CLARIN-IT

the Italian Common Language Resources and Technology Infrastructure



CLARIN-IT: Language Resources and Infrastructures for Digital Humanities

Monica Monachini
CLARIN Italian National Coordinator

ILC-CNR
CLARIN Italian National Representative

Summer School "New trends in quantitative and computational linguistics"
Universidad de Castilla-La Mancha, Ciudad Real, 28 September 2016

1 October 2015

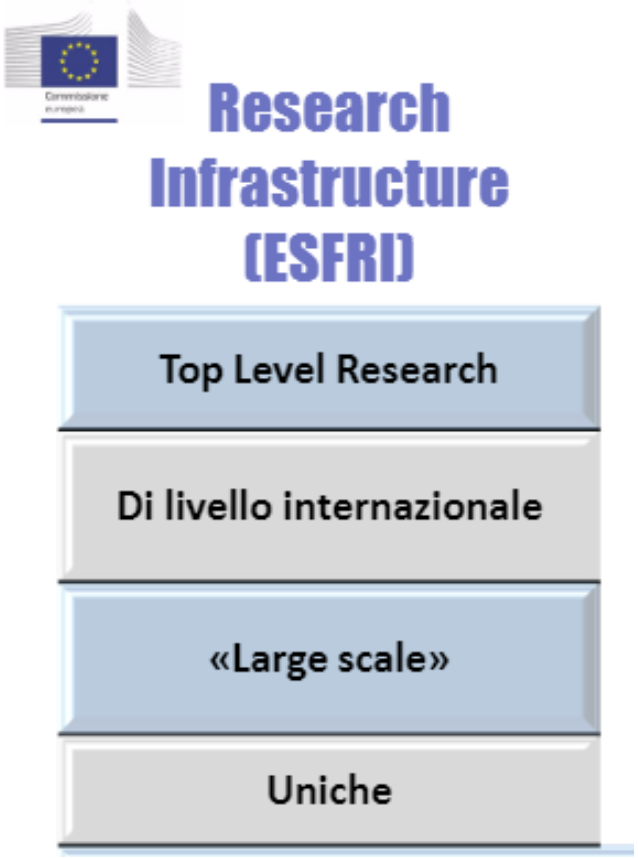
- Italy became a Full Member of **CLARIN ERIC**, the governance and coordination body of the Common Language Resources and Technology Infrastructure at the European level



- This is a great opportunity for those who are involved in **Humanities and Social Sciences**

What is meant by RI

- *A set of resources, tools, competencies, data, services and functions specialized and dedicated to feed the necessary streams to scientific and technological research on a continuous basis*
- CLARIN
  - is an infrastructure of type **ERIC** (European Research Infrastructure Consortium) formally acknowledged
  - has completed la **ESFRI** (European Strategy Forum on Research Infrastructures) **roadmap**
  - has achieved the status of **«land-mark»**

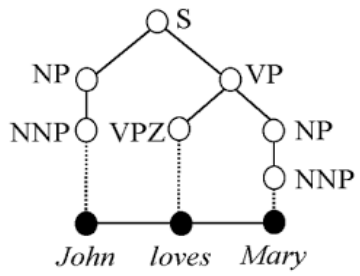


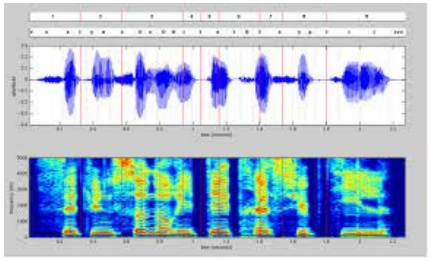
What LRs are

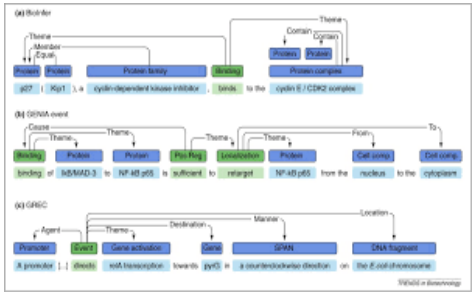
- Digital language resources
  - corpora (annotated and not), lexica, ontologies
  - formal grammars, statistical models
  - tools for (written, spoken, multimodal) Natural Language Processing

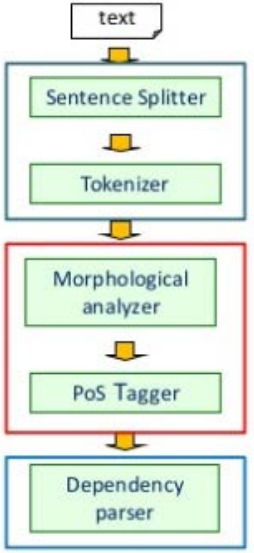






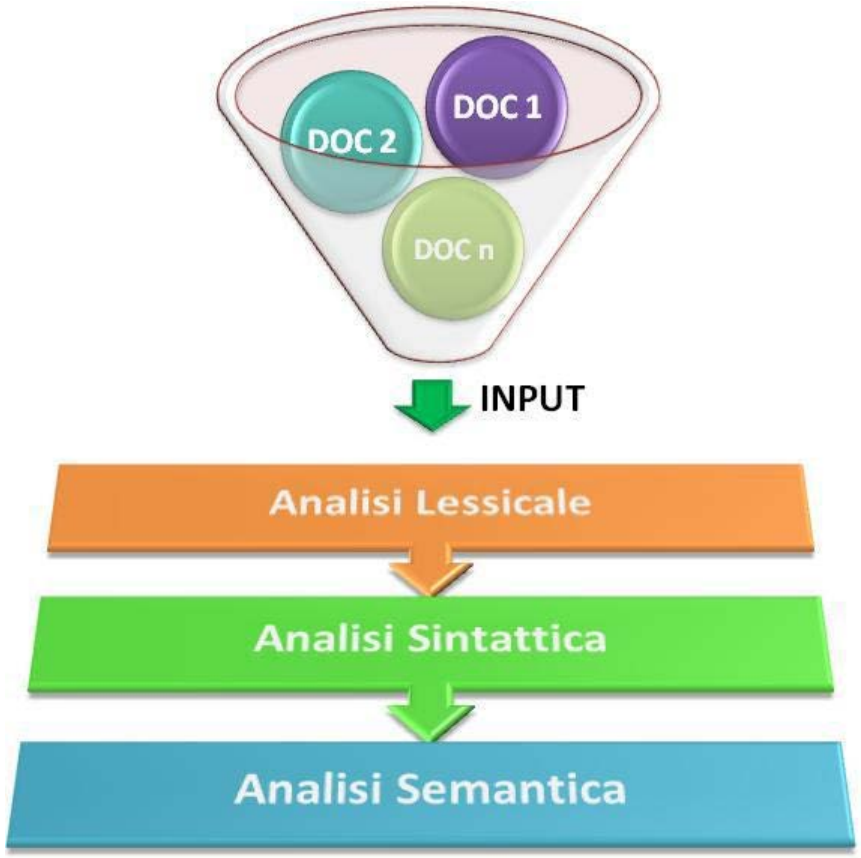






What LRs are for

- Language resources can be useful in
  - linguistic studies
  - analysis of textual sources
  - language processing
  - automatic annotation of texts
  - …



Current situation and needs

- Language resources often
  - are hosted by the institution that has produced them
  - are intended for the scientific purposes and environments of a project
  - are subject to restrictions
- Building language resources requires a good deal of time and money
  - it is crucial to ensure an easy access to resources to researchers of different institutions and countries
  - it is fundamental to ensure that such resources remain accessible over time
- A change of paradigm is essential
  - sharing, access and replicability of results







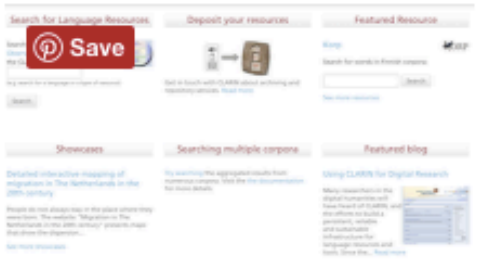


CLARIN ERIC



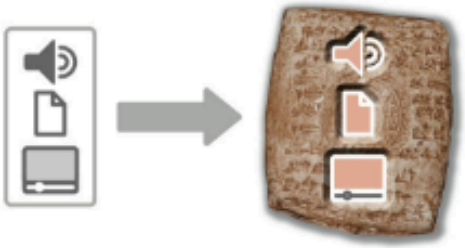
- CLARIN is an intangible digital infrastructure aiming at providing
  - a system for the **description** and **cataloguing** of existing language resources
  - a **repository** for the long-term electronic storage
  - an easy and sustainable **access** to scholars of member countries
- CLARIN furthermore provides
  - advanced tools for **identifying** language resources suitable to one's research wherever they are hosted and **combining** them in order to analyse, explore and annotate digital data

CLARIN services



CLARIN portal

Get an example-based impression of what's currently available



Depositing services

Store language resources in a sustainable repository at a CLARIN centre

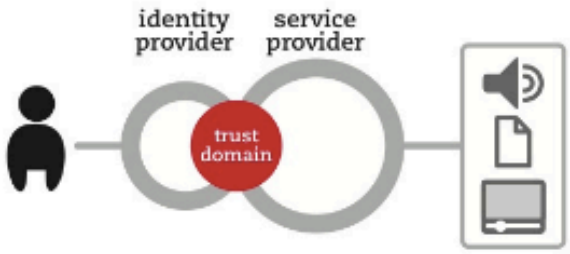


Virtual Language Observatory

Discover language resources using a faceted browser or a map

https://www.clarin.eu/content/services

CLARIN services



Easy access to protected resources

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



Web services and applications

Explore and analyze language data with a wide variety of tools

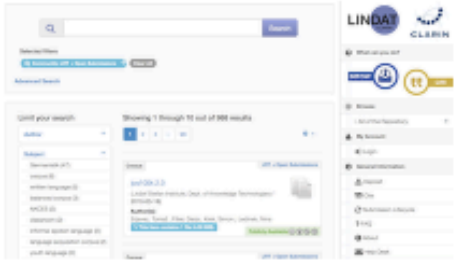


Virtual Collections

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

https://www.clarin.eu/content/services

CLARIN services



Language Resource Inventory

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



Content Search (prototype version)

Search different corpora with a single search engine



Consulting Services

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

https://www.clarin.eu/content/services

How CLARIN ERIC works



- CLARIN ERIC is organized as a consortium of nations (and supranational bodies)
- Each nation
  - pays an annual membership fee
  - offers in-kind services for the maintenance of the consortium
  - make resources and services available through a **CLARIN Centre** or more

How a CLARIN National Consortium works



- CLARIN includes 19 member countries (plus a few observer countries)
- CLARIN National Consortia include various institutions (research centres, universities, libraries, institutions supporting research, …) that
  - provide data, tools and services as CLARIN Centres
  - are financially supported by one's nation through national strategic projects

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
Francia
Ungheria
Romania
Croazia
Serbia
Mar Nero
Roma
Bulgaria
Georgia
Portogallo
Madrid
Spagna
Mar Tirreno
Istanbul
Grecia
Turchia
Azerbaijan
Siria
Tunisia
Mar Mediterraneo
Libano
Iraq
Google

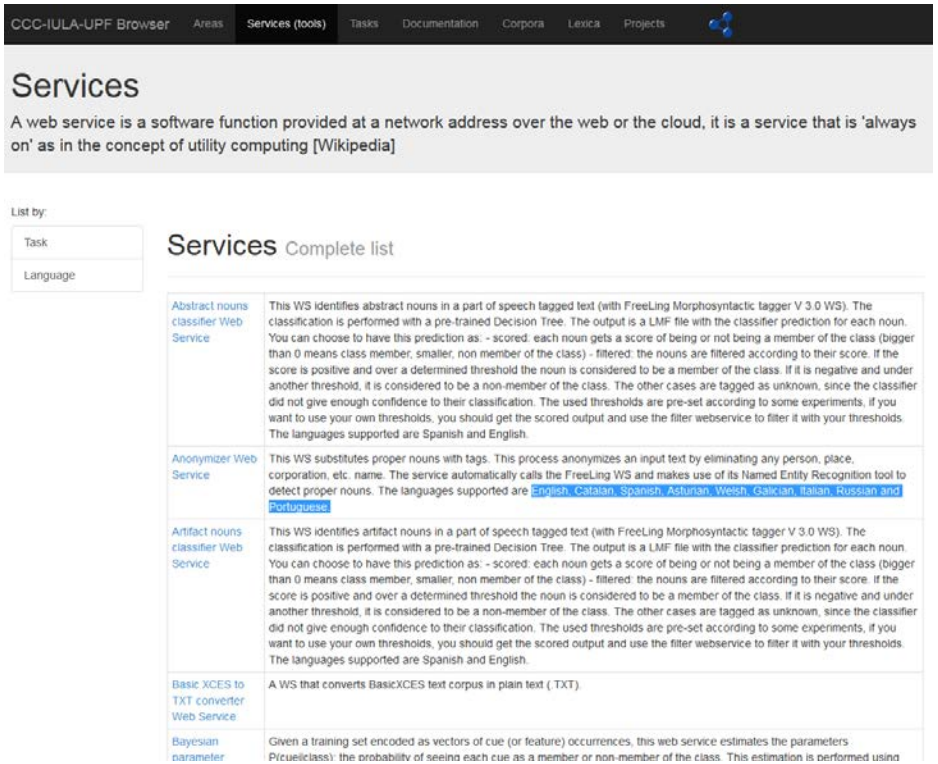
CLARIN in Spain

- Centre K
  http://clarin-es-lab.org/index-en.html



CLARIN in Spain

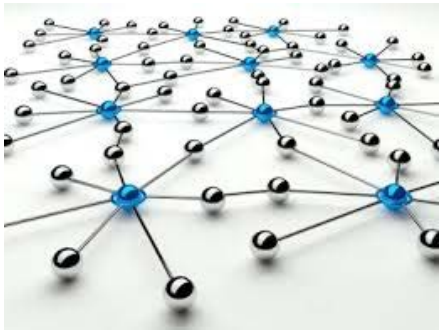
- Services
  http://lod.iula.upf.edu/types/Service



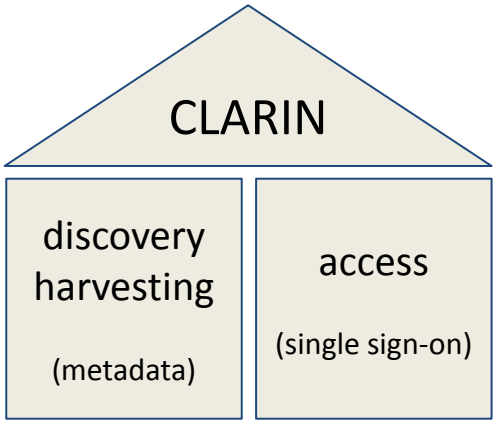
CLARIN ERIC pillars



- CLARIN is a network of different centres that are the pillars of the infrastructure



- The fundamental pillars are
  - **documentation**
    - language resources have to be documented with a standard format of metadata (CMDI = Component MetaData Infrastructure)
  - **exposition** of metadata to be **harvested** in a **centralized meta-catalogue** (VLO = Virtual Language Observatory)
  - **actual access** to language resources
    - **totally free**
    - **through a single sign-on unified authentication system** (which is particularly important for all those digital language resources that are available for academic purposes only)



CLARIN ERIC governance

- **General Assembly**
  - Managers appointed by the Ministry for Research of each nation
  - *highest decision-making body*
- **CLARIN Office** + **Board of Directors**
  - *executive management*
- **National Coordinators' Forum**
  - Coordinators of CLARIN National Consortia
    +
    Standing Committee for Technical Centers
  - *integration and interoperability at a trans-national level*

https://www.clarin.eu/content/governance



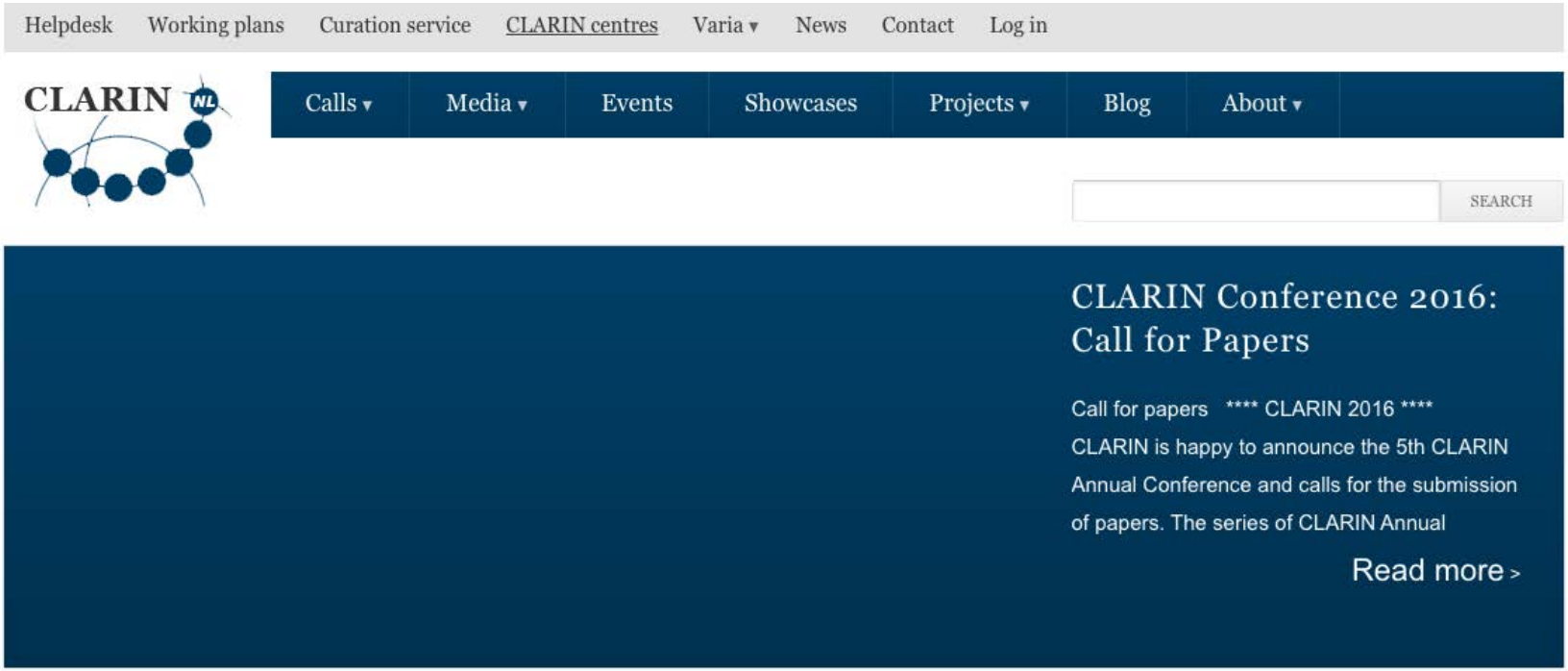




CLARIN National Consortia



- CLARIN ERIC
  - does not impose a model of governance to its various National Consortia
  - requires
    - the creation of a community by involving the main actors of the research sector
    - the establishment of a centre (or more) to make various types of data, tools and services available
    - the participation in centralized activities, committees etc.


