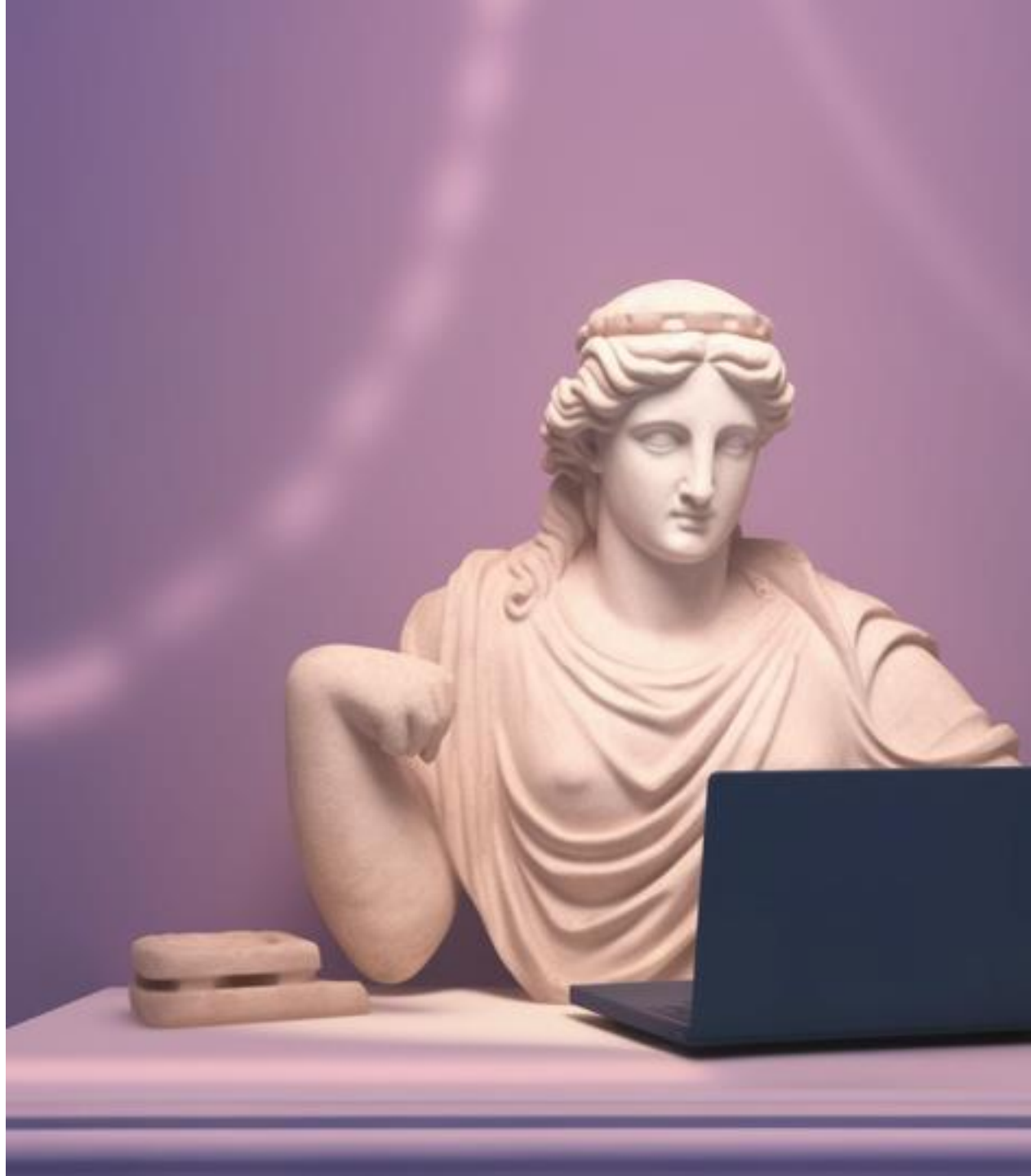
Humit – senter for digital utvikling på HF

