CLARIN - DK
Common Language Resources and Technology Infrastructure

CLARIN

The Role of ERIC Research Infrastructures in Advancing FAIR and Open Science: CLARIN

Costanza Navarretta
Centre for Language Technology,
Department of Nordic Studies
and Linguistics

costanza@hum.ku.dk

UNIVERSITY OF COPENHAGEN

CLARIN

- a distributed digital infrastructure offering data, tools and services to support research based on language resources
- an ESFRI landmark
- a network of researchers and IT specialists

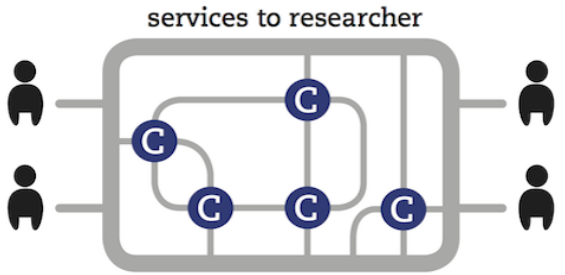




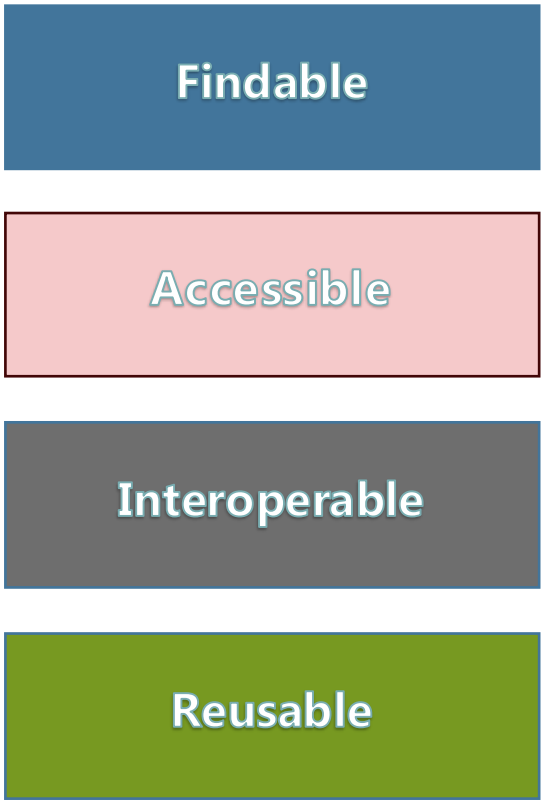
CLARIN

offers:

- expertise from 26+ member countries and 75+ CLARIN centres
- digital and FAIR language resources
- language processing tools
- archiving and depositing services
- training and course materials
- partner in EU funded projects
- contributes to EOSC



FAIRNESS: key elements



Key elements

- **Persistent Identifiers (PIDs)**
- **Data Management Plan**
- **Metadata**
- **Licenses**
- **Certified Repositories**

FAIR and CLARIN

Findable

- a search portal for language data across centers, based on harvested metadata (Virtual Language Observatory, Virtual Collection Registry)
- search in multiple corpus search engines at the same time (Federated Content Search)
- search in Resource Families

Accessible

- CLARIN B centers provide a certified repository including support for Persistent Identifiers and metadata
- federated login system, when password-protected access is needed (CLARIN-DK uses WAYF)

FAIR and CLARIN cont.

Interoperable:

- Help discovering which tools can process a language data set (Language Resource Switchboard and knowledge centres)
- rich vocabularies
- Standards for Language Resource Technologies
- Standards Information System (SIS) with overview of standards and formats for data deposition at CLARIN centres

Reusable:

CLARIN centres (hereunder CLARIN-DK) provide:

- recommendations on licenses (Creative Common and other), formats and standards – license type given for each resource
- metadata curation and continuous metadata quality and link checks

All resources are findable, but not all are open

- More than 350,000 resources in CLARIN (more than 50% of resources in CLARIN-DK) are open to everybody
- The rest (approx. 50%) are open for academic use (institutional login) or have restricted use (contact the provider).

CLARIN-DK
Common Language Resources and Technology Infrastructure

CORE TRUST SEAL

- the Danish part of CLARIN
- hosted (and the past years funded) by The Faculty of Humanities at the University of Copenhagen
- manpower and expertise at the Centre for Language Technology (CST) part of the Department of Nordic Studies and Linguistics
- a certified B-centre that provides depositing service and online tools
- part of 3 certified knowledge centres (in Denmark: DANSK with the Danish Language Council)