Helpdesk
Working plans
Curation service
CLARIN centres
Varia
News
Contact
Log in
CLARIN
Calls
Media
Events
Showcases
Projects
Blog
About
SEARCH
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
CLARIN
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

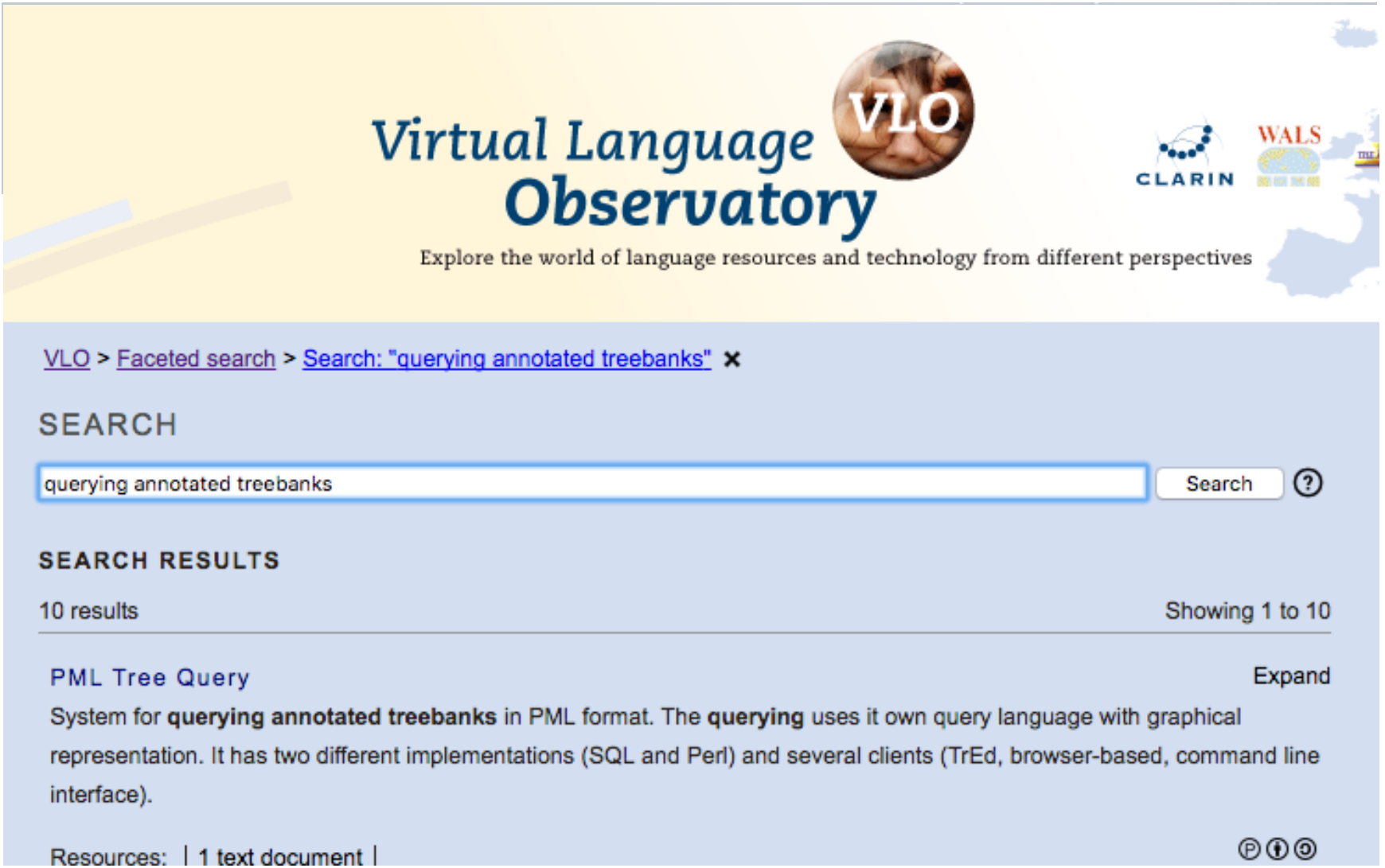
CENTRE	FOCUS ON RESOURCES	RESOURCE EXAMPLES
MEERTENS INSTITUTE	relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.	typological databases 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

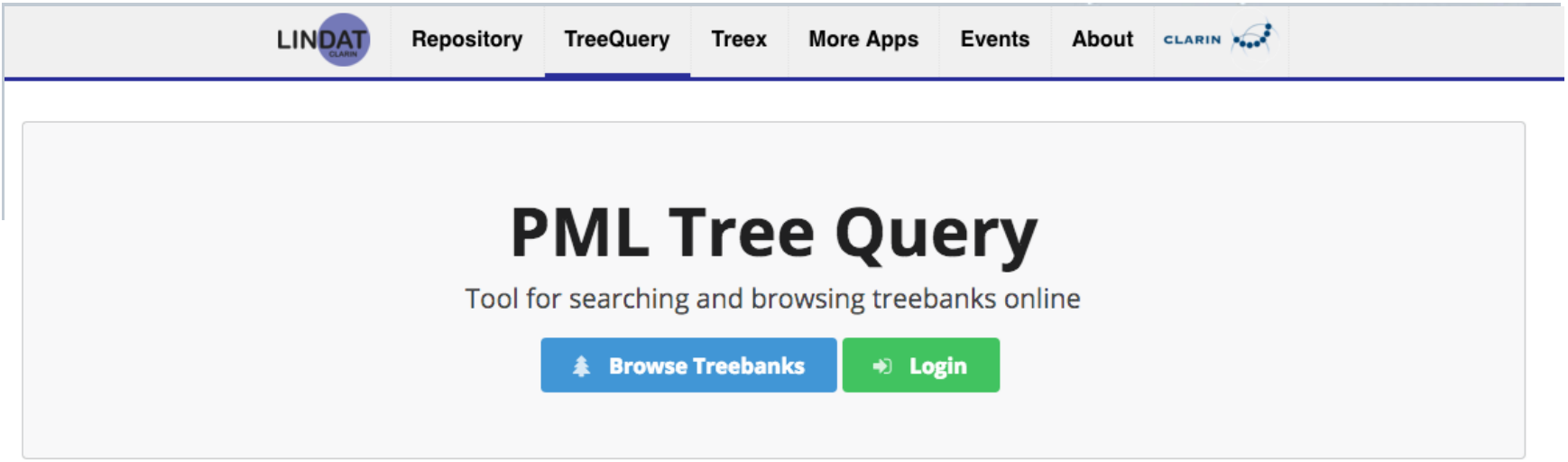
How it works



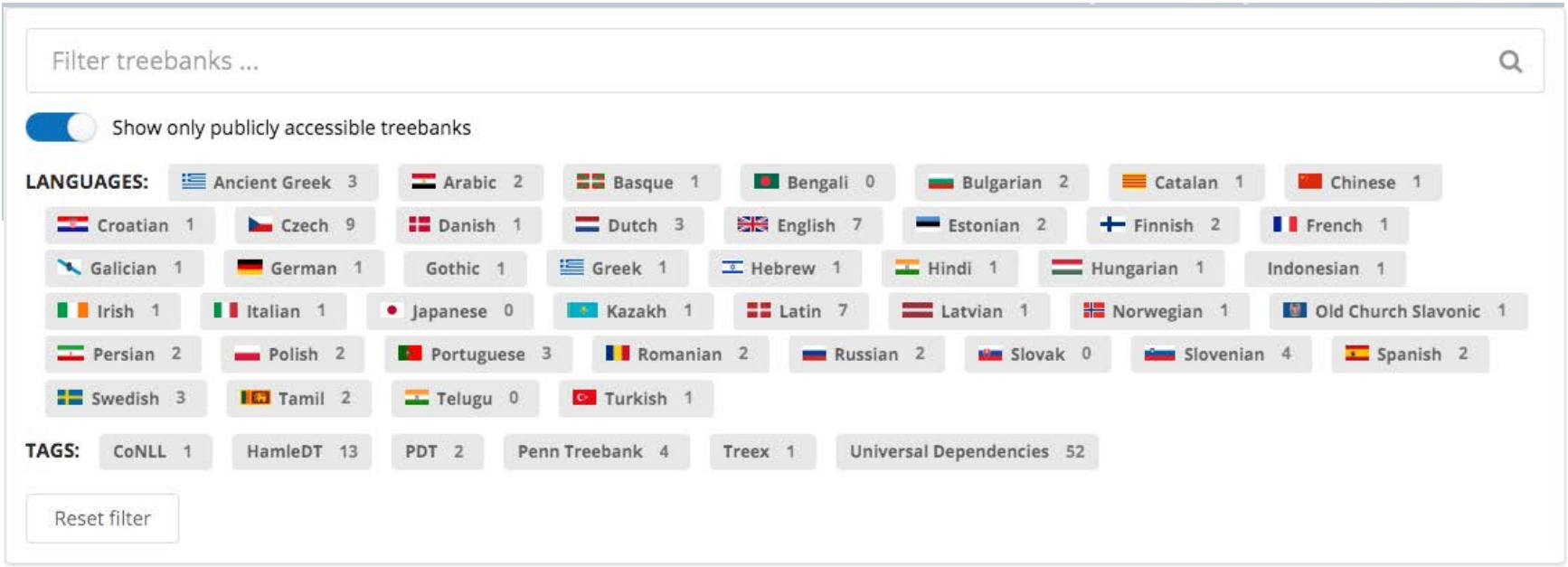
- A researcher is interested in the study of syntax
- In particular, he/she is interested in the Spanish language(s)
- First, he/she checks the availability of syntactically annotated corpora (treebanks)



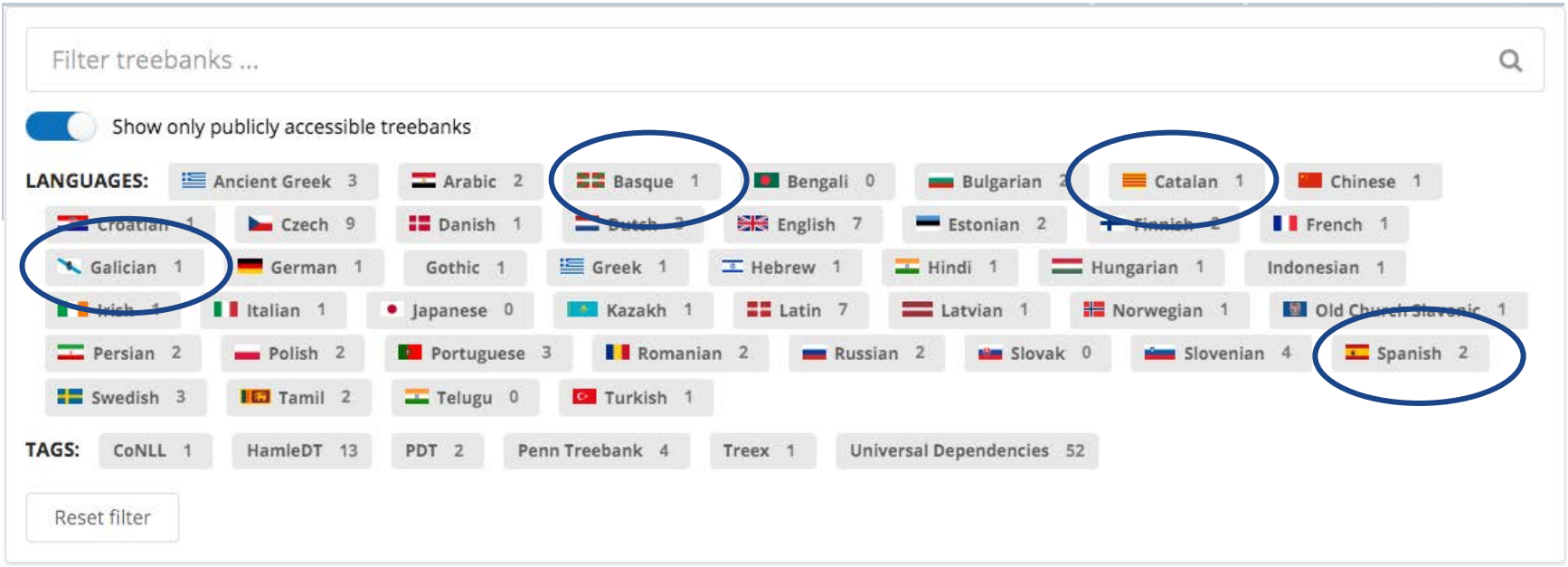
- The central catalogue: VLO
- The search and PML system for querying annotated treebanks



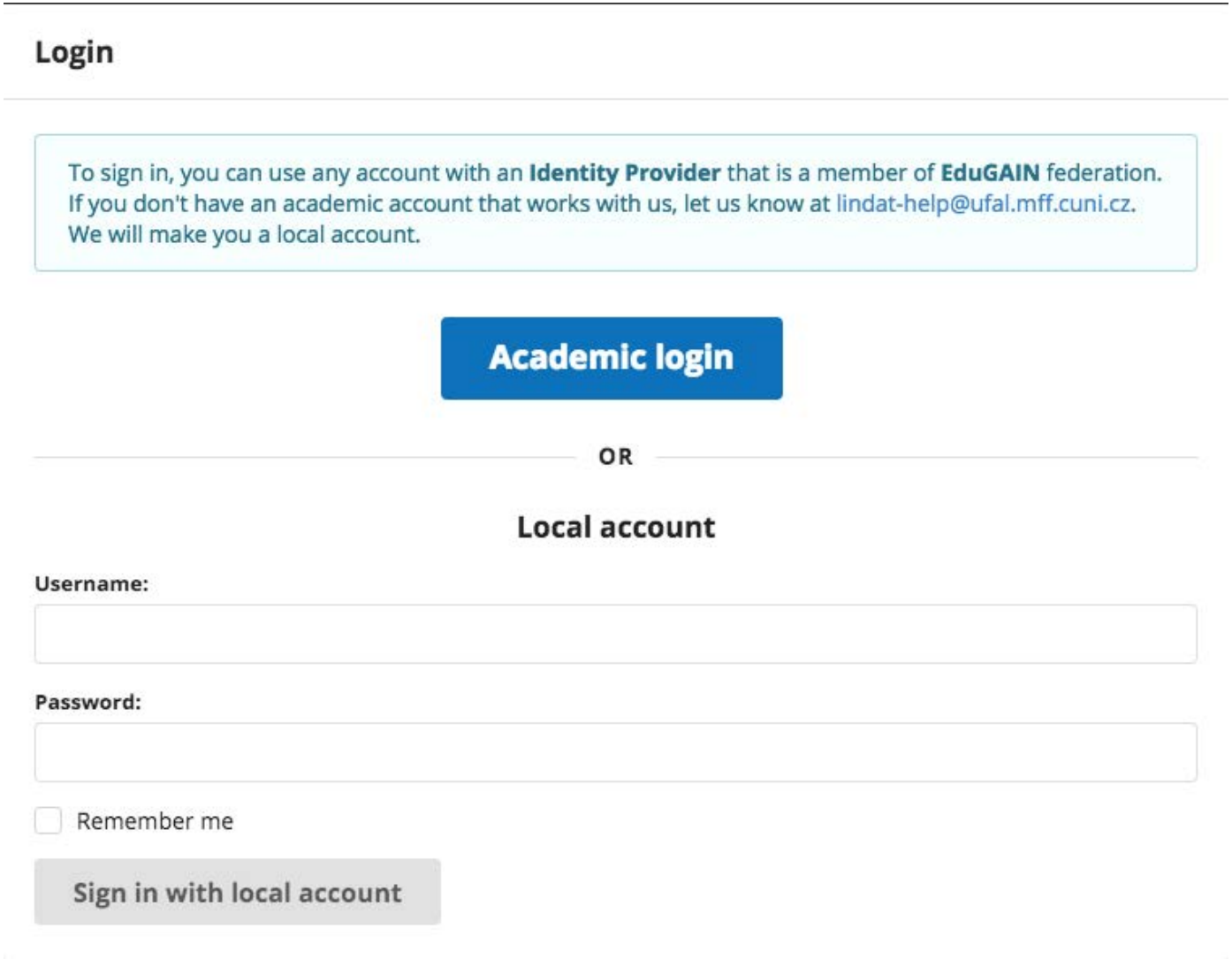
- The system is available at the Lindat CLARIN Centre repository (Czech Republic)



