# Italian Language Resources. From CLARIN-IT to the VLO and Back: Sketching a Methodology for Monitoring LRs Visibility

**Dario Del Fante**
ILC-CNR - Italy
dario.delfante@ilc.cnr.it

**Francesca Frontini**
ILC-CNR - Italy & CLARIN ERIC
francesca.frontini@ilc.cnr.it

**Monica Monachini**
ILC-CNR - Italy
monica.monachini@ilc.cnr.it

**Valeria Quochi**
ILC-CNR - Italy
valeria.quochi@ilc.cnr.it

## Abstract

This paper sketches a user-oriented, qualitative methodology for both (i) monitoring the existence and availability of language resources relevant for a given CLARIN national community and language and (ii) assessing the offering potential of CLARIN, in terms of Language Resources provided to national consortia. From the user perspective, the methodology has been applied to investigate the visibility of language resources available for Italian within the CLARIN central services, in particular the Virtual Language Observatory. As a proof-of-concept, the methodology has been tested on the resources available through the CLARIN-IT data centres, but, ideally, it could be applied by any national data centre aiming to assess the existence of LRs in CLARIN for any given languages and check their accessibility for the interested users. We thus argue that such an assessment might be a useful instrument in the hands of national coordinators and centre managers for (i) bringing to the fore both strengths and critical issues about their data providing community and (ii) for planning targeted actions to improve and increase both visibility and accessibility of their LRs.

## 1 Introduction

With a distributed network of over 70 centres, CLARIN ERIC's principal aim is to ensure easy access to language resources and tools by researchers from all over Europe and beyond, independently of the original producers, and of the centre or consortium physically hosting them. Therefore, a lot of effort has always been put by CLARIN ERIC into developing and operating central functionalities that would serve this key purpose, to the point that today one does not need to know where a given resource is deposited or even be aware of its existence to be able to find, access and use it. This is achieved also thanks to the CLARIN portal, which acts as a gateway to the whole network's offerings.

The first and foremost central service, the CLARIN virtual *shop window*, is the Virtual Language Observatory, the VLO, (Uytvanck et al., 2010)[1], which makes language resources (LRs) searchable via a unified interface offering faceted search, on the basis of common standardised metadata descriptions. The VLO harvests metadata from all of the official CLARIN data providing centres, as well as from other affiliated catalogues and repositories, e.g. the Europeana catalogue (Eskevich et al., 2017)[2]. Other interesting and useful central discovery services are the Federated Content Search (FCS)[3], the Language Resources Switchboard (SB)[4], and the CLARIN Resource Families[5]. Visibility and usability are directly proportional not only to the quality of the data itself but also, more importantly, of its metadata descriptions. While the central services offer key discovery and data inspection functionalities, because of the



[1] https://vlo.clarin.eu

[2] See also https://pro.europeana.eu/post/clarin-and-europeana-make-discovery-and-processing-quick-and-easy-for-135-000-cultural-heritage-objects

[3] https://contentsearch.clarin.eu/

[4] https://switchboard.clarin.eu/

[5] https://www.clarin.eu/resource-families

Dario Del Fante, Francesca Frontini, Monica Monachini and Valeria Quochi 2022. Italian Language Resources. From CLARIN-IT to the VLO and Back: Sketching a Methodology for Monitoring LRs Visibility. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 10–22. DOI: https://doi.org/10.3384/9789179294441

distributed nature of the CLARIN infrastructure, the responsibility of the quality of both data and metadata descriptors ultimately lies within the official repositories and data providing centres of each national consortium and partner.

In this paper we argue that, in order to maximise the visibility of LRs within the CLARIN central services, a good practice for national consortia or national data centres would be to regularly monitor these four "points of access" and analyse how the language resources hosted at their centres or relevant for their research community show up from a user perspective. The activities we propose are complementary to the automated metadata checks that centres can carry out thanks to the CLARIN curation dashboard[6]. What we suggest here is a methodology of a more qualitative nature: an assessment aimed at ensuring that any researcher/end-user can effortlessly find the resources she needs and easily use them as intended. As pointed out by Sugimoto (2016), 'despite the wide array of useful services for (digital) research in linguistics and the humanities [. . . ] it is unclear whether the community is thoroughly aware of the status-quo of the growing infrastructure'. Such an analysis could thus prove useful for bringing to the fore both strengths and critical issues of their data providing community and for planning targeted actions to improve and increase LRs visibility and accessibility: e.g. (meta-)data curation activities; training events on best practices for data and metadata representation, publication and management; specific outreach and communication activities. For reasons of space, in this work we focus on the VLO and attempt to sketch an analysis of the visibility and searchability of language resources (broadly intended as both data and tools/services) as a useful instrument in the hands of repository managers, consortium managers, user-involvement referents and/or national coordinators for planning recovery or improvement actions, as well as targeted communication and engagement strategies. As a case study, we will look at the resources hosted within the CLARIN-IT consortium as well as resources in or about the Italian language hosted elsewhere, under the assumption that they would be of special interest for the Italian CLARIN user communities.

