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

Parthenos wp3 meeting Roma

1st October 2015

- Italy became a Full Member of **CLARIN ERIC**, the **C**ommon **La**nguage **R**esources and Technology **In**frastructure at the European level



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

What is meant by Infrastructure and RI

- Infrastructure is the basic physical systems of a nation; transportation, water and electric systems



- … but also transimission media, networks for **communication,** for moving data…



- A Research Infrastructure is a set of resources, tools, data, services and functions but also **competencies**, dedicated to the needs of a scientific and technological research community



CLARIN: legal frameworks

CLARIN is

- an infrastructure of **ERIC** type, has completed la **ESFRI** roadmap and has now achieved the status of «land-mark»

- **ESFRI** is a body created in 2002 to support a coherent and strategy-led approach to policy-making on research infrastructures to facilitate the better use and development, at EU and international level.
- 2006: publication of a **roadmap**

- **ERIC** is a legal framework born in 2009 with the specific aim to facilitate the establishment and operation of RI overcoming barriers arising from different legal systems in the various countries



Current situation and needs

- Language resources are often “closed”
  - hosted by the institution that has produced them
  - intended for the scientific purposes of a project
  - subject to restrictions





- Building language resources is time and money consuming
  - it is crucial to ensure easy access to resources
  - it is fundamental to ensure that such resources remain accessible over time



- A change of paradigm is essential
  - sharing, access and replicability of results



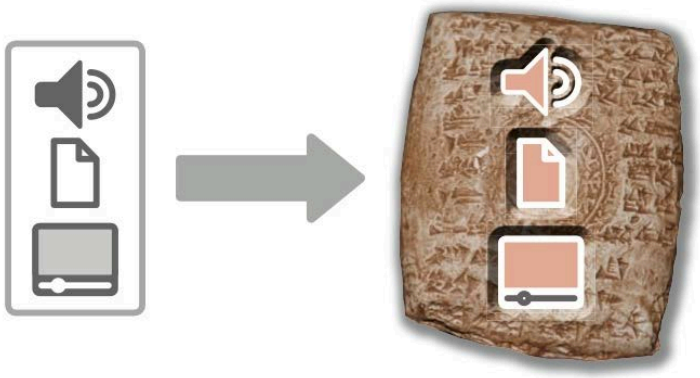
CLARIN: the solution



- CLARIN is a virtual digital infrastructure aiming at providing
  1. a system for the **deposit, description**, **cataloguing** and long term archiving of language resources
  2. a service for easy **discovering** of language resources
  3. an easy and sustainable **access** to scholars of member countries
  4. advanced tools and services for **combining** them in order to perform computational tasks
  5. Functionalities for creating virtual collections of digital objects of interest for scholars
  6. Advanced search functions across collection of data

1. CLARIN: depositing services

- **Long-term archiving**: a storage guarantee can be given for a long period
- **Persistent identifiers:** resources can be cited easily.
- **Search and enrich:** once integrated into the infrastructure, it is possibe to easily combine and analyze them with various **linguistic tools**
- **Institutional login:** password-protected resources can be made available



2. CLARIN: virtual language observatory

- **Discover** language resources with a facet browser or a map

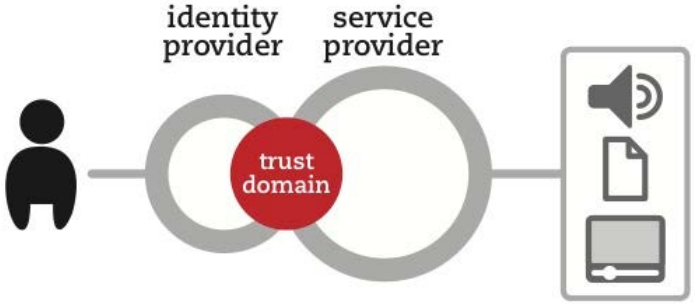


3. CLARIN: federated login

- **Federated login**: easy access to protected resources, with your institutional username and password
- Applications and data sets are available to anyone **with academic credentials** from many European countries.
- Accessing CLARIN applications becomes as **easy and simple** as logging in your university's webmail system

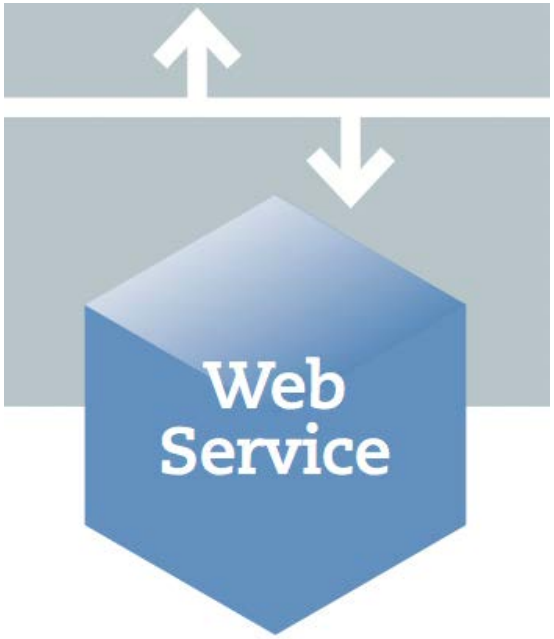


- There is a protocol of Authorization, Authentication, Identification that matches the credentials of the user ("Identity Providers") with those required by the service ("Service Providers")



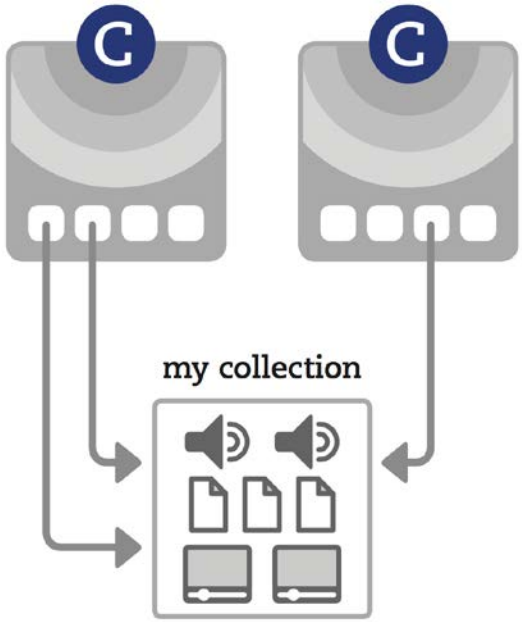
4. CLARIN: web services

- **Web services** available for users: e.g.
- Weblicht: execution environment for automatic annotation of text corpora. Linguistic tools are encapsulated as web services, which can be combined by the user into custom processing chains. The results can then be visualized in a table or tree format.
- CLARIN-DK offers automated creation and execution of workflows.
- TTNWW - a webservice workflow environment to process spoken and written Dutch.
-



