CLARIN - Infrastructural support for the study of language as social and cultural data

Franciska de Jong
CLARIN ERIC
f.m.g.dejong@uu.nl

Workshop *Linguistique & Big Data*, Paris
30 November 2017

Overview

- Introduction into CLARIN
- CLARIN's data architecture
- The turn to big as incentive for multidisciplinary work
- *CLARIN pour les chercheurs*

CLARIN in seven bullets

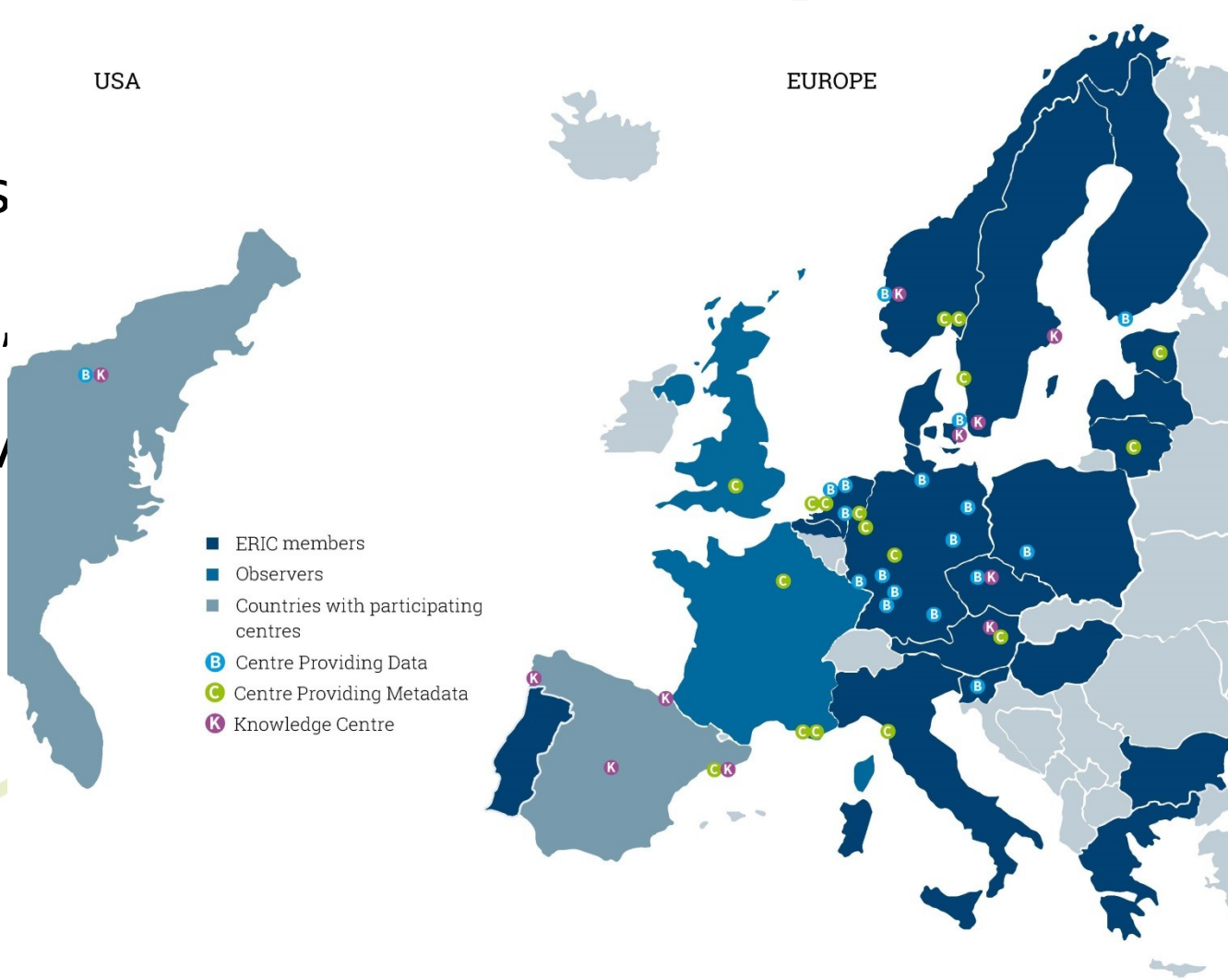
- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** environment
- and that serves as an ecosystem for **knowledge sharing**.