The paper is organised as follows. Section 2 will briefly present the background and related works we considered. Section 3 outlines the methodology devised as an instrument for monitoring the visibility and searchability of LRs in the VLO. In section 4 we will apply the methodology to CLARIN-IT and to Italian as a use case and proof-of-concept. The insights and take-home messages are discussed in section 5. Finally, the conclusions will briefly anticipate how a monitoring of the other points of access, an issue that will have to be tackled in the near future, might be implemented and how it can be useful.

## 2 Background Context and Related Works

### 2.1 The CLARIN-ERIC Four Points of Access

**Virtual Language Observatory.** The VLO (Uytvanck et al., 2010; Goosen and Eckart, 2014) constitutes the main CLARIN central discovery service, i.e. the principal means of finding and exploring Language Resources, broadly intended as both data and tools/services, which exploits the potential of CMDI metadata (Broeder et al., 2010; Broeder et al., 2012) harvested from all the CLARIN B and C centres and other affiliated repositories. The more data providers make use of shared CMDI metadata profiles for describing their LRs, the more easily findable metadata would be in the VLO, because centrally constant work is dedicated to refining the mapping of metadata onto VLO facets and to improve VLO functionalities. This huge harmonisation effort of course comes at the cost of losing some specificity (that can still be maintained locally) and of a certain degree of fallacy. This is why periodic monitoring and metadata curation also on the side of local repositories would be advantageous.

There are two ways the VLO can be searched: it can be queried, first, with a classic search by keyword terms and the results further filtered using predefined facets. Keyword search also supports a pretty expressive advanced syntax (Goosen and Eckart, 2014) that allows the expert user to perform quite specific searches[7].

[6] https://curation.clarin.eu/

[7] For details see https://www.clarin.eu/blog/vlo-updated-advanced-search-facilities

The second one offers faceted browsing: as facets can also be used independently, a user can filter metadata records according to the available categories - the facets - and carry out targeted searches. There are 12 different categories, plus two other useful facets, that can be selected in order to narrow down the selection of displayed records:

1. *Language* - the object language relevant for the resource or tool;
2. *Collection* - the collection to which the resource or tool belongs;
3. *Resource Type* - the type of the language resource (e.g. tool, lexicon, grammar, corpus);
4. *Modality* - the modality of the content of the resource or intended for the tool (e.g. spoken or speech);
5. *Format* - the mime type used in the resource or by the tool;
6. *Keyword* - keywords describing the resource or tool;
7. *Genre* - the genre of the content of the resource (e.g. narrative or conversation);
8. *Subject* - the subject or topic of the content of the resource;
9. *Country* - the country of origin of the source material of the resource;
10. *Organisation* - the organisation currently responsible for the resource or tool, i.e. the holder of distribution rights;
11. *Data provider* - the repository in which the resource is actually deposited and that makes it available;
12. *National project* - the CLARIN national consortium to which the resource pertains.

**The Federated Content Search (FCS).** Most textual and corpus resources hosted by CLARIN centres can be searched and inspected via dedicated query interfaces or applications run by the centres themselves or by the institutions that own the resources (Odijk, 2017). However, such search interfaces are not always easily usable by first-time users as a vast array of query languages and different implementations can be found, which require time and effort to be learned. In order to spare researchers from learning several new query languages before even being certain that the resource actually meets their needs, CLARIN offers a Federated Content Search (FCS) service[8], "an integrated search facility to make these unrelated and partly overlapping content search engines available to the research community" from one single point of access and by using a common syntax (Odijk, 2017, p.41). FCS enables a user to enter a single query that is sent simultaneously to multiple search engines at different federated CLARIN centres, which in turn search in the corpora they host. FCS therefore gives users the possibility to conduct a full-text search across all federated resources, or a selection of them. FCS is thus thought of as a first-level exploration of CLARIN corpus- and textual data, which allows the identification of relevant resources.

**The Language Resource Switchboard (LRS).** Many of the resources that can be discovered through the VLO can be used in many ways. On the one hand, a user can directly download the resources from the local CLARIN repository where these are stored and analyse them offline with her own preferred tools. On the other hand, these resources may be fit to be processed directly by CLARIN tools and services which are available online and distributed over various technical centres in Europe. CLARIN has streamlined this process, thus allowing the users to easily access tools that can be applied to a specific resource by immediately selecting it from the VLO or by using a specific tool: the Language Resource Switchboard (henceforth LRS). LRS therefore is a tool that helps a user find a matching language processing web application for a given resource and process it directly.

[8] `https://www.clarin.eu/content/federated-content-search-clarin-fcs`

**The CLARIN Resource Families** (CRF, Fišer et al. (2018)) is an initiative aimed at providing a user-friendly overview per data type of the available language resources in the CLARIN infrastructure. The listings for each family are meant to facilitate comparative research and are designed for researchers from the digital humanities, social sciences and human language technologies.
Each family is briefly described and the metadata and the links to the respective download pages and concordancers are displayed. Currently, there are 12 corpora families, 5 families of lexical resources, and 4 tool families.[9]

### 2.2 Related Works

Considering the variability of the CMDI metadata framework (Haaf et al., 2014) and the fact that the needs of the users may change over time[10], the VLO represents an asset that needs a frequent scrutiny in terms of visibility and searchability of LRs.