5. CLARIN: virtual collections

- Coherent set of links to digital objects that can be easily created, accessed and cited by scholars. The links can originate from different archives,).
- CLARIN provides a registry where scholars can create and publish their virtual collections



- Some examples:
  - *data as mentioned in an article's footnotes gathered in a single virtual collection*
  - *a virtual collection with links to video or sound recordings illustrating a book*

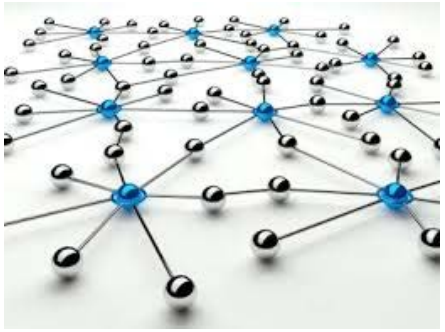
6. CLARIN: federated content search

- **Federated content search** enables researchers to search for specific patterns across collections of data
- **search engine** that connects to the local data collections that are available in the centres.



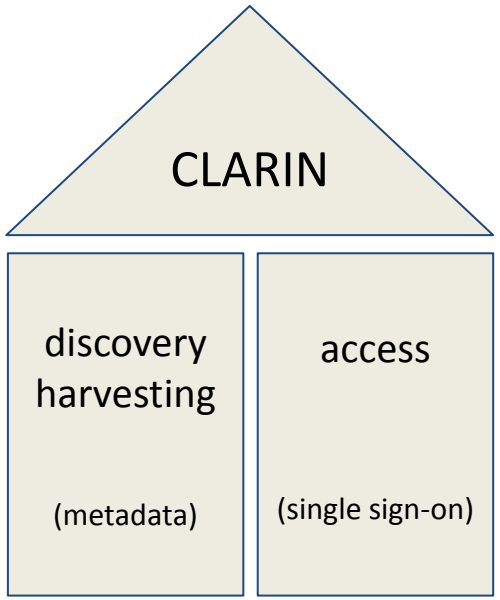
How a CLARIN National Consortium works

- CLARIN National Consortia include various institutions (research centres, universities, libraries, institutions supporting research, …) that
  - offer in-kind services for the maintenance of the consortium
  - make resources and services available through a **CLARIN Centre** (or more)
  - are financially supported by through national strategic projects
- CLARIN is a network of centres that are the pillars of the infrastructure



CLARIN ERIC: centre pillars

- Fundamental pillars of centres are
  - **Documentation/cataloguing**
    - language resources have to be documented with metadata (CMDI)
  - **exposition** of metadata
  - **harvesting** in a **central meta-catalogue** (VLO)
  - **access** to language resources
    1. **totally free**
    2. **through a single sign-on unified authentication system** (particularly important for all those language resources that are available for academic purposes only)




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

CLARIN National Consortia



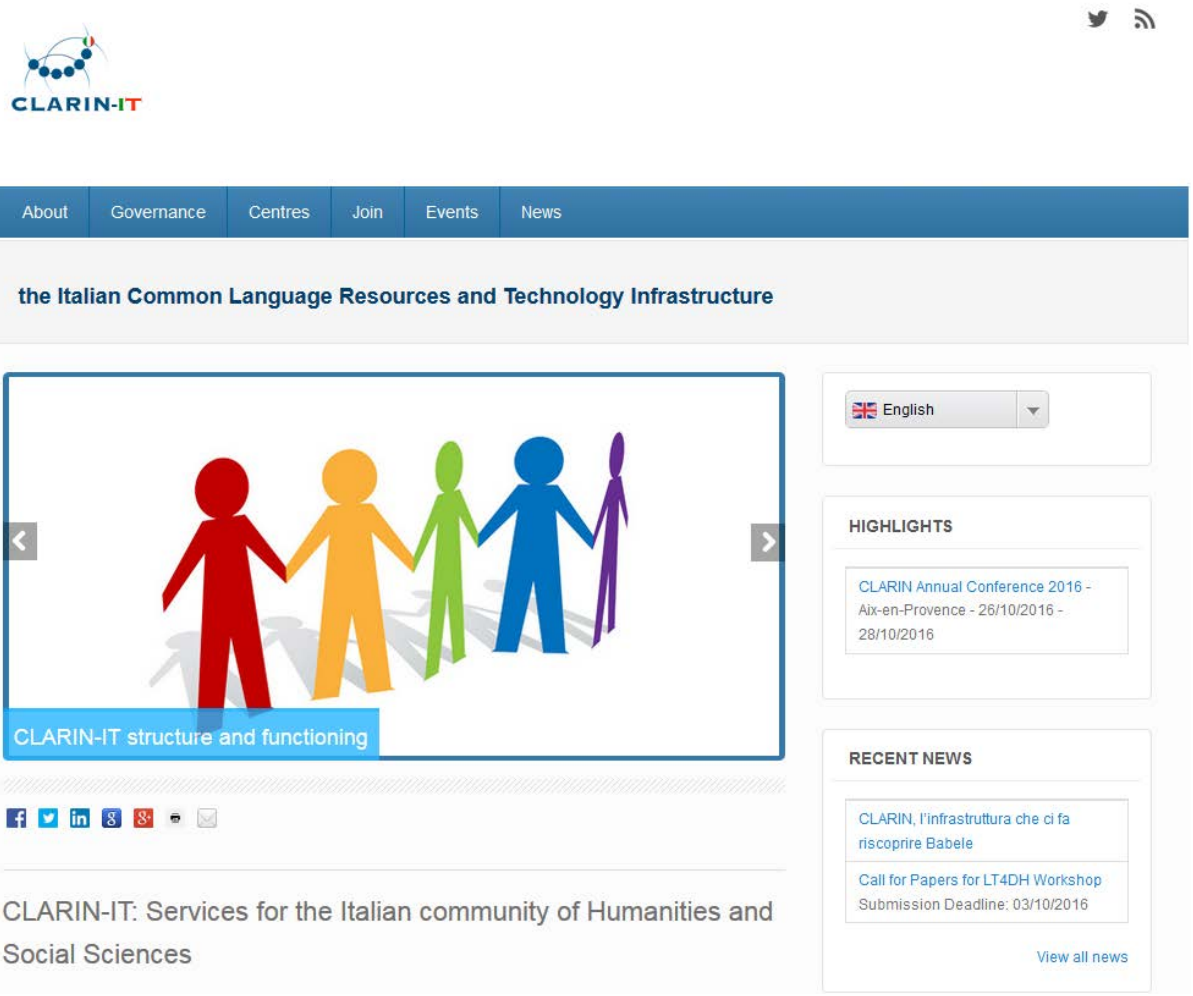
- CLARIN ERIC
  - does not impose a model of governance to its various National Consortia
  - requires
    - the creation of a community by involving the main actors of the research sector
    - the implementation of a centre (or more) to make various types of data, tools and services available
    -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

