the right toolbox
Index

▸ ACDH-OEAW introduction
▸ The right toolbox
  ▹ Reference model
  ▹ Use cases
▸ What else is happening?
  ▹ DARIAH
  ▹ PARTHENOS
ACDH-OEAW - The New Centre

- 2014: ACDH Project; Planning & Preparation
- 1 January 2015 official foundation as 29th institute of the Academy
- April 2015 operational
- Team expansion (100 applications, 40 interviews, ~20 new colleagues)
- Current number of staff: ~50
- 12-15 “technical” (sys admins, software development, data curation)

- Research & Service!
The Four Pillars

- DH research
- Technical services & expertise
- Social Infrastructure
- Knowledge Transfer
4 Working groups

- Networks & Outreach
- Tools, Service, Systems
- Data, Resources and Standards
- eLexicography
Position inside Academy
ACDH-OEAW coordinating the national consortium for CLARIN and DARIAH

Focus on commonalities and synergies right from the start

Some 13 institutes from 6 institutions

Contributing In-Kinds

Engaging on EU-level

Knowledge Sharing/Exchange

http://digital-humanities.at/ as common „platform“
CLARIAH-AT - European Involvement

DARIAH
- Co-head VCC1
- Task leaders in VCC 1 + 2
- Participation in NCC, JRC, SMT
- Participation in 10 working groups

CLARIN
- Involvement in the development of core infrastructures (CMDI, FCS)
- Participation in all the bodies (NCF, SCCTC, TFs: CMDI, FCS, Curation, CCR ...)
- Running a CLARIN Centre
- Hosting/Providing data
- Developing/Providing tools
Research Infrastructures

Work of ACDH-OEAW deeply embedded in & contributing to RIs

▸ **DARIAH** (Working groups)
  ▷ thesaurus maintenance - Backbone Thesaurus
  ▷ guidelines and standards, ...

▸ **CLARIN**
  ▷ CLARIN Centres ([CLARIN Centre Vienna](#))
  ▷ Depositing service, NLP services
  ▷ Metadata Catalogue - [VLO](#)

▸ **PARTHENOS**
  ▷ Cluster of Humanities and Cultural Heritage RIs
  ▷ Common semantic framework
  ▷ Architecture / Framework for service provisioning

▸ **EUDAT 2020**
  ▷ storage for research data
DHA website - relaunch

http://digital-humanities.at/ (still beta)

See especially 5+7 DH projects from 2 calls in 2014:
  ▶ Go!digital
  ▶ Long-term projects on cultural heritage

ACDH-OEAW involved in 3 (PI) + 2 (SP)
DHA website - relaunch

▶ beta version / work in progress
▶ Expand to a comprehensive information resource (*Knowledge Hub*)
  ▶ Overview of DHA projects
  ▶ Bibliography (maintained as a Zotero group)
    - Reading
    - Resources & Services
  ▶ Glossary - working definitions of common terms (use for content tagging as well)
▶ Collaborative: inviting contributions from the consortium and the public
AG-1 Networks & Outreach

- Knowledge transfer
  - ACDH Lectures, DH Tool Gallery, DHA-Days

- ESFRI-Liaison
  - CLARIN, DARIAH, PARTHENOS, dariahTeach
  - CLARIAH-AT - coordination, platform digital-humanities.at, DHA-Days
AG-2 Tools, Services & Systems

Provide technical infrastructure

- Provision of services (tools, applications, ...)
- Develop/adopt software
  
  “beg, steal & borrow” approach (quote Marc)

- Maintain the systems
  - Run servers (in cooperation with ARZ – Academy’s computing centre)
  - Virtualisation
  - Databases
  - Low-level data management (backups, access)
  - AAI (SP)
  - Monitoring
AG-3 Data, Resources, Standards

- Data Modelling
  - TEI
  - Relational DBs
  - (RDF)
- Data/Knowledge management
  - Reference resources, Vocabularies
- Content Curation
  - Annotation
- Metadata Curation
- Digital Editions

AG-2 and AG-3 tightly intertwined
AG-4 eLexicography

- Data management and production (Open Access)
  - explore!AT
- Corpus-Dictionary interface
  - AMC
  - TUNICO
- Semantic technologies
  - RDF, SKOS, LEMON
- PPPs
  - LDL4HELT
  - ÖWB
- DictGate (Research platform)
  - DictObserver (Registry, resource locator)
  - ViDi
- Hoist of inter/national cooperations
ACDH projects, resources, services

http://www.oeaw.ac.at/acdh/projects

www.oeaw.ac.at/acdh/resources

https://github.com/acdh-oeaw
the right toolbox
Tools & Services

- Tool:
  *something (as an instrument or apparatus) used in performing an operation* (Merriam Webster)