Different researchers have approached the analysis of the VLO from different perspectives (among others see Lušicky and Wissik (2017) and King et al. (2016)). Particularly, two works have influenced the development of the current methodology. Odijk (2014) approaches the VLO from a critical perspective. He examines the searchability of the resources in the VLO with a special attention to the analysis of the structure of their metadata. He shows that finding data of which it is unknown whether they exist is very difficult and in practice in most cases even impossible, given the widely varying granularity of the metadata descriptors and the fact that metadata are often made in isolation ending up in unnecessary differences. His idea of outlining a method aimed at regularly checking the state of the resources in the VLO has proved to be really important given its dynamic nature and its continuous update with new resources.

On the same line, Odijk (2019) further investigates how to enable an easy discovery of LRs and focuses on tools because the search was not easy and because there were no facets dedicated to software for refining a search. He implements a specific faceted search and proposes a curation procedure to secure the uniformity of descriptors and to make sure that the descriptions based on other profiles correctly contain the relevant information and use the right vocabularies. Odijk fundamentally highlights the necessity for a coordination of a national metadata creation and stresses that every national consortium must reserve economic effort for active participation in the metadata curation task force.

## 3 Methodology

In this paper, we take into account and complement previous related works by proposing a user-oriented methodology for a qualitative assessment of the visibility and findability of LRs, which can be applicable to every national project. We suggest the following checks should be carried out not only by data managers of new consortia after the registration of at least one B or C centre, but also, periodically, for any national consortium, especially when new centres are registered or new large collections are injected. The aim is to determine the extent to which the resources are correctly and adequately described in terms of metadata descriptors associated to them.

In general, the idea is to explore and test an assessment procedure that may assist repository managers, national coordinators or even the CLARIN central office in harmonising the content of each repository and consequently of the VLO, to the benefit of end-users.

The methodology is composed of two phases. The first phase deals with the inspection of the LRs available from the data centres of a given national consortium, as they show up in the VLO. This is performed by exploiting various facets. The second phase deals with the investigation of the existence in the VLO of the main language(s) of the national project under scrutiny, by means of a systematic analysis of LRs distribution among organisations, collections and data providers outside the national consortium. In what follows we display the structure of this methodology.

[9] https://www.clarin.eu/resource-families

[10] Hence the urgency to periodically assess also user needs, as done by Monachini et al. (2018), or Lušicky and Wissik (2017)

### 3.1 Phase 1: Check for a National Project

At first, the national consortium of interest should be selected from *the National Project facet*, in order to highlight only the LRs that are provided by its centres. Successively, the results are to be filtered in order to classify the retrieved LRs and check how they are described using the VLP facets, according to four different steps:

1. Languages: check the languages present and calculate the number of LRs for each language;
2. Organisations and Collections - check the Organisations and Collections involved and calculate the number of LRs for each;
3. Resource Type and Data Providers - classify the type of LRs deposited by each data provider;
4. Formats and Availability - Check if:
   (a) The information on availability is clearly and correctly provided;
   (b) The items deposited and marked as available have an actionable resource, i.e. a resource that can be downloaded and potentially analysed with offline tools or processed directly e.g. via the LRS.

### 3.2 Phase 2: Check for a Specific Language

As mentioned in 1, under the assumption that the resources for national language(s) are of particular interest for the community of the corresponding national consortium, we deem it useful to monitor also for the existence, visibility and accessibility of LRs for that specific language(s) outside the consortium centres. This might give national coordinators a useful overview of the presence of their language(s) internationally, and help them plan specific actions in the interest of their communities. In this phase, we thus foresee the following steps:

1. Select the national language(s) from the *Language* tab in order to show only the LRs of interest;
2. Calculate the number of Collections hosting these LRs;
3. Calculate the number of Organisations responsible for these LRs;
4. Calculate the number of Data Providers who deposited these LRs;
5. Compare the results with the information in possession of the national coordination team.

## 4 Applying the Methodology: the Case of CLARIN-IT and Italian Resources

### 4.1 CLARIN-IT

At present[11] CLARIN-IT has 8 member institutions and two data centres:

- ILC4CLARIN[12], the national CLARIN B data centre, hosted and managed by the CNR Institute for Computational Linguistics "A. Zampolli" in Pisa, the founding member of CLARIN-IT; and
- ERCC[13], a CLARIN C centre, hosted by the Institute for Applied Linguistics (IAL) at EURAC Research in Bolzano

Both centres offer the whole community to deposit services for the long-term preservation of data in certified repositories that comply with CLARIN requirements and which are regularly harvested by the VLO. The consortium also comprises two K-centres:

[11] We last updated data on 15 January 2022. As new resources may be constantly deposited, the figures are likely to be different at the time of reading.