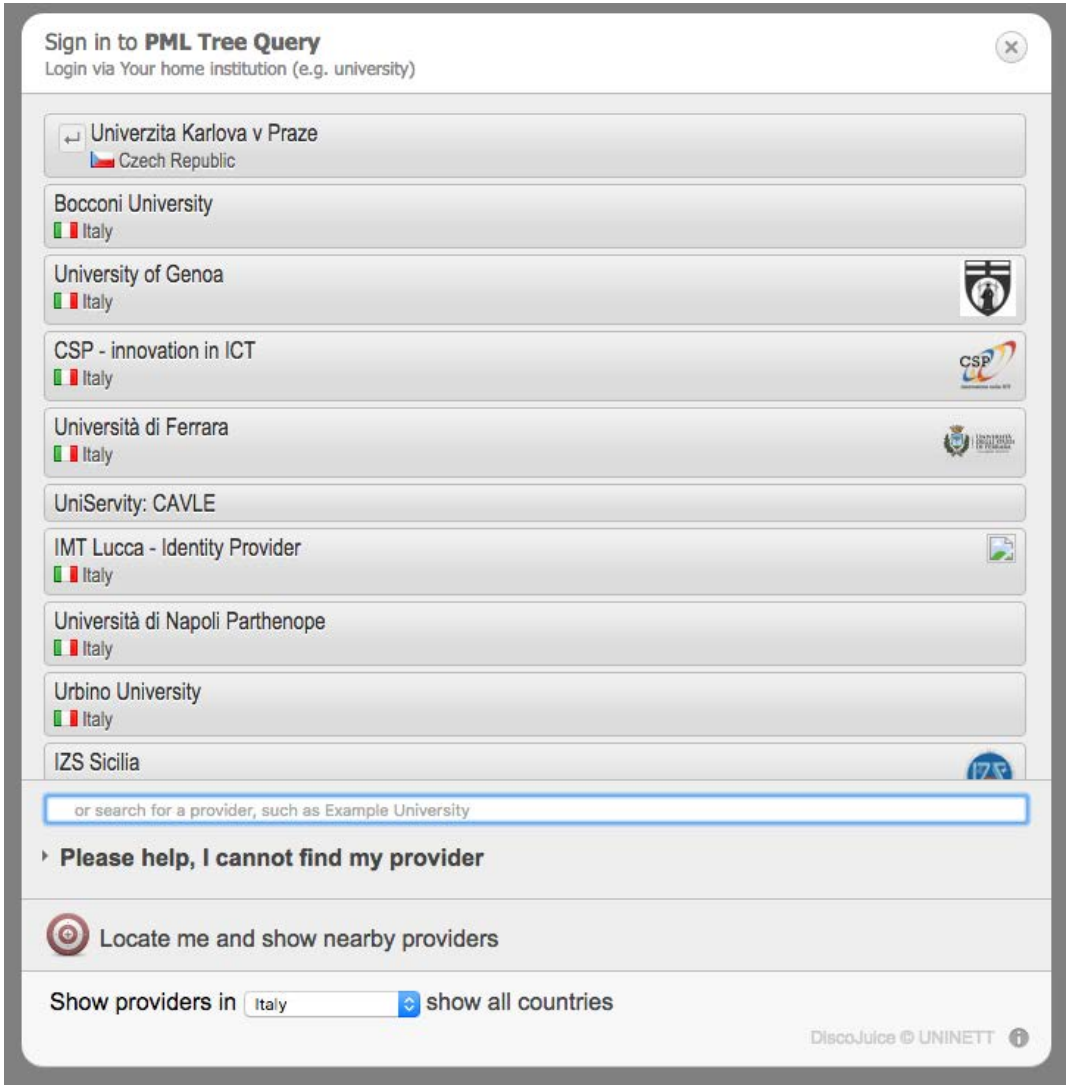
- The tool is freely accessible



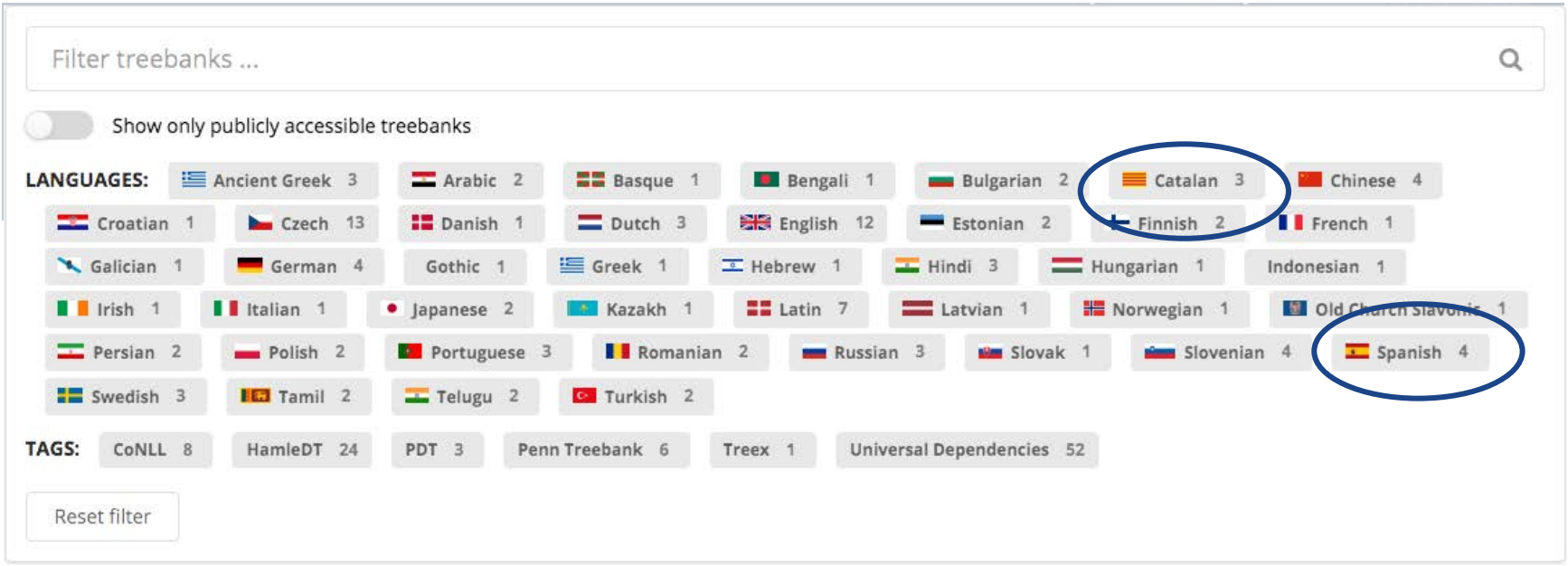
- Some corpora are freely available



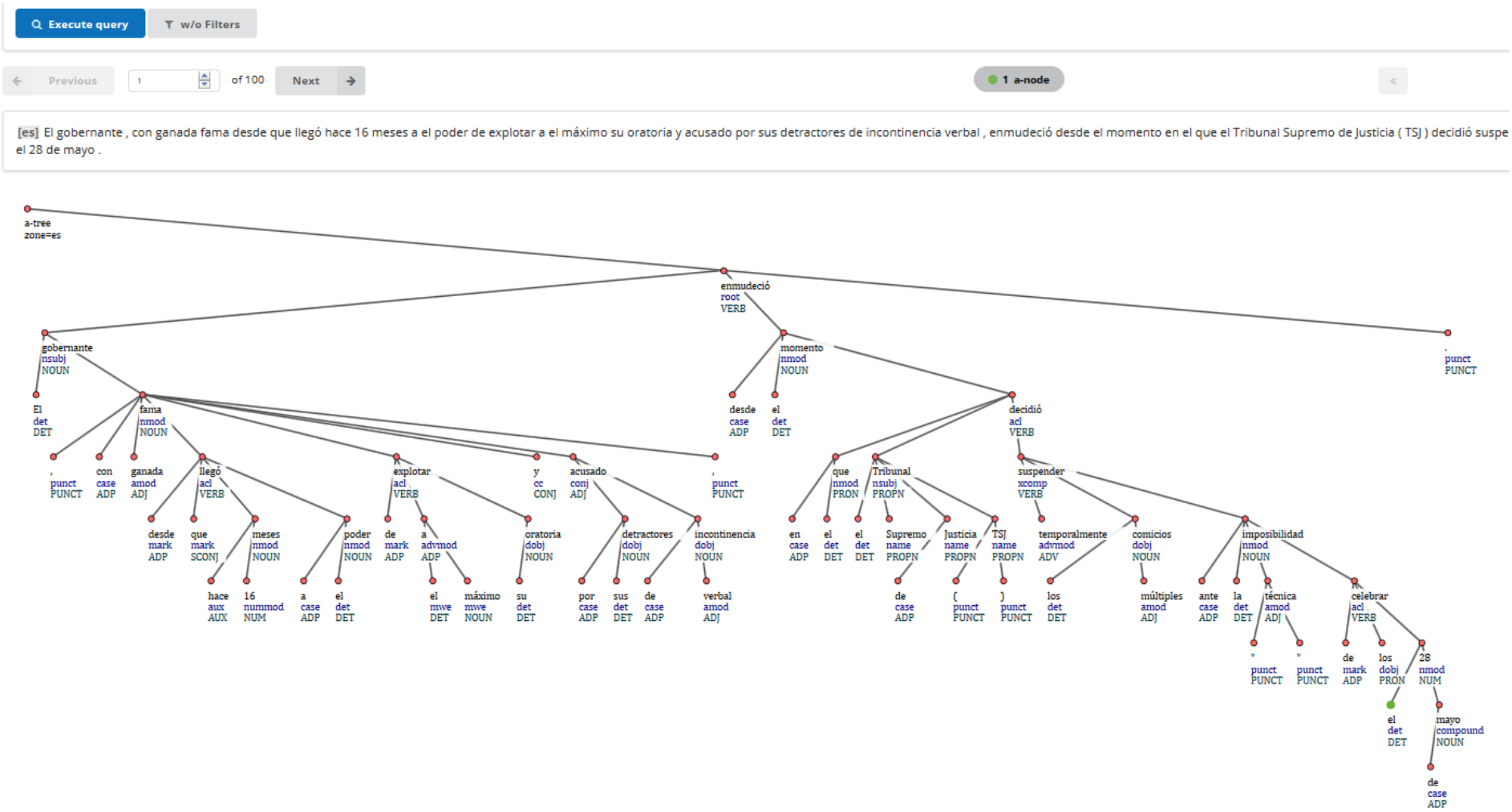
- A researcher can access the system with an institutional login if his/her institution is part of the EduGAIN Federation



- For instance, here is the list of the Italian institutions which are part of the EduGAIN Federation

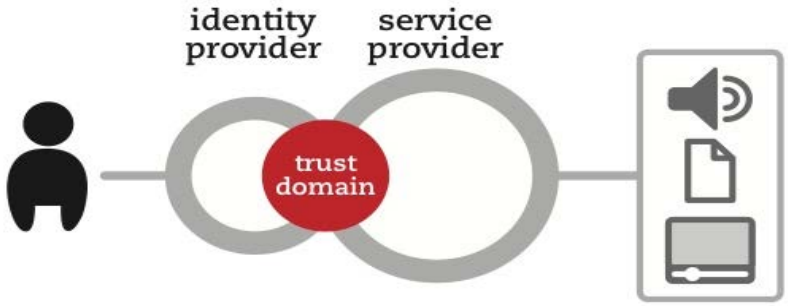


- After accessing the system, the number of available treebanks grows (because not only those ones freely available are shown)


Execute query
w/o Filters
Previous
1
of 100
Next
1 a-node
[es] El gobernante , con ganada fama desde que llegó hace 16 meses a el poder de explotar a el máximo su oratoria y acusado por sus detractores de incontinencia verbal , enmudeció desde el momento en el que el Tribunal Supremo de Justicia (TSJ) decidió suspe
el 28 de mayo .
a-tree
zone=es
enmudeció root VERB
gobernante nsubj NOUN
momento nmod NOUN
. punct PUNCT
El det DET
fama nmod NOUN
desde case ADP
el det DET
decidió acl VERB
, punct PUNCT
con case ADP
ganada amod ADJ
llegó acl VERB
explotar acl VERB
y cc CONJ
acusado conj ADJ
, punct PUNCT
que nmod PRON
Tribunal nsubj PROPN
suspender xcomp VERB
desde mark ADP
que mark SCONJ
meses nmod NOUN
poder nmod NOUN
de mark ADP
a advmod ADP
oratoria dobj NOUN
detractores dobj NOUN
incontinencia dobj NOUN
en case ADP
el det DET
el det DET
Supremo name PROPN
Justicia name PROPN
TSJ name PROPN
temporalmente advmod ADV
comicios dobj NOUN
imposibilidad nmod NOUN
hace aux AUX
16 nummod NUM
a case ADP
el det DET
el mwe DET
máximo mwe NOUN
su det DET
por case ADP
sus det DET
de case ADP
verbal amod ADJ
de case ADP
(punct PUNCT
) punct PUNCT
los det DET
múltiples amod ADJ
ante case ADP
la det DET
técnica amod ADJ
celebrar acl VERB
" punct PUNCT
" punct PUNCT
de mark ADP
los dobj PRON
28 nmod NUM
el det DET
mayo compound NOUN
de case ADP

Federated identity

- Making it easier for academic users to get access to password-protected language resources is an important goal of CLARIN
- Academic users should be able to login with their existing institutional credentials
- Accessing CLARIN applications becomes as simple as logging in one's university's webmail system
- The user' stores from universities and academic institutions ("Identity Providers") are connected to password-protected web applications ("Service Providers")
  - These trust networks ("Identity Federations") already exist at the national level in the EU countries
  - CLARIN is working at crossing the country borders when logging in, so that e.g. a Spanish researcher can access language resources hosted in Prague



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.		

Easy access to protected LRs

- Thanks to a federated login, the following applications and data sets are available to anyone with an academic computer account from many European countries 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.

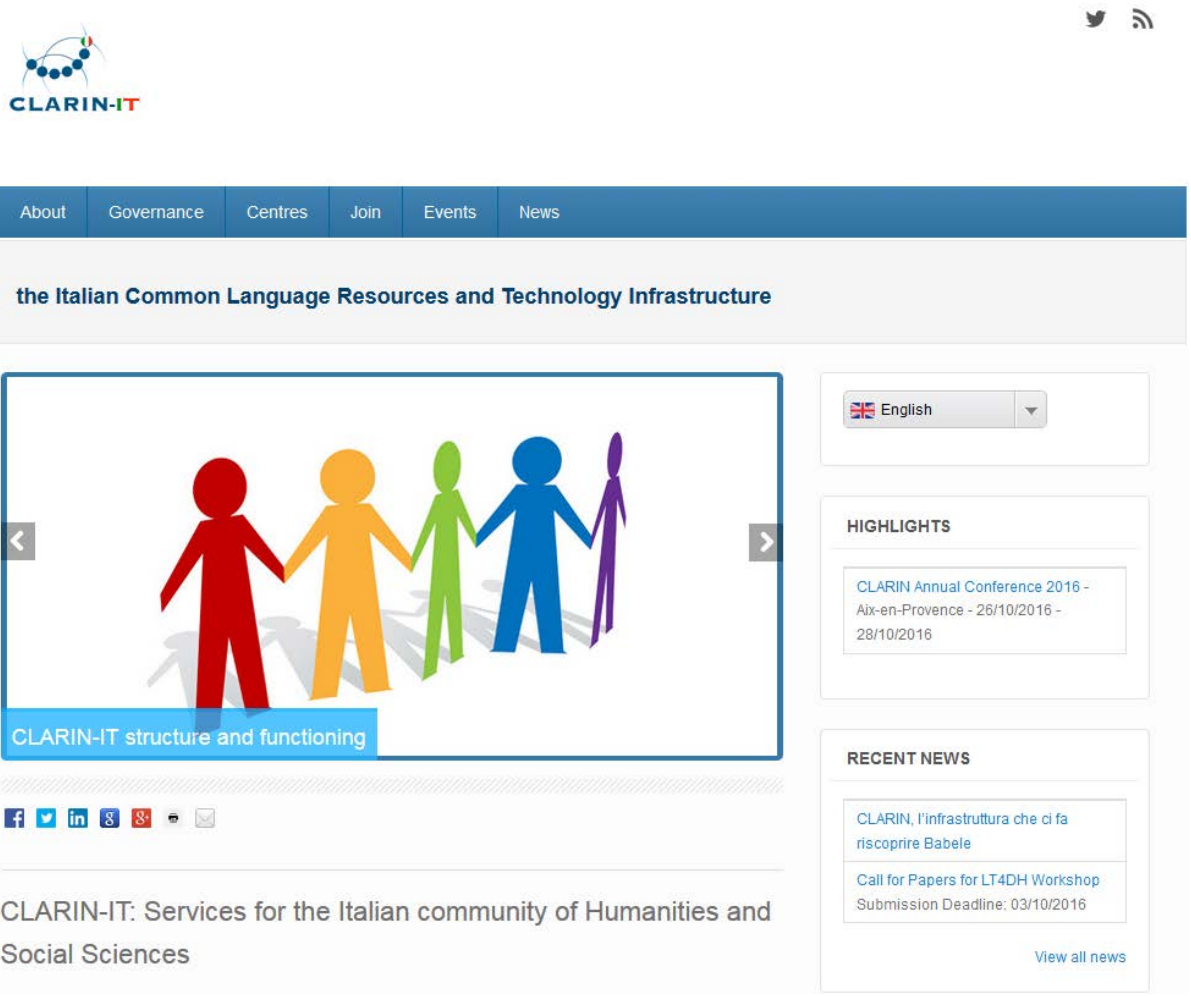
Resource or Tool	Description	Provided by
OpenSoNaR	over 500 million word Dutch reference corpus	INL (Instituut voor Nederlandse Lexicologie)
Corpus Hedendaags Nederlands	written corpus of contemporary Dutch	INL (Instituut voor Nederlandse Lexicologie)
VU-DNC	diachronic Dutch newspaper corpus	INL (Instituut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

CLARIN-IT: who does what



- CNR (Department of SSH) is the Representing Entity
- The National Coordinator
  - builds the National Consortium
  - is the interface between the research community and the ERIC
- ILC-CNR is the Executor
  - builds an Italian centre for documenting and accessing language resources and providing services

CLARIN-IT: Web portal

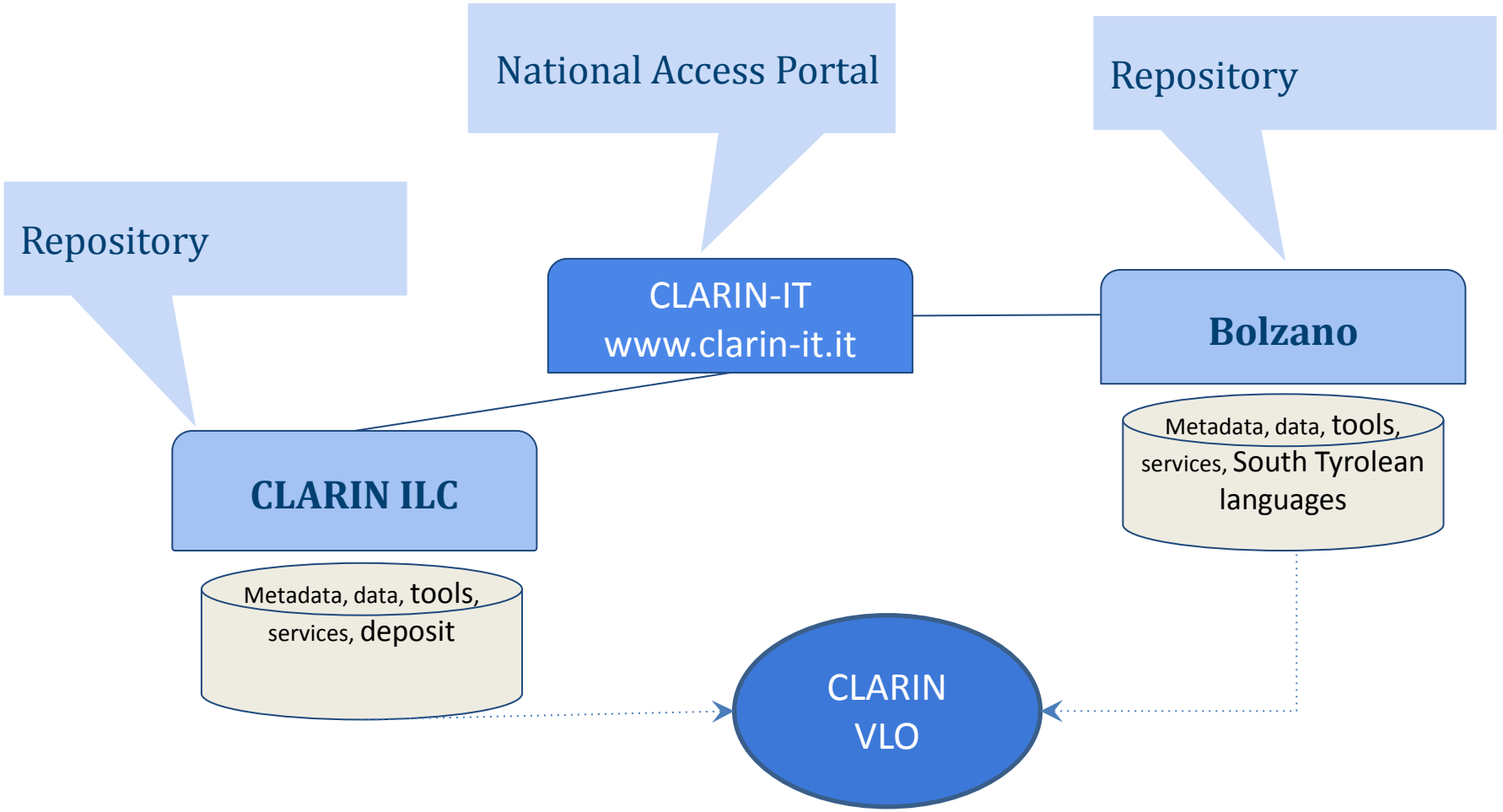


CLARIN-IT: consortium



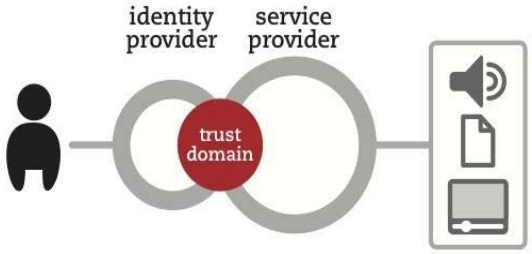
- **IDEM-GARR** Federated Identity
- Insitutions that have already joined or are about to join CLARIN-IT
  - **Siena University**
  - **EURAC Bozen**
  - **FBK Trento**
  - **Cattolica Milano University**
  - **Venezia University**
  - **Pisa University**
  - **Pisa Scuola Normale Superiore**
  - **Parma University**
  - **Roma University**

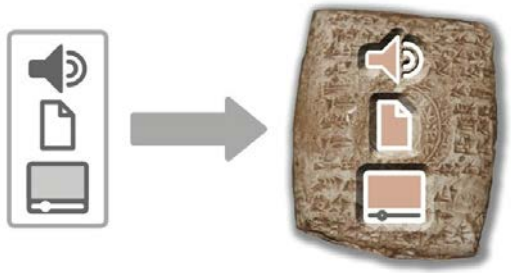
CLARIN-IT: structure

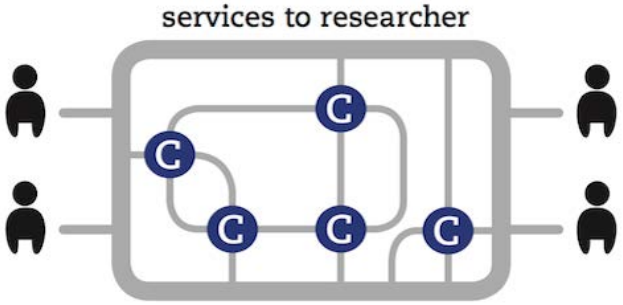


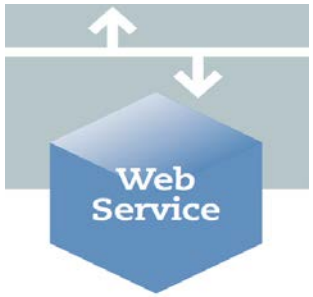
CLARIN ILC centre

- Access to language resources and services of the Institute
- Depositing services by third parties
- National collector of data and tools
- Italian language and classical languages
- Long term preservation
- Single sign-on