- Service:
  *the continued, declared willingness and ability of an actor to execute on demand by a client certain activities of specific benefit to the client* (Martin Doerr et al., PARTHENOS Entities)
Reference workflow

new project
- plan
- manage

data management plan

connect & interpret
processing

present
publishing

- explore
- visualize

print
- discover
- cite
- access

online
- publish
- share

new “cycle”

reuse

new project 2
...
new project n

collect
pre-processing

analog material
- digitize

quarantine

legacy data
- acquire
- generate

new data
- discover
- cite
- access

existing data

working data
- control
- document
- prevent

repository

article
- deposit
- licence

store & preserve

quality assurance

Wissik / Durco 2016
Methods flower

django-apps (APIS, DEFC), cfdb, tokenEditor, VLE, CorpusEditor, OpenSKOS, (oXygen)
tagging pipeline, stanbol, reconcile, django-apps, cs-xsl, tokenizer
corpus_shell, django-apps, GraphViewer GeoBrowser
Islandora/Fedora Databases Virtuoso - triple store (git, shares)
Services of ACDH-OEAW

- **~ 200 Services** (in our internal inventory)
  - In very broad meaning: “Anything running on a server”
- **Of different types**
  - Databases, (Repository!)
  - Web services (machine interaction)
  - Web applications (user interaction)
- **Human Services**
  - Data modelling for new projects
  - Consultancy for drafting project proposals (Data Management Plans)
  - Converting legacy data
- **In the (not so far) future:**
  - Preservation in a dedicated “Digital Humanities” Repository (CCV/acdh-repo)
  - Digitisation Centre (in cooperation with the Library and Archive of the Austrian Academy of Sciences)
Priorities/Topics

- NLP-pipeline / Corpus Search / Analysis
- Semantic Technologies & Knowledge Management (‘from data to knowledge’)
- Repositories & Data Management (‘save the data’)
- Online-Publication & Digital Editions
- Visualization (esp. networks/graphs)
- Geocoding / Geovisualization
- Virtual Research Environments (or “collaborative spaces”) 
- Digitisation
use cases
Use Cases

- amc-extract - corpus analysis
- Dictionary framework – collaborative lexicographic environment
- corpus_shell - online publishing environment (for digital editions)
- Repository solutions
- Metadata curation
- Lexical & Semantic Services
amc-extract
Postprocess aggregated results from a corpus

Workflow:

1. source data → nlp-pipeline (tagging, NER) → indexing (sketch engine)
2. query → result (word lists, KWICs, metadata)
   
   [3a: results go to researcher (as excel sheets/text fragments/KWICS etc. depending on research question) → publication ]
3. Import results to an editor
4. (collaboratively) perform semi-automatic corrections (e.g. lemma or tagging)
5. Use manual corrections in the (next iteration of the) nlp-pipeline

Tools:

- Tree-tagger, RF-tagger
- Stanbol
- Sketch Engine
- tokenEditor
What is amc

- a digital collection of journalistic prose
- created via contract with the Austria Press Agency
- Unique: offers almost complete coverage of Austria’s media landscape over more than the last two decades

Contents:
Austrian newspapers and magazines 1990-2015
TV-news & interviews 2003-2015
Original Text Service 1989-2015

Counts: 9 billion token, 37 million articles
Access Tool: SketchEngine

- specialised search engine for text corpora
- complex queries on linguistic content
- collocation extraction (word sketches)
- statistical analysis

Current Applications

- Empirical base in several lexicographic and linguistic research cooperations
- academic research
<table>
<thead>
<tr>
<th>Word sketch “Flüchtling”</th>
</tr>
</thead>
</table>

### Flüchtling

**Constructions**

<table>
<thead>
<tr>
<th>Regular</th>
<th>395,585</th>
<th>2.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casus_Nom</td>
<td>155,155</td>
<td>1.20</td>
</tr>
<tr>
<td>Casus_Acc</td>
<td>111,119</td>
<td>2.00</td>
</tr>
<tr>
<td>Casus_Dat</td>
<td>78,460</td>
<td>1.20</td>
</tr>
<tr>
<td>Casus_Gen</td>
<td>51,355</td>
<td>2.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modifier</th>
<th>92,290</th>
<th>2.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;unknown&gt;</td>
<td>11,380</td>
<td>1.48</td>
</tr>
<tr>
<td>bosnisch</td>
<td>4,862</td>
<td>0.85</td>
</tr>
<tr>
<td>syrisch</td>
<td>4,462</td>
<td>0.16</td>
</tr>
<tr>
<td>palästinensisch</td>
<td>4,408</td>
<td>0.71</td>
</tr>
<tr>
<td>minderjährig</td>
<td>4,047</td>
<td>0.96</td>
</tr>
<tr>
<td>afganisch</td>
<td>2,810</td>
<td>0.80</td>
</tr>
<tr>
<td>politisch</td>
<td>2,627</td>
<td>0.63</td>
</tr>
<tr>
<td>anerkannt</td>
<td>2,455</td>
<td>0.76</td>
</tr>
<tr>
<td>welt</td>
<td>2,209</td>
<td>0.23</td>
</tr>
<tr>
<td>kurdisch</td>
<td>2,079</td>
<td>0.79</td>
</tr>
<tr>
<td>afrikanisch</td>
<td>1,999</td>
<td>0.62</td>
</tr>
<tr>
<td>serbisch</td>
<td>1,709</td>
<td>0.57</td>
</tr>
<tr>
<td>albanisch</td>
<td>1,560</td>
<td>0.73</td>
</tr>
<tr>
<td>jung</td>
<td>1,541</td>
<td>2.19</td>
</tr>
<tr>
<td>irakisch</td>
<td>1,497</td>
<td>0.58</td>
</tr>
<tr>
<td>andere</td>
<td>1,370</td>
<td>1.40</td>
</tr>
<tr>
<td>jüdisch</td>
<td>1,341</td>
<td>0.31</td>
</tr>
<tr>
<td>erst</td>
<td>1,286</td>
<td>0.53</td>
</tr>
<tr>
<td>Tausend</td>
<td>1,154</td>
<td>5.00</td>
</tr>
<tr>
<td>illegal</td>
<td>1,142</td>
<td>5.07</td>
</tr>
<tr>
<td>moslemisch</td>
<td>892</td>
<td>6.70</td>
</tr>
</tbody>
</table>