CMDI – metadataformat for språkressurser

Kristin Hagen

23. april 2024



UNIVERSITETET
I OSLO

Humit, CLARIN, CLARINO og CMDI

- Humit – senter for digital utvikling på HF:
  - Ny enhet fra 1.1 2023 der Tekstlaboratoriet inngår.
  - Kompetansesenter for IT innen humaniora som tilbyr skreddersydde digitale løsninger for humanistisk forsking. https://www.hf.uio.no/humit/
- CLARIN:
  - Europeisk digital infrastruktur som bl.a tilbyr språkressurser, språkdata og verktøy.
- CLARINO:
  - Den norske delen av CLARIN som har fått infrastrukturstøtte av NFR i to omganger (2012-2016 og 2020-2023). CLARINO er ledet av UiB med bl.a Nasjonalbiblioteket, NHH, UiT og UiO ved Tekstlaboratoriet som partnere. Tekstlaboratoriet er et såkalt C-senter i CLARIN.
- CMDI:
  - Metadataformatet i CLARIN. Metadata høstes fra de ulike senterne til søkbare kataloger:

https://vlo.clarin.eu/?2

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or continue to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

See all records | Take a quick tour

Search through 1,282,329 records

https://www.nb.no/sprakbanken/ressurskatalog/

Nasjonalbiblioteket | Språkbanken

English

Søk i ressurskatalogen ...

1-10 av 29 treff 1 2 3 Neste

Per side

Sist oppdatert

ORIGIN: CLARINO Tekstlaboratoriet

Type

- Leksikon 1
- Tale 10
- Tekst 23
- Verktøy 5
- Video 7

Opphav

- CLARINO Bergen 0
- CLARINO Tekstlaboratoriet 29

VERKTØY 11.01.2024

Glossa

Glossa is a tool for researchers who want to search linguistically annotated corpora. Glossa is designed to make it easy for researchers to: - create complex searches - explore the result via e.g. ...

Språk:
Opphav: CLARINO Tekstlaboratoriet
Lisens: MIT license

TALE, TEKST 01.12.2022

LIA-trebanken

LIA-trebanken består av 7536 talemålssegment og 77 701 ord/token frå talespråkskorpuset LIA norsk. Trebanken er annotert morfologisk og syntaktisk og manuelt korrigert. LIA-trebanken er ...

Språk: norsk, nynorsk
Opphav: CLARINO Tekstlaboratoriet
Lisens:
Creative_Commons-BY-NC-SA (CC-BY-

Hvorfor bruke CMDI

- Utviklet for språkressurser
- Fleksibelt
- Egne norske moduler (som Tekstlaboratoriet var med på å utvikle)
- Arbeidet ble finansiert av CLARINO-prosjektet
- Enkelt å redigere metadata med editoren COMEDI utviklet av CLARINO-prosjektet ved UiB

CLARINO COMEDI CMDI Records Metadata Editor Groups Documentation Admin

https://clarino.uib.no/comedi/home

Resource type [1/1] [CCR]

Indicates the type of the resource

corpus

Identification info [1/1] | ref: - | **Remove content** | **Hide content**

Groups together information needed to identify the resource. Derived from clarin.eu:cr1:c_1349361150743 (identificationInfo by META-SHARE 3.0)

Resource name [1/2] [CCR] +

The full name by which the resource is known; the element can be repeated for the different language versions using the "lang" attribute to specify the language.

The BigBrother Corpus lang: en

BigBrother-korpuset lang: nb

Description [1/2] [CCR] +

Provides the description of the resource in prose

BigBrother-korpuset er et talespråkskorpus som består av den første sesongen av realityserien BigBrother som ble sendt på TVNorge våren lang: nb

The BigBrother Corpus is a speech corpus with recordings from the first season of the BigBrother show, sent on Norwegian television by lang: en

Resource short name [0-1/1] [CCR] +

The short form (abbreviation, acronym etc.) used to identify the resource; the element can be repeated for the different language versions using the "lang" attribute to specify the language.

BigBrother lang:

Url [0-∞/1] [CCR] +

A URL used as homepage of an entity (e.g. of a person, organization, resource etc.) and/or where an entity (e.g. LR, document etc.) is located

http://www.tekstlab.uio.no/nota/bigbrother/index.html

description:

PID [0-1/1] [CCR]

A globally unique, persistent identifier for the resource

http://hdl.handle.net/11538/0000-0005-E7C1-C

Mer om CMDI (Component Metadata Infrastructure)

- Utviklet ved Max Planck Institute og er valgt som standard for metadata i CLARIN.
- Fleksibelt: kan velge mellom eksisterende profiler og komponenter som ligger i et søkbart arkiv. Eller man kan lage sine egne og laste opp til arkivet. Også felles begreper (shared concepts) som man bør bruke.
- CLARINO-prosjektet har lagd egne profiler og komponenter blant annet for:
  - Korpus
  - Leksikografiske ressurser
  - Språkteknologiske verktøy
  - TEI-dokumenter
- Alle profilene har en hovedkomponent for «Resource common info» som skal fylles ut.