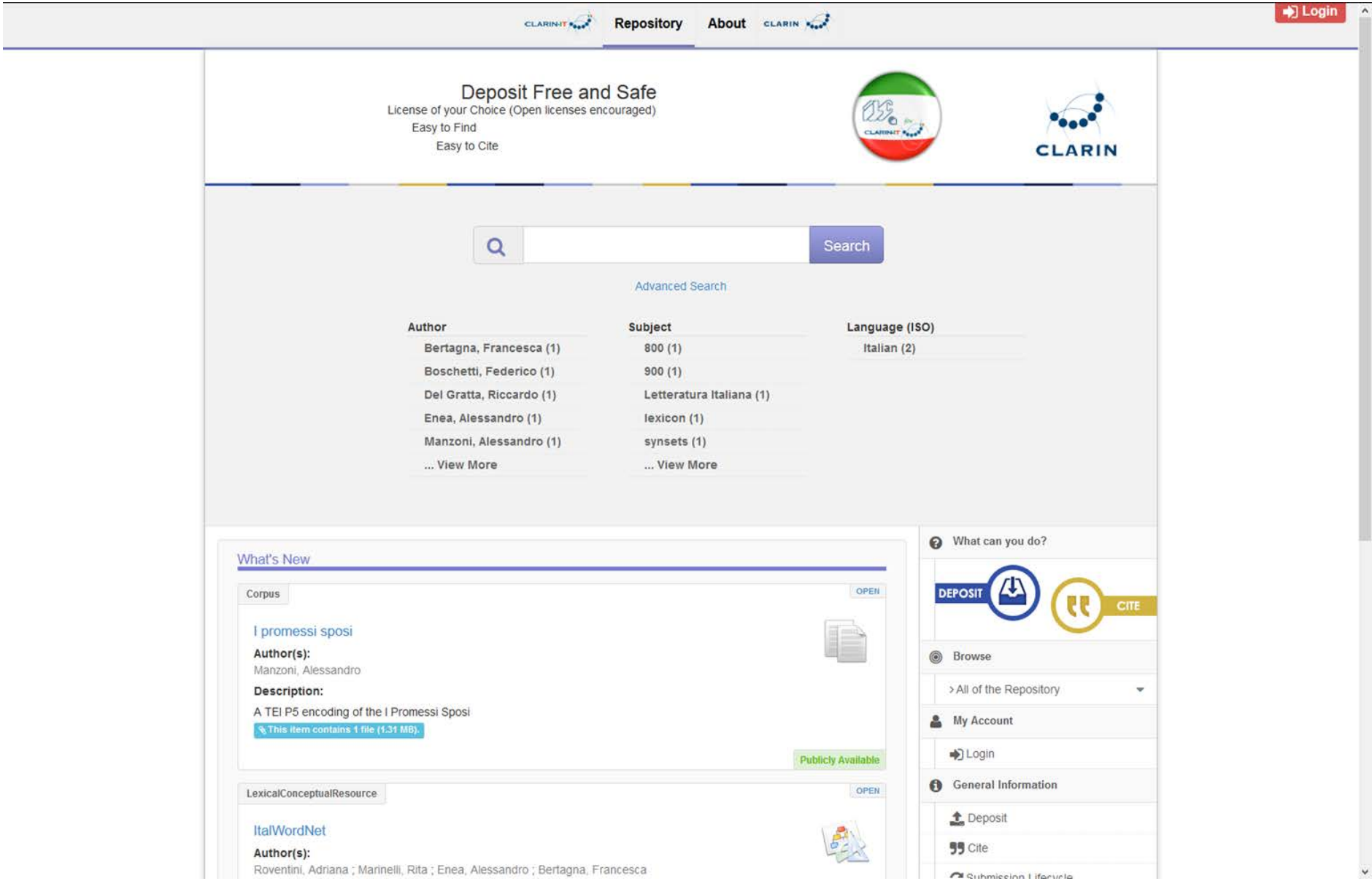


CLARIN ILC centre: documentation and deposit

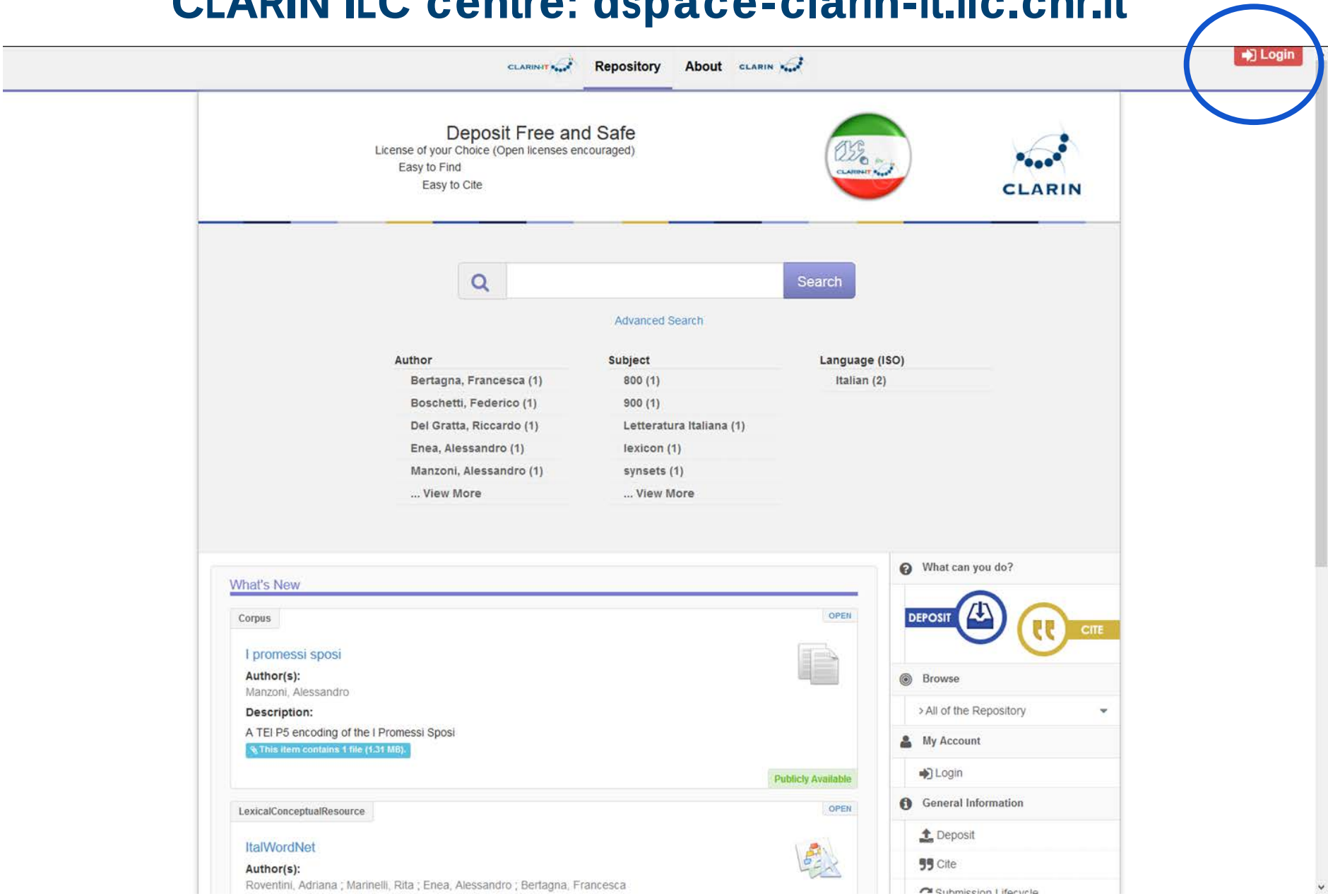


- Users can document and deposit their language resources in the CLARIN ILC repository
- That guarantees
  - availability
  - citation
  - usability

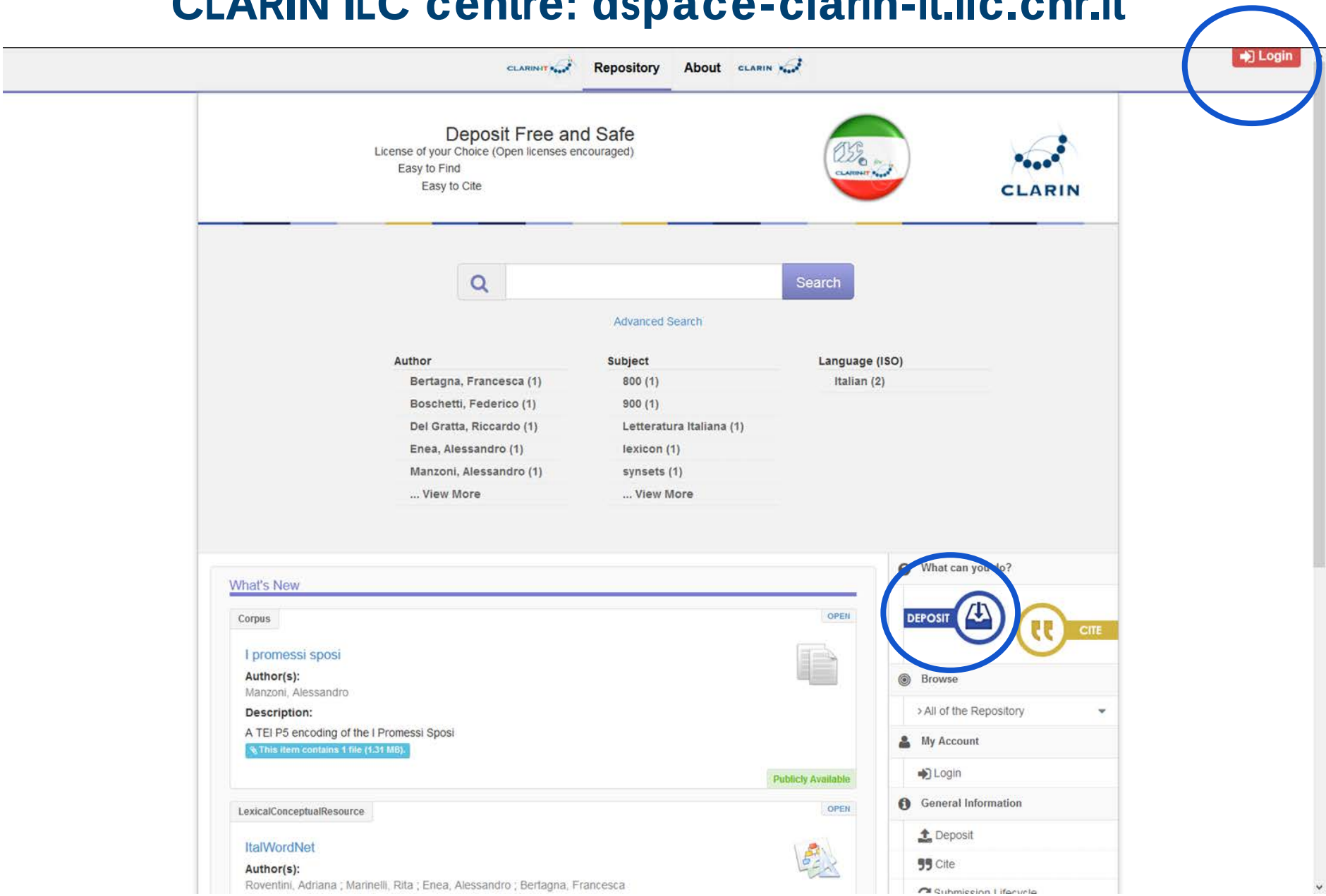
CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: 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.

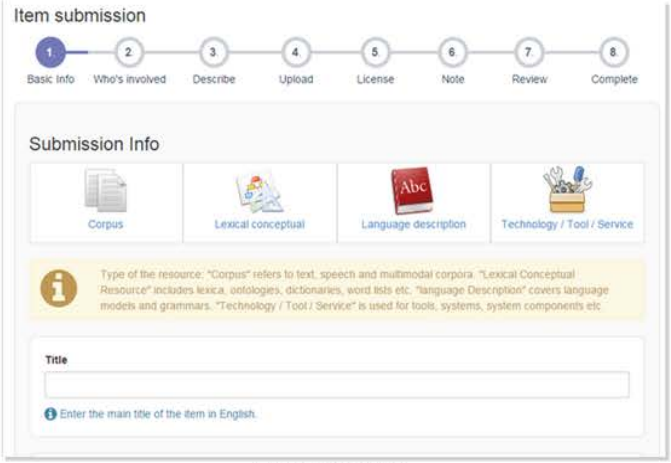
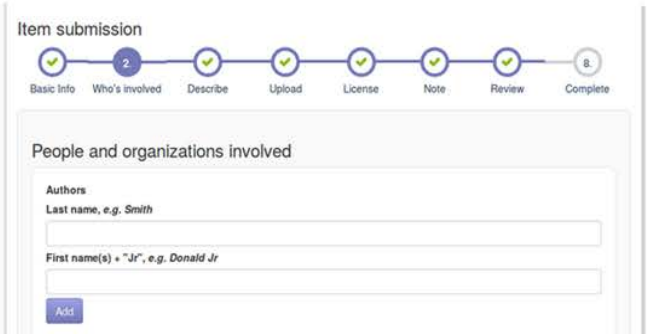


Fig4. Submission info

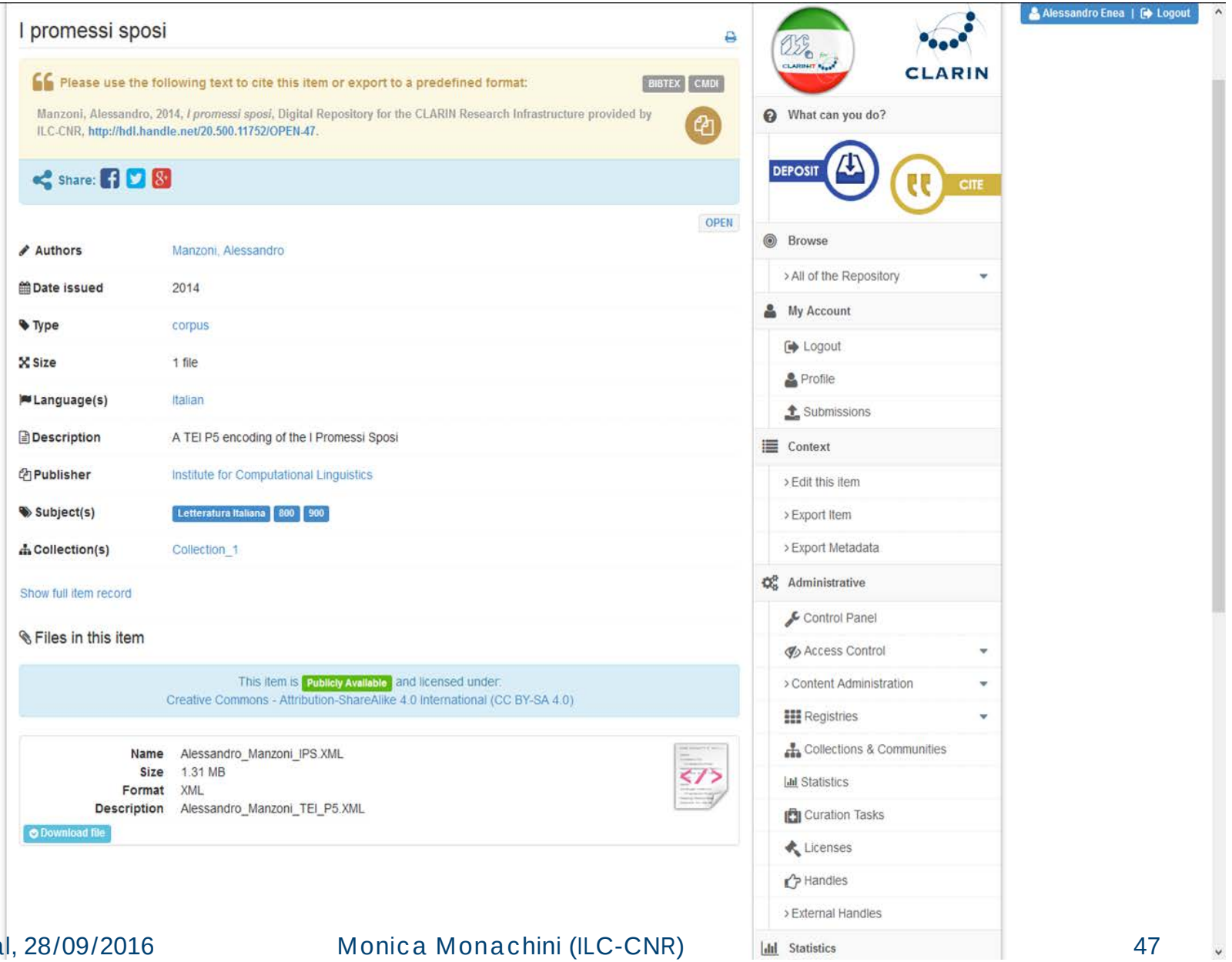
Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as the title. Click Next to continue the following step.

Step 4: Describe your item

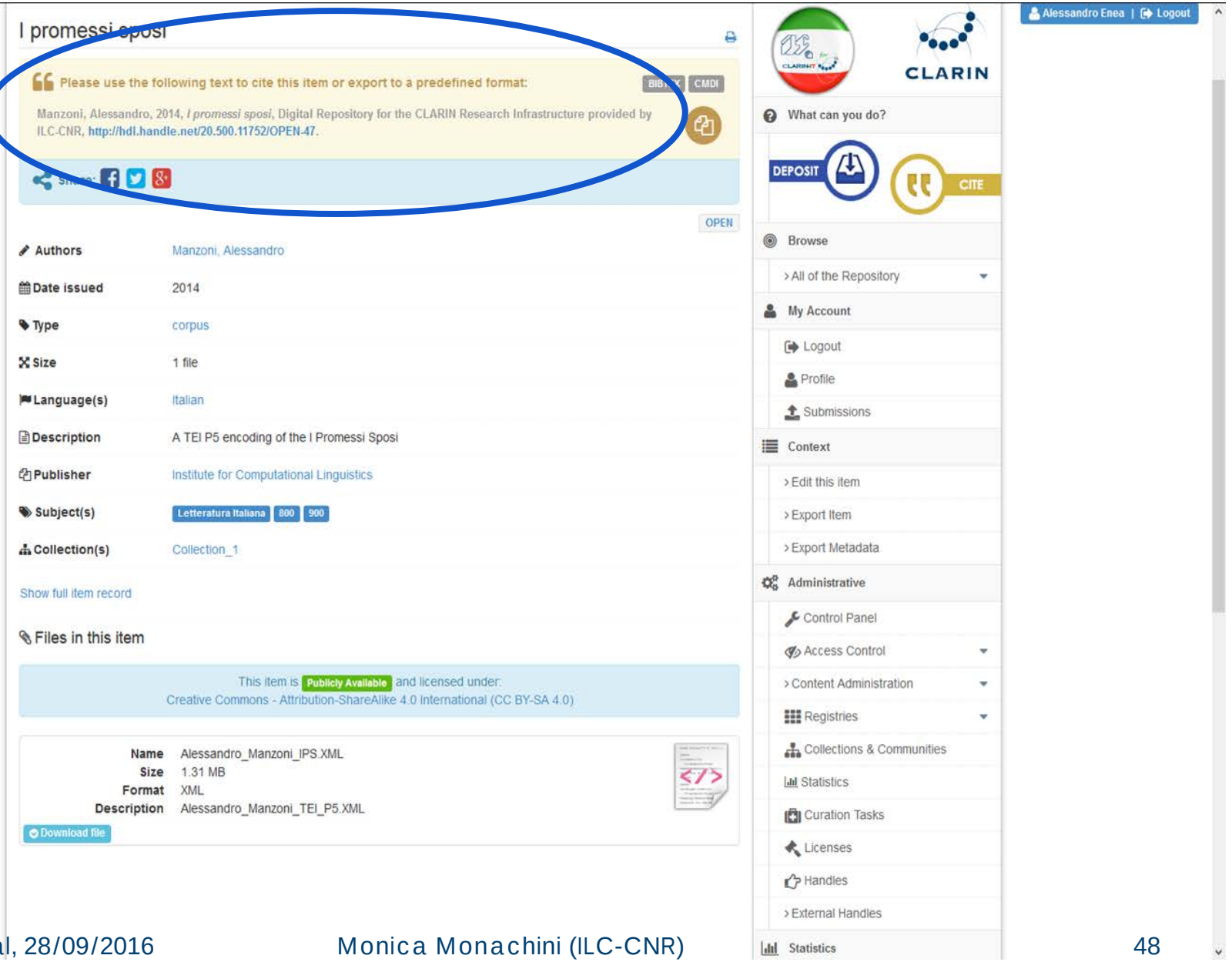
In the following two steps you will provide more details for your item. First describe the people, organization and projects involved with the item.



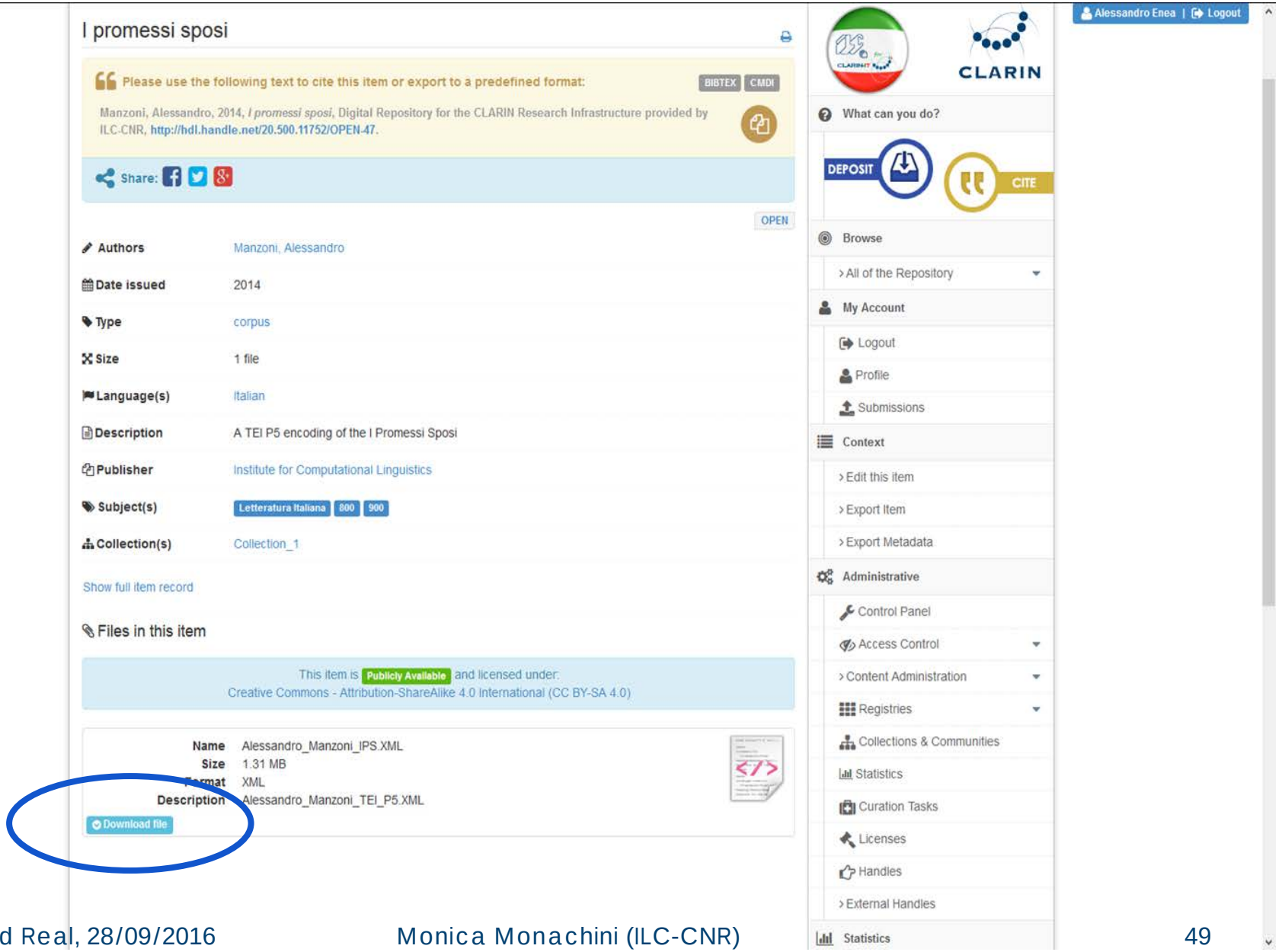
CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



