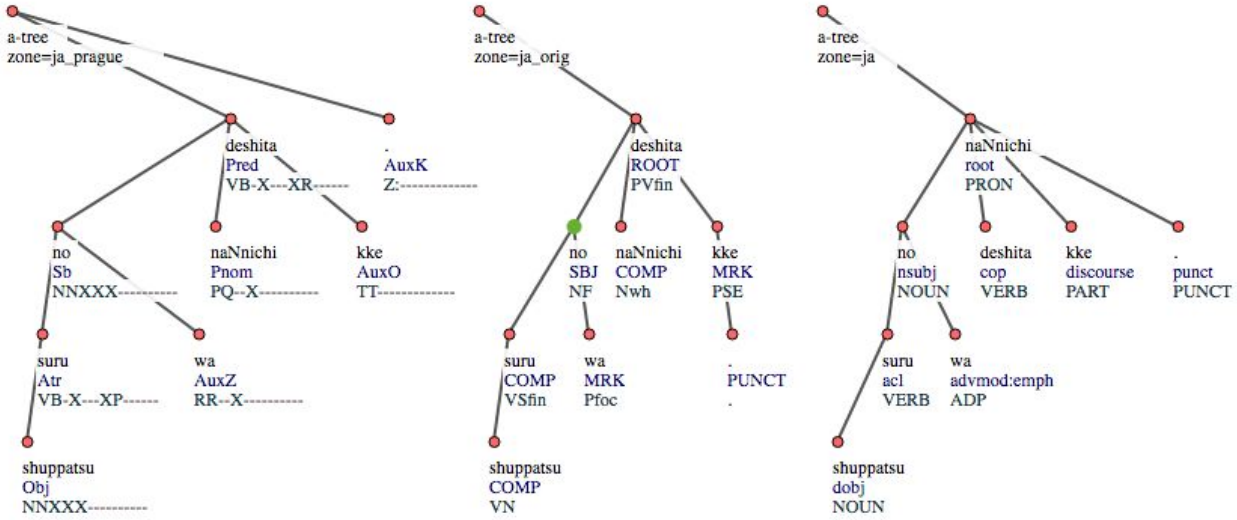
Relations | Node Types | Attributes | Operators | Functions

```
a-node [ ]
```

Execute query | w/o Filters

Previous | 1 of 100 | Next

[ja_prague] shuppatsu suru no wa naNnichi deshita kke .
[ja_orig] shuppatsu suru no wa naNnichi deshita kke .
[ja] shuppatsu suru no wa naNnichi deshita kke .



Servizi protetti

Una lista di servizi accessibili agli utenti italiani grazie all'interfederazione IDEM / CLARIN

https://www.clarin.eu/content/easy-access-protected-resources

Easy access to protected resources

Thanks to federated login, the following applications and data sets are available to anyone with an academic account from many European countries -- for a complete list, see Participating Identity Federations. We are working hard to extend this to the rest of Europe. In the meanwhile, you can request a CLARIN account if you want to access these services from another country or from an institution that does not participate in these identity federations.

Resource or Tool	Description	Provided by
OpenSoNaR	over 500 million word Dutch reference corpus	INL (Instituut voor Nederlandse Lexicologie)
Corpus Hedendaags Nederlands	written corpus of contemporary Dutch	INL (Instituut voor Nederlandse Lexicologie)
VU-DNC	diachronic Dutch newspaper corpus	INL (Instituut voor Nederlandse Lexicologie)
Tündra	treebank search application	Eberhard Karls Universität Tübingen

Verso CLARIN-IT

- Come National Coordinator ho il compito:
  - creare il consorzio nazionale CLARIN-IT
  - far da tramite tra la comunità italiana e il CLARIN-ERIC