CLARIN-IT: who does what



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

CLARIN-IT: Web portal

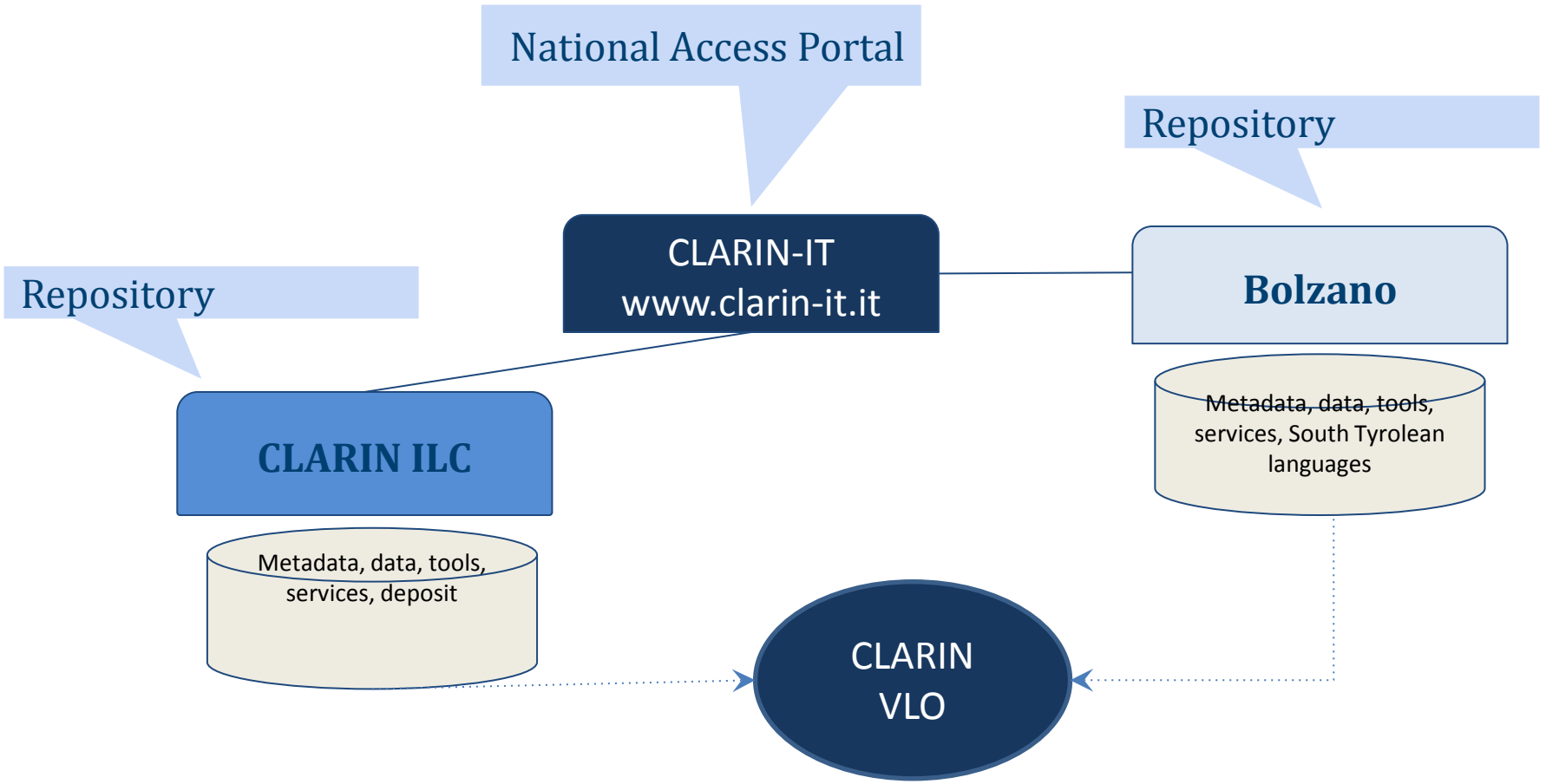


CLARIN-IT: Consortium



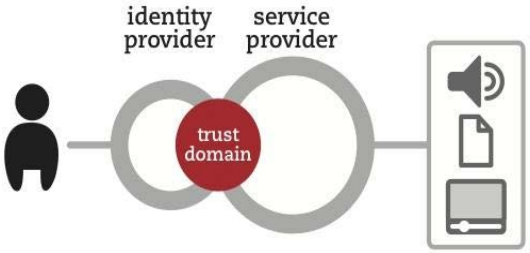
- **IDEM-GARR** Federated Identity
- Institutions that have already joined or are about to join CLARIN-IT
  - **Siena University** (oral archives)
  - **Scuola Normale Superiore** (oral archives)
  - **EURAC Bozen** (multilingualism and regional languages)
  - **FBK Trento** (language resources)
  - **Cattolica Milano Univ.** (services for Latin)
  - **Venezia Univ.** (thematic annotation in DH)
  - **Pisa Univ;** (NLP lab)
  - **Parma Univ.** (tools for Ancient Greek)
  - **Roma Univ.** (ontologies for DH)