CLARIN ILC centre: harvesting

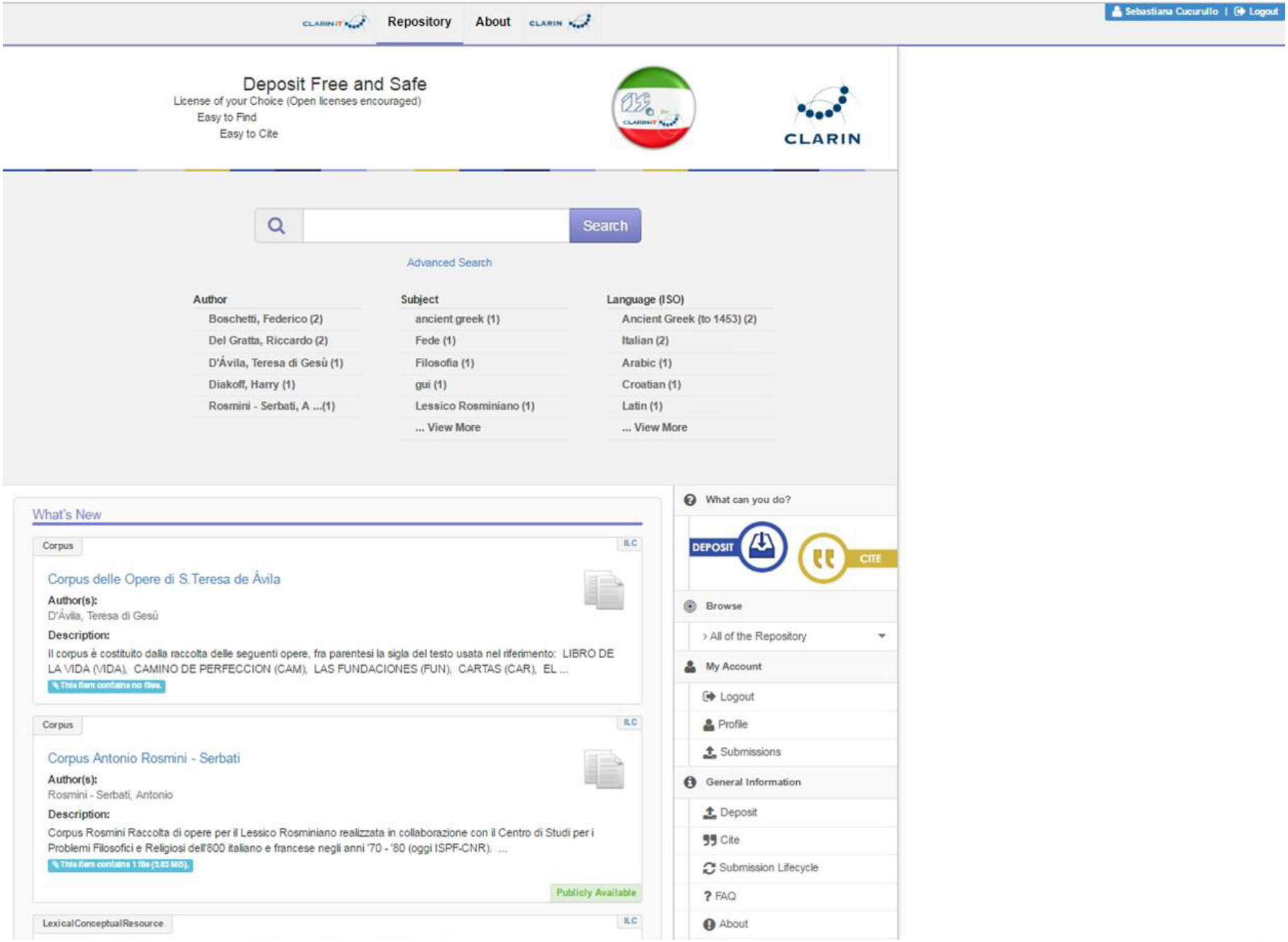


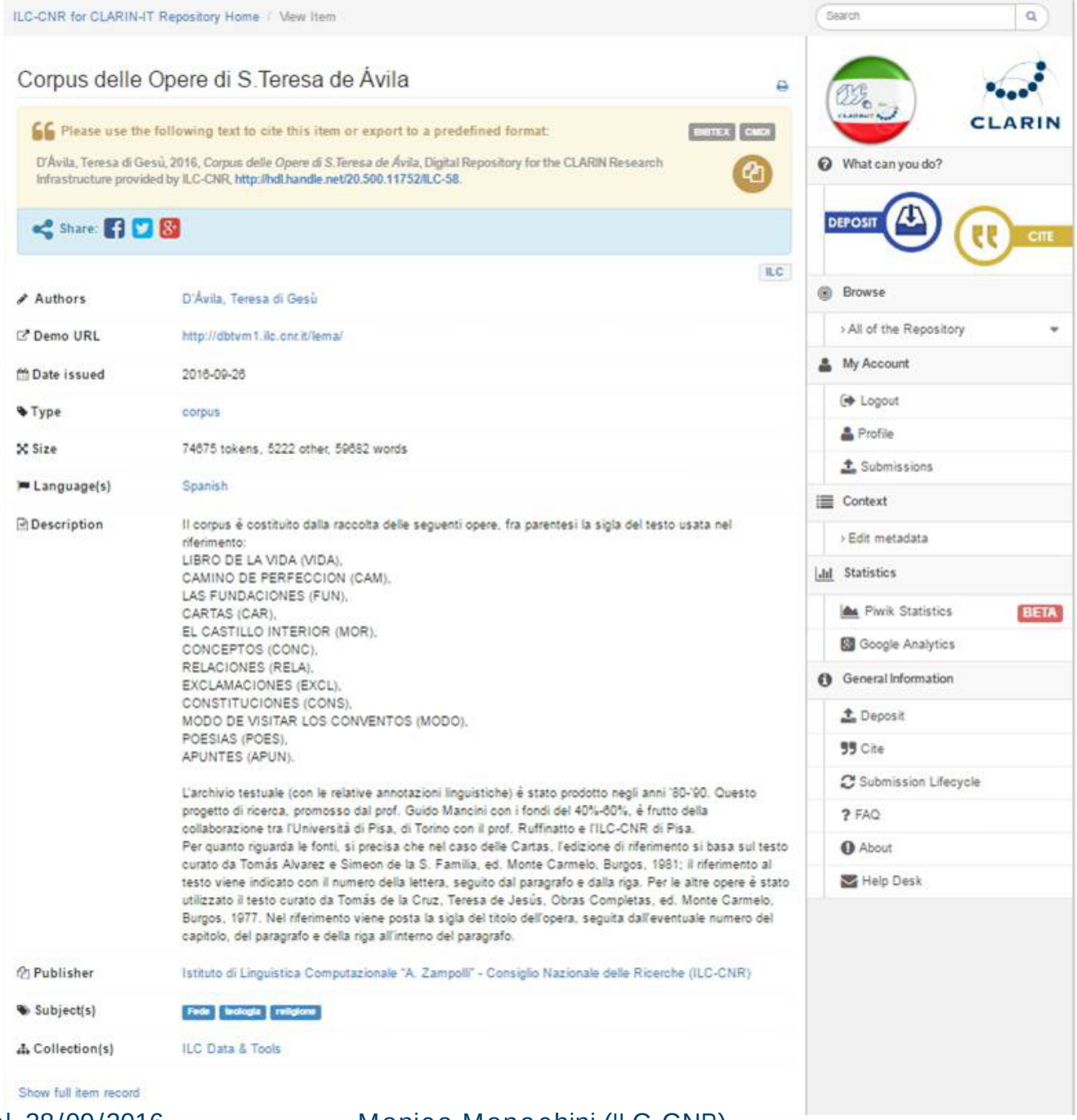
- CLARIN ILC metadata, as long as the metadata of the other European centres, are harvested in the central VLO
- Language resources are searchable in the central VLO

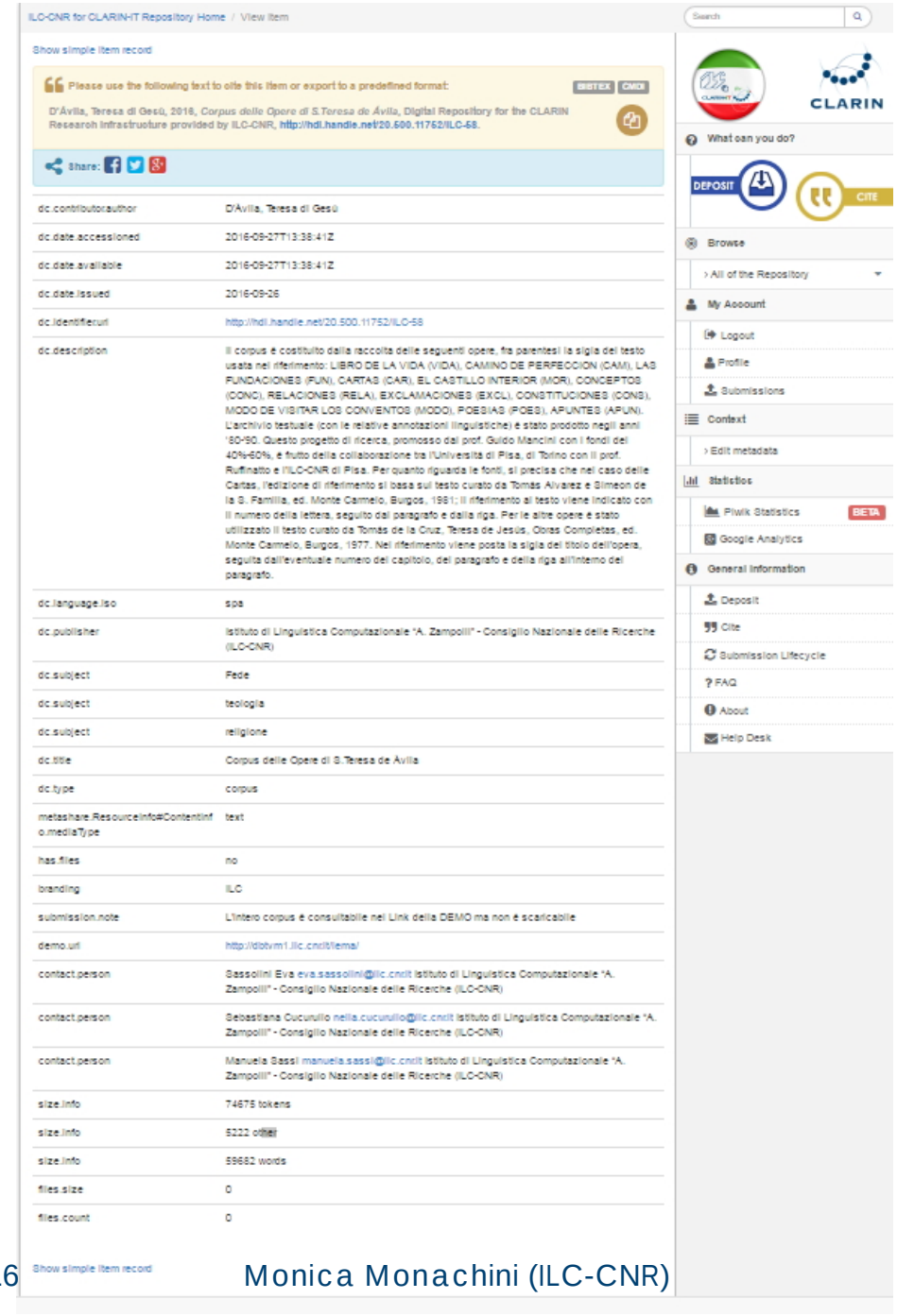
CLARIN ILC centre: Spanish LRs



Author	Texts
Santa Teresa d'Avila	Cartas - Camino Opemin - Fundacio Castillo
Cristoforo Colombo	Giornale di bordo - Lettere (Ediz. Varela) Lettere (parte I e II - Ediz. De Lollis)
Pedro Salinas	Poesias completas
Neruda	Poesie (volume I e II)
Miguel de Unamuno y Jugo	San Manuel bueno martir y tres histori
Gabriel García Márquez	Cronica de una muerte anunciada
Antonio Machado	Raccolta di Poesie
Juan del Encina	Raccolta delle opere teatrali (Teatro completo)
El Pais	El servilismo de las ciencias sociales


CLARIN-IT
Repository
About
CLARIN
Sebastiana Cucurullo | Logout
Deposit Free and Safe
License of your Choice (Open licenses encouraged)
Easy to Find
Easy to Cite
CLARIN
Search
Advanced Search
Author
Boschetti, Federico (2)
Del Gratta, Riccardo (2)
D'Ávila, Teresa di Gesù (1)
Diakoff, Harry (1)
Rosmini - Serbati, A ...(1)
Subject
ancient greek (1)
Fede (1)
Filosofia (1)
gui (1)
Lessico Rosminiano (1)
... View More
Language (ISO)
Ancient Greek (to 1453) (2)
Italian (2)
Arabic (1)
Croatian (1)
Latin (1)
... View More
What's New
Corpus
ILC
Corpus delle Opere di S.Teresa de Ávila
Author(s):
D'Ávila, Teresa di Gesù
Description:
Il corpus è costituito dalla raccolta delle seguenti opere, fra parentesi la sigla del testo usata nel riferimento: LIBRO DE LA VIDA (VIDA), CAMINO DE PERFECCION (CAM), LAS FUNDACIONES (FUN), CARTAS (CAR), EL ...
This item contains no files.
Corpus
ILC
Corpus Antonio Rosmini - Serbati
Author(s):
Rosmini - Serbati, Antonio
Description:
Corpus Rosmini Raccolta di opere per il Lessico Rosminiano realizzata in collaborazione con il Centro di Studi per i Problemi Filosofici e Religiosi dell'800 italiano e francese negli anni '70 - '80 (oggi ISPF-CNR). ...
This item contains 1 file (3.83 MB).
Publicly Available
LexicalConceptualResource
ILC
What can you do?
DEPOSIT
CITE
Browse
› All of the Repository
My Account
Logout
Profile
Submissions
General Information
Deposit
Cite
Submission Lifecycle
FAQ
About


ILC-CNR for CLARIN-IT Repository Home / View Item
Search
Corpus delle Opere di S.Teresa de Ávila
Please use the following text to cite this item or export to a predefined format:
BIBTEX
CMDI
D'Ávila, Teresa di Gesù, 2016, Corpus delle Opere di S.Teresa de Ávila, Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR, http://hdl.handle.net/20.500.11752/ILC-58.
Share:
ILC
Authors
D'Ávila, Teresa di Gesù
Demo URL
http://dbtvm1.ilc.cnr.it/lema/
Date issued
2016-09-26
Type
corpus
Size
74675 tokens, 5222 other, 59682 words
Language(s)
Spanish
Description
Il corpus è costituito dalla raccolta delle seguenti opere, fra parentesi la sigla del testo usata nel riferimento:
LIBRO DE LA VIDA (VIDA),
CAMINO DE PERFECCION (CAM),
LAS FUNDACIONES (FUN),
CARTAS (CAR),
EL CASTILLO INTERIOR (MOR),
CONCEPTOS (CONC),
RELACIONES (RELA),
EXCLAMACIONES (EXCL),
CONSTITUCIONES (CONS),
MODO DE VISITAR LOS CONVENTOS (MODO),
POESIAS (POES),
APUNTES (APUN).
L'archivio testuale (con le relative annotazioni linguistiche) è stato prodotto negli anni '80-'90. Questo progetto di ricerca, promosso dal prof. Guido Mancini con i fondi del 40%-60%, è frutto della collaborazione tra l'Università di Pisa, di Torino con il prof. Ruffinatto e l'ILC-CNR di Pisa.
Per quanto riguarda le fonti, si precisa che nel caso delle Cartas, l'edizione di riferimento si basa sul testo curato da Tomás Alvarez e Simeon de la S. Familia, ed. Monte Carmelo, Burgos, 1981; il riferimento al testo viene indicato con il numero della lettera, seguito dal paragrafo e dalla riga. Per le altre opere è stato utilizzato il testo curato da Tomás de la Cruz, Teresa de Jesús, Obras Completas, ed. Monte Carmelo, Burgos, 1977. Nel riferimento viene posta la sigla del titolo dell'opera, seguita dall'eventuale numero del capitolo, del paragrafo e della riga all'interno del paragrafo.
Publisher
Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)
Subject(s)
Fede
teologia
religione
Collection(s)
ILC Data & Tools
Show full item record
CLARIN
What can you do?
DEPOSIT
CITE
Browse
› All of the Repository
My Account
Logout
Profile
Submissions
Context
› Edit metadata
Statistics
Piwik Statistics
BETA
Google Analytics
General Information
Deposit
Cite
Submission Lifecycle
FAQ
About
Help Desk


ILC-CNR for CLARIN-IT Repository Home / View item
Search
Show simple item record
Please use the following text to cite this item or export to a predefined format:
BIBTEX
CMDI
D'Ávila, Teresa di Gesù, 2016, Corpus delle Opere di S.Teresa de Ávila, Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR, http://hdl.handle.net/20.500.11752/ILC-58.
Share:
dc.contributor.author D'Ávila, Teresa di Gesù
dc.date.accessioned 2016-09-27T13:38:41Z
dc.date.available 2016-09-27T13:38:41Z
dc.date.issued 2016-09-26
dc.identifier.uri http://hdl.handle.net/20.500.11752/ILC-58
dc.description Il corpus è costituito dalla raccolta delle seguenti opere, fra parentesi la sigla del testo usata nel riferimento: LIBRO DE LA VIDA (VIDA), CAMINO DE PERFECCION (CAM), LAS FUNDACIONES (FUN), CARTAS (CAR), EL CASTILLO INTERIOR (MOR), CONCEPTOS (CONC), RELACIONES (RELA), EXCLAMACIONES (EXCL), CONSTITUCIONES (CONS), MODO DE VISITAR LOS CONVENTOS (MODO), POESIAS (POES), APUNTES (APUN). L'archivio testuale (con le relative annotazioni linguistiche) è stato prodotto negli anni '80-'90. Questo progetto di ricerca, promosso dal prof. Guido Mancini con i fondi del 40%-60%, è frutto della collaborazione tra l'Università di Pisa, di Torino con il prof. Rufinatto e l'ILC-CNR di Pisa. Per quanto riguarda le fonti, si precisa che nel caso delle Cartas, l'edizione di riferimento si basa sul testo curato da Tomás Alvarez e Silmeon de la S. Familia, ed. Monte Carmelo, Burgos, 1981; il riferimento al testo viene indicato con il numero della lettera, seguito dal paragrafo e dalla riga. Per le altre opere è stato utilizzato il testo curato da Tomás de la Cruz, Teresa de Jesús, Obras Completas, ed. Monte Carmelo, Burgos, 1977. Nel riferimento viene posta la sigla del titolo dell'opera, seguita dall'eventuale numero del capitolo, del paragrafo e della riga all'interno del paragrafo.
dc.language.iso spa
dc.publisher Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)
dc.subject Fede
dc.subject teologia
dc.subject religione
dc.title Corpus delle Opere di S.Teresa de Ávila
dc.type corpus
metashare.ResourceInfo#ContentInfo.mediaType text
has.files no
branding ILC
submission.note L'intero corpus è consultabile nel Link della DEMO ma non è scaricabile
demo.url http://dobvm1.ilc.cnr.it/tema/
contact.person Sassolini Eva eva.sassolini@ilc.cnr.it Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)
contact.person Sebastiana Cucurullo nella.cucurullo@ilc.cnr.it Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)
contact.person Manuela Sassi manuela.sassi@ilc.cnr.it Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)
size.info 74675 tokens
size.info 5222 other
size.info 59682 words
files.size 0
files.count 0
Show simple item record
CLARIN
What can you do?
DEPOSIT
CITE
Browse
All of the Repository
My Account
Logout
Profile
Submissions
Context
Edit metadata
Statistics
Piwik Statistics
BETA
Google Analytics
General Information
Deposit
Cite
Submission Lifecycle
FAQ
About
Help Desk