CLARIN ERIC in members and centres

A consortium of:

- 19 members AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, IT, LT, LV NL, NO, PL, PT, SE, SI
- 2 observers:

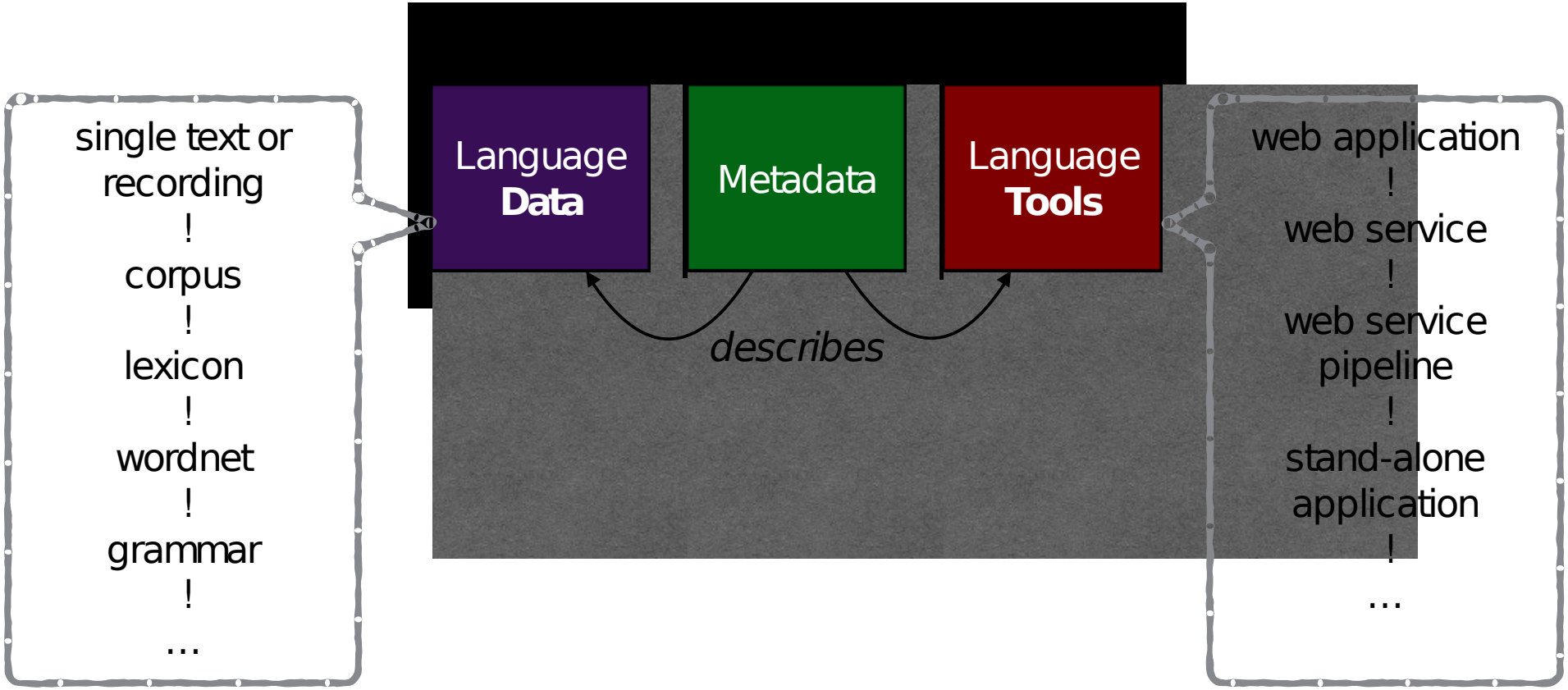
- FR, UK;
