The CLARIN-D Centre Infrastructure

Thomas Zastrow
Introduction

• The decentralized character of CLARIN-D
• Integration of the technical pillars of CLARIN
• The elements of the CLARIN-D centre infrastructure
CLARIN-D: 9 Centres

- University of Tübingen - Annotated corpora, ling. knowledge components and web services
- University of Leipzig - Web services and specific reference corpora
- BBAW Berlin - German language, lexica, diachronic corpora
- University of Stuttgart - Corpora and corpus tools, adaptable tools and web services
- IDS Mannheim - German language, large corpora for German
- LMU Munich - German speech and multimodal data, phonetic tools and services
- MPI Nijmegen - Minority languages, multimedia and multimodal data
- University of Hamburg - Multilingual speech corpora, transcription tools
- Saarland University - Multilingual corpora and corpus tools
3 Computing Centres

- Rechenzentrum der Max-Planck-Gesellschaft am Max-Planck-Institut für Plasmaphysik (RZG)
- Forschungszentrum Jülich (FZJ)
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)
Decentralized Infrastructure
What does it mean to be a CLARIN-D centre?

• From a technical perspective: implementing the pillars of the infrastructure
• Differentiation between several types of CLARIN centres:
  - **B**: These centres are full members of the CLARIN federation; however, they still act as individual centres not taking over responsibilities for the federation
  - **A**: These centres take over responsibility for helping to manage the federation in one way or another. These centres normally extend their services described under Type B Centres
Repositories

- Archives for digital data
- Metadata in form of CMDI files is offered via OAI-PMH
- PIDs are used to identify individual data sets

One resource, one digital object (can contain many files), one PID, one CMDI file
Index Thomisticus Treebank

[en] Index Thomisticus Treebank

Class: Corpus
PID: http://hdl.handle.net/11858/00-1778-0000-0008-C024-C
Legal Owner: Centro Interdisciplinare di Ricerche per la Computarizzazione dei Segni dell’Espressione (CIRCSE)

General Info

[en] The Index Thomisticus Treebank is the syntactically annotated subset of the Index Thomisticus corpus. The Index Thomisticus corpus features the opera omnia by Thomas Aquinas in Latin, for a total of around 11 million words. The Index Thomisticus Treebank is a dependency-based treebank, whose annotation style resembles that of the Prague Dependency Treebank of Czech.

Creators

Dr. Marco Passarotti

Access

[en] Free for academic use

Please go to the following page to find out more: http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/index-thomisticus-baumbank.html

Files

Type: Resource
PID: http://hdl.handle.net/11858/00-1778-0000-0008-C024-C@ds1

Technical Information

Encoding: ISO-8859-1

Metadata by: Marco Passarotti, 2012-11-07
CMDI & PIDs

<xml version="1.0" encoding="UTF-8"?>
  <xsl:output method="xml" indent="yes" encoding="UTF-8"/>
  <xsl:strip-space elements="*"/>
  <xsl:template match="/">
    <xsl:element name="cmd">
      <xsl:element name="cmdheader">
        <xsl:element name="cmdcreator">Marco Passarotti</xsl:element>
        <xsl:element name="cmdcreationdate">2012-11-07</xsl:element>
        <xsl:element name="cmdinternalid">http://hdl.handle.net/11858/00-1778-0000-0008-C024-C</xsl:element>
        <xsl:element name="cmdprofile">clarin.eu:crp_12904316945804</xsl:element>
        <xsl:element name="cmdcollectiondisplayname">Tübingen Curated Resource</xsl:element>
      </xsl:element>
      <xsl:element name="cmdresources">
        <xsl:element name="cmdresourceproxylist">
          <xsl:element name="cmdresourceproxy" id="cmdlcp">
            <xsl:element name="cmdresourcetype">LandingPage</xsl:element>
          </xsl:element>
        </xsl:element>
        <xsl:element name="cmdresourceproxy" id="ds1">
          <xsl:element name="cmdresourcetype">Resource</xsl:element>
          <xsl:element name="cmdresourceref">http://hdl.handle.net/11858/00-1778-0000-0008-C024-C#ds1</xsl:element>
        </xsl:element>
      </xsl:element>
      <xsl:element name="cmdjournalfileproxylist">
        <xsl:element name="cmdjournalfileproxy"/>
        <xsl:element name="cmdjournalfileproxy"/>
      </xsl:element>
      <xsl:element name="cmdresourcerelationlist">
        <xsl:element name="cmdresourcerelation" type="Res1"/>
      </xsl:element>
      <xsl:element name="cmdispartoflist">
        <xsl:element name="cmdispartof"/>
      </xsl:element>
    </xsl:element>
    <xsl:element name="cmdcomponents">
      <xsl:element name="cmdgenerallinfo">
        <xsl:element name="cmdresourcename" xml:lang="en">Index Thomisticus Treebank</xsl:element>
        <xsl:element name="cmdresourcetitle" xml:lang="en">Index Thomisticus Treebank</xsl:element>
        <xsl:element name="cmdresourceclass">Corpus</xsl:element>
        <xsl:element name="cmdlifecyclestatus">development</xsl:element>
        <xsl:element name="cmdstartyear">2012</xsl:element>
        <xsl:element name="cmdtimecoverage" xml:lang="en">Lifespan of Thomas Aquinas (1225-1274)</xsl:element>
        <xsl:element name="cmdlegalowner" xml:lang="it">Centro Interdisciplinare di Ricerca per la Computerizzazione dei Sogni dell’Espressione (CIRCEE)</xsl:element>
        <xsl:element name="cmdlocation"/>
      </xsl:element>
    </xsl:element>
  </xsl:template></xsl:stylesheet>
The CLARIN Center Assessment is done on the European level by the Centre Assessment Committee:

The Centre Assessment Committee (CAC) is a permanent Committee that reports to the Board of Directors (BoD). Its main responsibility is to assess the level and quality of CLARIN Centres of type A, B and E, as described in the Centre Types Document.
Center Assessment

• All 9 CLARIN-D centres applied for and where successfully accessed as Type B centres in May 2013.
Part of the CLARIN Center Assessment is also an external evaluation: the Data Seal of Approval (DSA):

The Data Seal of Approval ensures that in the future, research data can still be processed in a high-quality and reliable manner, without this entailing new thresholds, regulations or high costs. The Data Seal of Approval and its quality guidelines may be of interest to research institutions, organizations that archive data and to users of that data.
Data Seal of Approval

• The Data Seal of Approval Guidelines consists of 16 requirements
• They cover the technical, but also the formal requirements for a sustainable data repository
Both the operator of the repository as well as the users are involved:

- The data producer deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms.
- The data repository enables the users to utilize the research data and refer to them.
Data Seal of Approval

• Legal issues are covered:
  – The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

• In CLARIN-D: depositor agreement
Data Seal of Approval

- All 9 CLARIN-D centres made an application for a Data Seal of Approval and were successfully assessed in May 2013
Computing Centres

• In addition to the nine centres, three computing centres are involved in CLARIN-D:
  – Basic services for the CLARIN-D infrastructure are hosted
  – Large computing facilities and data storage are provided
Monitoring: checks the availability of all CLARIN-D services and resources and informs the responsible administrator of any problems.
Personal Workspaces: online storage for the CLARIN-D community (FZJ)
PID services through EPIC consortium (GWDG)
Computing Centres - RZG

- Web service deployment
- „Drop Off“ – temporary storage for audio/video data (RZG)

<table>
<thead>
<tr>
<th>Stuttgart Dependency Parser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desc: Stuttgart Dependency Parser</td>
</tr>
<tr>
<td>Creator: IMS: University of Stuttgart</td>
</tr>
<tr>
<td>Contact: <a href="mailto:andre.blessing@ims.uni-stuttgart.de">andre.blessing@ims.uni-stuttgart.de</a></td>
</tr>
<tr>
<td>URL: <a href="http://ws1-clarind.esc.rzg.mpg.de/webservice-parser">http://ws1-clarind.esc.rzg.mpg.de/webservice-parser</a></td>
</tr>
<tr>
<td>ID: <a href="http://hdl.handle.net/11858/00-247C-0000-0007-3734-F">http://hdl.handle.net/11858/00-247C-0000-0007-3734-F</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Features</th>
<th>Output Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>type: text/tcf+xml</td>
<td>lemmas</td>
</tr>
<tr>
<td>lang: de</td>
<td>postags.tagset: stts</td>
</tr>
<tr>
<td>tokens</td>
<td>depparsing.tagset: tiger</td>
</tr>
<tr>
<td>version: 0.4</td>
<td>depparsing.multigovs: false</td>
</tr>
<tr>
<td>sentences</td>
<td>depparsing.emptytoks: false</td>
</tr>
</tbody>
</table>
The Centre Registry

- For finding services in a decentralized environment, a centralized registry is necessary.
- For the internet, this is the DNS
- In CLARIN-D: the Centre Registry
The Centre Registry

• Implemented and hosted at RZG
• A database with information about the CLARIN-D centres and the services they provide
• Two interfaces:
  − A web based, human readable one
  − Based on RESTStyle web services for automatic access
# Center Registry