[12] https://ilc4clarin.ilc.cnr.it/

[13] https://clarin.eurac.edu/

- the CLARIN KNOWLEDGE CENTRE FOR DIGITAL AND PUBLIC TEXTUAL SCHOLARSHIP (DIPTEXT-KC)[14], jointly maintained by University Ca' Foscari in Venice and ILC-CNR; and
- the transnational CLARIN KNOWLEDGE CENTRE FOR COMPUTER-MEDIATED COMMUNICATION AND SOCIAL MEDIA CORPORA (CKCMC), a transnational K centre hosted in Italy by the Institute for Applied Linguistics, Eurac Research (IAL) in Bolzano.[15]

Currently, CLARIN-IT offers seven different digital collections, which are deposited in one of the two data centres. It thus serves different research sub-communities, particularly oral archives, computer-mediated communication, and digital classics. For the latter it is worth mentioning the effort carried out to facilitate the integration and deposit of important textual collections, such as for instance the *Archivio della Latinità Italiana del Medioevo* (ALIM).

### 4.2 Phase 1: CLARIN-IT in the VLO - the National Project

Following our methodology, the CLARIN-IT resources in the VLO are easily extracted by filtering the results for *CLARIN-IT* within the national project facet[16], as shown in fig. 1.

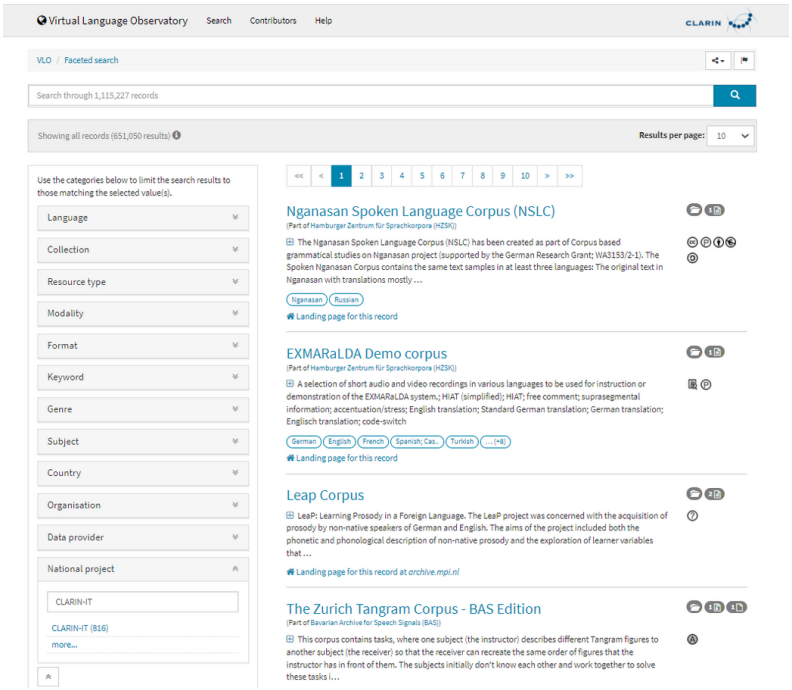


Figure 1: VLO facets - searching for National project.

The query returns 890 different LRs, of which 72 are hidden because they are considered duplicates, which leaves us with 818 distinct resources, as displayed in Table 1.

Duplicate records are automatically hidden in the VLO main search results on the basis of their naming (i.e. title), and are listed under each affected record instead. After a careful examination, most of the apparently duplicate items from our query result in being false duplicates. Within the *ALIM Literary Sources* for instance, all of the 50 hidden items are in fact different critical editions of the same texts made by different editors. Since they have exactly the same title, the system considers them as duplicates. For example, the *Summa Dictaminis* corresponds to three records, one for each editor (Matteo de' Libri, M. Thumser, Emil Polak). While a possible strategy to avoid this could be to add "by EDITOR" to the

[14] https://diptext-kc.ilc4clarin.ilc.cnr.it/

[15] https://cmc-corpora.org/ckcmc

[16] The executed query is https://vlo.clarin.eu/search?fqType=nationalProject:or&fq=nationalProject:CLARIN-IT

| *Facets* | *LRs* |
|---|---|
| Languages | 27 |
| Organisations | 16 |
| Collections | 12 |
| Format | 10 |
| Resource type | 6 |
| Data providers | 2 |
| | 818 + 72 duplicates |

Table 1. CLARIN-IT resources in the VLO along the main dimensions of analysis.

'title' (e.g. *Summa Dictaminis BY MATTEO DE' LIBRI*), this practice would conflict with the decisions taken by the data depositors.

#### 4.2.1 Step-1 - Checking the Languages of the National Project

The first step involves the analysis of the distribution of the languages covered by the resources deposited or described in the CLARIN-IT centres: we identified 27 different languages.

| *Language* | *n. LRs* | *Language* | *n. LRs* |
|---|---|---|---|
| Latin | 734 | Cimbrian | 1 |
| English | 44 | Croatian | 1 |
| Italian | 38 | Karelian | 1 |
| Arabic | 32 | Ladin | 1 |
| German | 12 | Ladino | 1 |
| Ancient Greek | 10 | Mòcheno | 1 |
| Ancient Greek (to 1453) | 8 | Sardinian | 1 |
| Dutch | 4 | Saurano | 1 |
| French | 4 | Slovenian | 1 |
| Czech | 2 | Spanish; Castilian | 1 |
| Modern Greek | 2 | Trentino | 1 |
| Modern Greek (1453-) | 2 | Tyrolean | 1 |
| Basque | 1 | Veneto | 1 |
| Breton | 1 | | |

Table 2. Languages in CLARIN-IT.