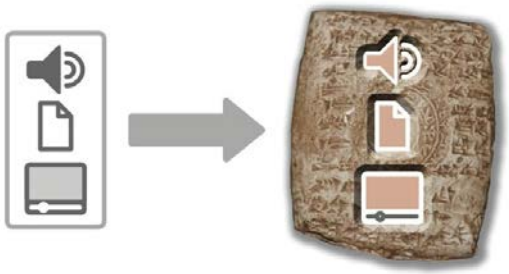
CLARIN-IT: structure

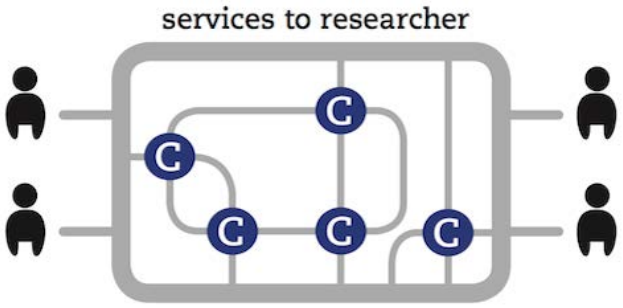


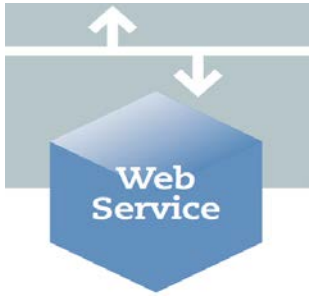
CLARIN ILC centre

- National **collector** of data and tools
- Italian language and classical languages
- **Access** to language resources and services of our Institute
- **Single sign-on**
- **Depositing** services by third parties: guarantees availability, citation, usability.
- Long term **preservation**
- **Harvesting**: our data are searchable in the central VLO and gain international visibility