## Group by: National Consortium - Country - Center Type

<table>
<thead>
<tr>
<th>Center</th>
<th>Repository</th>
<th>Status</th>
<th>National Consortium</th>
<th>Metadata</th>
<th>AAM</th>
<th>PID</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLARIN Center Vienna</td>
<td>Fedra based, being setup, planned to go online 2013-06</td>
<td>Aiming for A+B</td>
<td>CLARIN-AT</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>Accession of National Federation (assess) pending</td>
</tr>
<tr>
<td>LINDAT-Clarin</td>
<td>Depaco</td>
<td>Aiming for A+B</td>
<td>LINDAT-CLARIN</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>One Service Provider available</td>
</tr>
<tr>
<td>ASV Leipzig</td>
<td>Fedra Commons</td>
<td>Aiming for B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>One service provider available</td>
</tr>
<tr>
<td>Bayerisches Archiv für Sprachdaten</td>
<td>own system</td>
<td>Aiming for B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>Currently no Service Providers available</td>
</tr>
<tr>
<td>Berliner-Brandenburg Academy of Sciences and Humanities</td>
<td>Fedra Commons</td>
<td>Aiming for B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>Service Provider available</td>
</tr>
<tr>
<td>Eberhard Karls Universität Tübingen</td>
<td>Fedra Commons</td>
<td>Aiming for A+B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>One Service Provider available</td>
</tr>
<tr>
<td>FZJ</td>
<td>E (computing center)</td>
<td>CLARIN-D</td>
<td></td>
<td></td>
<td></td>
<td>PID</td>
<td>ok</td>
</tr>
<tr>
<td>GWDG</td>
<td>E (computing center)</td>
<td>CLARIN-D</td>
<td></td>
<td></td>
<td></td>
<td>PID</td>
<td>own idP</td>
</tr>
<tr>
<td>IMS, Universität Stuttgart</td>
<td>Fedra Commons</td>
<td>Aiming for B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>Currently no Service Providers available</td>
</tr>
<tr>
<td>Institut für Deutsche Sprache</td>
<td>own system (COSMAS II, OWID, Subversion)</td>
<td>Aiming for A+B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>IDP and SPs available</td>
</tr>
<tr>
<td>RZG</td>
<td>E (computing center)</td>
<td>CLARIN-D</td>
<td></td>
<td></td>
<td></td>
<td>PID</td>
<td>ok</td>
</tr>
<tr>
<td>Universität des Saarlandes</td>
<td>Fedra Commons</td>
<td>Aiming for B</td>
<td>CLARIN-D</td>
<td>CMDI</td>
<td></td>
<td>PID</td>
<td>Not available yet</td>
</tr>
</tbody>
</table>
The Centre Registry

- At the moment, 22 CLARIN centres are already listed: not only CLARIN-D centres
- A user or a machine can find information about
  - Administrative information (contacts, addresses...)
  - Web services
  - OAI-PMH endpoints
  - SRU/CQL endpoints
  - ...

The Centre Registry

Center Detail view

Details for center: Eberhard Karls Universität Tübingen

Center Basic Information

Name: Eberhard Karls Universität Tübingen
Type: A
Status: Aiming for A+B
Description: A center in CLARIN-D
Member of: CLARIN-D
Assessment: DSA granted: http://www.sfs.uni-tuebingen.de/ascl/clarin-center/repository.html

Center Extended information

Website: http://www.uni-tuebingen.de/
Repository system: Fedora Commons
PSD Status: handle via EPIC
Strict versioning: True
Web references
  Web Site: http://weblicht.sfs.uni-tuebingen.de/rws/ eru/
  Description: CQL

AAT
  AAT status: One Service Provider available
  CLARIN Infrastructure services: True
  Member of CLARIN service Provider Federation: True
  Member Of National Federation: DE

Metadata
  Metadata Schema: CMDI
  OAI/AccessPoint: http://weblicht.sfs.uni-tuebingen.de/oaiprovider/?verb=Identify
  Metadata Web Service Type: WebData
  Metadata Web Service Type: SOAP
  WebService Set: WeblichtWebService

  <Header/>
  <Resources>
    <ResourceProxyList/>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <CenterProfile>
      <CenterBasicInformation>
        <Name>Eberhard Karls Universität Tübingen</Name>
        <Description>A center in CLARIN-D</Description>
        <Status>Aiming for A+B</Status>
        <Country>
          <Code>DE</Code>
        </Country>
      </CenterBasicInformation>
      <CenterExtendedInformation>
        <Website>http://www.uni-tuebingen.de/</Website>
        <RepositorySystem>http://repository.tuebingen.de</RepositorySystem>
        <PSDStatus>http://weblicht.sfs.uni-tuebingen.de/oaiprovider/?verb=Identify</PSDStatus>
        <MetadataWebServiceType>WebData</MetadataWebServiceType>
        <MetadataWebServiceType>SOAP</MetadataWebServiceType>
        <WebServiceSet>WeblichtWebService</WebServiceSet>
      </CenterExtendedInformation>
    </CenterProfile>
  </Components>
</CMD>
## Building Blocks

### (Web) Applications
- TCF Visualiz.
- TLA
- WebLicht
- FCS
- WebAnno
- VLO
- Annex
- WebMaus
- PWS

### Protocols and Specifications
- OAI-PMH
- PID
- AAI
- SRU/CQL
- ISOCat

### Dataformats
- CMDI
- TEI
- TCF

---

Nine CLARIN-D Data Repositories

---

CLARIN-D M24 Workshop, Nijmegen, 27. June 2013
Thank you.