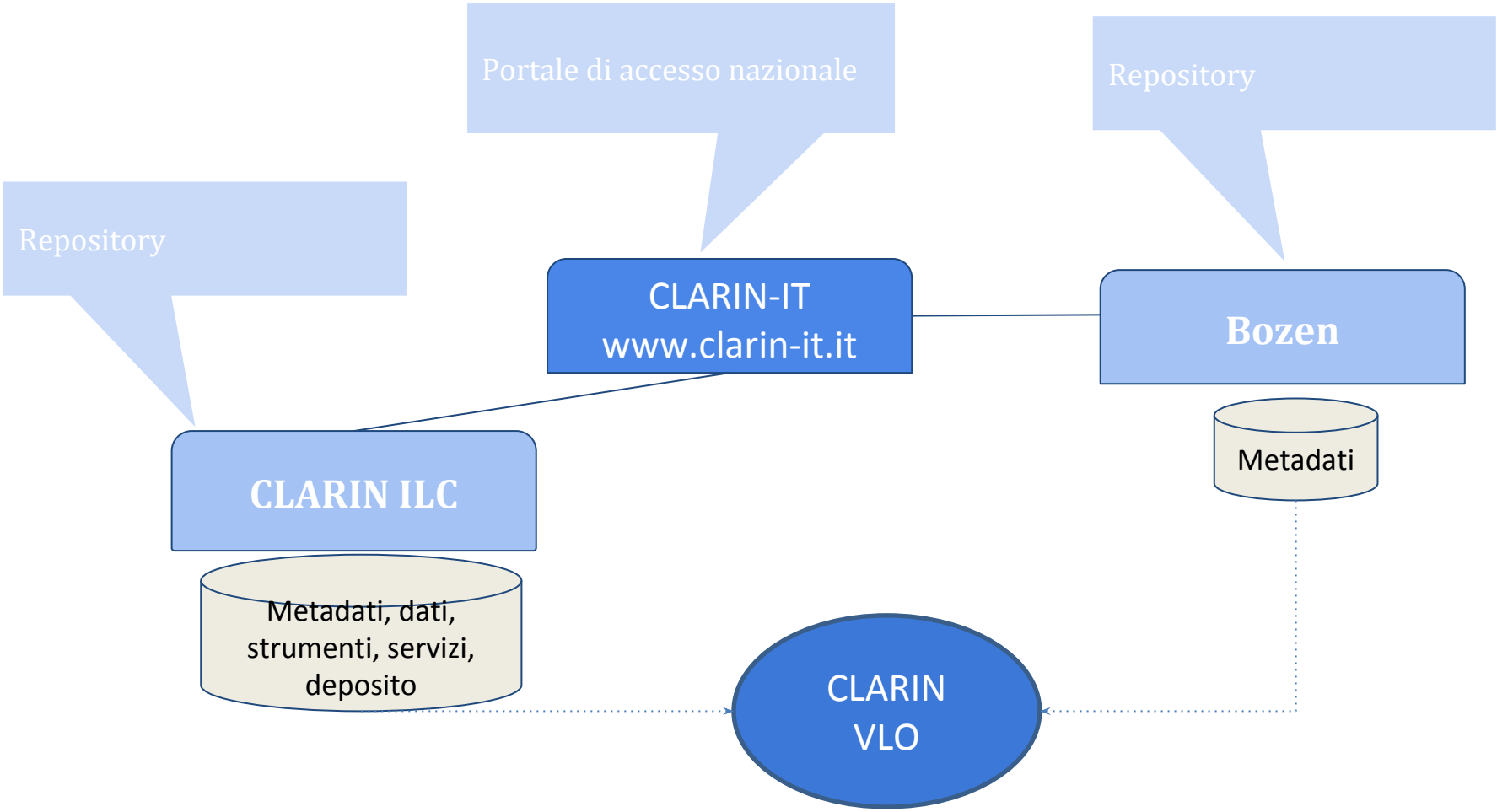
- ILC-CNR è l'istituto esecutore con il compito di
  - creare un centro italiano per la documentazione accesso e consultazione delle risorse linguistiche

Consorzio CLARIN-IT

CLARIN-IT: Consorzio

- Per ora le istituzioni interessate sono:
  - **ILC-CNR**
  - **EURAC+ University of Bozen**
  - **University of Siena**
  - **Universtiy of Venice**
  - **Universtiy of Pisa**
  - **FBK Trento**
- Da subito, forte collaborazione con **IDEM-GARR**