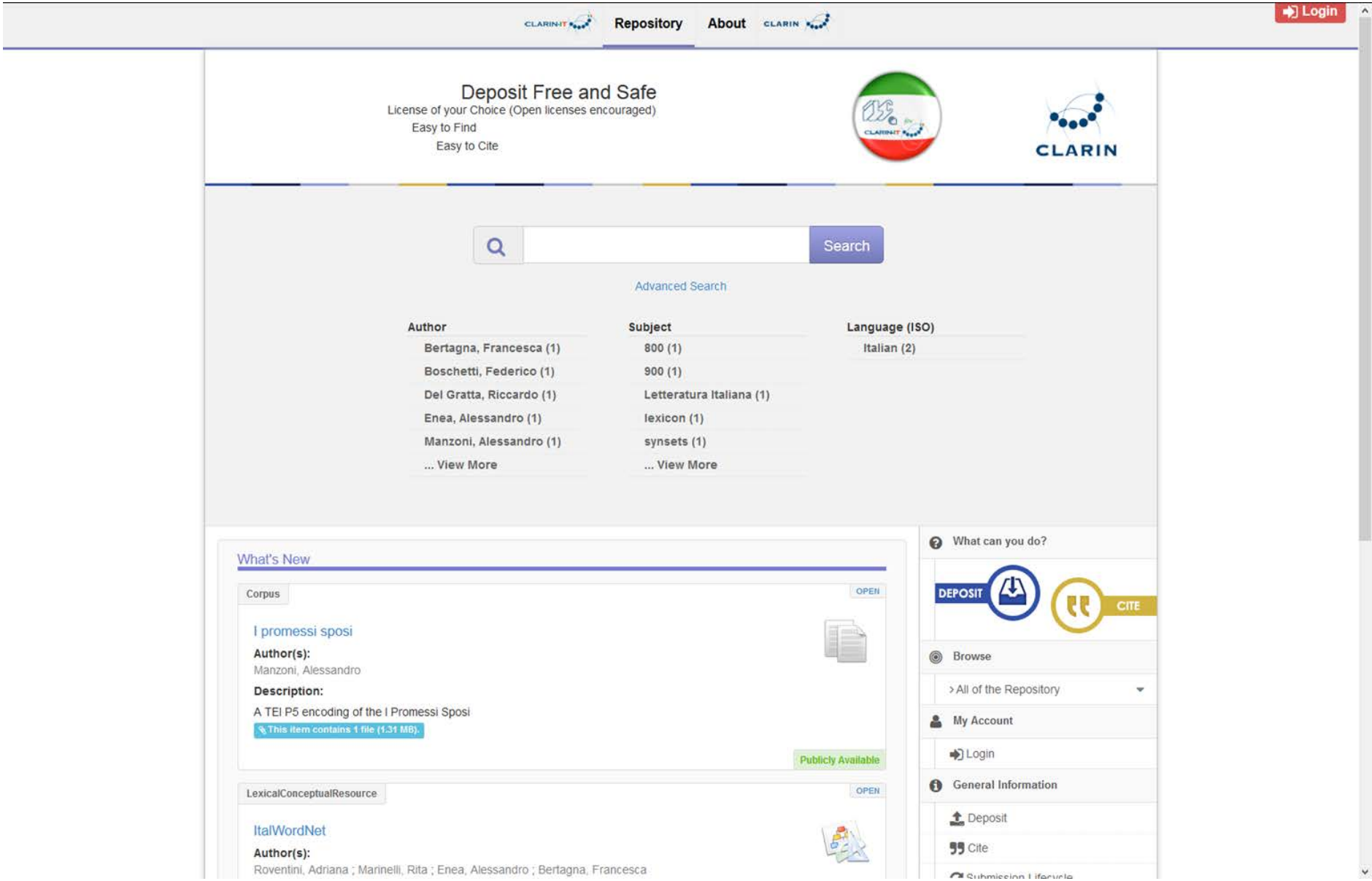




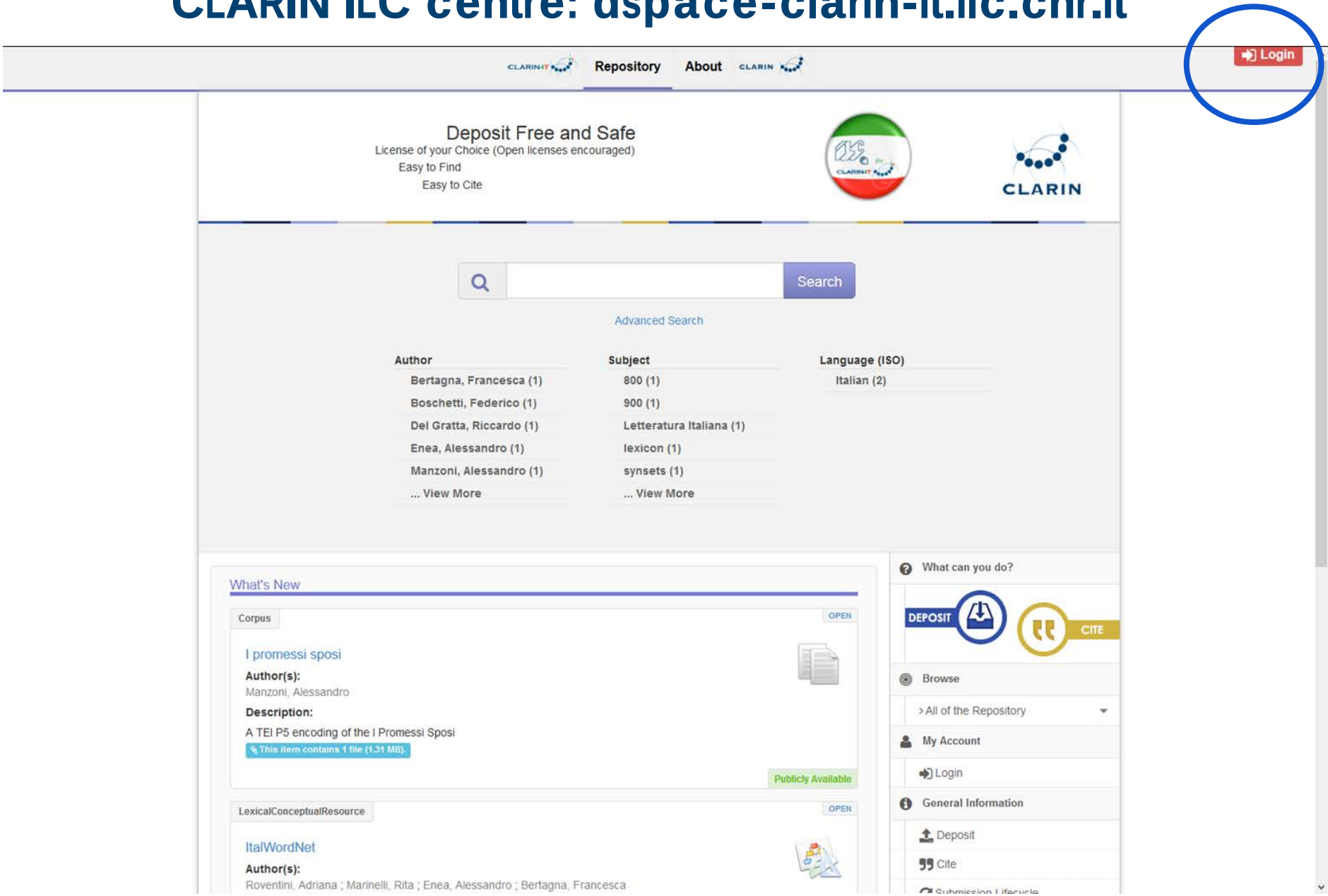




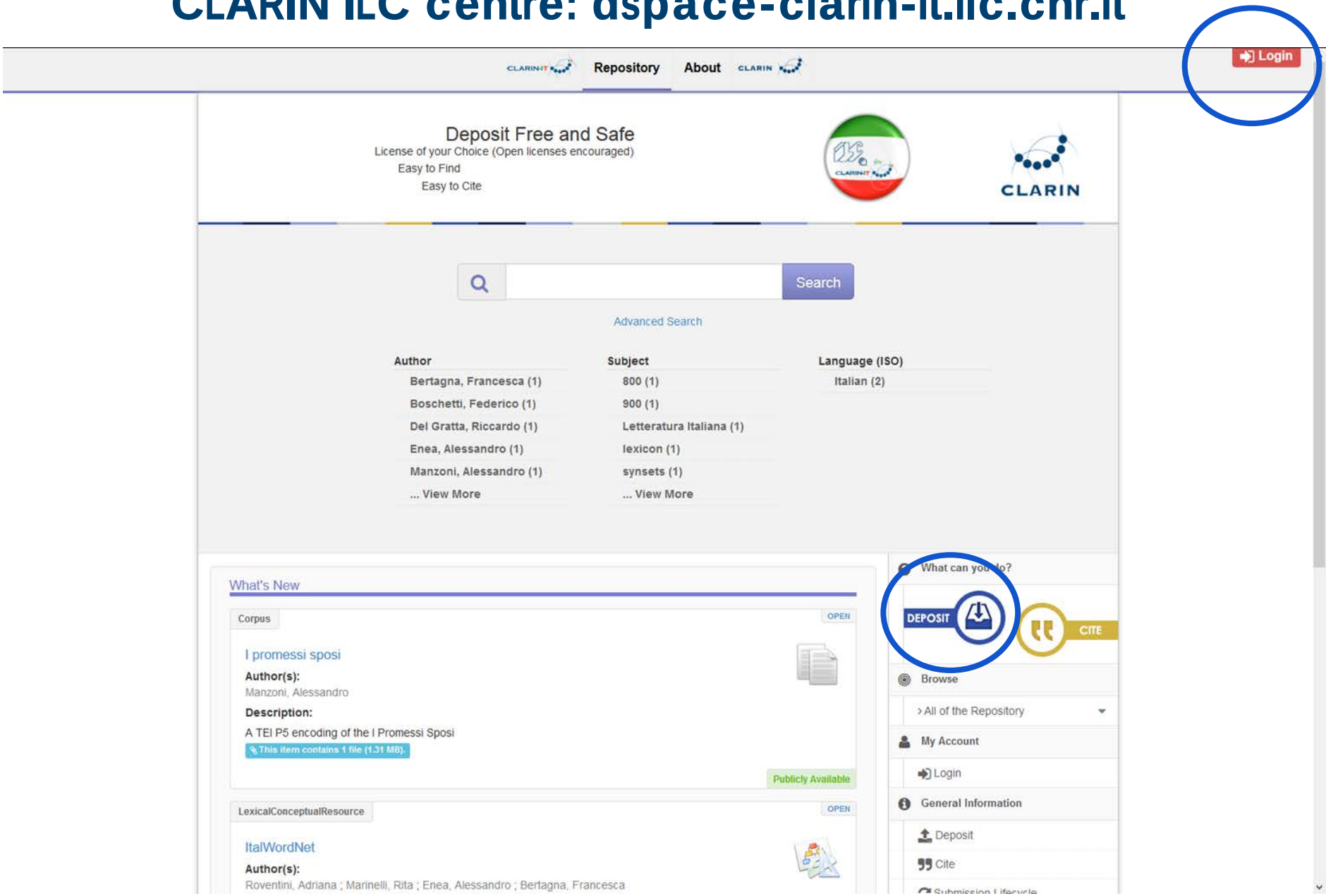
CLARIN ILC centre: https://dspace-clarin-it.ilc.cnr.it



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



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



FAQ
About
Help Desk
Alessandro Enea | Logout

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.