As Table 2 shows, CLARIN-IT offering is not restricted to Italian only, but it presents LRs in a variety of languages. We acknowledge the substantial presence of LRs in various other languages like English, German, Dutch, French, in addition to a smaller number of LRs in other important European languages like Czech, Modern Greek, Slovenian and Spanish. The wide variety of languages deposited in CLARIN-IT is also evidenced by different regional and minority languages spoken in the North of Italy such as Tyrolean, Trentino, Saurano, Ladin, Cimbrian and Mòcheno[17].

Among the languages, a conspicuous share is represented by Latin and Ancient Greek LRs, and the ILC4CLARIN centre appears to be specialised in hosting them. A closer look, though, reveals an over-representation of Latin LRs, which appear to be more numerous even than the Italian ones. This is due

[17]For Sardinian, Karelian, Basque and Breton, there is actually no LR available. In fact, the record refers to a survey run in 2016 by the Digital Language Diversity Project, a dataset containing the original responses to a questionnaire about the online use and usability of these four regional and minority languages.

an organisational choice of the ALIM corpus. As described in Boschetti et al. (2020), every text of that corpus is deliberately deposited as a separate 'corpus' resource, so that it reflects the organisation of the original archive; this, however, contrasts with common corpus linguistics practices. Indeed, the peculiar choice of depositing the ALIM corpus as a collection with every document described as a single corpus resource has obvious drawbacks in terms of visibility and searchability, in that it skews the counts in the searches and unnecessarily overloads the search result pages in the VLO. At the same time this structure makes the texts directly and easily actionable by means of NLP services or corpus management tools, available for instance via the Language Resource Switchboard; thus, in terms of accessibility and usability, such an organisation may prove advantageous.

Finally, a similar issue can be observed for Lexical Resources, among which Arabic appears to be over-represented. A closer inspection reveals that the high amount of Arabic LRs - in total 32 - is also due to the way in which the *Al Qamus al Muhit - the Medieval Arabic Lexicon* (Nahli et al., 2016) has been deposited: each letter of the lexicon has been treated as a single deposited resource. Therefore, as in the ALIM archive case, the 30 entries are actually all parts of the same dictionary, i.e. the *Al Qamus al Muhit*.

### 4.2.2 Step 2 - Checking the Organisations and the Collections Involved

The second step allows us to identify the organisations and consortium members which are actively contributing their resources to CLARIN-IT and analyse how the collections are represented in the VLO for each organisation.

| ***Organisations*** | ***Collections*** | ***n. LRs*** | ***Organisations*** | ***Collections*** | ***n. LRs*** |
|---|---|---|---|---|---|
| ALIM Archivio della Latinità Italiana del Medioevo | ALIM Literary Sources;<br>ALIM Documentary Sources | 344 - 10 | Institute of Information Science and Technologies "Alessandro Faedo"<br>ISTI CNR | ILC4CLARIN : ILC Data & Tools | 1 |
| DigiLibLT | DigilibLT | 364 | Escuela Universitaria de Turismo "Felipe Moreno"<br>Universitat de les Illes Balears | ERCC Open: CMC & WaC | 1 |
| Istituto di Linguistica Computazionale "A. Zampolli" - ILC-CNR | ILC4CLARIN : ILC Data & Tools<br>MQDQ Galaxy | 40 - 2 | Università del Piemonte Orientale | ILC4CLARIN : OPEN Data & Tools | 1 |
| Institute for Applied Linguistics, Eurac Research | Eurac Research: Learner Language;<br>Eurac Research: CMC & WaC | 12 | CNR Edizioni | ILC4CLARIN : OPEN Data & Tools | 1 |
| CIRCSE Research Centre<br>Università Cattolica del Sacro Cuore | CIRCSE | 8 | Università di Pisa | ILC4CLARIN : OPEN Data & Tools | 1 |
| Università di Parma | ILC4CLARIN : OPEN Data & Tools | 3 | Università di Salerno | ILC4CLARIN : OPEN Data & Tools | 1 |
| Ghent University | ERCC: Various;<br>ERCC Open: Learner Language | 2 - 1 | University of Verona | ERCC: Various | 1 |
| Gruppo di ricerca BIA<br>Bibliotheca Iuris Antiqui | BIA-Net FONTES | 1 | Venice Centre for Digital and Public Humanities (VePDH) | ILC4CLARIN : OPEN Data & Tools | 1 |

Table 3. Organizations and Collections in CLARIN-IT.