CLARIN-IT

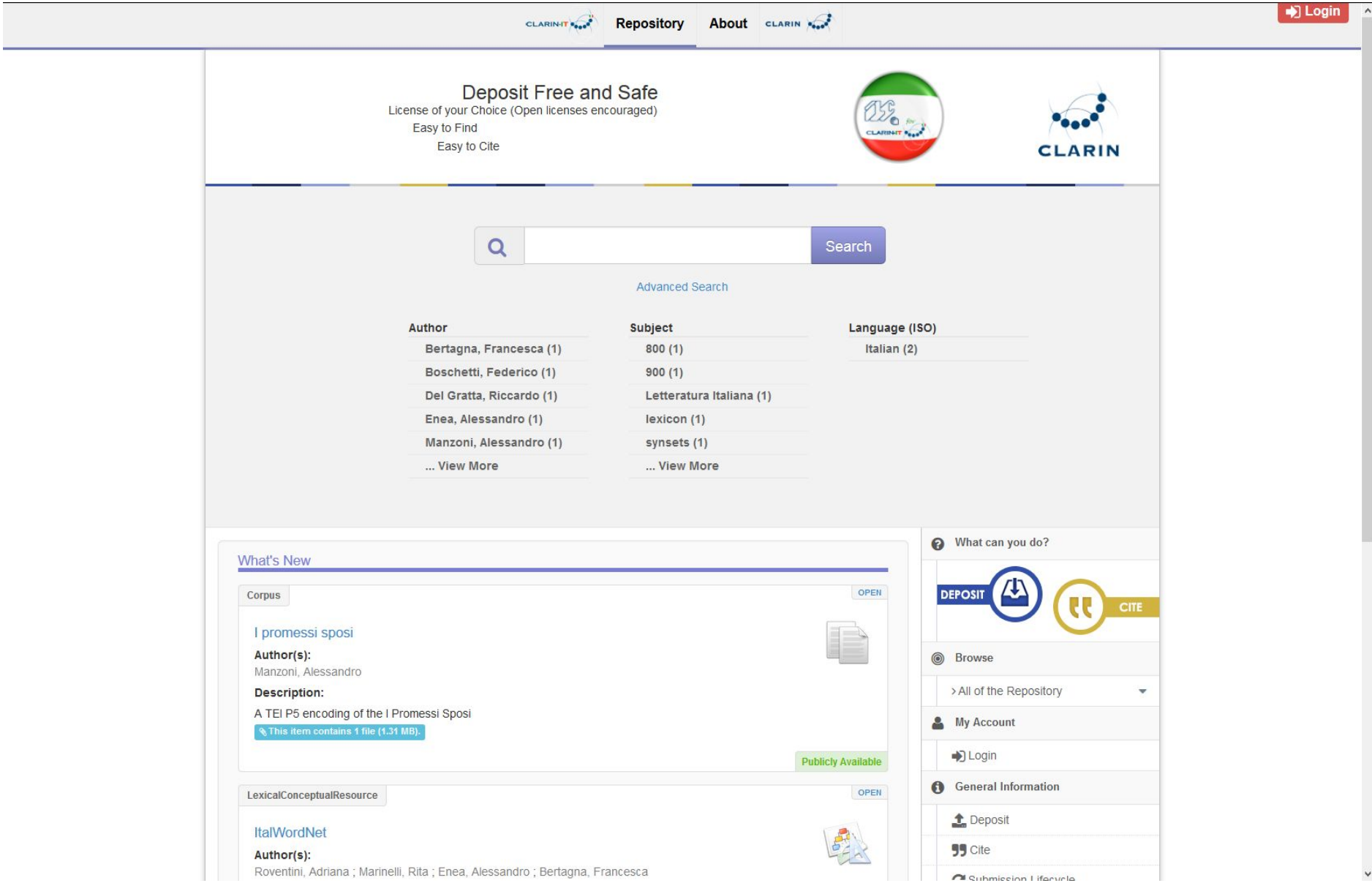

Portale di accesso nazionale
Repository
Repository
CLARIN-IT
www.clarin-it.it
Bozen
CLARIN ILC
Metadati
Metadati, dati, strumenti, servizi, deposito
CLARIN VLO