- >40 centres



CLARIN in resource types

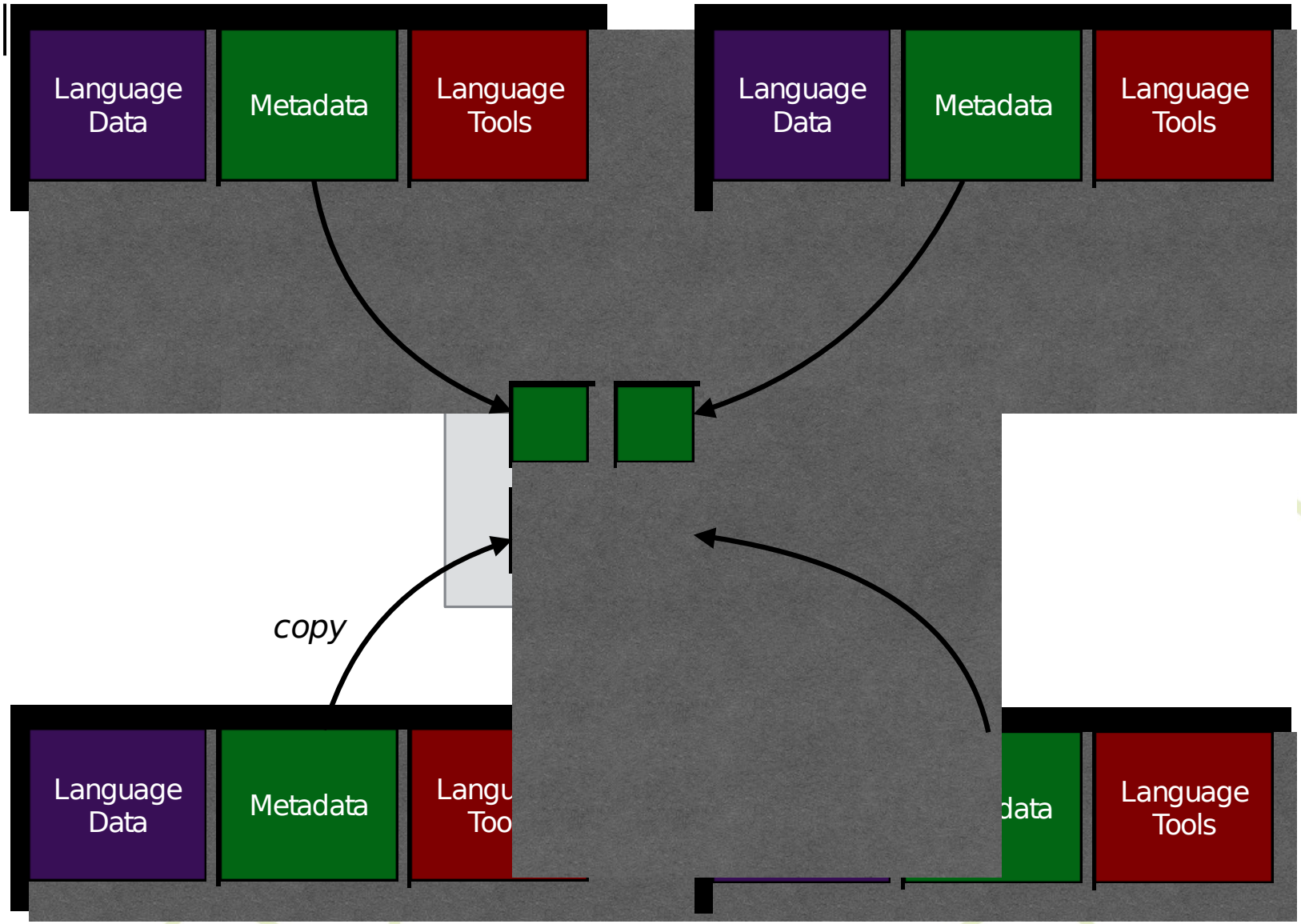
- Parliamentary records
- Literary texts
- Social Media data
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- Newspaper archives
- …

