

Felles infrastruktur for norske og europeiske språkdata-baser

Med CLARINO realiseres en felles infrastruktur for norske språkdata-baser som kan kobles opp mot europeiske baser i ESFRI-prosjektet CLARIN. Dette åpner for store forskningsmuligheter innenfor språkvitenskap og er nyttig for flere fag innenfor humaniora og samfunnsvitenskap.



Prosjekt:
CLARINO – Common Language Resources and Technology Infrastructure Norway

Mer informasjon:
www.clarin.eu
<http://clarin.b.uib.no>

Infrastrukturen er et viktig verktøy for forskning på språk som studieobjekt og som bærer av innhold. Den vil ha betydning for forskning, utvikling og undervisning innenfor en rekke felt som ikke bare omfatter språkteknologi og språkvitenskap, men også andre fagfelt. CLARINO er et viktig ledd i arbeidet med å bygge opp historiske og nåtidige elektroniske språkkressurser.

Norske forskere vil få tilgang til et mangfold av norske og europeiske språkdata-baser. Data som tidligere var vanskelige

å få tak i, blir dermed tilgjengelige for mange forskningsmiljøer. Norge har allerede sterke fagmiljøer innenfor språkvitenskap, og denne infrastrukturen vil kunne bidra til internasjonalt ledende forskning.

Når norske språkdata-baser kobles opp mot europeiske baser, åpner det seg store muligheter for interessante komparative studier. Man kan sammenligne språkbruk både historisk og over landegrensene. Store tekstbaser vil bli søkbare. Når man søker på norske ord, vil man kunne få svar på betydning i andre språk til ulike tider. Samtidig vil de norske dataene bli tilgjengelige for det internasjonale forskningssamfunnet.

CLARINO vil ha betydning for forskningsmiljøer og bedrifter som lager språkteknologiske systemer for å organisere kunnskap og utvikle terminologi. Slik kan man utvikle programvare beregnet på det norske markedet og sørge for at norske hjelpemidler og løsninger for universell utforming blir tilgjengelige. Miljøer som utvikler flerspråklige teknologier, som f.eks. maskinoversettelse, vil også kunne dra nytte av infrastrukturen.

Gjennom et felles kodespråk vil infrastrukturen kunne legge til rette for såkalt semantisk web. Det betyr at store datamengder fra svært mange kilder kan knyttes sammen tematisk og bli allment søkbare på nettet.

ESFRI-node

Prosjektet utgjør den norske noden i ESFRI-prosjektet CLARIN-ERIC.

Nasjonal koordinator

Universitetet i Bergen

Partnere

Alle sentrale språkvitenskapelige miljøer i Norge er partnere i prosjektet, og bidrar med årsverk og utstyr. Nasjonalbiblioteket og UNINETT AS inngår også i samarbeidet.

Finansiering

Forskningsrådet har gitt 25 millioner kroner til utvikling av CLARINO, mens konsortiet bidrar med 16,3 millioner kroner i egenfinansiering. 1,6 millioner kroner kommer fra andre kilder. Dette dekker etableringen og medlemskostnader i den europeiske infrastrukturen CLARIN-ERIC. Kostnadene dekkes deretter av konsortiet.

Tidsplan

CLARINO skal etableres i løpet av fem år og stå ferdig i 2017.

Single sign-on access to language data repositories and tools to explore, exploit, analyse, enrich or combine them

TYPE: distributed
COORDINATING COUNTRY: NL
MEMBER COUNTRIES/ENTITY: AT, BG, CZ, DE, DK, EE, EL, FI, IT, LT, NL, NO, PL, PT, SE, SI, (DLU)

PARTICIPANTS: ES, UK, US

TIMELINE

- ESFRI Roadmap entry: 2006
- Preparation phase: 2008-2011
- Construction phase: 2011-2015
- Operation start: 2012
- Legal status: ERIC, 2012

ESTIMATED COSTS

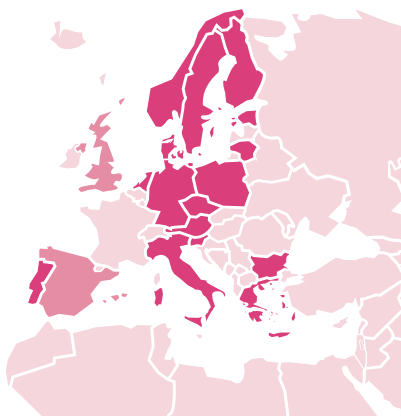
- Capital value: Not Available
- Operation: 12 M€/year

HEADQUARTERS

CLARIN ERIC
 Utrecht University
 Utrecht
 The Netherlands

WEBSITE

<http://www.clarin.eu>



THE NETHERLANDS

CLARIN ERIC

Common Language Resources and Technology Infrastructure



Description

The Common Language Resources and Technology Infrastructure (CLARIN) provides easy and sustainable access for scholars in the humanities and social sciences to digital language data — in written, spoken or multimodal form — and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of their location. To this end CLARIN is building a networked federation of language data repositories, service centres and centres of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. CLARIN integrates existing data and service centres, without major capital investments.

Currently the CLARIN infrastructure is under construction, but a growing number of participating centres is already offering services to data, tools and expertise. Since 2012, CLARIN is a European Research Infrastructure Consortium (CLARIN ERIC).

Activity

The operation, maintenance and continuous expansion of the infrastructure is carried out in the participating countries at the individual CLARIN centres (currently over 30), which have to meet clearly defined technical and organizational criteria, ensuring the coherence of the whole infrastructure. CLARIN also works closely together with research communities in creating and expanding a knowledge infrastructure that can support developers of language resources and tools, as well as the end-users of the available data and services. In total over 100 institutes across Europe are involved. CLARIN has been offering operational access and services since 2012, and the offer will grow as centres expand and new countries join. To stimulate the uptake and increase the insight in the usability of the services, CLARIN participates in the development of courseware and organises workshops and data camps focussing on the processing and analysis of specific data types, such as parliamentary data, social media data, and historical texts. The collaboration with other RIs that support scholarly communities in Social Sciences and Humanities (SSH) takes places via the collaboration in H2020 consortia. CLARIN promotes open access to research data within the present legislation and ethical standards and will develop outreach activities to stimulate the reuse and repurposing of the data resources integrated.

Impact

CLARIN stimulates the reuse and repurposing of available research data, thereby enabling scholars in SSH (including digital humanities) to increase their productivity and, more importantly, open new research avenues in and across disciplines that address one or more of the multiple societal roles of language: as a carrier of cultural content and information, both synchronically and diachronically, as a reflection of scientific and societal knowledge, as an instrument for human communication, as one of the central components of the identity of individuals, groups, cultures or nations, as an instrument for human expression, or as an object of study or preservation. Through the access and discovery services, CLARIN increases the potential impact of data and tools produced with publicly funded projects. Working with CLARIN data and tools will increase the skill levels for data analysis tasks among the new generations of SSH students, which is likely to be welcomed by the data science sector.