Mer om CMDI

- Hovedkomponenten «Resource common info» inneholder komponenter som beskriver:
  - Navn på ressursen
  - Beskrivelse av ressursen
  - Type ressurs
  - Kontaktinfo
  - Distribusjon
  - Lisens (CLARIN har også egne lisenser)
  - PID
  - Versjon
  - Prosjekt

Mer om CMDI

- CLARINO-profilene har i tillegg til hovedkomponenten «Resource common info», komponenter som beskriver ressursen mer spesifikt, for eksempel:
  - Type korpus, type tekster, antall tekster, formater, annotering, antall informanter osv
- Hver profil har mange komponenter og mange felt man kan fylle ut. Men mange felt er ikke obligatoriske, så det er mulig å gjøre metadataregistreringen forholdsvis raskt.
- CMDI-editoren COMEDI kan også klone tidligere registrerte ressurser slik at man bare endrer de feltene som er forskjellige.
- COMEDI gir også forslag til utfylling basert på hva man har registrert tidligere.
- CMDI-filene kan vises i COMEDI, som XML og som OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting).

CMDI som XML

CMDI 1.2 Metadata

Header

- **cmd:MdCreator**: Kristin Hagen
- **cmd:MdCreationDate**:
- **cmd:MdSelfLink**:
- **cmd:MdProfile**: clarin.eu:cr1:p_1407745711925
- **cmd:MdCollectionDisplayName**: Clarino - Textlab

Resources

- **cmd:ResourceProxyList**:
  - **cmd:ResourceProxy** [id='bigbrother-lp']:
    - **cmd:ResourceType** [mimetype='']: LandingPage
    - **cmd:ResourceRef**: http://www.tekstlab.uio.no/nota/bigbrother/index.html
  - **cmd:ResourceProxy** [id='bb-transcriptions']:
    - **cmd:ResourceType** [mimetype='']: Resource
    - **cmd:ResourceRef**: http://www.tekstlab.uio.no/nota/bigbrother/index.html#transkripsjon
  - **cmd:ResourceProxy** [id='bb-corpus']:
    - **cmd:ResourceType** [mimetype='']: Resource
    - **cmd:ResourceRef**: https://tekstlab.uio.no/glossa3/bb
- **cmd:JournalFileProxyList**:
- **cmd:ResourceRelationList**:
  - **cmd:ResourceRelation**:
    - **cmd:RelationType**: transcriptions
    - **cmd:Resource**:
      - **cmd:Role**:
        - **cmd:Resource**:
          - **cmd:Role**:

Components

- **cmdp:corpusProfile**:
  - **cmdp:resourceCommonInfo**:
    - **cmdp:resourceType** [cmd:ref='bb-corpus']: corpus
    - **cmdp:identificationInfo**:
      - **cmdp:resourceName** [cmd:ref='bb-corpus'] [xml:lang='en']: The BigBrother Corpus
      - **cmdp:resourceName** [cmd:ref='bb-corpus'] [xml:lang='nb']: BigBrother-korpuset
      - **cmdp:description** [cmd:ref='bb-corpus'] [xml:lang='nb']: BigBrother-korpuset er et talespråkskorpus som består av den
      - **cmdp:resourceShortName**: BigBrother
      - **cmdp:url**: http://www.tekstlab.uio.no/nota/bigbrother/index.html
      - **cmdp:PID**: http://hdl.handle.net/11538/0000-0005-E7C1-C
    - **cmdp:distributionInfo**:
      - **cmdp:licenceInfo** [cmd:ref='bb-corpus']:
        - **cmdp:userCategory**: Academic
        - **cmdp:distributionAccessMedium**: accessibleThroughInterface
        - **cmdp:executionLocation** [cmd:ref='bb-corpus']: https://tekstlab.uio.no/glossa3/bb
        - **cmdp:licence**:
          - **cmdp:licenceFamily**: CLARIN
          - **cmdp:licenceName**: CLARIN_ACA-NC-LOC-PRIV-ND-*
          - **cmdp:licenceURL**: https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEulaAca?ID=1&AFFIL=EDU&BY=1&NC=1&LOC=1&PRIV=1&NORED=1&ND=1
          - **cmdp:conditionsOfUse**: *
          - **cmdp:conditionsOfUse**: BY
          - **cmdp:conditionsOfUse**: ID
          - **cmdp:conditionsOfUse**: LOC
          - **cmdp:conditionsOfUse**: NC
          - **cmdp:conditionsOfUse**: ND
          - **cmdp:conditionsOfUse**: NORED
          - **cmdp:conditionsOfUse**: PRIV
          - **cmdp:nonStandardConditionsOfUse**: The corpus has audio and video recordings classified as personal data.The production company Nordic Entertainment has generously given their consent to the usage of the videos as a speech corpus. Every individual researcher is responsible for treating the participants with resepct and sincerity. Furthermore, the informants in the corpora should be anonymized, e.g. by changing their names, in every published paper or other output.
        - **cmdp:licensor**:
          - **cmdp:actorInfo**:
            - **cmdp:actorType**: organization
            - **cmdp:organizationInfo**:
              - **cmdp:organizationName** [xml:lang='nb']: Universitetet i Oslo
              - **cmdp:organizationName** [xml:lang='en']: University of Oslo
              - **cmdp:organizationShortName** [xml:lang='nb']: UiO
              - **cmdp:organizationShortName** [xml:lang='en']: UoU
              - **cmdp:departmentName** [xml:lang='en']: Department of Linguistics and Scandinavian Studies
              - **cmdp:departmentName** [xml:lang='nb']: Institutt for lingvistiske og nordiske studier
            - **cmdp:communicationInfo**:
              - **cmdp:email**: tekstlab-post@iln.uio.no
              - **cmdp:url**: http://www.hf.uio.no/iln/om/organisasjon/tekstlab/
              - **cmdp:address**: Box 1102 Blindern
              - **cmdp:zipCode**: 0317

CMDI som OAI-PMH


```
−<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2024-04-22T11:45:37Z</responseDate>
  <request verb="GetRecord" identifier="oai:clarino.uib.no:bigbrother" metadataPrefix="cmdi">https://clarino.uib.no/oai</request>
 −<GetRecord>
  −<record>
   −<header>
      <identifier>oai:clarino.uib.no:bigbrother</identifier>
      <datestamp>2024-01-05T12:30:53Z</datestamp>
      <setSpec>Tekstlab</setSpec>
    </header>
   −<metadata>
    −<cmd:CMD CMDVersion="1.2" xsi:schemaLocation="http://www.clarin.eu/cmd/1 https://infra.clarin.eu/CMDI/1.x/xsd/cmd-envelop.xsd http://www.clarin.eu/cmd/1/profiles/
     clarin.eu:cr1:p_1407745711925 https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1407745711925/1.2/xsd">
     −<cmd:Header>
        <cmd:MdCreator>Kristin Hagen</cmd:MdCreator>
        <cmd:MdCreationDate/>
        <cmd:MdSelfLink/>
        <cmd:MdProfile>clarin.eu:cr1:p_1407745711925</cmd:MdProfile>
        <cmd:MdCollectionDisplayName>Clarino - Textlab</cmd:MdCollectionDisplayName>
      </cmd:Header>
     −<cmd:Resources>
      −<cmd:ResourceProxyList>
       −<cmd:ResourceProxy id="bigbrother-lp">
          <cmd:ResourceType mimetype="">LandingPage</cmd:ResourceType>
         −<cmd:ResourceRef>
            http://www.tekstlab.uio.no/nota/bigbrother/index.html
          </cmd:ResourceRef>
        </cmd:ResourceProxy>
       −<cmd:ResourceProxy id="bb-transcriptions">
          <cmd:ResourceType mimetype="">Resource</cmd:ResourceType>
         −<cmd:ResourceRef>
            http://www.tekstlab.uio.no/nota/bigbrother/index.html#transkripsjon
          </cmd:ResourceRef>
        </cmd:ResourceProxy>
       −<cmd:ResourceProxy id="bb-corpus">
          <cmd:ResourceType mimetype="">Resource</cmd:ResourceType>
          <cmd:ResourceRef>https://tekstlab.uio.no/glossa3/bb</cmd:ResourceRef>
        </cmd:ResourceProxy>
      </cmd:ResourceProxyList>
      <cmd:JournalFileProxyList/>
     −<cmd:ResourceRelationList>
      −<cmd:ResourceRelation>
```