The CLARIN data architecture*: repositories

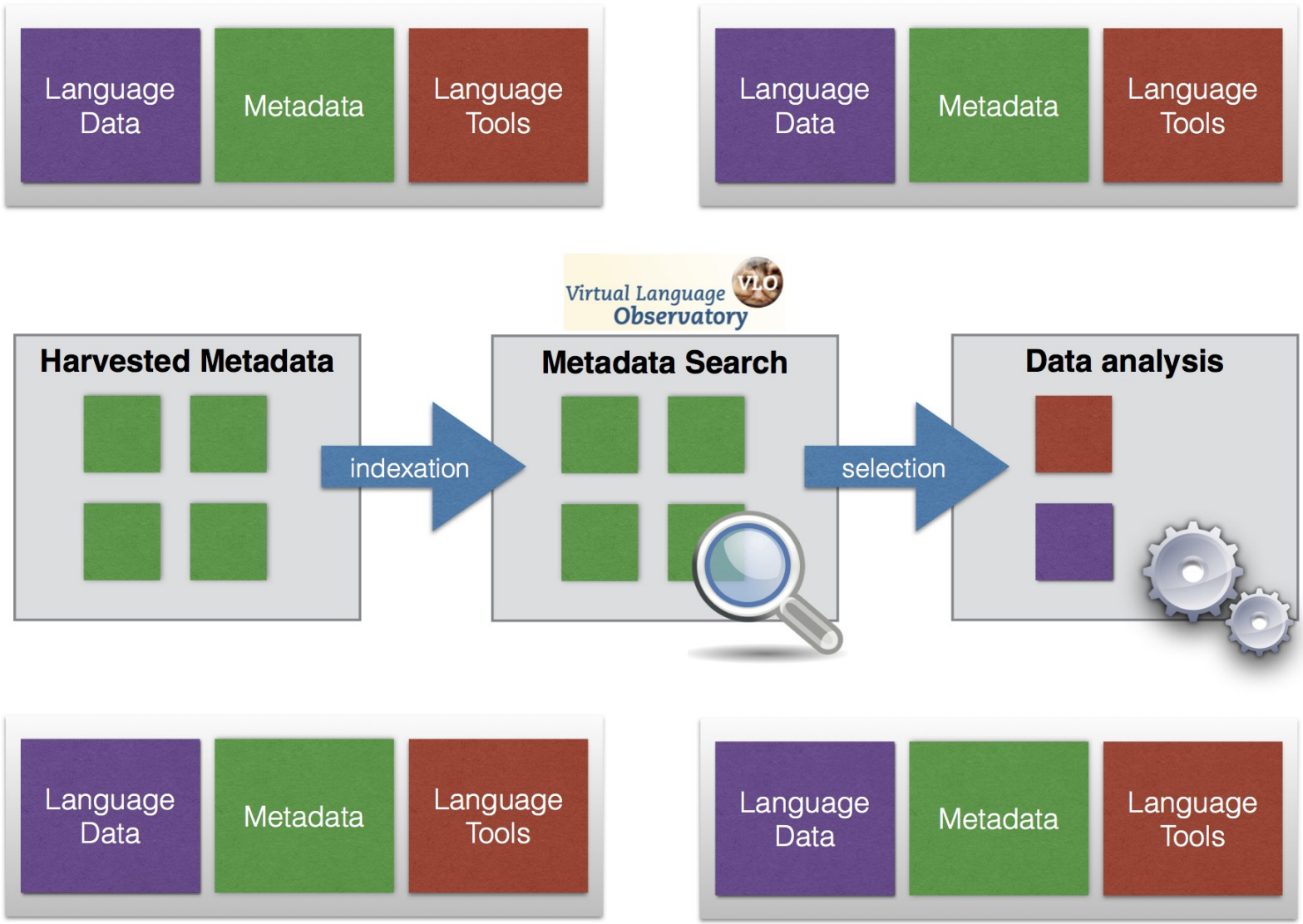


* slides by Dieter Van Uytvanck

The CLARIN data architecture:



The CLARIN data architecture:



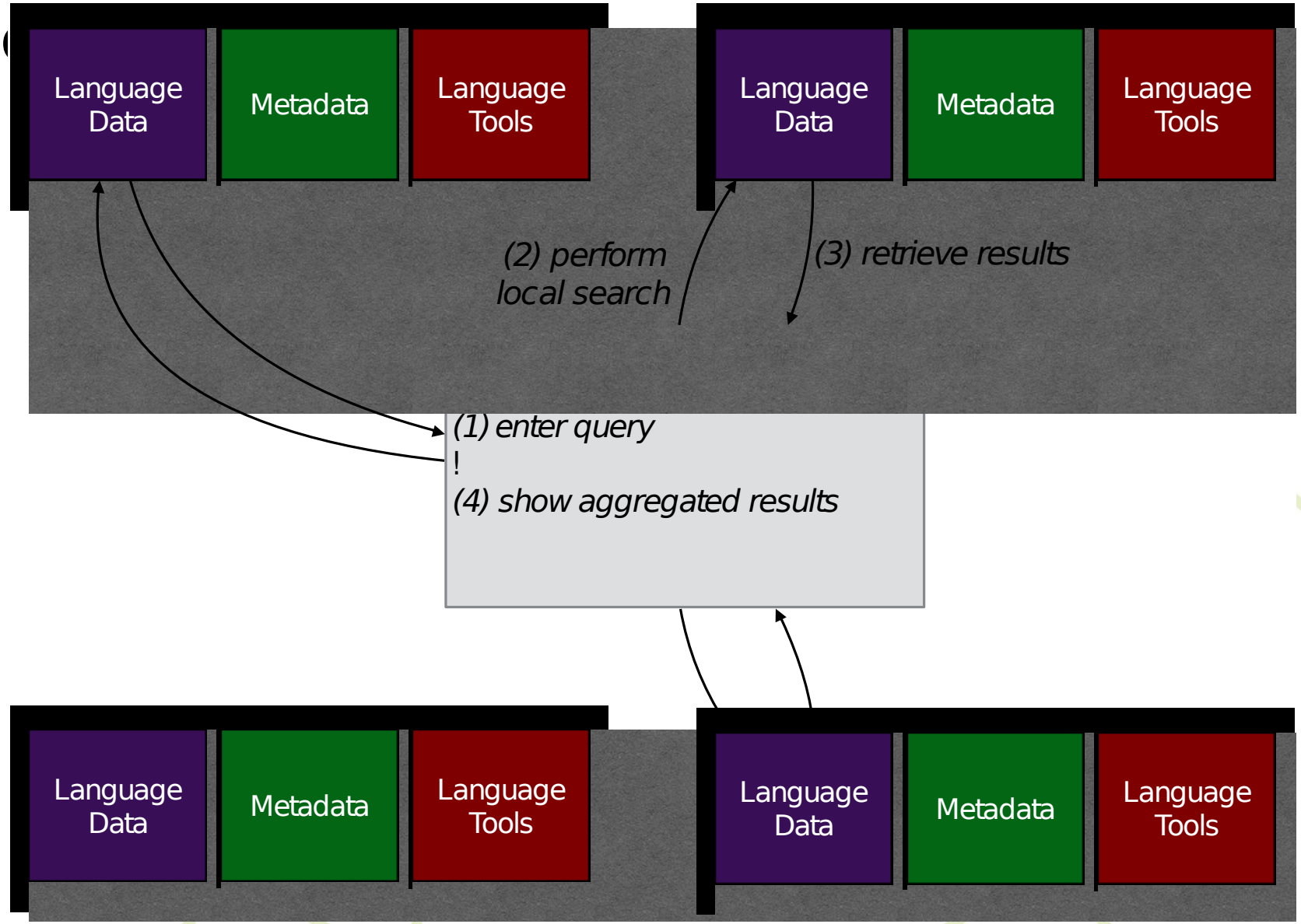
Virtual Language Observatory*

- Language Resource facet browse: vlo.clarin.eu
- >1,500,000 records (dont 2% francais)
- Qui peuvent être filtrés
  par: Language | Collection |
- Resource type | Modality | Genre | Other »