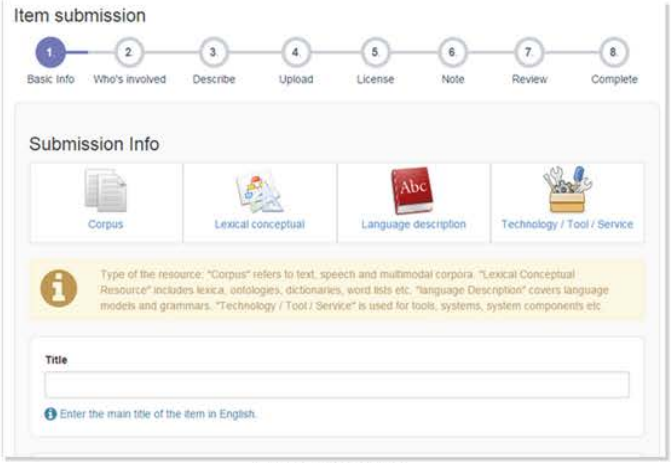
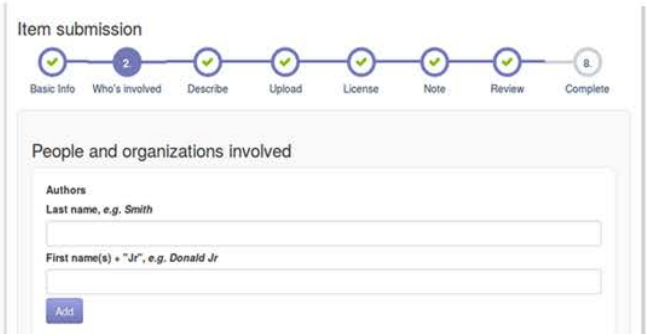


Fig4. Submission info

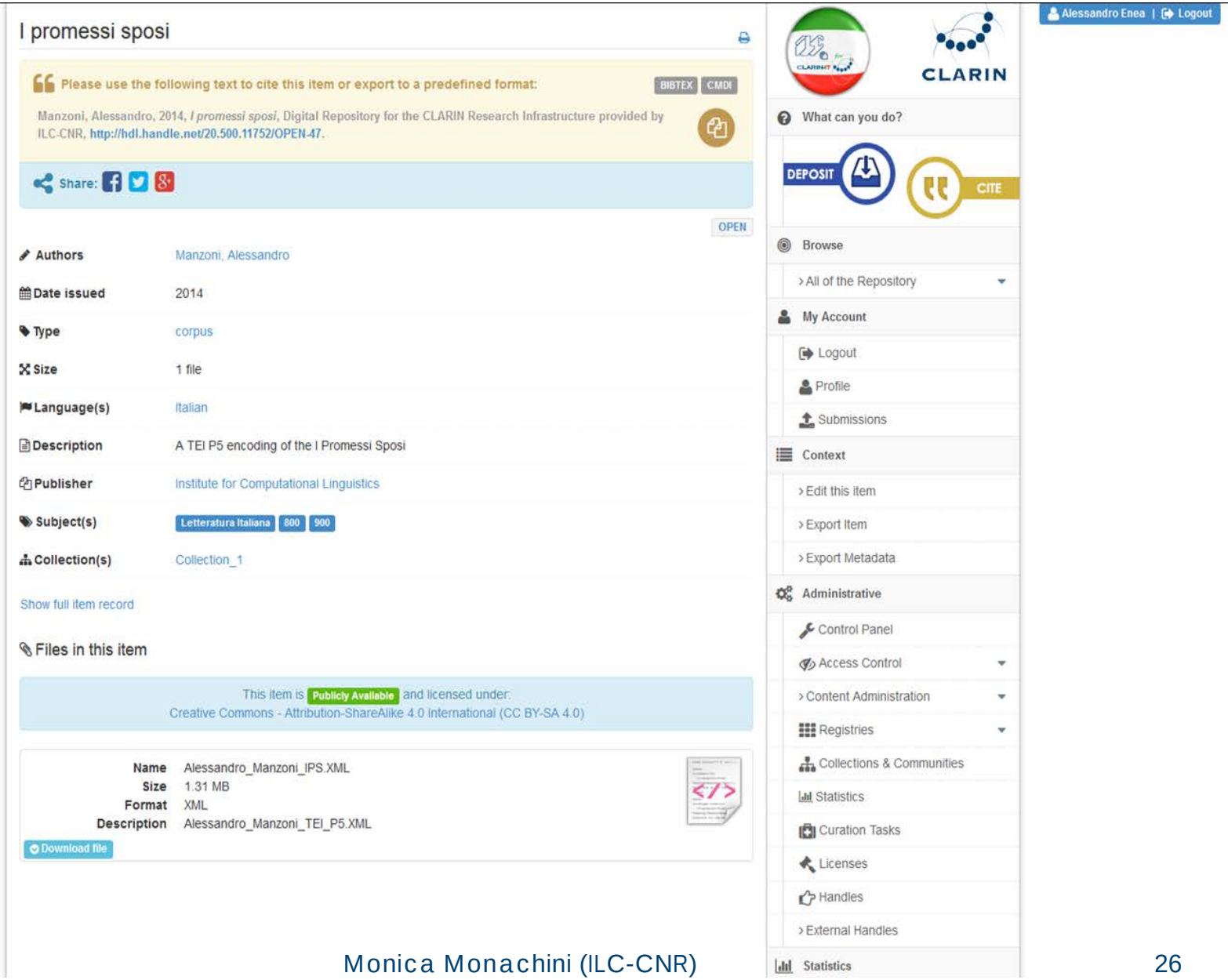
Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as the title. Click Next to continue the following step.