CENTRO ILC

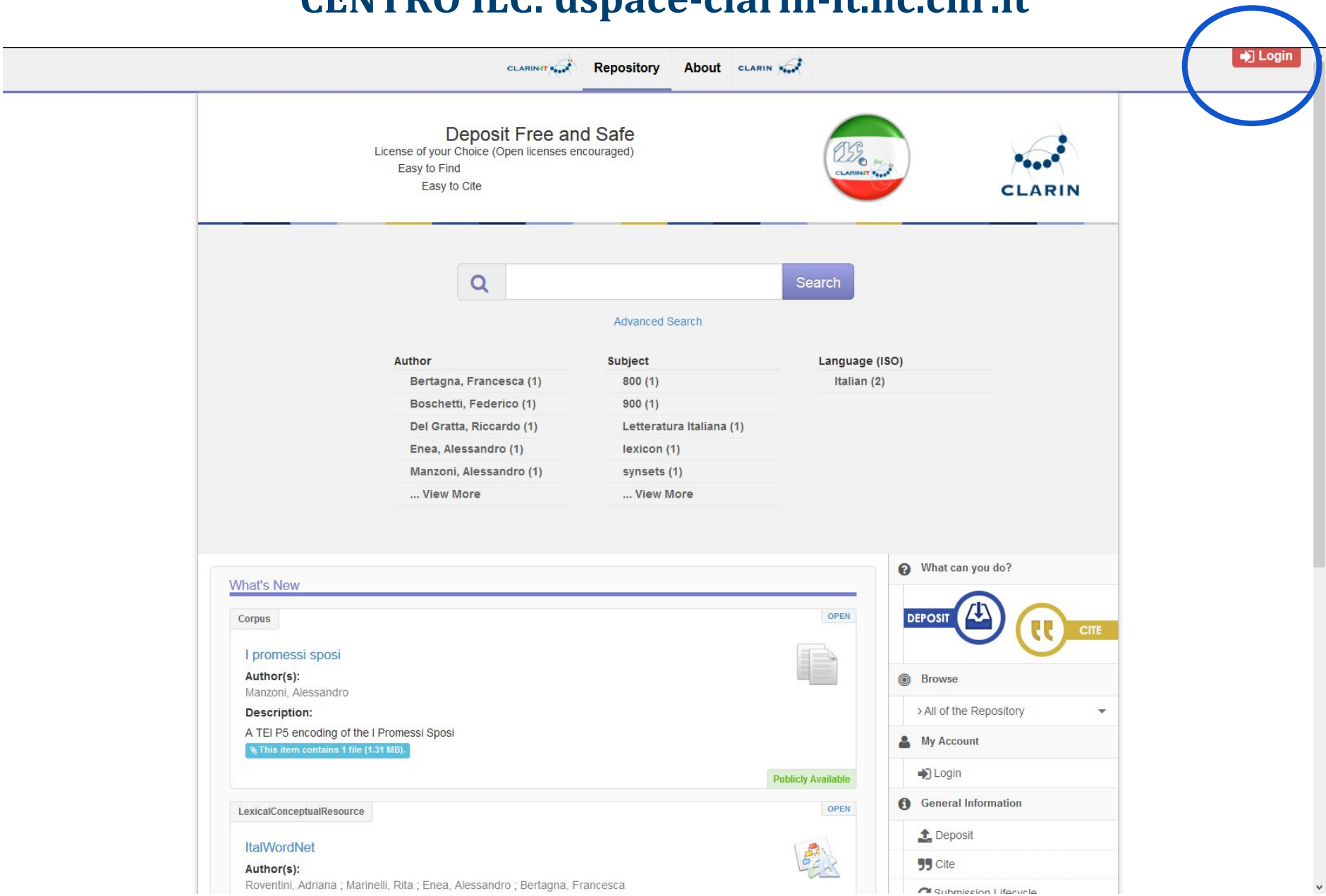
- Il centro ILC offrirà accesso a risorse e servizi dell'Istituto di Linguistica Computazionale.
- Inoltre offrirà un servizio di **deposito** di risorse di terze parti, al fine di garantire che risorse e strumenti sviluppati nell'ambito di progetti ormai finiti possano essere preservati a lungo termine.
- Il servizio di deposito è gestito tramite single sign-on.

- Per fare questo si avvale del repository LINDAT DSPACE, creato da CLARIN della Repubblica Ceca.