In the case of the Italian network, for instance, we identify 16 different organisations currently responsible for 13 collections, as Table 3 shows. Among these, the *Archivio della Latinità Italiana del Medioevo* (ALIM)[18] and the *Digital Library of late antique Latin texts* (DigiLibLT)[19] are responsible for the highest number of Latin LRs. The former is responsible for the *ALIM Literary Sources* and *ALIM Documentary Sources* collections. The latter controls the *DigilibLT* collection. Similarly, Università Cattolica del Sacro Cuore, and more specifically the CIRCSE Research Centre, is responsible for the Latin collection *CIRCSE*, which is composed of Latin lexical resources, corpora and dictionaries, as well as tools for processing Latin texts. ILC-CNR and EURAC, the two CLARIN-IT data providers (cfr. Section 4.1), are directly responsible for 42 and 12 LRs, respectively. ILC-CNR is responsible for two collections: *ILC4CLARIN: ILC Data & Tools*, containing about 40 resources, and the *MQDQ Galaxy* collection, which contains 2 resources[20]. EURAC is responsible for the *Eurac Research: Learner Language* collection, with 2 resources, and for the *EURAC Research: CMC & WaC* collection, with 4 resources.

As it is clear from Table 3 above, the majority of the organisations are from Italy. Surprisingly, there are, however, two foreign organisations depositing data in Italy: the Ghent University from Belgium, responsible for three annotated learners' corpora, two in English, French and Dutch, and one in German; and the Universitat de les Illes Balears from Spain, which is responsible for an English lexical resource.

[18] `http://en.alim.unisi.it/`

[19] `https://digiliblt.uniupo.it/`

[20] *Musique Deoque* is a project exploring texts of Latin poetry composed in Italy between 1250 and 1550. `http://mizar.unive.it/poetiditalia/public/`

At close examination of the collection *ILC4CLARIN: ILC Data & Tools*, we noticed the presence of additional 17 LRs for which the organisation is not visible from the VLO, although this information is encoded in the full metadata records stored in in the local repository.

This discrepancy may be due to mapping issues between CMDI profiles and VLO facets (already mentioned by Odijk (2019)).

#### 4.2.3 Step 3 - Checking the Resource Types and the Data Providers

As regards the third step, we shall examine how LRs are distributed among the data providers with a focus on *Resource type* in order to get some information on their specialisation. As Table 4 shows, there are two Data Providers in CLARIN-IT: The *ILC4CLARIN Centre* at the Institute for Computational Linguistics and the *EURAC Research CLARIN Centre*.

| ***ILC4CLARIN*** | ***n. LRs*** | ***Eurac Research*** | ***n. LRs*** |
|---|---|---|---|
| Corpus | 375 | corpus | 18 |
| Text | 362 | Lexical Resource | 1 |
| Lexical Resource | 47 | | |
| Software, webservice | 14 | | |
| Language Description | 1 | | |
| Total | 799 | | 19 |

Table 4. Resource type for each Data provider.

A rich array of LRs is deposited within both of the CLARIN-IT centres. The majority of them are described under the label *corpus* and *text*. The type *lexical resource* corresponds to a broad category and includes lexicons, ontologies, terminologies, e-dictionaries, wordlists etc. . Lastly, the *software* and *webservice* categories include on-line applications, off-line tools and (web)services, which can be used to perform different kinds of analyses on language data.

#### 4.2.4 Step 4 - Checking the Formats and Availability

The last step concerns the query on the available *Formats* and *Subjects* of CLARIN-IT resources. This step allows us to assess whether all resources are correctly deposited and whether they have been further described with suitable and harmonised subject keywords. In the Italian case, the coverage for the latter seems to be incomplete (only 18 LRs are mapped onto VLO subjects keywords, whereas many of the keywords present in the national repositories are not visible in the VLO) and harmonisation could be increased by using controlled vocabularies. One important final check relates to the *Availability* facet, which indicates the "degree to which resources and tools are publicly accessible". In the case of CLARIN-IT, most of the LRs are publicly available; however, the filter also returned 29 resources with unspecified availability . A closer inspection shows that these correspond to corpora from the ERCC repository and webservices from ILC4CLARIN which are in fact available. This finding might be helpful and lead to amendments of the records.

### 4.3 Phase 2: Other Resources for Italian in the VLO - an Overview

This second phase aims at investigating the existence, visibility and availability of Italian language resources within the CLARIN ERIC network outside CLARIN-IT [21]. The idea behind this examination is to shift the perspective from a specific National Project and to focus, instead, on resources of potential interest for the national community, but residing somewhere else.

As indicated in section 3, this phase is composed of 4 steps.

1. Selection of the language. In this use case, we select *Italian* from the language facet and this search gave 9774 LRs present in the VLO.

[21] The query relative to Phase 2 is `https://vlo.clarin.eu/?0&fq=languageCode:code:ita&fqType=languageCode:or`

2. Check for collections: we determined that 94 different collections contain Italian LRs.

3. Check for organisations: there are altogether 109 organisations responsible for the distribution of Italian resources.

4. Lastly, check for data providers: we identified 31 data providers distributed over 15 national projects depositing Italian resources in CLARIN.