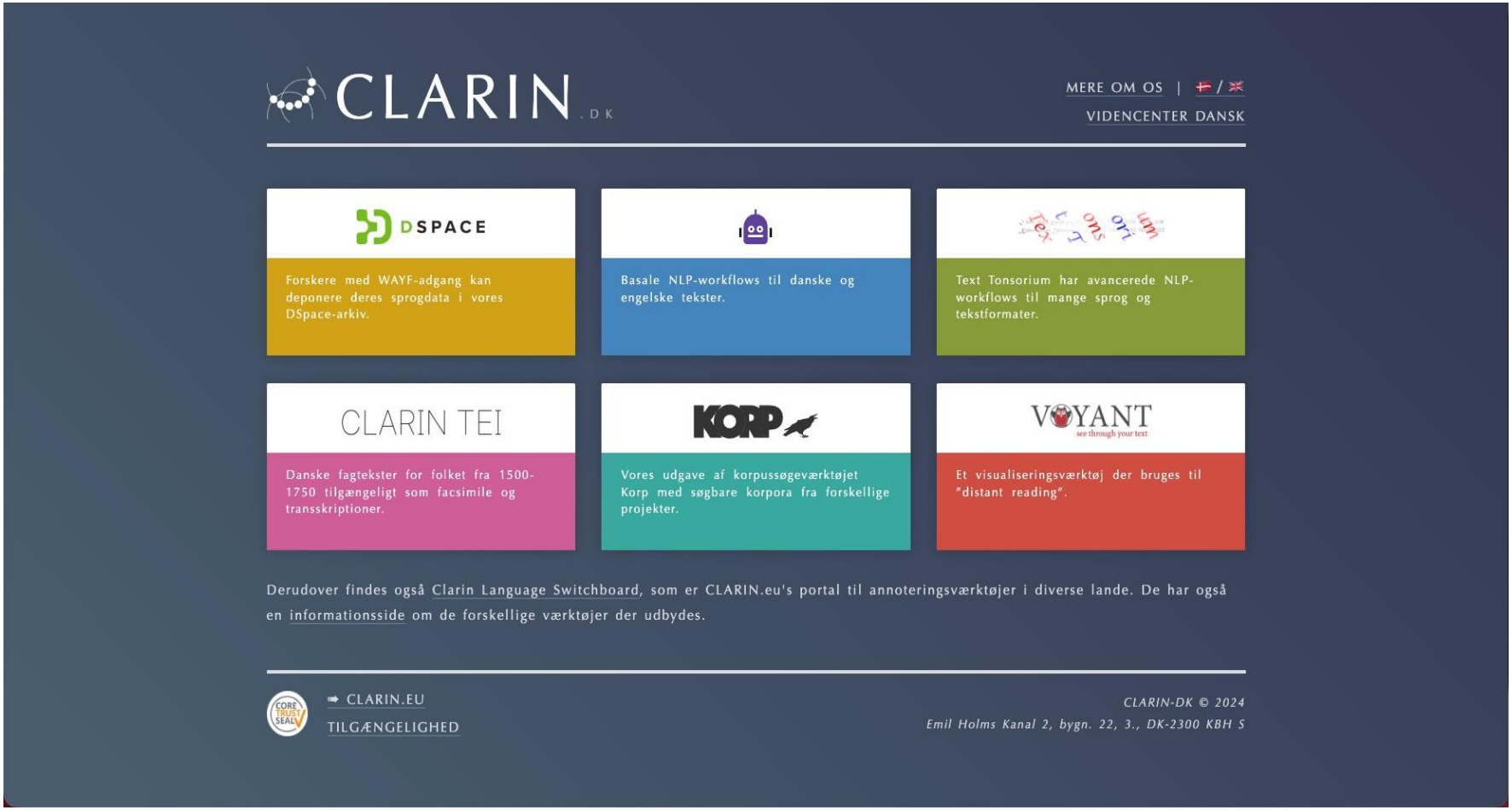
CLARIN K-centre

- part of CLARIN expert groups (e.g. technical, legal, users)

History of CLARIN-DK

- DK-CLARIN project (2008-2011) with University of Copenhagen, Aalborg University, Aarhus University, The Royal Library, The Danish Literature Society, Copenhagen Business School, Danish Language Council, National Museum
- DK in CLARIN 2012 with the DigHumLab project (2012-2022) Aarhus University (DARIAH), University of Copenhagen (CLARIN), University of Southern Denmark, Aalborg University (VILA), The Royal Library
- 2023- funded by the University of Copenhagen

The CLARIN-DK services (www.clarin.dk)



Data Management in CLARIN

European level

- A data management knowledge center
- Assessment (Core Seal)
- Wizard suggesting how to create a data management plan
- Cooperation with e.g., EOSC and ISO.

Country level:

- Different strategies
- In Denmark, most institutions have their own data management systems/procedures.

CLARIN for Open Science

- Promotion of the sharing and re-usage of language resources through sustainable data registries
- Enhancement and deployment of interoperability of language data and services:
  - common metadata framework
  - distributed network of FAIR certified data repositories for language data.
- Promotion of:
  - comparative perspectives
  - multidisciplinary collaboration
  - transnational research
  - responsible data science
- Support for linguistic diversity:
  - data and tools covering many languages
  - language resources in all modalities
  - discipline- and language-agnostic
- Participates in Open Science projects

Some lessons from CLARIN

- Metadata must be as simple as possible and relevant to the research community
- The infrastructure must continuously adapt to new (versions of) standards, formats, tools, methods
- It is essential to take into consideration the individual researchers and their specific needs and requirements - variety of research fields in SSH
- At present, focus on evaluation and benchmarking of resources produced with LLMs and guidance on their use

Funds to existing infrastructures in Denmark

- The Danish Ministry of of Higher Education and Science is formally responsible for the ERIC infrastructures but only funds them the first five years + five years of self-funding.
- In other countries, the responsible ministries continuously fund the ERIC infrastructures totally or partially (after an evaluation or new applications).
- Therefore, the economy of CLARIN in the different countries varies.
- Some activities in CLARIN-DK are done under projects.