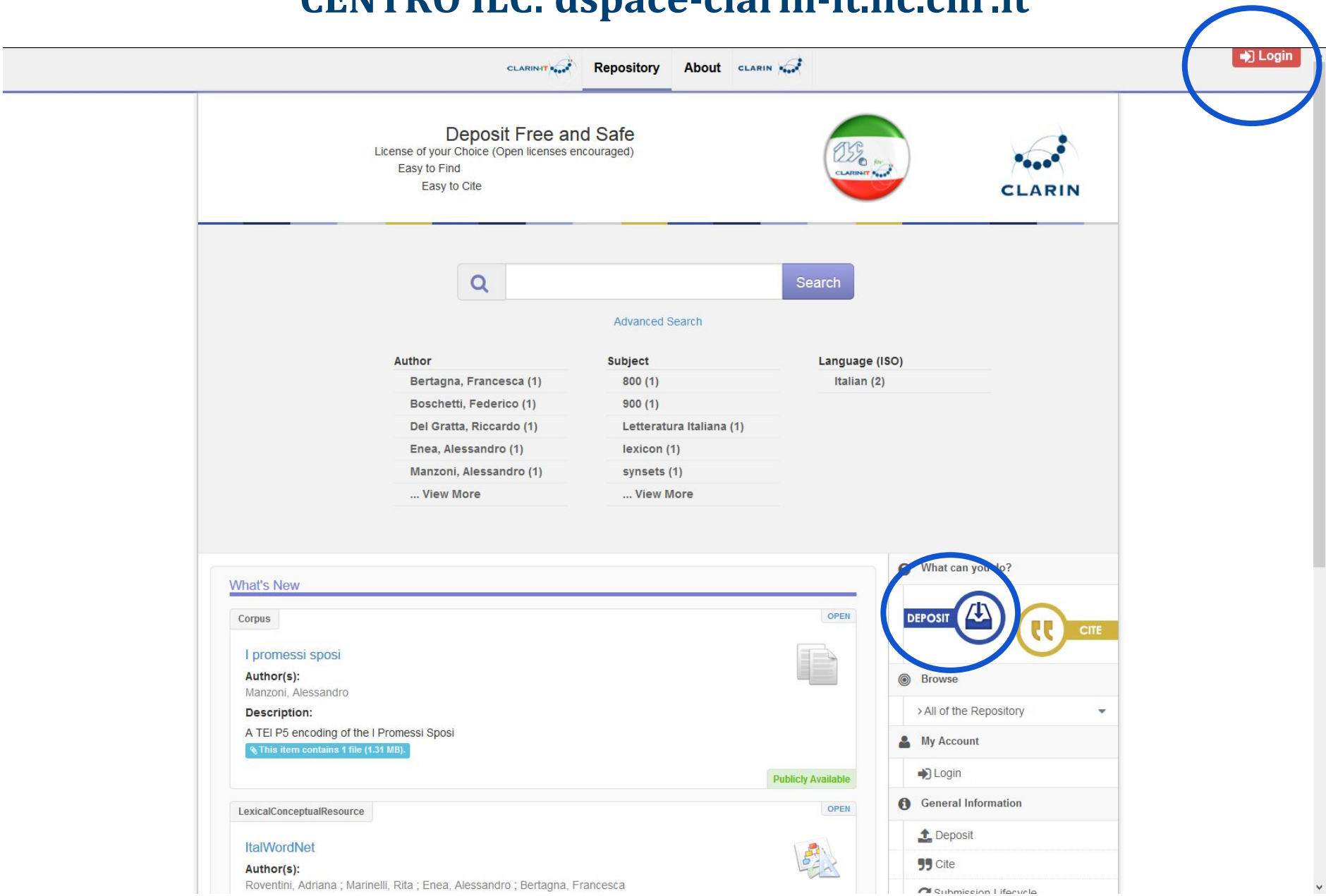
CENTRO ILC: dspace-clarin-it.ilc.cnr.it



CENTRO ILC: dspace-clarin-it.ilc.cnr.it



CENTRO ILC: dspace-clarin-it.ilc.cnr.it



Step 3: Select type of your submission

You have initiated a new workflow item. In the next few steps you will provide the details about the item and upload content files. First select the type of the resource you are about to submit.

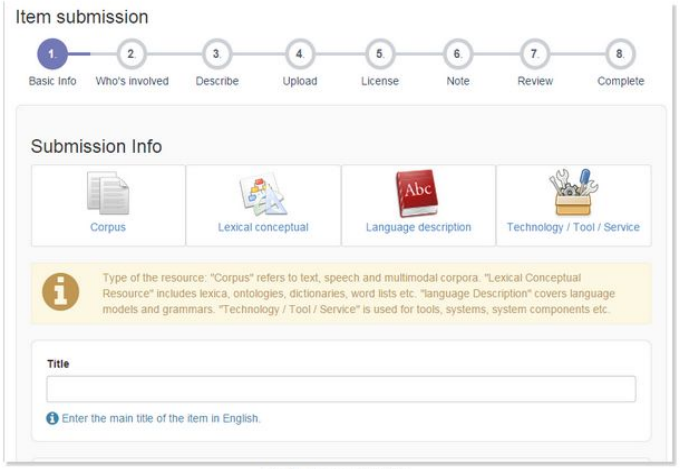
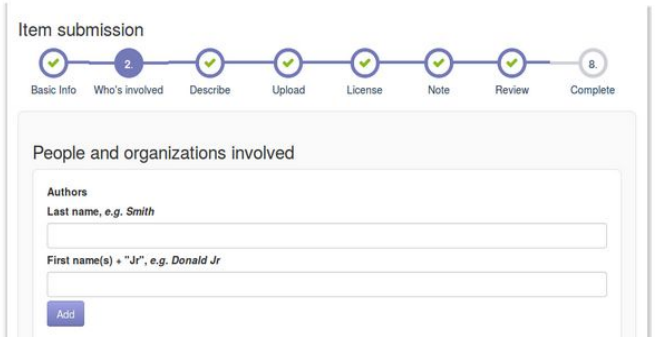