The following Table 5 summarises these pieces of information.

| **Italian in the VLO** | |
|---|---|
| Number of LRs | 9774 |
| Collections | 94 |
| Organisations | 109 |
| Data Provider | 31 |
| National Project | 15 |

Table 5. Italian resources in the VLO, provided by organisations outside CLARIN-IT.

By comparing the list of the member organisations of CLARIN-IT with the results obtained from the first test, we end up finding some interesting aspects of the presence of the Italian language in the VLO. First, in addition to Italy and CLARIN-IT, fourteen other countries manage and deposit Italian related LRs. Secondly, and more interesting, there are some institutions located in Italy, not (yet) members of CLARIN-IT, who have deposited some LRs in other CLARIN ERIC centres. For example, the organisation *CNR OVI* appears to be the provider of 5498 resources which are catalogued by *Europeana*[22], a European Initiative not directly linked to CLARIN-IT which also catalogues the collections of the *Plutei della Biblioteca Laurenziana*, with 1966 LRs, and the collection of the Biblioteca Riccardiana, with 10 LRs. The Italian *Archivio Storico Civico e Biblioteca Trivulziana* located in Milan, which is responsible for the manuscript *Concetti amorosi, cioè lettere giovenili, et amorose* by Vinetia Compagni, and the *Biblioteca Bertoliana* located in Vicenza, responsible for the *Lettere Amorose* by Ferrante Pallavicino Luca Assarino, have both deposited these resources to The Language Bank of Finland (FIN-CLARIN). As another example, the University of Naples L'Orientale is responsible for the *MPI EVA corpora: Jakarta Field Station*, a collection of 251 recorded conversations of bilingual Indonesian/Italian children, and deposited it into the *Max Planck Institute for Psycholinguistics* (CLARIAH).

The overview offered by this second test gives us some important insights that can help to enhance the visibility of the Italian language from the perspective of the national consortium, by also highlighting those resources which are not under its management.

## 5 Lesson Learnt

After having sketched and applied a step-wise methodology for monitoring the availability and visibility in the VLO of LRs, in this section we concentrate on and sum up the lessons we learn from this exercise. While there certainly is ample margin for improvement, we see that the methodology already provides useful insights: on the one hand, it gives useful indications on possible actions that national coordination teams and centre managers may put in place to enhance and promote the consortium offerings; on the other hand, it brings to the fore problematic or controversial issues and thus indicates possible directions for improvements. The results emerging from the different search dimensions may thus help devise targeted communication and engagement actions, or plan recovery strategies. Here below we provide a few examples.

The 'Organisation' dimension in Phase 1 helps us monitor the consortium members' activity as LR providers, and may help coordinators set up actions targeted, for instance, at understanding the reasons

[22] `https://www.europeana.eu/en/about-us`

behind a low activity and plan opportune recovery actions, such as specific training events, incentives, focused workshops, or other UI events. At the same time, low activity in providing LRs may instead indicate the given institution has a different profile and might be better involved in other kinds of actions, e.g. in training and UI activities.

The 'Organisation' dimension in Phase 2 helps monitor the activity in depositing data of (national or international) non-member institutions. This might be valuable information for coordinators interested in enlarging their consortium, or in strengthening its international collaborations, as it brings to light the most prominent external providers of LRs of interest for the consortium communities. This may lead to fostering, for instance, mobility grants, joint initiatives, or even new joint K-centres.

The 'Language' dimension in Phase 1 provides interesting indications on the main interests of the active national community and may lead to targeted communication and dissemination actions aimed at maximising national visibility and findability of relevant resources and technology. For instance, national and local websites may be restructured in order to highlight the key resources for the target communities (something similar to the LRFs initiative, but tailored to the national scene). Dedicated info- or training days may be organised at various selected locations, for increasing awareness within the national communities.

The 'Language' dimension in Phase 2 instead provides useful insights on the availability of LRs for the target language(s) from other CLARIN projects and also informs about the interest for one's national language in the rest of the pan-European network. This again may lead to the setting-up of new mobility grants and joint initiatives.

On a more technical, day-by-day operational level, we have seen how the proposed monitoring procedure may also bring to light inconsistencies and problematic issues, for which repository managers may want to plan recovery actions. Some of the identified problematic issues may also be of interest at central level and worth discussing in the appropriate CLARIN ERIC committees, as they may lead to improvements of the central services, as the work by Odijk (2014) already demonstrated. For instance, both the issues of false duplicates and of the granularity of text collections that emerged from Phase 1 (cfr. section 4.2 and 4.2.1) may be an interesting point of discussion both locally and at central level.
Locally, it might lead to a confrontation with the responsible of the collection to explore the possibility of aligning their choices to the current majority practices, as well as to targeted data curation activities and metadata refinement.
Centrally, it could raise awareness on peculiar needs of local communities and either lead to an enhancement of central services, or foster the development and dissemination of more stringent and clear common best-practices. Hence, they might be interesting topics for discussion for the Standing Committee for CLARIN Technical Centres and the curation task force.

