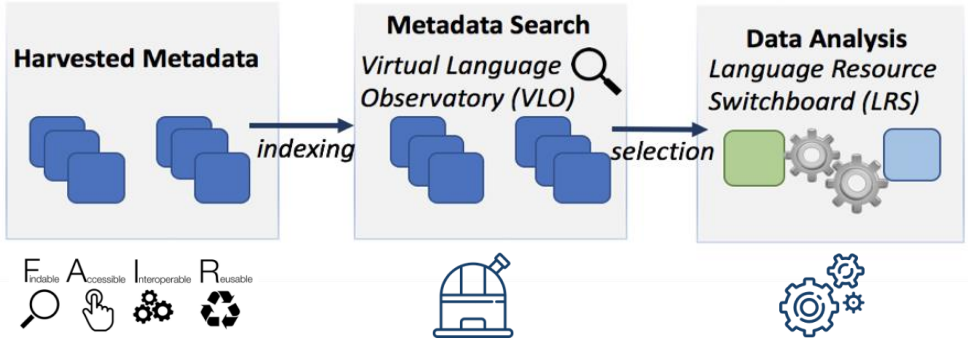
CLARIN-IT Resources in CLARIN ERIC
A Bird's-Eye View

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

CLARIN Conference 2021

CLARIN -The technical infrastructure



Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Spoken corpora in the CLARIN infrastructure

Corpora with transcriptions and audio recordings

Corpus	Language	Description	Availability
Arabic Speech Corpus **Licence:** CC BY 4.0	Arabic	The corpus is available for download from a dedicated webpage. For a relevant publication, see Halabi (2016).	Download
DIALEKT v1: dialectal corpus with multi-tier transcription **Size:** 100,000 words **Annotation:** orthographically and phonetically (dialect features) transcribed, MSD-tagged, lemmatised **Licence:** Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains traditional dialectological material, mostly unprepared monologue-type speech. The corpus is available download (upon request) and through the concordancer KonText. For a related publication, see Komrsková et al. (2018).	Concordancer Download

https://www.clarin.eu/resource-families

Research aims

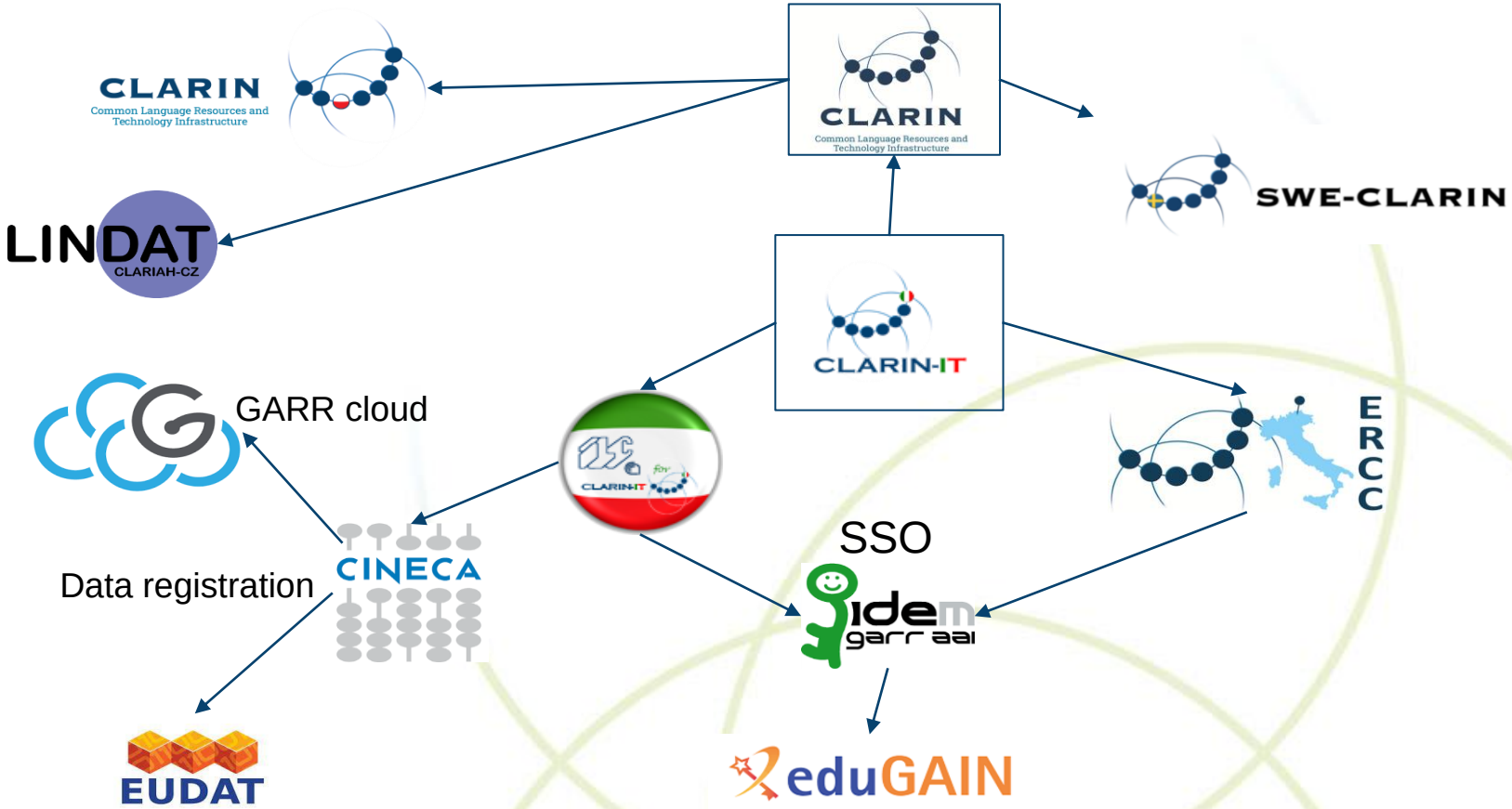
National consortia should monitor regularly these four “points of access”