Fig4. Submission info

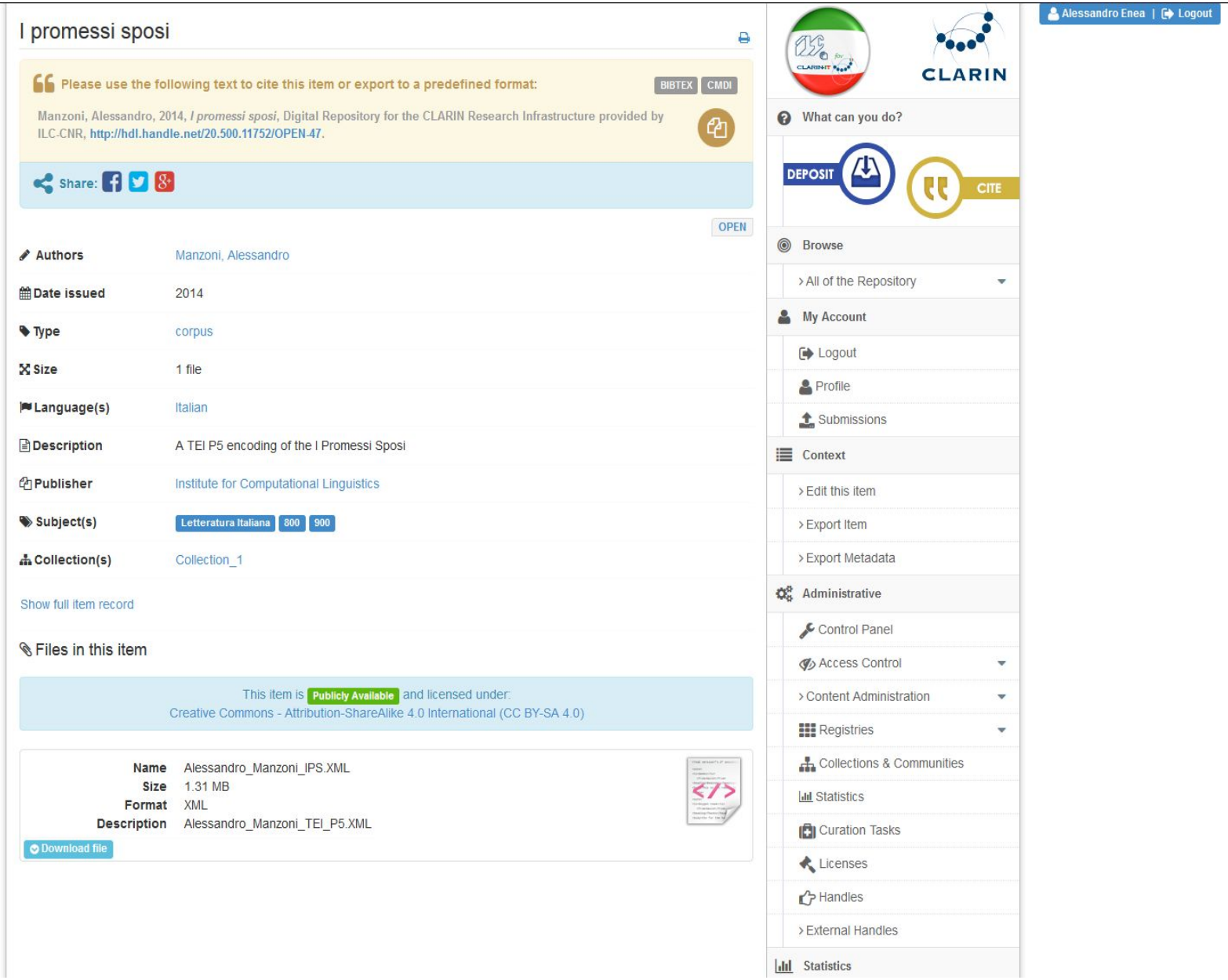
Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as the title. Click Next to continue the following step.