ABaC:us – Austrian Baroque Corpus

* slide by Bente Maegaard

The CLARIN data architecture: c



Interoperability is key

- to the exhar
- to the excha
  output of analytic tools
- to the options for supporting comparative research

The digital turn and the turn to big*

close reading
word, sentence,
dialogue, text, corp

distant reading
…, text, corpus, collection, web

* image from: https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-hum

The digital turn and the turn to ‘more’

Language is reflecting

- the underlying linguistic system

-> linguistics

- cultural and societal dynamics

-> multidisciplinary work

CLARIN and data science (1)

- Analytics for **text and speech data** as a pillar for data science
- Integration of language **data in context**: heterogeneous data
- Contribution to the development of
  - **new methodological** frameworks for the integrated processing of multiple datatypes, and
  - **multidisciplinary research agendas**.

Europe's multilinguality is key*

- to our understa
  language affects identity, culture, society
- to our understanding of diversity across boundaries of time and regions
- and therefore for comparative studies

* image from

CLARIN and data science (2)

- Europe's **multilinguality** as a basis for **comparative research** of societal and cultural phenomena, that are reflected in language use; some examples:
  - Migration patterns
  - Intellectual history
  - Language variation across period and region
  - Dynamics in mental health conditions
  - Parliamentary discourse
- Text and speech as **social** and

Case 1

Multidisciplinary collaboration with social scientists on CMC and social media data (content + context)

- Computational Sociolinguistics*
- Computational Social Sciences
- Domain experts (political science, finance, religion studies, …)
- ….

- Methods
  - Text classification
  - Sentiment analysis
- Auxiliary language resources
  - Annotated corpora of CMC data for training
  - Lexica

* Dong Nguyen et al, Computational Sociolinguistics: A survey. In: Computational Linguistics, Vo

Case 2

- Multidisciplinary collaboration with cultural history on literary data to determine the authorship of texts (content + context)
  - Cultural historians
  - Literary studies
  - Digital Humanities
  - ….
- Methods
  - Text classification
- Auxiliary language resources
  - Text corpus with data from multiple authors