Step 4: Describe your item

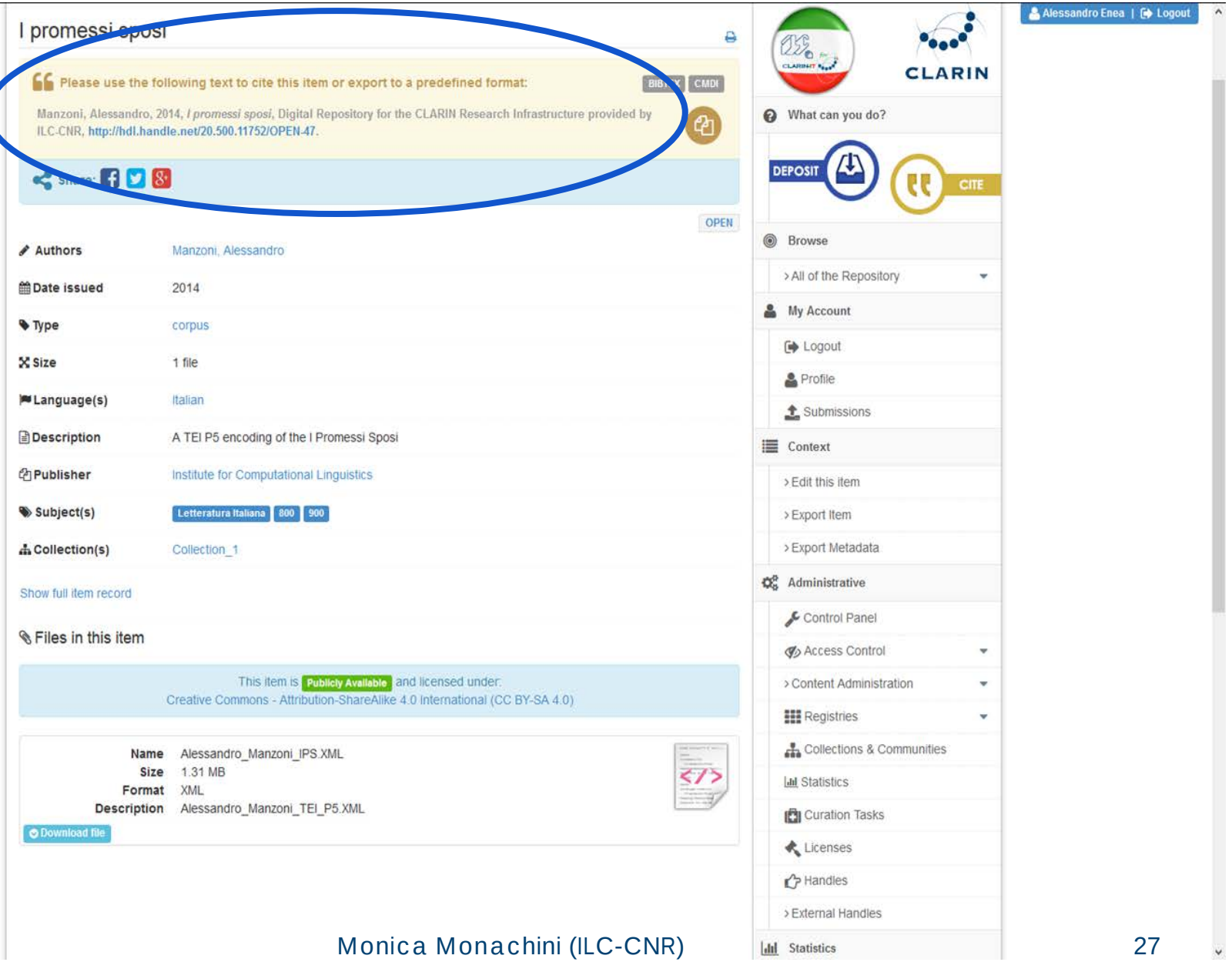
In the following two steps you will provide more details for your item. First describe the people, organization and projects involved with the item.



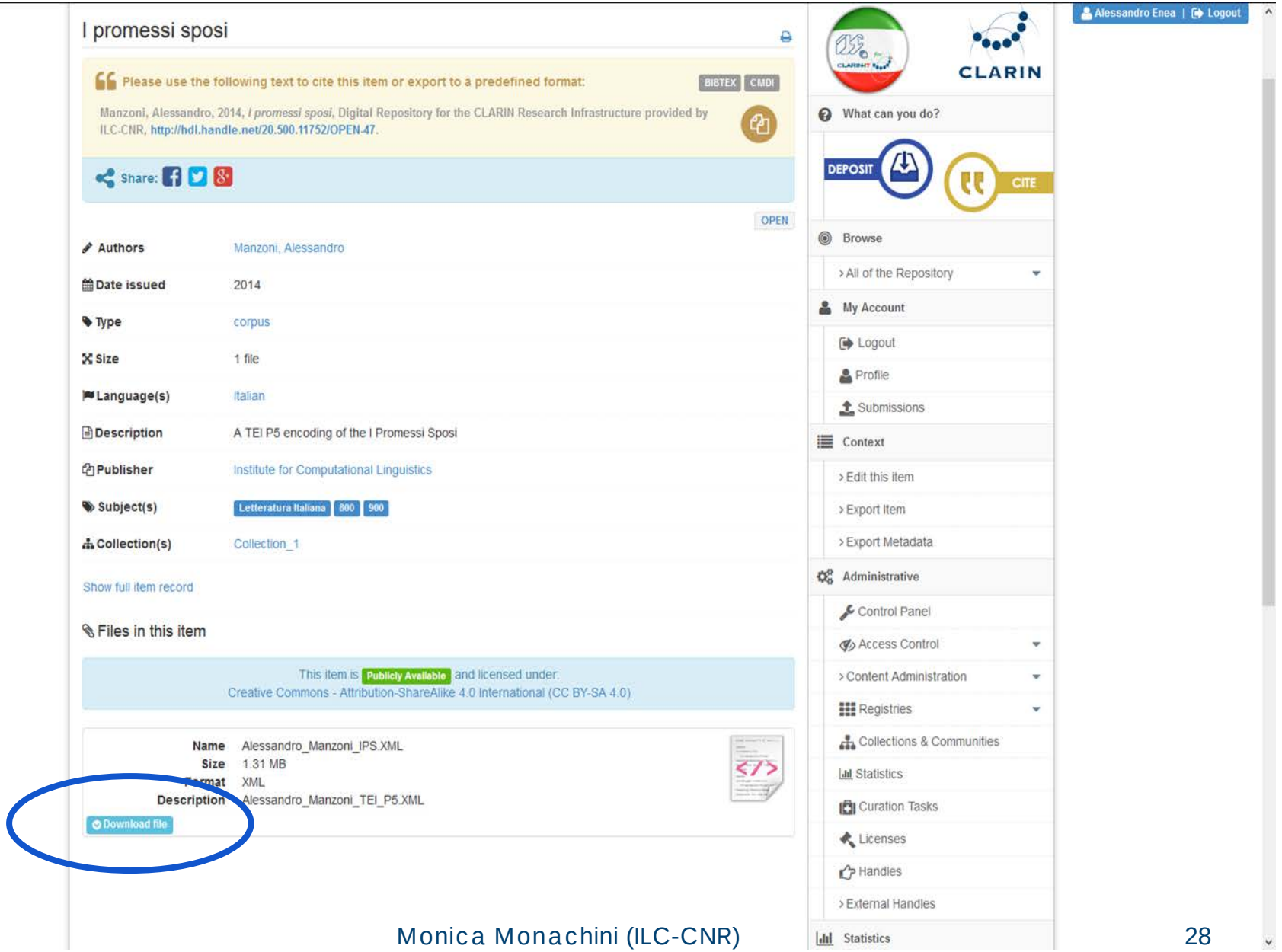
CLARIN ILC centre: dspace-clarin-it.ilc.cnr.it