Step 4: Describe your item

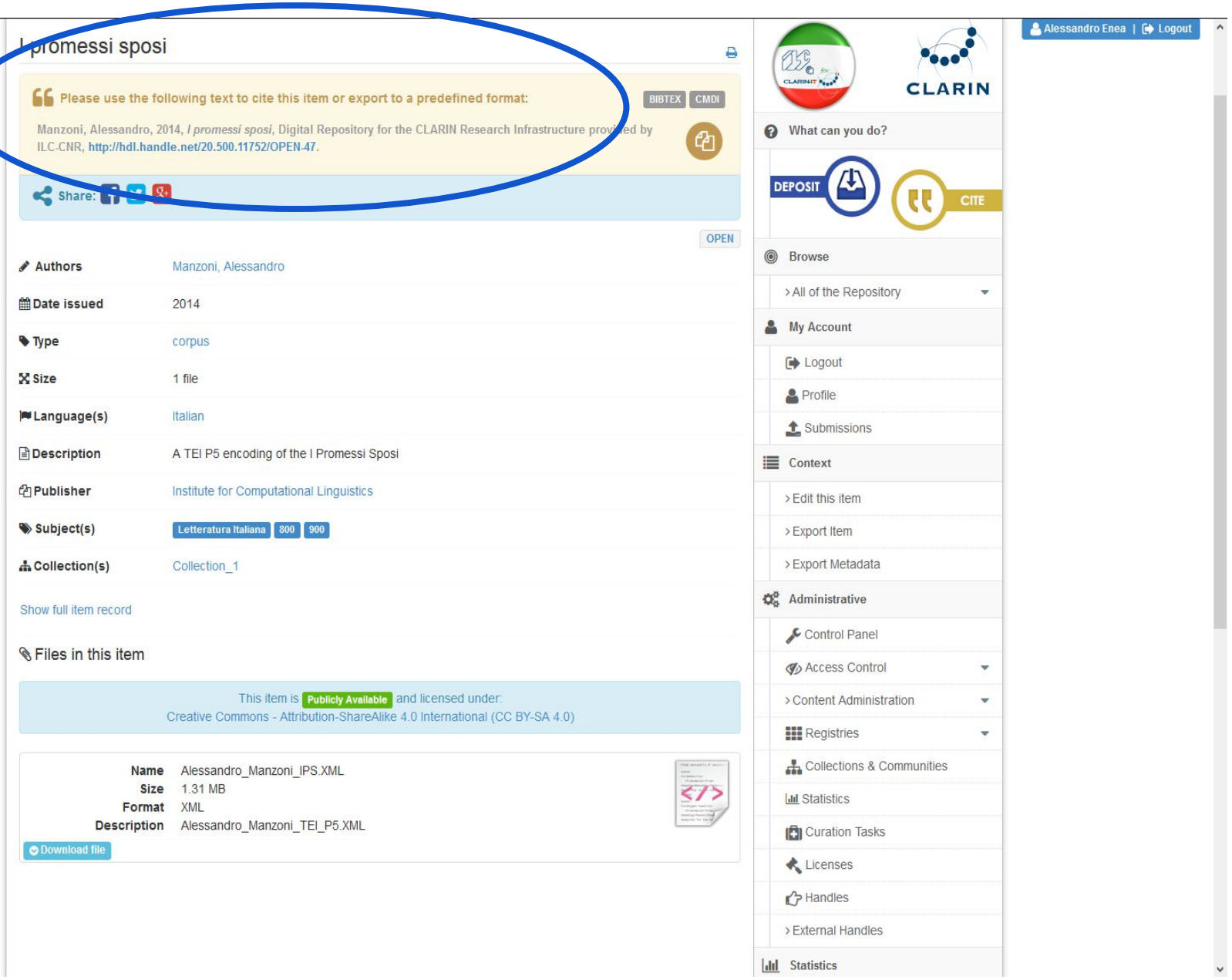
In the following two steps you will provide more details for your item. First describe the people, organization and projects involved with the item.