### Subst+Subst

<table>
<thead>
<tr>
<th>SubstX + SubstY</th>
<th>234,845</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubstX in-i SubstY</td>
<td>54,044</td>
</tr>
<tr>
<td>SubstX für-i SubstX</td>
<td>34,631</td>
</tr>
<tr>
<td>SubstX aus-i SubstY</td>
<td>32,210</td>
</tr>
<tr>
<td>SubstY von-i SubstX</td>
<td>29,415</td>
</tr>
<tr>
<td>SubstY mit-i SubstX</td>
<td>8,706</td>
</tr>
<tr>
<td>SubstX nach-i SubstY</td>
<td>8,234</td>
</tr>
<tr>
<td>SubstX auf-i SubstY</td>
<td>7,633</td>
</tr>
<tr>
<td>SubstX an-i SubstX</td>
<td>6,710</td>
</tr>
<tr>
<td>SubstX von-i SubstY</td>
<td>5,012</td>
</tr>
<tr>
<td>SubstX zu-i SubstY</td>
<td>3,764</td>
</tr>
<tr>
<td>SubstX mit-i SubstY</td>
<td>3,568</td>
</tr>
<tr>
<td>SubstY als-c SubstX</td>
<td>3,500</td>
</tr>
<tr>
<td>SubstY von-i SubstX</td>
<td>2,541</td>
</tr>
</tbody>
</table>

### PräpY SubstX Dat

<table>
<thead>
<tr>
<th>PräpY</th>
<th>58,499</th>
<th>1.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>von</td>
<td>32,954</td>
<td>3.77</td>
</tr>
<tr>
<td>mit</td>
<td>12,418</td>
<td>2.38</td>
</tr>
<tr>
<td>zu</td>
<td>2,704</td>
<td>0.52</td>
</tr>
<tr>
<td>bei</td>
<td>2,143</td>
<td>0.55</td>
</tr>
<tr>
<td>unter</td>
<td>2,048</td>
<td>2.47</td>
</tr>
<tr>
<td>gegenüber</td>
<td>1,070</td>
<td>4.14</td>
</tr>
<tr>
<td>zwischen</td>
<td>713</td>
<td>2.25</td>
</tr>
<tr>
<td>neben</td>
<td>136</td>
<td>0.58</td>
</tr>
<tr>
<td>mit</td>
<td>91</td>
<td>0.16</td>
</tr>
<tr>
<td>pro</td>
<td>84</td>
<td>0.27</td>
</tr>
<tr>
<td>zugunsten</td>
<td>52</td>
<td>2.63</td>
</tr>
<tr>
<td>angesichts</td>
<td>38</td>
<td>0.53</td>
</tr>
<tr>
<td>statt</td>
<td>25</td>
<td>0.19</td>
</tr>
<tr>
<td>zufolge</td>
<td>18</td>
<td>0.85</td>
</tr>
<tr>
<td>inmitten</td>
<td>11</td>
<td>1.26</td>
</tr>
</tbody>
</table>

### Subst+Verb

<table>
<thead>
<tr>
<th>SubstX + Verb</th>
<th>57,731</th>
</tr>
</thead>
<tbody>
<tr>
<td>VerbY + für-i</td>
<td>19,473</td>
</tr>
<tr>
<td>SubstXAcc (pp_obj_of_acc)</td>
<td>15,050</td>
</tr>
<tr>
<td>VerbY + von-i</td>
<td>6,830</td>
</tr>
<tr>
<td>SubstXDat (pp_obj_of_dat)</td>
<td>1,794</td>
</tr>
<tr>
<td>VerbY + mit-i</td>
<td>1,638</td>
</tr>
<tr>
<td>SubstXAcc (pp_obj_of_acc)</td>
<td>1,395</td>
</tr>
<tr>
<td>VerbY + zu-i</td>
<td>1,318</td>
</tr>
<tr>
<td>SubstXDat (pp_obj_of_dat)</td>
<td>1,304</td>
</tr>
<tr>
<td>VerbY + um-i</td>
<td>1,294</td>
</tr>
<tr>
<td>SubstXAcc (pp_obj_of_acc)</td>
<td>1,194</td>
</tr>
<tr>
<td>VerbY + auf-i</td>
<td>1,183</td>
</tr>
<tr>
<td>SubstXDat (pp_obj_of_dat)</td>
<td>1,178</td>
</tr>
<tr>
<td>VerbY + unter-i</td>
<td>1,167</td>
</tr>
<tr>
<td>SubstXAcc (pp_obj_of_acc)</td>
<td>1,157</td>
</tr>
<tr>
<td>VerbY + bei-i</td>
<td>1,147</td>
</tr>
<tr>
<td>SubstXDat (pp_obj_of_dat)</td>
<td>1,137</td>
</tr>
<tr>
<td>VerbY + gegen-i</td>
<td>1,127</td>
</tr>
<tr>
<td>SubstXAcc</td>
<td>1,117</td>
</tr>
</tbody>
</table>