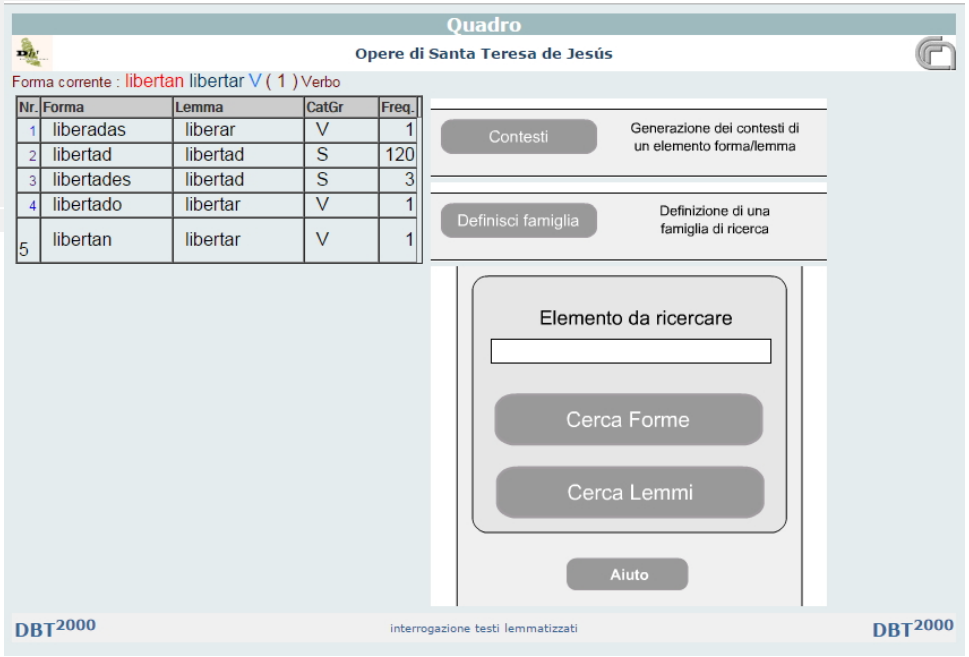
Query system

- Santa Teresa's corpus is morpho-syntactically analysed and queried by a dedicated query system http://dbtvm1.ilc.cnr.it/lema/
- The system will be available soon through different platforms and devices and also provided as a CLARIN service

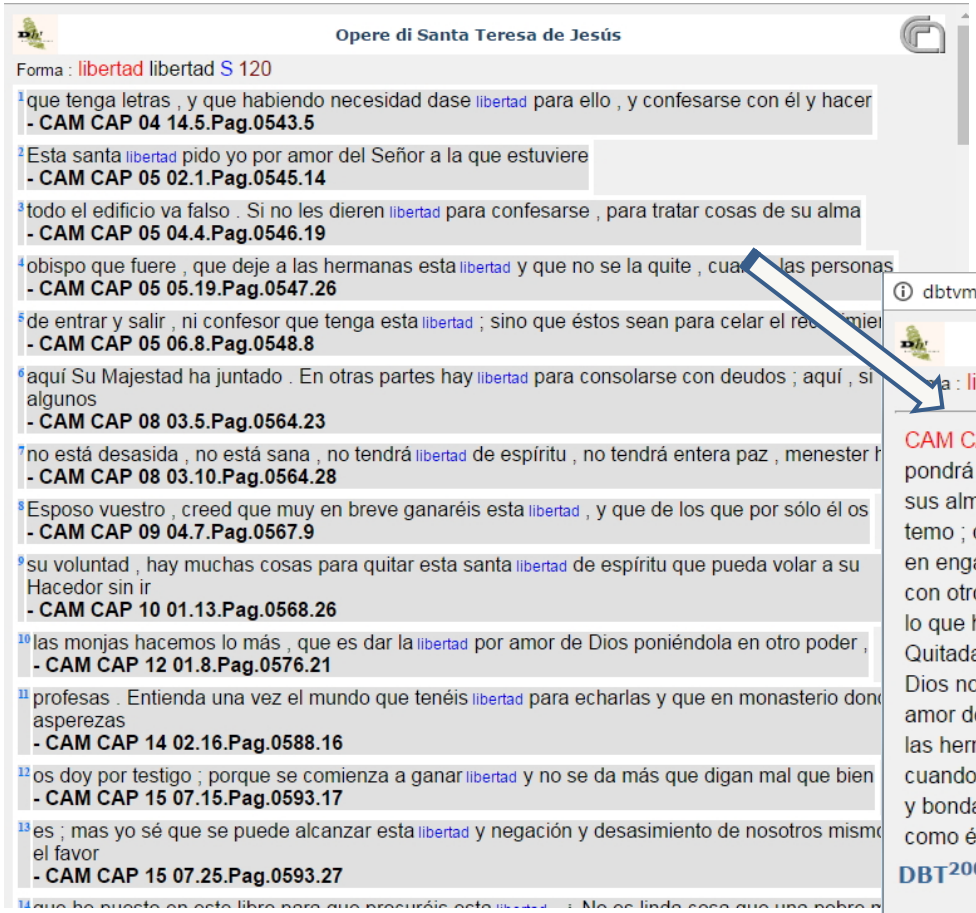
Query system



The corpus can be browsed by lemma or form

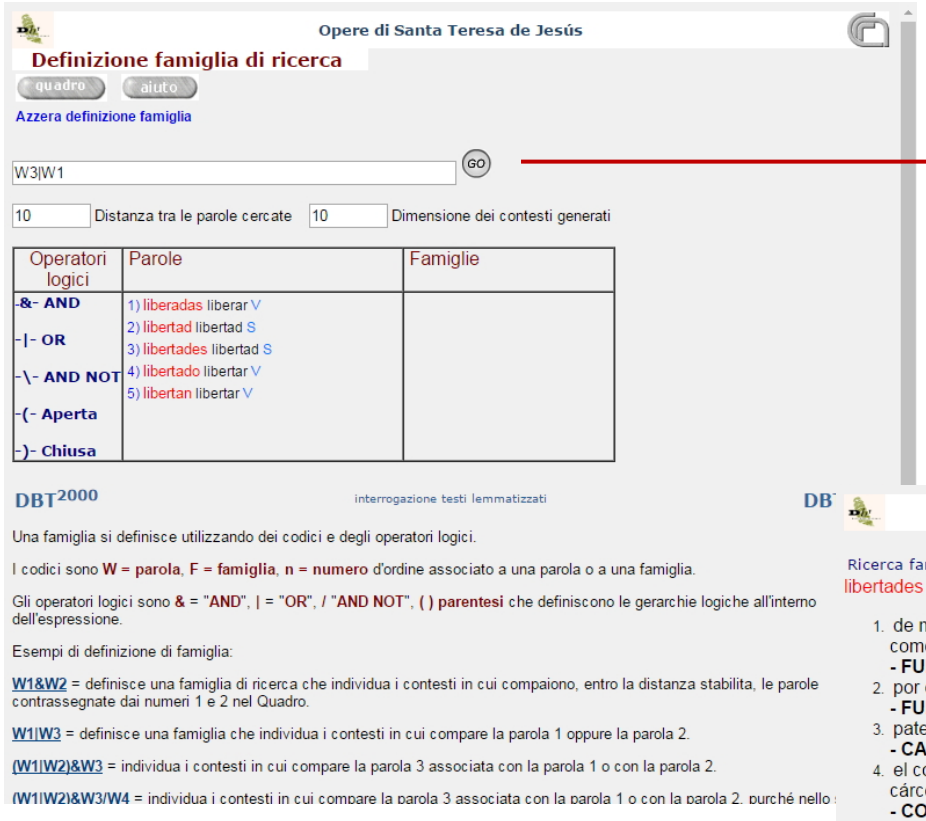


Query system

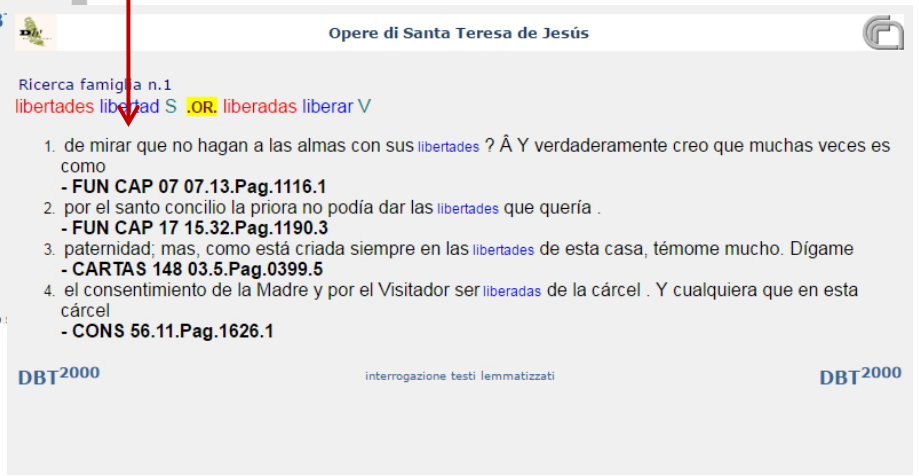


From the form to the context and from the context to the full sentence

Query system

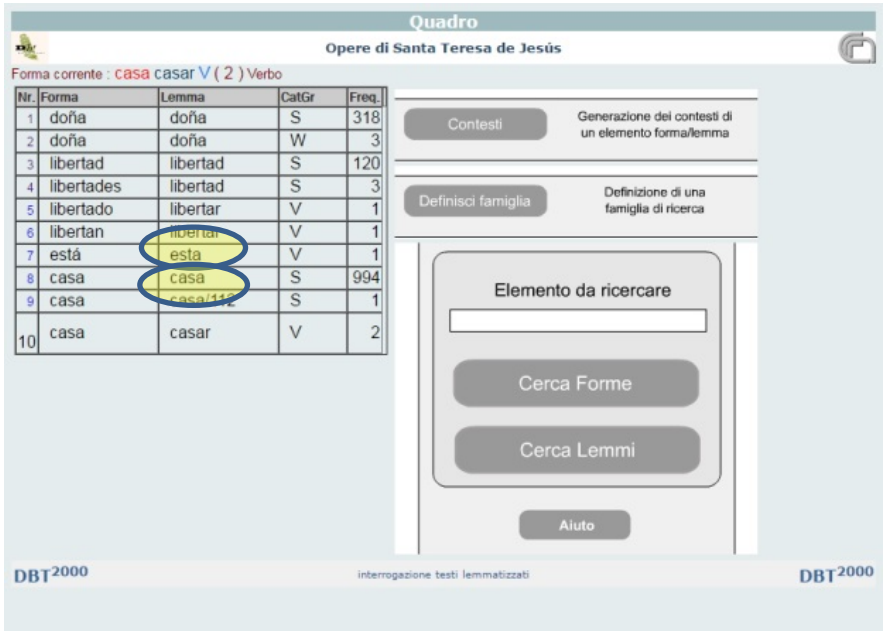


Search for groups of words (logical operators AND, OR, NOT)



Query system

Some examples



Search for «casa» and «esta»

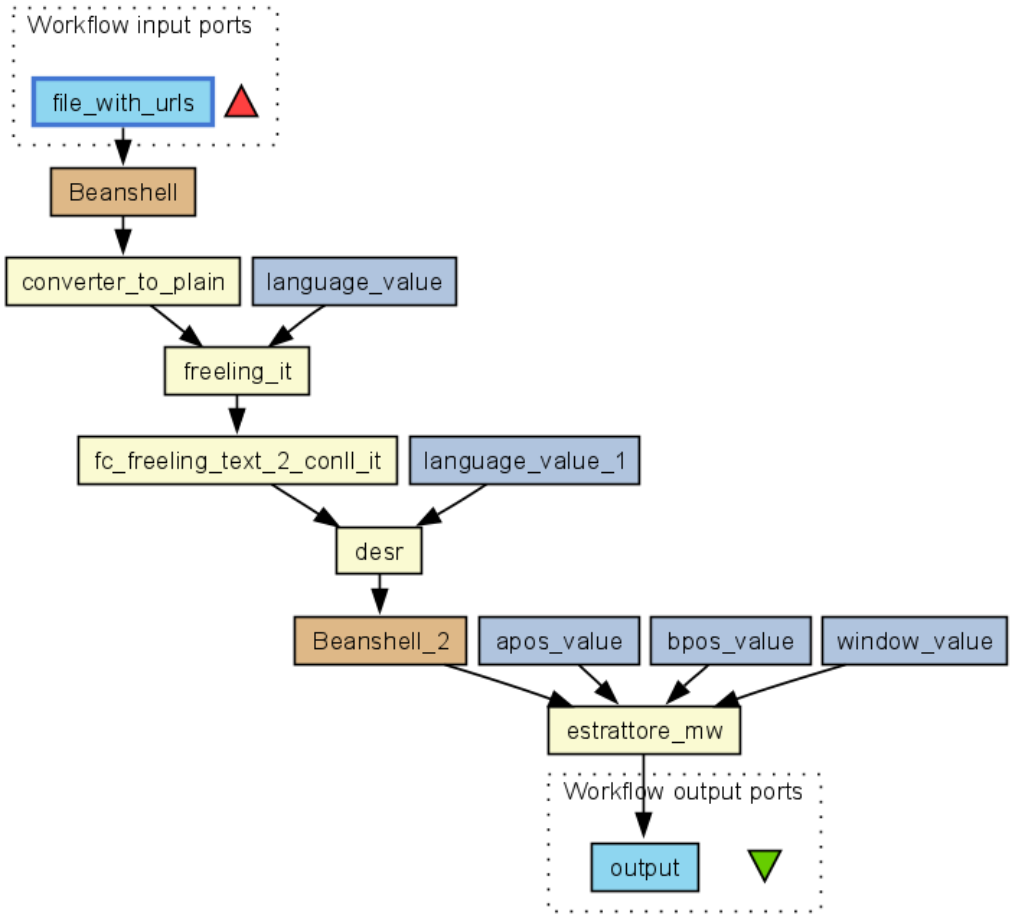


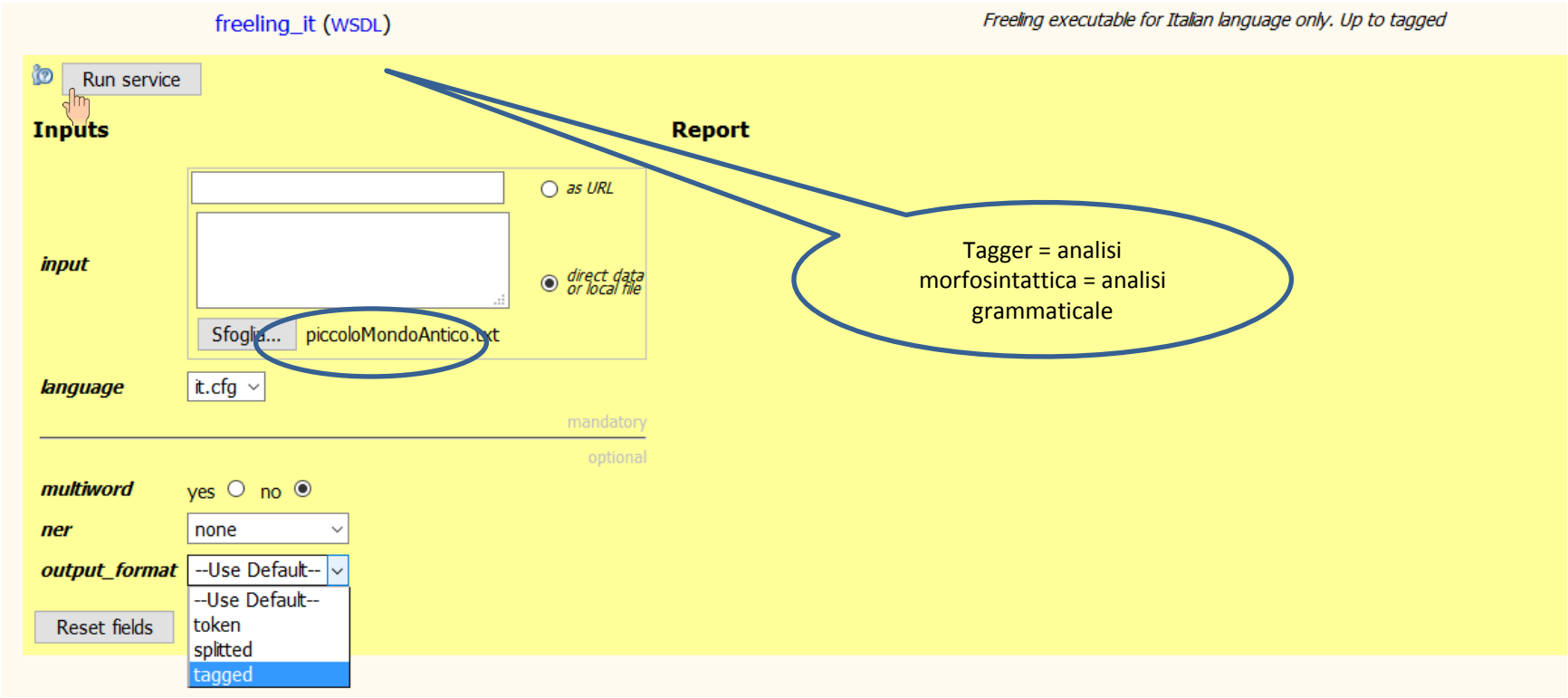
Query system

Ricerca famiglia n.4
esta este A .AND. casa casa S

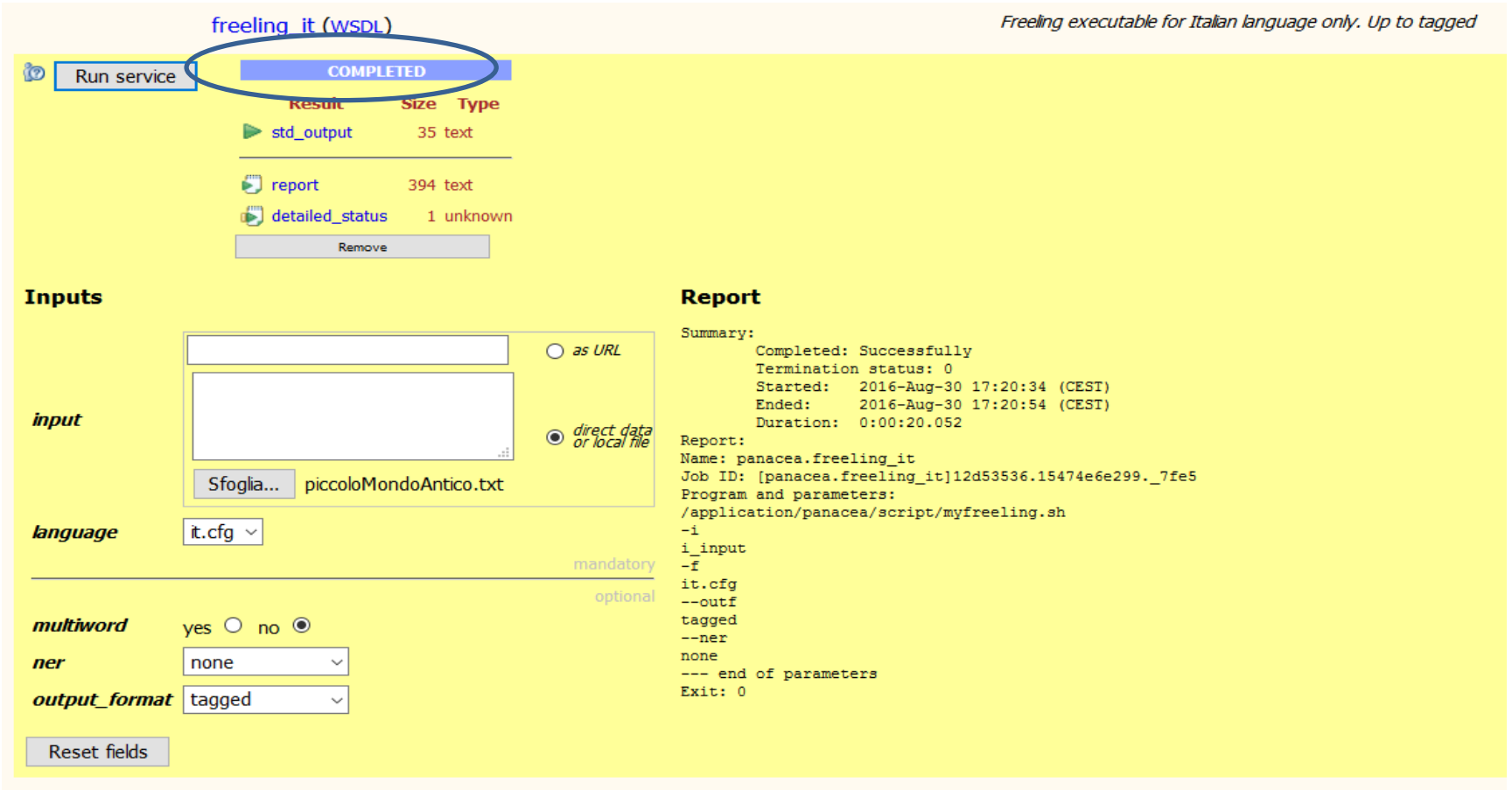
1. al modo y manera de vivir que se lleva en esta casa conviene . Y si fuere mal acertado el Padre
   - CAM JHS 01.16.Pag.0517.3
2. que dio a entender se había mucho de servir en esta casa no era mi intención hubiera tanta aspereza en lo
   - CAM CAP 01 01.5.Pag.0519.7
3. la más chiquita , cuando esto entendiese alguna vez en esta casa clame a Su Majestad y acuérdelo a la
   - CAM CAP 02 04.4.Pag.0523.30
4. mientras esto hicieren , no hayan miedo la religión de esta casa , con el favor de Dios ; que ,
   - CAM CAP 02 08.5.Pag.0526.1
5. lo principal para lo que el Señor nos juntó en esta casa y por lo que yo mucho deseo algo para
   - CAM CAP 03 01.2.Pag.0527.22
6. yo muchas , donde hay gran convento , que en esta casa " que no son más de trece , ni
   - CAM CAP 04 07.2.Pag.0538.13
7. y pues éste ha de ser el cimiento de esta casa , es menester traer estudio en aficionarnos a
   - CAM CAP 04 09.13.Pag.0540.7
8. espero en Su Majestad siempre la habrá en las de esta casa . Así que en esto no hay que encomendar
   - CAM CAP 04 10.11.Pag.0540.19
9. No dé el Señor a probar a nadie en esta casa el trabajo que queda dicho , por quien Su
   - CAM CAP 05 01.1.Pag.0544.19
10. demonio , yo espero en Dios no le tendrá en esta casa ; y así pido por el amor del Señor
    - CAM CAP 05 05.17.Pag.0547.24
11. santas , que han mirado lo que más convenía a esta casa para que la perfección de esta casa fuese adelante
    - CAM CAP 05 06.3.Pag.0548.3
12. santas , que han mirado lo que más convenía a esta casa para que la perfección de esta casa fuese adelante
    - CAM CAP 05 06.3.Pag.0548.3