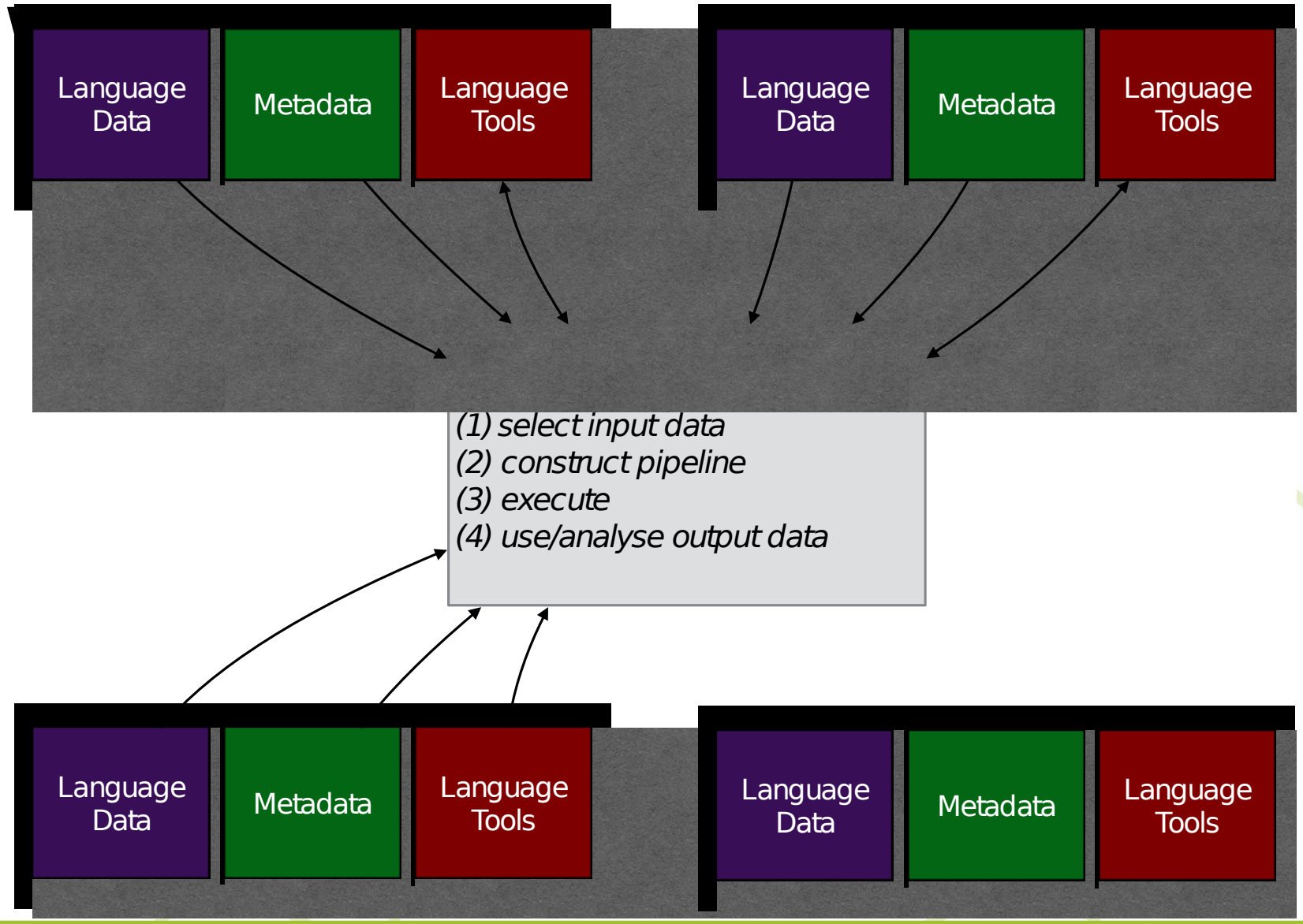
Case 3

Multidisciplinary collaboration with social scientists on written and spoken language to develop detection tools for early symptoms of dementia (content + context)

- Computational Sociolinguistics
- Computational Social Sciences
- Psycholgists
- Specialists in health and ageing
- ...

- Methods
  - Speech recognition
  - Sentiment analysis
- Auxiliary language resources
  - Annotated corpora with spoken material from elderly people to train speech and language models

The CLARIN data architecture:



From big data to big decisions

Uptake of the results of data analysis is dependent on

- **transparency** of the algorithms applied
- **explainability / interpretability** of the results is a precondition for integration
- reproducability
- insight in the **validity** of outcomes

CLARIN pour les chercheurs*

- Trouver et déposer des ressources
  - Chercher par métadonnées, chercher par contenu
  - Conservation à long terme, PID (Persistent Identifier)
- Outils avancés et installations informatiques
- Connexion fédérée - accès facile
- Expertise pour les chercheurs
  - Helpdesks, Centres de connaissances, Comités
- Collaboration internationale
- Tutoriels en ligne et exemples d'utilisation
  - Toujours en cours
- Ateliers et séminaires, bourses de mobilité
- Prise en charge des plans de gestion des données

* slide by Bente Maegaard

see you @

www.clarin.e
or
f.m.g.dejong

*

CLARIN

* prints can be ordered here: https://www.clarin.eu/value-pro