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



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



CLARIN Licences



- Three main licence categories are available in CLARIN to deposit Language Resources:
  - CLARIN PUB public use
  - CLARIN ACA academic use
  - CLARIN RES restricted use
- In addition, there are several minor categories with more specific variants of the general licenses.

PUB language resources



- PUB language resources are not restricted by copyright or personal data protection issues; distributed publicly

PUB licences and CLARIN cats.

- ***Public domain***
- CLARIN PUB CC-ZERO
- CLARIN PUB Public Domain
- CLARIN PUB Princeton Wordnet

PUB licences and CLARIN cats.

- ***Software***
- CLARIN PUB+BY ApacheLicense 2.0
- CLARIN PUB+BY BSD 2 or BSD 3
- CLARIN PUB+BY MIT license
- CLARIN PUB+BY Microsoft Public License (MS-PL)
- CLARIN PUB+BY LGPL
- CLARIN PUB+BY+SA AGPL
- CLARIN PUB+BY+SA GFDL
- CLARIN PUB+BY+SA Open Data Commons Open Database
- CLARIN PUB+BY+SA Microsoft Reciprocal License (MS-RL)
- CLARIN PUB+BY+SAGPL

PUB licences and CLARIN cats.

- ***Other works***
- CLARIN PUB+BY
- CC-BY
- CLARIN PUB+BY+SA
- CC-BY-SA
- CLARIN PUB+BY+NC
- CC-BY-NC
- CLARIN PUB+BY+NC+SA
- CC-BY-NC-SA
- CLARIN PUB+ID+BY+LRT+NORED
- META-SHARE Commercial No Redistribution
- CLARIN PUB+ID+BY+NC+LRT+NORED
- META-SHARE Noncommercial No Redistribution

ACA language resources



- **ACA language resources** can be accessed only for research purposes.
- The end-user does not need to ask for usage permission but can access the resources via e.g. federated login.

ACA licences and CLARIN cats.

- CLARIN ACA+BY+NORED CLARIN ACA
- CLARIN ACA+BY+NC+NORED CLARIN ACA-NC
- CLARIN ACA+META+BY META-SHARE Commons-BY
- CLARIN ACA+META+BY+NCMETA-SHARE Commons-BY-NC
- CLARIN ACA+META+BY+NC+SAMETA-SHARE Commons-BY-NC-SA
- CLARIN ACA+META+BY+SAMETA-SHARE Commons-BY-SA

RES language resources



- have additional restrictions, which require permission from the rights holder.
- These resources may contain material whose usage is limited due to copyright and/or personal data protection issues.
- These language resources require both using federated login to authenticate the end-user and sending a separate application to the rights holder for authorization.

RES licences and CLARIN cats.

- CLARIN RES+BY+NORED CLARIN RES
- CLARIN RES+BY+NC+NORED CLARIN RES-NC
- CLARIN RES+FF+BY+LRT+NOREDMETA-SHARE Commercial No Redistribution For a Fee
- CLARIN RES+FF+BY+NC+LRT+NOREDMETA-SHARE Noncommercial No Redistribution For a Fee
- CLARIN RES+FF+BY+LOC+LRT+NORED+* ELRA licenses

CLARIN License Calculator

Identification and Access conditions		
	Does the user need to be authenticated, i.e. identified?	◯Yes ◯No
	Does the user need to be affiliated with some specific community, e.g. university researchers (EDU) or more generally language resource and technology researchers (META)?	◯EDU ◯META ◯No
	Can the user only be given permission to use the resource on a case-by-case basis, e.g., based on a mandatory fee or a research plan?	◯Yes ◯No
FF	Is a fee required to get access to the resource?	◯Yes ◯No
PLAN	Does the right holder require a research plan for granting access?	◯Yes ◯No
General use conditions		
BY	Is attribution, i.e. acknowledgement of authorship, required?	◯Yes ◯No
NC	Is the content available only for non-commercial purposes?	◯Yes ◯No
INF	Is informing the rights owner about the use of the resource required?	◯Yes ◯No
LOC	Is the content available only at a single location, center, or site?	◯Yes ◯No
LRT	Is the content available only for language research and technology development?	◯Yes ◯No
PRIV	Are there personal data in the resource?	◯Yes ◯No
Distribution conditions		
NORED	Can the user distribute the original resource to third parties?	◯Yes ◯No
ND	Can the user distribute derived works, i.e. works containing copyrighted parts of the original?	◯Yes ◯No
DEP	If the user cannot distribute derived works, is the user still allowed to distribute modified versions via CLARIN?	◯Yes ◯No
SA	If the user can distribute derived works, should the same license be used, i.e. is the license reciprocal?	◯Yes ◯No
Other conditions		
*	Are there other non-standard conditions in the license that the user should pay attention to?	◯Yes ◯No

Thank you for your attention!