---

**Notes:**

- The word sketch for “Flüchtling” includes various constructions, modifiers, and verbs related to the concept of a refugee or immigrant.
- The highest frequency words include “SubstX in-i SubstY,” “SubstX für-i SubstX,” and “SubstX aus-i SubstY.”
- “PräpY SubstX Dat” entries show common prepositions like “von,” “mit,” “zu,” “bei,” “unter,” and “gegenüber.”
- “Subst+Verb” combinations show various verbs such as “für-i,” “von-i,” “mit-i,” and “zu-i,” indicating the use of “Flüchtling” in different contexts.

---

**Additional Resources:**

- Concordance
- Word list
- Thesaurus
- Sketch diff
- Corpus info
- Save
- Change options
- Cluster
- Sort by score
- Hide grammars
- More data
- Less data
amc-extract – exported data

e.g. frequency lists from corpus
Nouns with unknown lemma sorted by frequency

<table>
<thead>
<tr>
<th>word/0</th>
<th>lemma/0</th>
<th>pos/0</th>
<th>posbase</th>
<th>postT</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Masc</td>
<td>N</td>
<td>NE</td>
<td>1578659</td>
</tr>
<tr>
<td>OTS</td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>NN</td>
<td>1452567</td>
</tr>
<tr>
<td>***</td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>ADJA</td>
<td>1367482</td>
</tr>
<tr>
<td>Austria</td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>NE</td>
<td>1214712</td>
</tr>
<tr>
<td>AUSSCHLIESSLICHER</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>1131512</td>
</tr>
<tr>
<td>ÖVP</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Fem</td>
<td>N</td>
<td>NN</td>
<td>1007602</td>
</tr>
<tr>
<td>Jänner</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>988952</td>
</tr>
<tr>
<td>SPÖ</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Fem</td>
<td>N</td>
<td>NE</td>
<td>856785</td>
</tr>
<tr>
<td>Rückfragehinweis</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>824386</td>
</tr>
<tr>
<td>AUSSENDERS</td>
<td>unknown</td>
<td>N.Name_Gen_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>776953</td>
</tr>
<tr>
<td>Mio.</td>
<td>unknown</td>
<td>N.Reg_Dat_Pl_Fem</td>
<td>N</td>
<td>ADJA</td>
<td>774478</td>
</tr>
<tr>
<td>-Jährige</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>753908</td>
</tr>
<tr>
<td>****</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>663604</td>
</tr>
<tr>
<td><a href="http://WWW.OTS.AT">WWW.OTS.AT</a></td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>NN</td>
<td>655606</td>
</tr>
<tr>
<td>Austria</td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>NN</td>
<td>654422</td>
</tr>
<tr>
<td>FPÖ</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Fem</td>
<td>N</td>
<td>NE</td>
<td>615681</td>
</tr>
<tr>
<td>Obmann</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NN</td>
<td>611861</td>
</tr>
<tr>
<td>OTS</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_Neut</td>
<td>N</td>
<td>NN</td>
<td>605600</td>
</tr>
<tr>
<td>Dornbirn</td>
<td>unknown</td>
<td>N.Name_Dat_Sg_Neut</td>
<td>N</td>
<td>NN</td>
<td>604217</td>
</tr>
<tr>
<td>Mio.</td>
<td>unknown</td>
<td>N.Reg_Nom_Pl_Fem</td>
<td>N</td>
<td>ADJA</td>
<td>595302</td>
</tr>
<tr>
<td>AUSSENDERS</td>
<td>unknown</td>
<td>N.Name**_Neut</td>
<td>N</td>
<td>NN</td>
<td>572139</td>
</tr>
<tr>
<td>Potten</td>
<td>unknown</td>
<td>N.Reg_Nom_Pl_Fem</td>
<td>N</td>
<td>NN</td>
<td>588276</td>
</tr>
<tr>
<td>Url</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NE</td>
<td>520684</td>
</tr>
<tr>
<td>OTS-ORIGINALTEXT</td>
<td>unknown</td>
<td>N.Name<strong>_</strong></td>
<td>N</td>
<td>NN</td>
<td>467020</td>
</tr>
<tr>
<td>PRESSEAUSSENDUNG</td>
<td>unknown</td>
<td>N.Name_Nom_Sg_*</td>
<td>N</td>
<td>NN</td>
<td>465964</td>
</tr>
<tr>
<td>Dr</td>
<td>unknown</td>
<td>N.Reg_Nom_Sg_Masc</td>
<td>N</td>
<td>NE</td>
<td>459412</td>
</tr>
<tr>
<td>Ralf</td>
<td>unknown</td>
<td>N.Name**</td>
<td>N</td>
<td>NF</td>
<td>454432</td>
</tr>
</tbody>
</table>
tokenEditor

- Web application for manual revision/curation of PoS-tags (or any annotation or any token lists)
- TEI (or any structured XML) as input (and output)
- Supports Federated Identity
- customizable annotation layers (soon)

http://www.oeaw.ac.at/acdh/en/tokenEditor
https://clarin.oeaw.ac.at/tokenEditor/
https://github.com/acdh-oeaw/tokeneditor
**tokenEditor**

Webinterface for manual corrections of text annotations.