A more qualitative assessment to ensure that any researcher/end-user can easily find the resources she/he needs and use them as intended.

A methodology to check and analyze the presence of the LRs available in the CLARIN-IT consortium:

- Assessing the Italian consortium presence
- Devising a reproducible qualitative methodology from the user perspective.
- Evaluating the visibility, reliability and searchability of CLARIN-IT LRs in the VLO.

CLARIN-IT - Technology infrastructure



The Methodology

Select in the National Project tab – focus on LRs of interest

Check which LRs are shown and how are presented in the VLO, filtering for:

- *Languages*
- Organisation Collections
- Resource type

Check the presence of *duplicates*

Check the status of activation for a sample of links to the original place

Register all the inconsistencies in terms of *accessibility* and *availability*

CLARIN-IT – A Bird's eye view

CLARIN-IT - a birds eye view	
Total Number of LR	439
Monolingual	388
Multilingual resources	46
Format	12
Languages	10
Organisations	8
Collections	7
Resource type	6
Data providers	2

Table 1: CLARIN-IT on VLO

Languages			
Latin	366	Italian	30
English	40	German	8
Arabic	32	Czech	2
Ancient Greek (to 1453)	6	Breton	1
Ancient Greek	8	Basque	1

Table 2: Languages in CLARIN-IT

Organisations			
Archivio della Latinità Italiana del Medioevo (ALIM)	354	CIRCSE - Università Cattolica Sacro Cuore	8
Istituto di Linguistica Computazionale - CNR	39	Ghent Universities	2
Institute for Applied Linguistic Research - EURAC	9	Università di Parma	2
Università di Salerno	8	Basque	1

Table 3: Organisations in CLARIN-IT

CLARIN-IT – A Bird's eye view

Collections			
ALIM Literary Sources	344	ILC4CLARIN : OPEN Data and Tools	7
ILC4CLARIN	54	ERCC Learner Corpora	8
Alim Documentary Sources	11	ERCC Web Corpora	4
CIRCSE	8	ERCC	1

Table 4: Collections in CLARIN-IT

ILC4CLARIN		**Eurac Research**	
Corpus	368	Corpus	13
Lexical Resource	43		
Software, webservice	12		
Webservice	2		
Text	1		

Table 5: Resource type for each Data provider

Concluding remarks