## 6 Conclusions

In this paper we have drafted a methodology for monitoring the existence and visibility of LRs of relevance to a national consortium and applied it to CLARIN-IT as a test case. With the growth of the Italian consortium, a thorough check of the LRs contributed by the various actors across the existing official repositories had also become necessary. This qualitative assessment exercise has proven extremely useful and, with adequate extensions and improvements, might become a model for other national projects. Future additions might be to include more dimensions to be assessed, for instance a template search for key resource types, such as reference corpora and lexicons, to ensure that they appear as expected.

While this first assessment focused on the first and most important CLARIN point of access – VLO – similar procedures should be devised also for the other ones. Centres offering NLP web services via the LRS should for example monitor which resources are considered a good match for their tools, and check the outcomes of the analysis, to see if they correspond to the expectations; centres having corpora in the FCS could run test queries to see if selected segments of their corpora actually appear as results in appropriate searches. As for the CRFs, manual curation and updating is currently carried out by a dedicated task force. However, with the growth of the initiative (in terms of families and described items), monitoring of own resources becomes all the more important, for instance to signal changes in resource

size, in the links to online access interfaces, etc. Also, via the CRFs initiative we may discover important resources that are not yet deposited in a CLARIN centre and take steps to encourage their authors to deposit them in a certified repository. Conversely, one may inform the CRFs task force of interesting resources that might be added to the appropriate RF. Thus, monitoring the CRFs and close collaboration with its team can be a source of mutual growth for national repositories and central services, also in terms of reaching out to resource creators that are not yet aware of what CLARIN has to offer them.

Finally, it is important to mention a further, useful central facility offered by CLARIN, which will need to be monitored once it becomes to be more widely used, namely the possibility of creating Virtual Collections[23] both by listing individual resources, and as the outcome of a specific VLO query. Creating a collection such as "Italian historical corpora" could allow users of the Italian national node to access distributed resources at a glance. Such collections, if created from a VLO query, will always be updated with all relevant resources, and in any case provided with persistent identifiers, thus making them easy to be cited in publications.

## Acknowledgements

This work has been supported by CLARIN-IT, a Project of International Significance, financed by the Italian MUR, Ministero dell'Università e della Ricerca (Ordinary fund for research institutes).

## References

Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. "Tea for Two": The Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. In Navarretta, C. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2020. Virtual Edition*.

Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. 2010. A Data Category Registry- and Component-based Metadata Framework. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Broeder, D., Windhouwer, M., Uytvanck, D. V., Goosen, T., and Trippel, T. 2012. CMDI: a Component Metadata Infrastructure. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Eskevich, M., Goosen, T., and Uytvanck, D. V. 2017. CLARIN CASE STUDY: Making Europeana's resources available for research purposes through the CLARIN infrastructure. Technical report, Europeana and CLARIN ERIC. `https://pro.europeana.eu/files/Images/Europeana_Research/CLARIN/CLARIN_case_study.pdf`.

Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN's key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Goosen, T. and Eckart, T. 2014. Virtual Language Observatory 3.0: What's new? In *Proceedings of the CLARIN annual conference 2014, 23-25 October 2014, Soesterberg, The Netherlands*.

Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., and Uytvanck, D. V. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny - the VLO taskforce of the CLARIN-D centres. In *CLARIN annual conference 2014*, Soesterberg, The Netherlands.

King, M., Ostojic, D., Ďurčo, M., and Sugimoto, G. 2016. Variability of the Facet values in the VLO – a case for metadata curation. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, pages 25–44. Linköping University Electronic Press, Linköpings universitet.

Lušicky, V. and Wissik, T. 2017. Discovering resources in the VLO: A pilot study with students of translation studies. In Borin, L., editor, *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016*, pages 63–75. Linköping University Electronic Press, Linköpings universitet.

[23] `https://collections.clarin.eu/`

Monachini, M., Nicolosi, A., and Stefanini, A. 2018. Digital classics and CLARIN-IT: What Italian scholars of ancient greek expect from digital resources and technology. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, pages 61–74.

Nahli, O., Frontini, F., Monachini, M., Khan, F., Zarghili, A., and Khalfi, M. 2016. Al qamus al muhit, a medieval Arabic lexicon in LMF. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 943–950, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Odijk, J. 2014. Discovering resources in CLARIN: Problems and suggestions for solutions. Unpublished paper: `https://dspace.library.uu.nl/handle/1874/303788`.

Odijk, J. 2017. Introduction to the CLARIN technical infrastructure. In Odijk, J. and van Hessen, A., editors, *CLARIN in the Low Countries*, pages 33–44. Ubiquity Press.

Odijk, J. 2019. Discovering software resources in CLARIN. In Skadina, I. and Eskevich, M., editors, *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, pages 121–132. Linköping University Electronic Press, Linköpings universitet.

Sugimoto, G. 2016. Number game - Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. In *Proceedings of the CLARIN Annual Conference 2016, 26-28 September 2016, Aix-en-Provence, France*.

Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P., and Gardellini, M. 2010. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, La Valletta, Malta. European Language Resources Association (ELRA).