---

**XML Example**

```xml
<item>
  <word>Jänner</word>
  <lemma>_unknown_</lemma>
  <pos>N.Reg.Non.Sg.Masc</pos>
  <posbase>N</posbase>
  <posTT>NN</posTT>
  <freq>988952</freq>
  <manual_lemma></manual_lemma>
  <manual_pos></manual_pos>
  <manual_posbase></manual_posbase>
  <manual_posTT></manual_posTT>
</item>
```
Dictionary framework
Dictionary Framework

Lexicography infrastructure (< Dictionary-in-a-box)

- Tools
  - Editing: Viennese Lexicographic Editor (VLE)
  - Publishing: corpus_shell
  - Storage: CCV, mysql, BaseX
  - Documentation: DictGate (standards, workflows, best practices, etc.)

- Data
  - VICAV, TUNICO, Dagaare, historical Viennese, etc.
VLE - Viennese Lexicographic Editor

- XML editor providing functionalities typically needed in editing lexicographic data (LMF, TEI, TBX, RDF ...)
- Making use of cognate technologies, e.g. XSLT, XPath, XML Schema
- Configurable keyboard layouts
- Validation
- Various editing modes
- Data visualisations
- Versioning
- Freely available
VICAV - Vienna Corpus of Arabic Varieties

- Research platform for variational linguists
- Cooperation with Institute for Oriental Studies, University of Vienna
- Wide range of different digital language resources
- Raw data available as TEI/XML
- Instance of corpus_shell (php drupal module)

https://minerva.arz.oeaw.ac.at/vicav2/
Short sample text (Cairo)

Sample Text

?ah-aan!
?ana ?ismi muna.
?ana min mašr il-7adīma.
?intu mnān?
?izzayyukum?
kuwāyisn in-ṣâ?-aš-ah.n
bi-tifmilū ?eh diwāti?
ṣawānī tīgu mašāya?
yalla nišrāb ?ahwāl

POS: Adjective
Lemma: 7adīm
English: old
viDi - historical Viennese Dictionaries

Collection of early lexicographic works on the german variety of Vienna. early beta version edited in VLE
Publishing (small/mid) dictionaries
Corpus_shell
online publication
Online publication - corpus_shell

- Modular framework for publishing a wide range of language resources designed to operate in a distributed and heterogeneous environment
- Distributed setup based on FCS
- Integration with CLARIN CMDI
- Open source: https://github.com/acdh-oeaw/corpus_shell
Ein lebendiger Entwurf
deß sterblichen Lebens / vnd daß
der Todt ein Regell ohne
Unterscheid
allen vorschreiben.

Nicht vmbsonst list man das
Wort Leben / zu ruck Nebel /
kauß daß ein Nobel dieser
trampische Sohn der morastigen Erz-
enden gehoben wird / so trohen ihme
schon die Sonnen-Strahlen den Gar-
auf: Also hat es ein gantz ähnliche
Beschaffenheit mit vonseren Leben / vix
orimur morimur: Unser erster Lebens
Athem ist schon ein Seufzer zum Todt / vnd
der erste Augenblicke des mensch-
lichen Lebden fällt schon er die Bott-
massigkeit des Knochenreichen Sensen-
tragers / auch den ersten Trunck an der
Sättig.
matches

placeName="Şam"

1 to 19 of 19 folia where the item was found.

1. AF268 046v
   ... de biz nice mali gideriz Şuhbš Şam u şamı Hind de éderız 596 Ve ...

2. AF268 055r
   ... li begler begiși ve şol geldı Şamı 'ancak ve Haleb begler be ...

3. AF268 071v
   ... e Miṣr sulınınına m-bûd edib Şamı a varดาqda şörna // qat-i ...

4. AF268 072r
   ... 'ağbet hållinden sa-ulé étûm Şamı i // öteye geçnez Şam da qal ...
   ... ál étûm Şam i // öteye geçnez Şamı da qahır dûd dûdûm kî yî b ...

5. AF268 072v
   ... -reb ber pâdîpa // Miṣr dan Şamı le teveccüh edûkde Halemi Î ...
   ... neme inta etmede bu bâl ile Şamı a varlığı Halemi Efendi nuit ...

6. AF268 073r
   ... 'hâlâyot gâhî bâ îlîrîmâm [?] Şamı da oldûramı evyânıva mêsîyi ...
   ... tî] kî kalla a-ğûd edib hava yî Şamı a hâlâyot kî kemân styledi ...