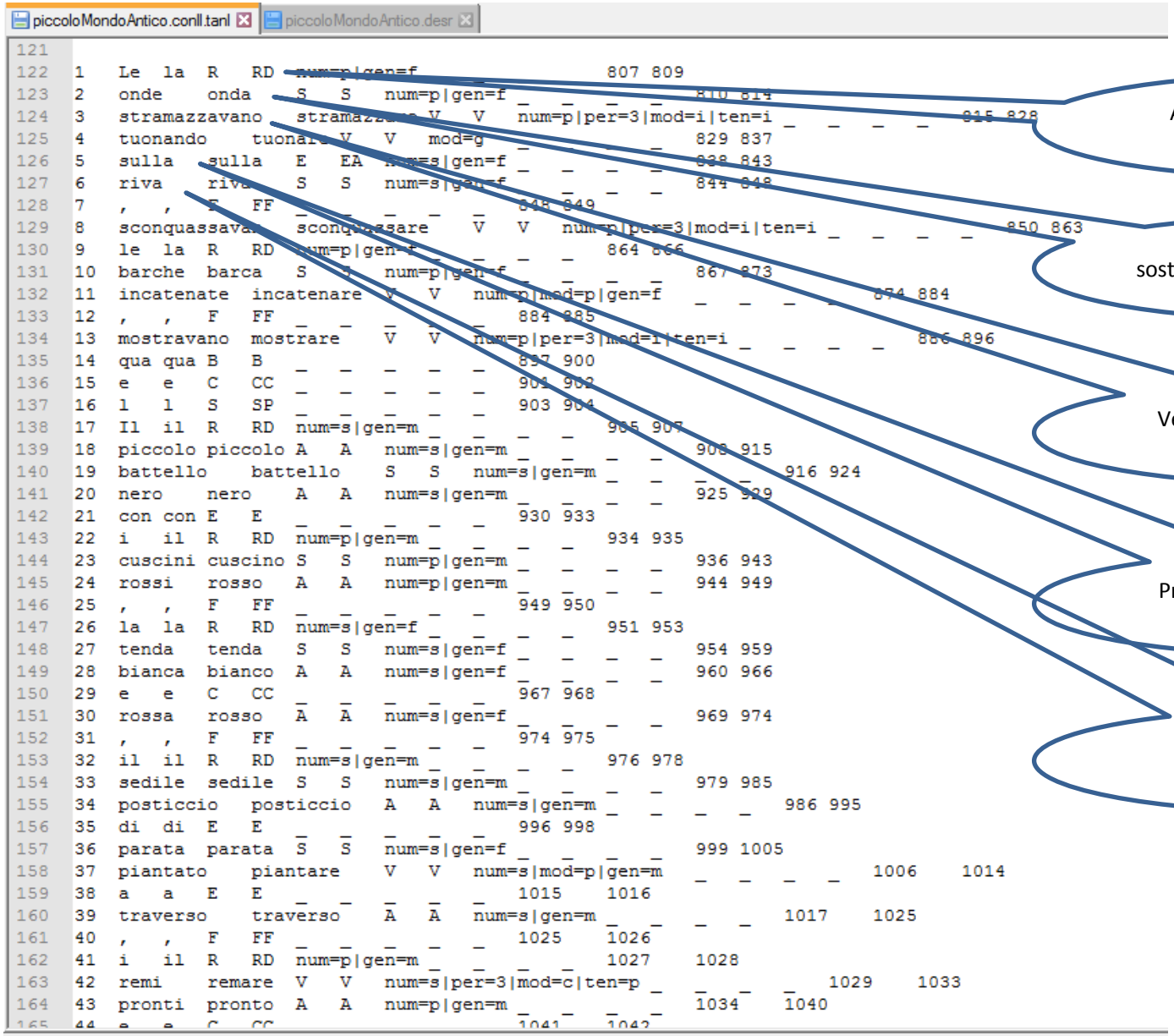
Pipelines

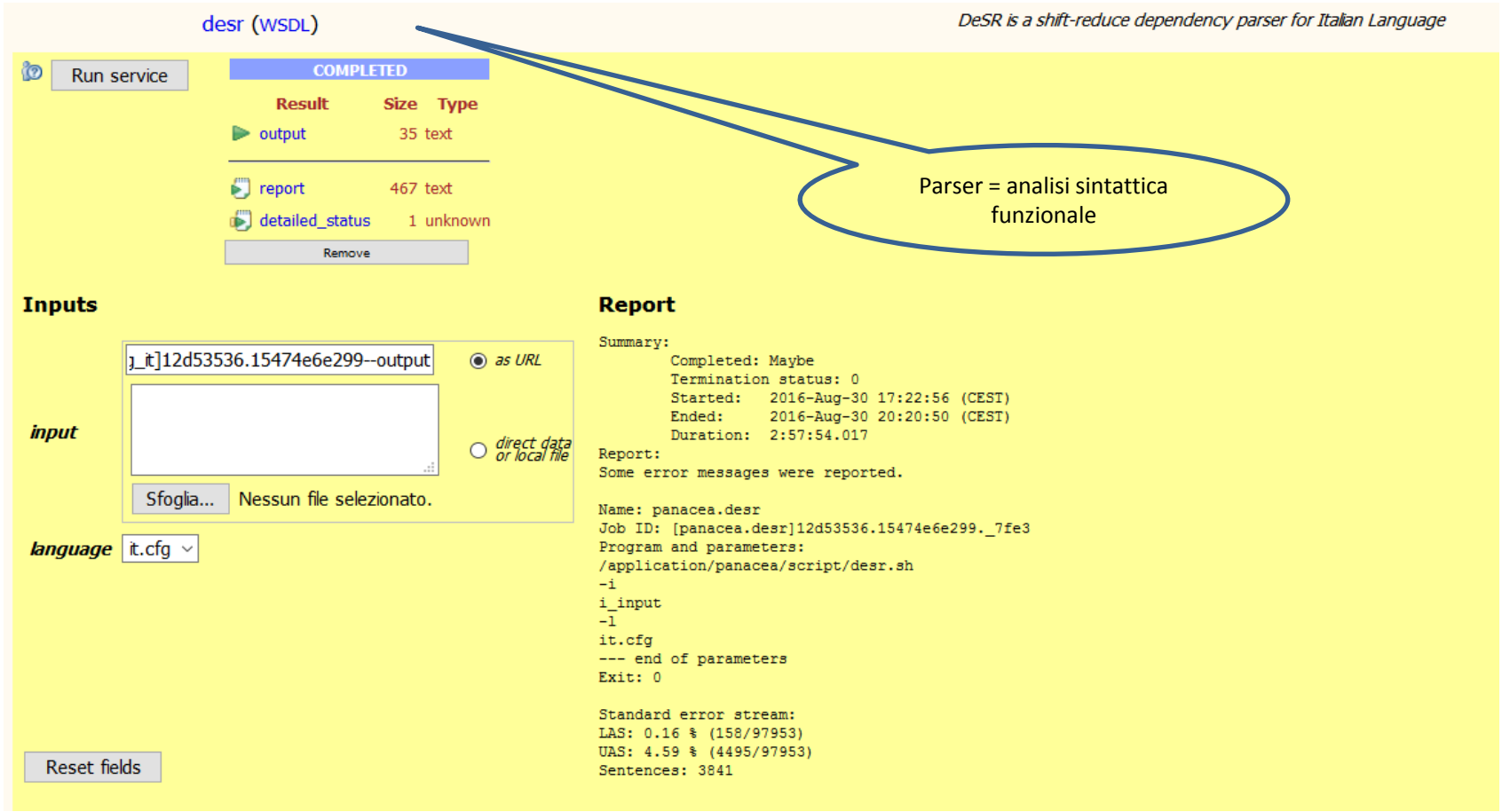

Workflow input ports
file_with_urls
Beanshell
converter_to_plain
language_value
freeling_it
fc_freeling_text_2_conll_it
language_value_1
desr
Beanshell_2
apos_value
bpos_value
window_value
estrattore_mw
Workflow output ports
output


freeling_it (WSDL)
Freeling executable for Italian language only. Up to tagged
Run service
Inputs
Report
input
as URL
direct data or local file
Sfoglia...
piccoloMondoAntico.txt
language
it.cfg
mandatory
optional
multiword
yes
no
ner
none
output_format
--Use Default--
--Use Default--
token
splitted
tagged
Reset fields
Tagger = analisi morfosintattica = analisi grammaticale


freeling_it (WSDL)
Freeling executable for Italian language only. Up to tagged
Run service
RUNNING
Status updated every 2 seconds
Result Size Type
report 254 text
Update now
Terminate
Inputs
Report
input
as URL
direct data or local file
Sfoglia...
piccoloMondoAntico.txt
language
it.cfg
mandatory
optional
multiword
yes
no
ner
none
output_format
tagged
Reset fields

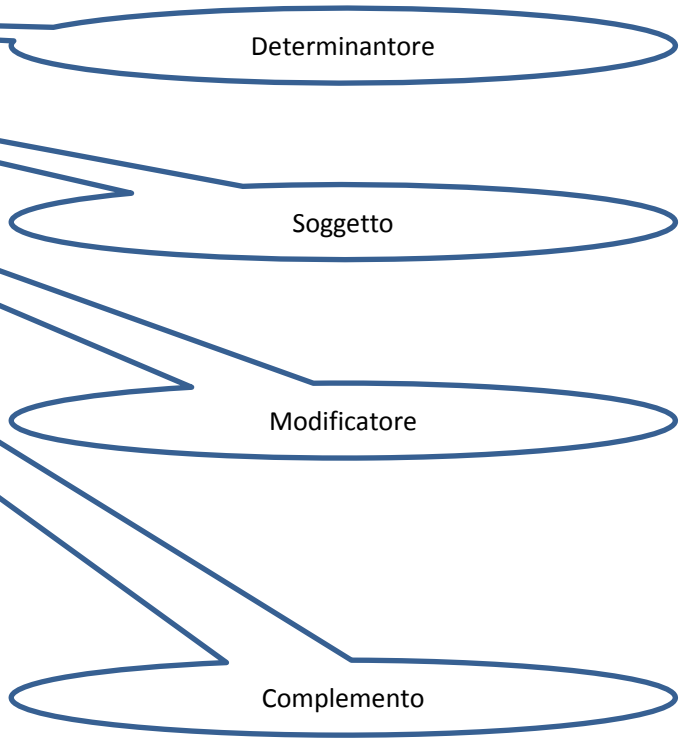

freeling_it (WSDL)
Freeling executable for Italian language only. Up to tagged
Run service
COMPLETED
Result Size Type
std_output 35 text
report 394 text
detailed_status 1 unknown
Remove
Inputs
input
as URL
direct data or local file
Sfoglia...
piccoloMondoAntico.txt
language
it.cfg
mandatory
optional
multiword
yes
no
ner
none
output_format
tagged
Reset fields
Report
Summary:
Completed: Successfully
Termination status: 0
Started: 2016-Aug-30 17:20:34 (CEST)
Ended: 2016-Aug-30 17:20:54 (CEST)
Duration: 0:00:20.052
Report:
Name: panacea.freeling_it
Job ID: [panacea.freeling_it]12d53536.15474e6e299._7fe5
Program and parameters:
/application/panacea/script/myfreeling.sh
-i
i_input
-f
it.cfg
--outf
tagged
--ner
none
--- end of parameters
Exit: 0

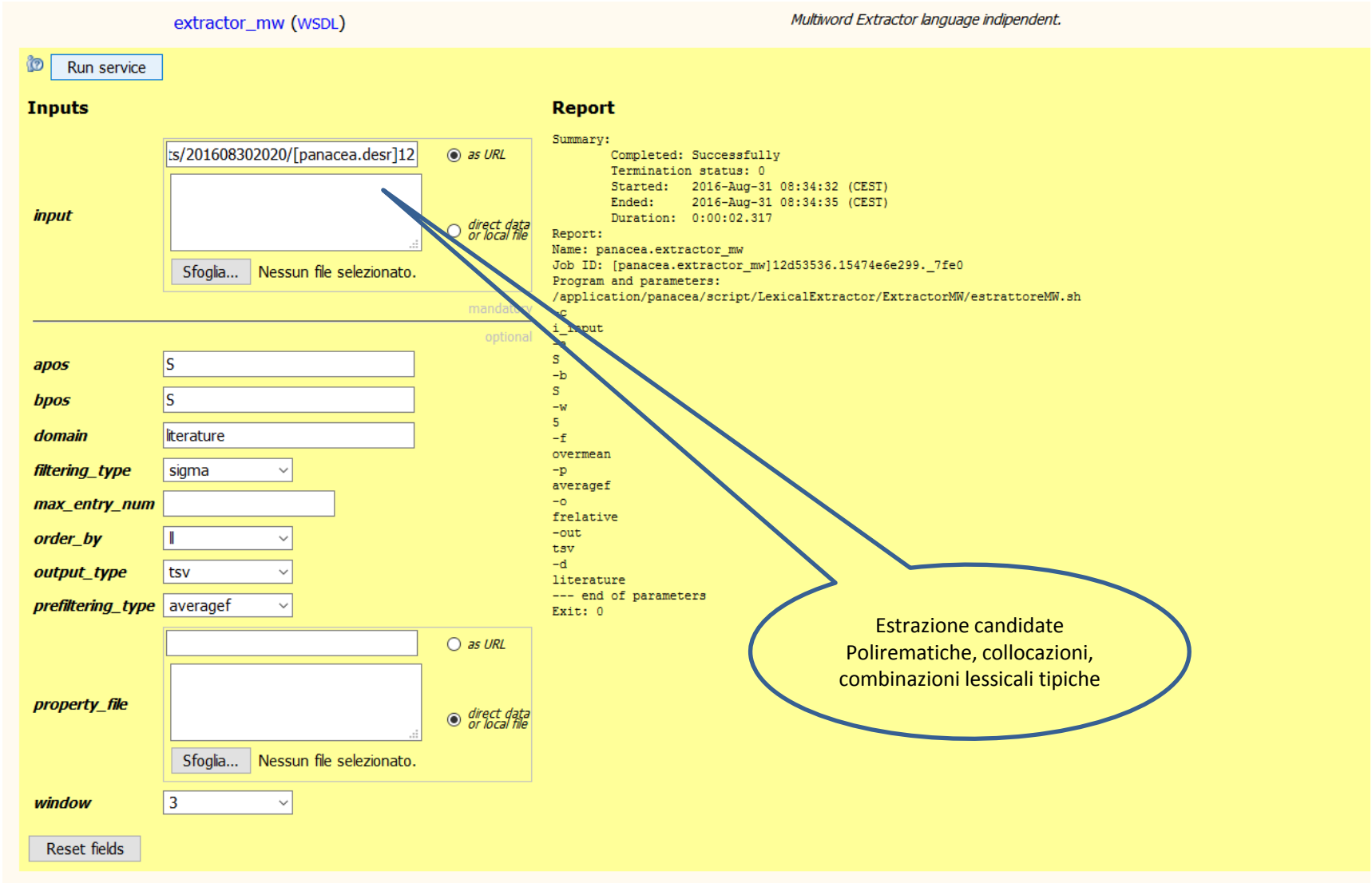

piccoloMondoAntico.conll.tanl
piccoloMondoAntico.desr
1 Le la R RD num=p|gen=f 807 809
2 onde onda S S num=p|gen=f _ _ _ _ 810 814
3 stramazzavano stramazzare V V num=p|per=3|mod=i|ten=i _ _ _ _ 815 828
4 tuonando tuonare V V mod=g _ _ _ 829 837
5 sulla sulla E EA num=s|gen=f _ _ _ 838 843
6 riva riva S S num=s|gen=f _ _ _ 844 848
7 , , F FF _ _ _ _ 848 849
8 sconquassavano sconquassare V V num=p|per=3|mod=i|ten=i _ _ _ _ 850 863
9 le la R RD num=p|gen=f _ _ _ 864 866
10 barche barca S S num=p|gen=f _ _ _ 867 873
11 incatenate incatenare V V num=p|mod=p|gen=f _ _ _ _ 874 884
12 , , F FF _ _ _ _ 884 885
13 mostravano mostrare V V num=p|per=3|mod=i|ten=i _ _ _ _ 886 896
14 qua qua B B _ _ _ _ 897 900
15 e e C CC _ _ _ _ 901 902
16 l l S SP _ _ _ _ 903 904
17 Il il R RD num=s|gen=m _ _ _ 905 907
18 piccolo piccolo A A num=s|gen=m _ _ _ 908 915
19 battello battello S S num=s|gen=m _ _ _ _ 916 924
20 nero nero A A num=s|gen=m _ _ _ 925 929
21 con con E E _ _ _ _ 930 933
22 i il R RD num=p|gen=m _ _ _ 934 935
23 cuscini cuscino S S num=p|gen=m _ _ _ 936 943
24 rossi rosso A A num=p|gen=m _ _ _ 944 949
25 , , F FF _ _ _ _ 949 950
26 la la R RD num=s|gen=f _ _ _ 951 953
27 tenda tenda S S num=s|gen=f _ _ _ 954 959
28 bianca bianco A A num=s|gen=f _ _ _ 960 966
29 e e C CC _ _ _ _ 967 968
30 rossa rosso A A num=s|gen=f _ _ _ 969 974
31 , , F FF _ _ _ _ 974 975
32 il il R RD num=s|gen=m _ _ _ 976 978
33 sedile sedile S S num=s|gen=m _ _ _ 979 985
34 posticcio posticcio A A num=s|gen=m _ _ _ _ 986 995
35 di di E E _ _ _ _ 996 998
36 parata parata S S num=s|gen=f _ _ _ 999 1005
37 piantato piantare V V num=s|mod=p|gen=m _ _ _ _ 1006 1014
38 a a E E _ _ _ _ 1015 1016
39 traverso traverso A A num=s|gen=m _ _ _ 1017 1025
40 , , F FF _ _ _ _ 1025 1026
41 i il R RD num=p|gen=m _ _ _ 1027 1028
42 remi remare V V num=s|per=3|mod=c|ten=p _ _ _ 1029 1033
43 pronti pronto A A num=p|gen=m _ _ _ 1034 1040
Articolo determinativo plurale femminile
sostantivo comune plurale femminile
Verbo stramazzare. 3° pers. Plur. Indicativo imperfetto.
Preposizione articolata singolare femminile
sostantivo comune singolare femminile


desr (WSDL)
DeSR is a shift-reduce dependency parser for Italian Language
Run service
COMPLETED
Result Size Type
output 35 text
report 467 text
detailed_status 1 unknown
Remove
Parser = analisi sintattica funzionale
Inputs
j_it]12d53536.15474e6e299--output
as URL
input
direct data or local file
Sfoglia...
Nessun file selezionato.
language
it.cfg
Reset fields
Report
Summary:
Completed: Maybe
Termination status: 0
Started: 2016-Aug-30 17:22:56 (CEST)
Ended: 2016-Aug-30 20:20:50 (CEST)
Duration: 2:57:54.017
Report:
Some error messages were reported.
Name: panacea.desr
Job ID: [panacea.desr]12d53536.15474e6e299._7fe3
Program and parameters:
/application/panacea/script/desr.sh
-i
i_input
-l
it.cfg
--- end of parameters
Exit: 0
Standard error stream:
LAS: 0.16 % (158/97953)
UAS: 4.59 % (4495/97953)
Sentences: 3841