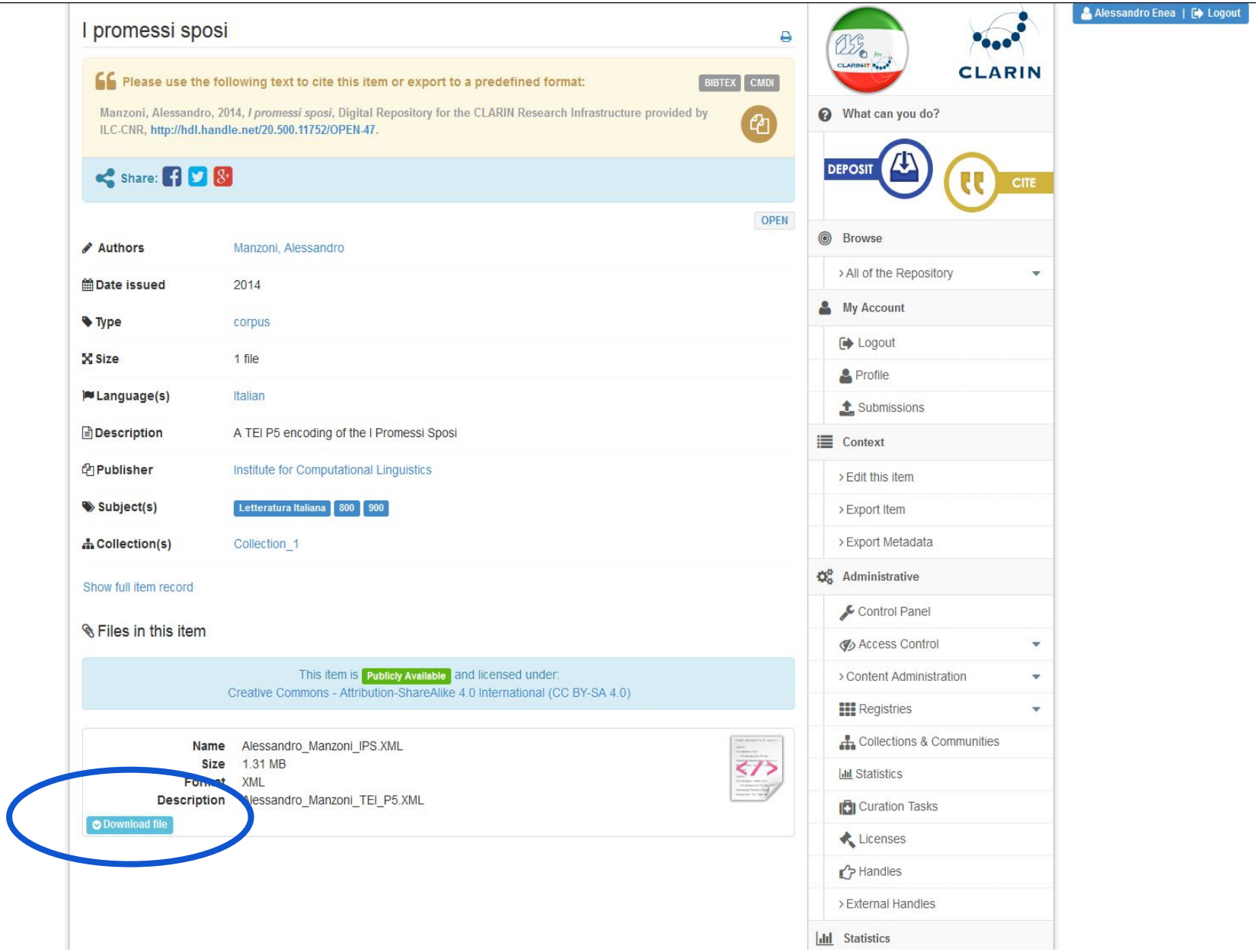
CENTRO ILC: dspace-clarin-it.ilc.cnr.it



CENTRO ILC: dspace-clarin-it.ilc.cnr.it



CENTRO ILC: dspace-clarin-it.ilc.cnr.it



CLARIN-IT & IDEM

- La CLARIN SPF include IDEM a garanzia della massima copertura degli utenti italiani.
- CLARIN-IT incoraggia le istituzioni italiane ad aderire
  - ad esempio, ingresso dell'Università di Siena in IDEM
- La collaborazione CLARIN-IT e IDEM consente la verifica degli attributi che i SP di CLARIN tipicamente chiedono per far in modo che IDEM li fornisca.

Alcuni link

https://www.clarin.eu/content/federated-identity

https://www.clarin.eu/content/service-provider-federation

https://www.clarin.eu/content/how-saml-metadata-about-sps-distributed-identity-federations

https://www.clarin.eu/content/participating-spf

CLARIN-IT: contatti

- www.clarin-it.it
- Twitter @CLARIN_IT
- coordination@clarin-it.it

Grazie per la vostra attenzione!!!!!!