7. AF268 073v
   ... diler bir qac // giûden şörna Şamı dan barza 'urâz [?] u mûkat ...
   ... [?] u mûkatib ile ulaq gelib Şamı dan ulaq gelib şeyhînî dah ...
   ... i lîhênarda a'yân u uşraf i Şamı a nûşu u vaşıfet edib se ...
   ... mehânâmın ta'dîn ve elî Şamı i lît'ât u îngâlînamra ir ...
   ... irâd edib demîş kî ey elî Şamı // bu sulün' i bâ-îlîrîmam ...

8. AF268 074r
   ... 'Arab da // nîçe bir şuvâs Şam u Haleb de cihâm ałqum qam ...

9. AF233a 008r
   AF233a 08 ve Şamı evliyâları ve bûrasım [1] pînleri birmüret-cî-

10. AF287b 006v
    ... devle sene 921 fet'h-i Haleb ve Şamı sene 921 // fet'h-i Belâqad ...

11. AF287b 010v
    ... n-vefâ-i Davûd 5 // bu gûn Şamı yelî eser // şefîr-i müste ...

12. AF287b 011v
    ... tîrîdan derler 22 // vilâyets Şamı da menû [?] yî sevûl vaqar d ...
save the data!
repository solutions
Requirements on online availability

Varying combinations of:

▸ full-text search

▸ semantic search (search for persons, places, concepts, search by classifications)

▸ full-view (e.g. text and facsimile of individual pages)

▸ specialized visualizations (temporal, spatial, graph, statistical data)

▸ raw data available for download

▸ stable references to resources and resource fragments

BUT before publication: collaborative editing → VRE !
Repositories

- CLARIN Centre Vienna / Language Resources Portal (FEDORA-based)
  - Instance of GAMS + client: Cirilo by Uni Graz

Reshaping ongoing:
- LR -> DH
- Collaboration with the institutional repository epub.oeaw
  - run by Academy Press
  - Mainly for publications, but also some structured data (lexicographic databases/apps)
- Cover whole research process (also working data)
- Integrate with the research process (rich API to connect the tools)
- Evaluating Fedora 4! (Fedora 3 is dead!)
Services & Repository interaction

1. R/W user
2. R/W user
3. R/W app
4. invoke
5. Rq
6. Rix
7. invoke
8. W d+md
9. Rmd
10. R user
11. ?
12. 

- VRE read/write
- local store
- storage plugin
- algorithmic processing
- external repos (B2DROP, DARIAH-DE, ...)
- data viewer
- web app read only
- local store
- publish repo (epub.oeaw)
- repo browser
- OAI-PMH
- .discover .cite
- .publish .share
- .deposit .licence
- .describe
- share
- git
- .curate .enrich
- .explore .visualize
- .compute .analyze
Metadata curation
Introduction - VLO

- 700 – 800 thousand metadata records of linguistic resources
- From 60 – 70 data providers
- CMDI as required input format
- Almost 100 CMDI profiles represented in VLO
  (out of some 200 defined in the Component Registry)
- The diverse information is mapped/reduced to 13(+) facets
  via data categories / concepts

From: King et al. (2015): Value Normalization - a Case for Metadata Curation, Presentation at CAC 2015, Wroclaw
Problems with Metadata / VLO

- Wide-scale variation of the metadata values
- Missing values
- Facet issues – definition/scope; selection of facets
- Curation workflow not fully functional

Stipulated reasons:
- Data providers act in isolation when creating their metadata
- CMDI allows too much flexibility
- Not enough guidance for data providers, insufficient coordination

From: King et al. (2015): Value Normalization - a Case for Metadata Curation, Presentation at CAC 2015, Wroclaw
Proposed Solutions

- Normalisation mapping to controlled vocabulary for each facet
- CLAVAS as common authoritative source for vocabularies
- Reconsider facets
- Feedback to data providers
- Best practices, guidelines (recommended profiles)

Requires:
- Sustainable workflow
  - Integrate human and automatic curation
- Continuous, collaborative process
  - Include national consortia (via Curation TF)

From: King et al. (2015): Value Normalization - a Case for Metadata Curation, Presentation at CAC 2015, Wroclaw
Facet/Concept: License/Availability

License Categories
https://trac.clarin.eu/wiki/Taskforces/Curation/ValueNormalization/License

- 4 – 6 concepts mapped to facets
- cover around 60,000 records! Records with missing value for availability

- 3 possible situations for missing values:
  - Profile does not have any information about licensing/availability (worst case)
  - Profile has information about L/A, but is not linked to a concept, or the concept is not in the facet mapping
  - Profile is well defined, with linking to one of the concepts in the facet mapping, but the information is simply not filled in the record.
Solutions: Prevention

Try to solve the problems at the source!

- Best practice guides, recommended profiles that cover all (VLO) facets and use controlled vocabularies (wherever defined)
- Provide **curation reports** back to data providers
- Offer curation module as a separate application that data providers can use to validate their metadata prior to harvesting
- Offer amended metadata records back to the data providers
- Availability of the controlled vocabularies via a simple API to be integrated as **autocomplete** with metadata authoring tools.