piccoloMondoAntico.conll.tanl | piccoloMondoAntico.desr

```
121
122  1   Le  la  R   RD  num=p|gen=f 2   det _   _
123  2   onde    onda    S   S   num=p|gen=f 3   subj    _   _
124  3   stramazzavano   stramazzare V   V   num=p|per=3|mod=i|ten=i 0   ROOT    _   _
125  4   tuonando    tuonare V   V   mod=g   3   mod _   _
126  5   sulla   sulla   E   EA  num=s|gen=f 4   comp    _   _
127  6   riva    riva    S   S   num=s|gen=f 5   prep    _   _
128  7   ,   ,   F   FF  _   4   con _   _
129  8   sconquassavan   sconquassare    V   V   num=p|per=3|mod=i|ten=i 4   conj    _   _
130  9   le  la  R   RD  num=p|gen=f 10  det _   _
131  10  barche  barca   S   S   num=p|gen=f 8   obj _   _
132  11  incatenate  incatenare  V   V   num=p|mod=p|gen=f   10  mod _   _
133  12  ,   ,   F   FF  _   4   con _   _
134  13  mostravano  mostrare    V   V   num=p|per=3|mod=i|ten=i 4   conj    _   _
135  14  qua qua B   B   _   13  mod _   _
136  15  e   e   C   CC  _   14  con _   _
137  16  l   l   S   SP  _   14  conj    _   _
138  17  Il  il  R   RD  num=s|gen=m 19  det _   _
139  18  piccolo piccolo A   A   num=s|gen=m 19  mod _   _
140  19  battello    battello    S   S   num=s|gen=m 13  obj _   _
141  20  nero    nero    A   A   num=s|gen=m 19  mod _   _
142  21  con con E   E   _   19  comp    _   _
143  22  i   il  R   RD  num=p|gen=m 23  det _   _
144  23  cuscini cuscino S   S   num=p|gen=m 21  prep    _   _
145  24  rossi   rosso   A   A   num=p|gen=m 23  mod _   _
146  25  ,   ,   F   FF  _   23  con _   _
147  26  la  la  R   RD  num=s|gen=f 27  det _   _
148  27  tenda   tenda   S   S   num=s|gen=f 23  conj    _   _
149  28  bianca  bianco  A   A   num=s|gen=f 27  mod _   _
150  29  e   e   C   CC  _   28  con _   _
151  30  rossa   rosso   A   A   num=s|gen=f 28  conj    _   _
152  31  ,   ,   F   FF  _   33  punc    _   _
153  32  il  il  R   RD  num=s|gen=m 33  det _   _
154  33  sedile  sedile  S   S   num=s|gen=m 19  mod _   _
155  34  posticcio   posticcio   A   A   num=s|gen=m 33  mod _   _
156  35  di  di  E   E   _   33  comp    _   _
157  36  parata  parata  S   S   num=s|gen=f 35  prep    _   _
158  37  piantato    piantare    V   V   num=s|mod=p|gen=m   33  mod _   _
159  38  a   a   E   E   _   37  comp    _   _
160  39  traverso    traverso    A   A   num=s|gen=m 38  prep    _   _
161  40  ,   ,   F   FF  _   33  punc    _   _
162  41  i   il  R   RD  num=p|gen=m 42  det _   _
163  42  remi    remare  V   V   num=s|per=3|mod=c|ten=p 19  mod _   _
164  43  pronti  pronto  A   A   num=p|gen=m 42  pred    _   _
165  44  e   e   C   CC  _   43  con
```




extractor_mw (WSDL)
Multiword Extractor language indipendent.
Run service
Inputs
Report
:s/201608302020/[panacea.desr]12
as URL
input
direct data or local file
Sfoglia...
Nessun file selezionato.
mandatory
optional
apos
S
bpos
S
domain
literature
filtering_type
sigma
max_entry_num
order_by
ll
output_type
tsv
prefiltering_type
averagef
property_file
window
3
Reset fields
Summary:
Completed: Successfully
Termination status: 0
Started: 2016-Aug-31 08:34:32 (CEST)
Ended: 2016-Aug-31 08:34:35 (CEST)
Duration: 0:00:02.317
Report:
Name: panacea.extractor_mw
Job ID: [panacea.extractor_mw]12d53536.15474e6e299._7fe0
Program and parameters:
/application/panacea/script/LexicalExtractor/ExtractorMW/estrattoreMW.sh
-c
i_input
-a
S
-b
S
-w
5
-f
overmean
-p
averagef
-o
frelative
-out
tsv
-d
literature
--- end of parameters
Exit: 0
Estrazione candidate
Polirematiche, collocazioni,
combinazioni lessicali tipiche

Lista candidati, formato testo
a campi tabbati

SCALINATA<->POMODORO	7	3	3	scalinata<_>s<_>scalinata<#>del<_>ea<_>del<#>pomodoro<_>s<_>pomodoro	3	6,05E+11	28.502.960.859.441.900	4.089.834.346.343.500	3.0	0.0
ZIO<->PIERO	91	47	47	zio<_>s<_>zio<#>piero<_>sp<_>piero	47	0.009479628882613958	17.363.527.336.373.600	4.054.085.254.546.460	47.0	0.0
SIGNORE<->GIACOMO	155	94	92	signor<_>s<_>signore<#>giacomo<_>sp<_>giacomo	92	0.01855586930213796	14.957.224.015.340.100	6.871.612.148.997.050	92.0	0.0
LAGO<->COMO	36	15	12	lago<_>s<_>lago<#>di<_>e<_>di<#>como<_>sp<_>como	12	0.0024203307785397336	20.421.816.121.831.100	10.773.145.943.338.000	12.0	0.0
INGEGNERE<->CAPO	28	49	10	ingegnere<_>s<_>ingegnere<#>in<_>e<_>in<#>capo<_>s<_>capo	10	0.002016942315449778	15.580.400.145.877.200	58.301.419.682.787.500	10.0	0.0
PUNTA<->PIEDE	11	28	8	punta<_>s<_>punta<#>di<_>e<_>di<#>piedi<_>s<_>piede	7	0.0016135538523598225	21.099.333.964.499.500	7.248.638.770.208.500	4.0	3.0
CAVAL<->SPADA	4	4	4	caval<_>s<_>caval<#>di<_>e<_>di<#>spade<_>s<_>spada	4	8,07E+11	30.933.341.346.304.900	649.764.788.358.962	4.0	0.0
SBRIO<->BACO	3	3	3	sbrio<_>s<_>sbrio<#>baco<_>s<_>baco	3	6,05E+11	32.182.728.712.387.900	5.045.905.692.924.160	3.0	0.0
CORPO<->BACCO	10	3	3	corpo<_>s<_>corpo<#>di<_>e<_>di<#>bacco<_>s<_>bacco	3	6,05E+11	2.695.394.125.958.450	38.241.770.888.143.700	3.0	0.0
NOTTE<->NATALE	21	5	4	notte<_>s<_>notte<#>di<_>e<_>di<#>natale<_>s<_>natale	4	8,07E+11	22.762.648.182.164.700	3.952.899.644.174.750	4.0	0.0
BERRETTO<->NOTTE	5	21	4	berretto<_>s<_>berretto<#>da<_>e<_>da<#>notte<_>s<_>notte	4	8,07E+11	22.762.648.182.164.700	3.952.899.644.174.750	4.0	0.0
MONTE<->CARONA	21	5	3	monti<_>s<_>monte<#>di<_>e<_>di<#>carona<_>s<_>carona	3	6,05E+11	21.513.260.816.081.700	26.518.615.069.091.800	3.0	0.0
SIGNORA<->TERESA	87	6	4	signora<_>s<_>signora<#>teresa<_>s<_>teresa	4	8,07E+11	15.797.836.142.841.500	2.495.579.002.063.350	4.0	0.0
PICCO<->CRESSOGNO	5	18	5	picco<_>s<_>picco<#>di<_>e<_>di<#>cressogno<_>s<_>cressogno	5	0.001008471157724889	2.440.121.620.855.140	5.771.783.308.402.180	5.0	0.0
GIACOMO<->PUTTINO	94	17	10	giacomo<_>sp<_>giacomo<#>puttini<_>s<_>puttino	10	0.002016942315449778	14.918.173.509.804.800	57.597.848.165.573.300	10.0	0.0
FANTE<->BASTONE	4	6	4	fante<_>s<_>fante<#>di<_>e<_>di<#>bastoni<_>s<_>bastone	4	8,07E+11	29.172.428.755.748.100	5.733.830.881.635.840	4.0	0.0
CASA<->RIGEY	78	18	7	casa<_>s<_>casa<#>rigey<_>sp<_>rigey	7	0.0014118596208148447	13.931.250.581.789.200	35.027.872.634.426.100	7.0	0.0
SIGNORA<->LUISINA	87	7	5	signora<_>s<_>signora<#>luisina<_>sp<_>luisina	5	0.001008471157724889	16.097.468.376.615.900	32.407.493.133.605.000	5.0	0.0
STRADA<->BOGLIA	18	12	4	strada<_>s<_>strada<#>di<_>e<_>di<#>boglia<_>s<_>boglia	4	8,07E+11	19.630.003.661.354.800	3.067.603.228.027.610	4.0	0.0
PIANO<->PIANO	18	18	4	piano<_>s<_>piano<#>piano<_>s<_>piano	4	8,07E+11	17.869.091.070.798.000	26.916.984.836.685.700	4.0	0.0
MONTE<->BISGNAGO	21	5	3	monte<_>s<_>monte<#>bisgnago<_>sp<_>bisgnago	3	6,05E+11	21.513.260.816.081.700	26.518.615.069.091.800	3.0	0.0
GRAZIA<->DIO	6	23	3	grazia<_>s<_>grazia<#>di<_>e<_>di<#>dio<_>s<_>dio	3	6,05E+11	203.263.629.427.688	2.435.379.583.739.040	3.0	0.0
TAVOLINO<->NOTTE	21	21	4	tavolino<_>s<_>tavolino<#>da<_>e<_>da<#>notte<_>s<_>notte	4	8,07E+11	16.530.155.278.185.700	24.192.886.619.078.100	4.0	0.0
SIGNORE<->INGEGNERE	155	28	8	signor<_>s<_>signore<#>ingegnere<_>s<_>ingegnere	8	0.0016135538523598225	0.9609943834378872	23.557.859.471.776.900	8.0	0.0
COMMISSARIO<->PORLEZZA	19	8	3	commissario<_>s<_>commissario<#>di<_>e<_>di<#>porlezza<_>s<_>porlezza	3	6,05E+11	19.906.717.927.333.400	23.332.235.972.039.900	3.0	0.0
SIGNORE<->MAIRONI	155	49	10	signor<_>s<_>signore<#>maironi<_>sp<_>maironi	10	0.002016942315449778	0.8148663477596492	22.686.354.386.872.300	10.0	0.0
DON<->MAIRONI	69	49	7	don<_>s<_>don<#>franco<_>s<_>franco<#>maironi<_>sp<_>maironi	6	0.0014118596208148447	1.011.446.995.206.940	214.423.820.213.908	3.5	2.5
AMORE<->CIELO	13	18	3	amor<_>s<_>amore<#>del<_>ea<_>del<#>cielo<_>s<_>cielo	3	6,05E+11	18.032.995.232.679.700	20.254.243.802.554.100	3.0	0.0
MATTINA<->SERA	17	16	3	mattina<_>s<_>mattina<#>e<_>cc<_>e<#>sera<_>s<_>sera	3	6,05E+11	17.379.464.766.439.100	19.246.262.042.230.500	3.0	0.0
MARITO<->MOGLIE	26	20	3	marito<_>s<_>marito<#>e<_>cc<_>e<#>moglie<_>s<_>moglie	3	6,05E+11	14.565.120.370.433.100	15.112.541.527.899.200	3.0	0.0
LAGO<->LUGANO	36	21	3	lago<_>s<_>lago<#>di<_>e<_>di<#>lugano<_>sp<_>lugano	3	6,05E+11	12.939.935.851.769.100	12.822.980.021.727.000	3.0	0.0
SIGNORE<->CONTROLLORE	155	6	3	signor<_>s<_>signore<#>controllore<_>s<_>controllore	3	6,05E+11	12.040.324.321.241.800	12.717.852.762.741.300	3.0	0.0
PORTA<->CASA	13	78	3	porta<_>s<_>porta<#>di<_>e<_>di<#>casa<_>s<_>casa	3	6,05E+11	11.664.774.256.807.900	1.128.741.257.407.750	3.0	0.0
SIGNORE<->CARLO	155	8	3	signor<_>s<_>signore<#>carlo<_>sp<_>carlo	3	6,05E+11	10.790.936.955.158.800	10.575.335.174.695.600	3.0	0.0
SIGNORE<->COMMISSARIO	155	19	4	signor<_>s<_>signore<#>commissario<_>s<_>commissario	4	8,07E+11	0.8283688181632963	9.196.806.974.836.170	4.0	0.0
VISO<->LUISA	39	156	3	viso<_>s<_>viso<#>di<_>e<_>di<#>luisa<_>s<_>luisa	3	6,05E+11	0.3883261752971559	19.235.202.837.561.900	3.0	0.0
SIGNORE<->PASOTTI	155	63	4	signor<_>s<_>signore<#>pasotti<_>sp<_>pasotti	4	8,07E+11	0.30778186966254356	17.032.703.638.689.400	4.0	0.0

SampleOutputTSV.txt | SampleOutputLMF.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "http://www.taqmatica.fr/lmf/DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
    <GlobalInformation>
    <Lexicon>
        <feat att="type" val="Panacea_MWE_Lexicon"/>
        <feat att="language" val="Italian"/>
            <LexicalEntry id="le_0f3c15d0cdaa8cc69b14db8fe32888c5">
            <feat att="entryType" val="Multiword"/>
            <feat att="MWEPattern" val="s+ea+s"/>
            <feat att="absoluteFrequency" val="3"/>
            <feat att="logLikelihood" val="6.050826946349334e-4"/>
            <feat att="writtenform" val="scalinata del pomodoro"/>
            <feat att="lemmaPair" val="scalinata-pomodoro"/>
            <feat att="domain" val="fiction"/>
            <Lemma>
            </Lemma>
            <ListOfComponents>
                <Component entry="le_9a0cf9fb3aa296e903ef11fde65bd8c4">
                    <feat att="rank" val="0"/>
                    <feat att="pos" val="s"/>
                    <feat att="lemma" val="scalinata"/>
                    <feat att="writtenform" val="scalinata"/>
                    <feat att="function" val="head"/>
                </Component>
                <Component entry="le_e19232a3a4f95cba523f09df6e1c2b73">
                    <feat att="rank" val="1"/>
                    <feat att="pos" val="ea"/>
                    <feat att="lemma" val="del"/>
                    <feat att="writtenform" val="del"/>
                </Component>
                <Component entry="le_f6d677cb2e6cd36f1cf7c8249e8167ee">
                    <feat att="rank" val="2"/>
                    <feat att="pos" val="s"/>
                    <feat att="lemma" val="pomodoro"/>
                    <feat att="writtenform" val="pomodoro"/>
                </Component>
            </ListOfComponents>
        </LexicalEntry>
        <LexicalEntry id="le_bac9ab8c65c2284225104f85b8a000fc">
            <feat att="entryType" val="Multiword"/>
            <feat att="MWEPattern" val="s+sp"/>
            <feat att="absoluteFrequency" val="47"/>
            <feat att="logLikelihood" val="0.009479628882613958"/>
            <feat att="writtenform" val="zio piero"/>
            <feat att="lemmaPair" val="zio-piero"/>
            <feat att="domain" val="fiction"/>
            <Lemma>
            </Lemma>
            <ListOfComponents>
                <Component entry="le_de7a6562a25b91e311978c3e7b878151">
                    <feat att="rank" val="0"/>
                    <feat att="pos" val="s"/>
                    <feat att="lemma" val="zio"/>
                    <feat att="writtenform" val="zio"/>
                    <feat att="function" val="head"/>
                </Component>
```

Lessico XML LMF

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title> <note>Versione elettronica di testi relativi al periodo 800 - 900 </note> Maria Stuarda </title>
      <author>Vittorio Alfieri</author>
      <editor></editor>
    <respStmt>
    <resp>Conversione in formato XML - TEI  curata dall'Istituto di Linguistica Computazionale "Antonio Zampolli" CNR Pisa <date when="2014">2014</date></resp>
    <name></name>
   </respStmt>

    </titleStmt>
    <publicationStmt>
    <sourceDesc>
  </fileDesc>
  <encodingDesc>
  <profileDesc>
    <langUsage>
        <language ident="it">Italiano</language>
    </langUsage>
</profileDesc>
</teiHeader>
<text>

 <body> <head type="Titolo_Testo">Maria Stuarda</head>
<div type="Tragedia"> <head type="Titolo">Maria Stuarda </head>

<div type="Capitolo"> <head type="Titolo"> i </head><p>

<pb n="0013"></pb>
Lamorre Se udire il vero osi, o regina, io l' oso
A te recar, poichè il tuo popol fido
Mi tien da tanto; e poichè al soglio intorno
Non è chi voglia o ardisca dirlo. In seno
Fiamma, cui non son esca umani affetti,
Ma che tutta arde in Dio, libera io nutro.

</p><sp><speaker>Maria</speaker> <p> Non lieve impulso è la licenza vostra
<sp><speaker>Lamorré</speaker> <p>  A te sgradito,
<sp><speaker>Maria</speaker> <p>  Schiavo Arrigo, o nemico, a me? Che parli?
<sp><speaker>Lamorré</speaker> <p>  Ei, da te lungi,
Tuoi veri sensi interpretar mal puote;
E men tu i suoi.
  </p> </sp>
<sp><speaker>maria</speaker> <p>  Lungi da me chi 'l tiene?
<sp><speaker>Lamorré</speaker> <p>  Da te in bando lo tien fredda accoglienza,
<sp><speaker>lamorré</speaker> <p>  E che? nol desti
Or per anco all' oblio? Straniero vile, In soverchio poter salito, ei spiacque
 Al tuo consorte; e al popol tuo... </p> </sp>
<sp><speaker>Maria</speaker> <p>  Ma farsi
<sp><speaker>maria</speaker> <p>  Io tel credo: ma basta. Or deggio in breve
<sp><speaker>Maria</speaker> <p>  Del volgo cieco instigator mendaci,
D' empia setta ministri, udrò sempr' io
Il favellar vostro arrogante? - ah! questo
```

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title> <note>Versione elettronica di testi patrimonio dell'Istituto di Linguistica Computazionale </note> proemio </title>
      <author>Giordano Bruno</author>
      <editor></editor>

  <respStmt>
  <resp>Conversione in formato XML - TEI  curata dall'Istituto di Linguistica Computazionale "Antonio Zampolli" CNR Pisa <date when="2014">2014</date></resp>
  <name></name>
  </respStmt>

    </titleStmt>
    <publicationStmt>
    <sourceDesc>
      <p>
<title></title>

      </p>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
  <profileDesc>
    <langUsage>
      <language ident="it">Italiano</language>
    </langUsage>
</profileDesc>
</teiHeader>
<text>

 <body> <head type="Titolo_Testo">proemio</head>
<div type="Tragedia"> <head type="Titolo">proemio </head>

<div type="Capitolo"> <head type="Titolo">proemio </head><p>
 </div>
<div type="Capitolo"> <head type="Titolo">sonetto i v. 01 </head><p>

 <div type="Capitolo"> <head type="Titolo">sonetto ii v. 01 </head><p>
 <div type="Capitolo"> <head type="Titolo">dial. 1 </head><p>
 <div type="Capitolo"> <head type="Titolo">dial. i explicit </head><p>

Fine del primo dialogo.
</p></div>
 <div type="Capitolo"> <head type="Titolo">dial. 2 </head><p>
<pb n="0225"></pb>
</p>
<sp><speaker>Dicsono</speaker> <p>. Di grazia, maestro Poliinnio, e tu, Gervasio,
<sp><speaker>Poliinnio </speaker> <p> <hi rend="italic"> Fiat</hi> .
</p> </sp>
<sp><speaker>Gervasio </speaker> <p> Se costui, che è il <hi rend="italic"> magister</hi> , parla, senza dubio
<sp><speaker>Dicsono</speaker> <p> . Sí che dite, Teofilo, che ogni cosa, che non è
<sp><speaker>Teofilo </speaker> <p> Senza dubio e senza controversia alcuna.
<sp><speaker>Dicsono</speaker> <p> . Credete per questo, che chi conosce le cose
<sp><speaker>Teofilo</speaker> <p> . Non facilmente la causa prossima e principio
<sp><speaker>Dicsono</speaker> <p>  Or come intendete che le cose, che hanno causa
```

CLARIN: networking



- Another important activity of CLARIN
  - user involvement
  - facilitation of collaborations
  - dissemination
  - education to the use of resources and services
  - organization of events, seminars
  - promotion of exchanges

CLARIN: networking



- CLARIN Annual Conference (Aix-en-Provence, 26-28/10/2016)
- CLARIN PLUS Workshops
- Mobillity grants
- Partecipation in research projects
- Interaction with other infrastructures and networks (DARIAH, …)

CLARIN-IT: survey



- Collecting requirements and desiderata of the community
- Go to the survey

CLARIN: searching for data

Home > Services

Virtual Language Observatory

From here you can explore the world of language resources and technology fom different perspectives:



Go to the Virtual Language Observatory, the CLARIN Language Resource facet browser (with CMDI and OLAC metadata)

- https://vlo.clarin.eu
- Demo

Summarizing



- Italy joined CLARIN
- My message today is about the aims and functioning of CLARIN, its potentialities, what is already in place, what was planned, how it can help advancing the sector
- Through CLARIN you can
  - appropriately document your resources
  - give them international visibility
  - find language resources and combine them with tools for further analysis
  - be in touch with scholars in the field
  - be informed about relevant events
  - apply for mobility grant to visit colleagues for specific joined research
  - be invited at CLARIN related events on strategic themes
  - take part to the CLARIN Annual Conference
- For the research community it is a great opportunity to let CLARIN-IT grow on the basis of concrete needs and expectations

CLARIN-IT: contacts



- Web: http://www.clarin-it.it
- Twitter: @CLARIN_IT
- eMail: coordination@clarin-it.it



Thank you for your attention!