Arbil (Withers, 2012) and COMEDI (Lyse et al., 2014) plan adoption?

From: King et al. (2015): Value Normalization - a Case for Metadata Curation, Presentation at CAC 2015, Wroclaw
Curation module

CLARIN-PLUS - T2.2.1: Metadata benchmarking and curation [CLARIN]

- Development of an application which can automatically analyze large amounts of CMDI metadata records and provide a quality score per record and collection, taking into account quality parameters such as technical (XML schema) validity, correctly resolving URLs and persistent identifiers, availability of required fields, size, sparseness and correct usage of multilingual fields. Results should be easily accessible for the metadata providers via a web application.

- [https://clarin.oeaw.ac.at/curate/](https://clarin.oeaw.ac.at/curate/)
- [https://trac.clarin.eu/wiki/Curation%20Module](https://trac.clarin.eu/wiki/Curation%20Module)
Metadata Workflow

Data Admin

User / Researcher

Third-party Application

Data Provider / Admin App

Stats

Curation Report

Content Provider

OAI-PMH

Harvesting

Raw MD records

Normalized MD records

MD catalog

MD2RDF

RDF Triple store

Curation
- Validate + Correct
- Normalize + Enrich
- Quality Assessment

cache

literal2entity

map

VocabularyService

SemanticMapping
<table>
<thead>
<tr>
<th>Name</th>
<th>Avg Score</th>
<th>Num Of Records</th>
<th>Size In Bytes</th>
<th>Avg Size</th>
<th>Num Of Profiles</th>
<th>Avg Num Of Res</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academia Sinica Balanced Corpus of Modern Chinese</td>
<td>10.851</td>
<td>3</td>
<td>25,946</td>
<td>8,648</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Academia Sinica Formosan Language Archive</td>
<td>10.851</td>
<td>3</td>
<td>25,916</td>
<td>8,638</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>African Language Materials Archive</td>
<td>10.666</td>
<td>53</td>
<td>215,818</td>
<td>4,072</td>
<td>1</td>
<td>0.981</td>
</tr>
<tr>
<td>Archive of the Indigenous Languages of Latin America</td>
<td>10.6</td>
<td>100</td>
<td>244,701</td>
<td>2,447</td>
<td>1</td>
<td>1.93</td>
</tr>
<tr>
<td>A Digital Archive of Research Papers in Computational Linguistics</td>
<td>10.557</td>
<td>696</td>
<td>1,269,305</td>
<td>1,823</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B2SHARE</td>
<td>10.709</td>
<td>16</td>
<td>40,918</td>
<td>2,557</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>BAS Repository</td>
<td>12.088</td>
<td>24,002</td>
<td>5,351,905,135</td>
<td>222,900</td>
<td>6</td>
<td>93.818</td>
</tr>
<tr>
<td>CHILDES</td>
<td>10.85</td>
<td>38,916</td>
<td>371,981,746</td>
<td>9,558</td>
<td>2</td>
<td>2.185</td>
</tr>
<tr>
<td>CLARINO provider National Library of Norway</td>
<td>9.688</td>
<td>37</td>
<td>430,663</td>
<td>11,639</td>
<td>2</td>
<td>3.351</td>
</tr>
<tr>
<td>CLARIN DK UCPH Repository</td>
<td>11.746</td>
<td>87,551</td>
<td>405,145,696</td>
<td>4,627</td>
<td>3</td>
<td>2.611</td>
</tr>
<tr>
<td>Claran D Repository Tubingen</td>
<td>11.252</td>
<td>58</td>
<td>591,578</td>
<td>10,199</td>
<td>4</td>
<td>1.672</td>
</tr>
<tr>
<td>CLARIN PL digital repository</td>
<td>11.538</td>
<td>140</td>
<td>4,851,970</td>
<td>34,656</td>
<td>5</td>
<td>97.75</td>
</tr>
<tr>
<td>CLARIN Virtual Collection Registry</td>
<td>10.384</td>
<td>4</td>
<td>25,935</td>
<td>6,483</td>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>CLEF ESRS Universite de Toulouse Le Mirail</td>
<td>10.709</td>
<td>7</td>
<td>13,680</td>
<td>1,954</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Collections de Corpus Oraux Numeriques CoCoON ex CRDO</td>
<td>10.852</td>
<td>10,299</td>
<td>48,911,012</td>
<td>4,749</td>
<td>1</td>
<td>2.265</td>
</tr>
<tr>
<td>Common Language Resources and Technology Infrastructure Norway CLARINO</td>
<td>10.903</td>
<td>36</td>
<td>247,481</td>
<td>6,874</td>
<td>5</td>
<td>2.139</td>
</tr>
<tr>
<td>DANS CMDI Provider</td>
<td>10.062</td>
<td>1,001</td>
<td>76,106,530</td>
<td>76,030</td>
<td>7</td>
<td>6.066</td>
</tr>
<tr>
<td>Deutsches Textarchiv</td>
<td>9.312</td>
<td>3,033</td>
<td>30,796,097</td>
<td>10,153</td>
<td>4</td>
<td>2.982</td>
</tr>
<tr>
<td>Ethnologue Languages of the World</td>
<td>10.624</td>
<td>7,413</td>
<td>15,262,857</td>
<td>2,058</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>European Language Resources Association</td>
<td>10.598</td>
<td>1,062</td>
<td>4,441,810</td>
<td>4,182</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Euskal Herriko Unibertsitatea</td>
<td>10.631</td>
<td>18</td>
<td>30,612</td>
<td>1,700</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
CLARIN VLO

ACDH’s Co-development of CLARIN VLO
Experimental instance on minerva [http://minerva.arz.oeaw.ac.at/vlo](http://minerva.arz.oeaw.ac.at/vlo)
Extra features:
- Multi-value selection
- Orig/curated facets
- Explicit [missing value]
- Normalization maps
SMC Browser

Visualizing the domain of CMDI: Profiles/Components/Elements/Data categories
Vocabularies & Taxonomies
Lexical & Semantic services
Vocabularies / Taxonomies

... reference resources are indispensable to categorize and enrich data in a consistent manner and represent a **conceptual backbone** of the (humanities) research.

- CLARIN Vocabularies – CLAVAS
- CLARIN Concept Registry

ACDH Vocabularies: [https://clarin.oeaw.ac.at/vocabs](https://clarin.oeaw.ac.at/vocabs)
- Backbone Thesaurus
  - top-level thesaurus – DARIAH-WG
- Instrumentensystematik Hornbostel-Sachs
- Professions
- VLO Resource Type vocabulary
- DHA taxonomy
- ...
BBT: Facets/Hierarchies

From the document by Martin Doerr and colleagues:

- **Activities**
  - Disciplines (Construction of material objects and installations, …)
  - Events (Social events, Confrontations, conflicts, Political, social and economic occurrences, …)
  - Intentional destructions
  - Functions

- **Natural processes**
  - Natural disasters
  - Geneses

- **Materials**

- **Material Objects**
  - Mobile objects
  - Built environment (Monuments, infrastructures)
  - Physical features
  - Structural parts of material objects

- **Epochs**

- **Conceptual Objects**
  - Symbolic objects (Information objects)
  - Propositional objects (Information objects)
  - Methods (Procedures, Techniques)

- **Groups and Collectivities**

- **Offices**
ACDH Vocabularies

Instance of [OpenSKOS](https://clarin.oeaw.ac.at/vocabs/) - a vocabulary repository for collaborative development and maintenance of taxonomies/thesauri.

[https://clarin.oeaw.ac.at/vocabs/](https://clarin.oeaw.ac.at/vocabs/)
Lexical & Semantic services

- **Motivation**
  - Extract structured information from unstructured data (text) (entails named entity recognition)
  - Enrich/link with external semantic reference resource (entity linking)
  - Collaboratively maintain own reference resources (vocabularies/thesauri)
  - *=> interlink heterogeneous data from different projects via reference resources*

- **Requirements**
  - Flexible selection of reference resource to match against
  - Customisable matching algorithm
  - Combination of automatic and manual processing
  - Programmatic access to the external resources (API)
  & User interface
  - publish as LOD
Lexical & Semantic services - Tools

- **Apache Stanbol** - flexible framework for semantic content management
- **OpenRefine** - tool for working with large datasets, RDF modelling
- **Reconcile** - UI service for matching scientific plant names to common names, works on top of Common Name Service’s API
- **OpenSKOS** - vocabulary repository, editor, service
Example: Stanbol

http://stanbol.herkules.arz.oeaw.ac.at/
Example: Reconcile  http://reconcile.eos.arz.oeaw.ac.at/

- Look up one scientific name
- Import and reconcile a list of names
- Get RDF file with response in SKOS
- Direct data post-processing online (filter, select, export)
Lexical & Semantic services - Data

External reference resources:

- GND - (Gemeinsame Normdatei): authority file for Persons, Corporate bodies, Conferences and Events, Geographic Information, Topics and Works
- GeoNames - Integrated geographical data source (contains: names of places in various languages, coordinates, population, location types, etc.)
- DBpedia - structured, machine-readable version of Wikipedia
- Common names service - federated search of common names for a given scientific name

Local resources:

- Dboe plants common names
- BBT - DARIAH BackBone Thesaurus
Example: GeoNames
Lexical & Semantic services - Workflow

- Unstructured Text -> NER + Entity Linking ->
  -> manual Curation/Disambiguation -> Publish as LOD
- Example of specific workflow in APIS project:
Lexical & Semantic services
Challenges & Future work

Challenges:
- Disambiguation
- False matches (precision)
- Missing matches (recall)

Plans:
- Establish the feedback loop from manual post-processing to automatic processing
- Integrated Portal for these Lexical & Semantic Services
- Make the resources available via Triple Store
What else is happening?
Other toolboxes
DARIAH(-DE) Services

- Basic services
  - AAI
  - Help Desk
  - *Kollaborative Arbeitsumgebungen*
  - Monitoring
  - PID Service
  - *Virtuelle Maschinen*

- Specialized services
  - Digivoy
  - Geo-Browser & Datasheet-Editor
  - MEI Score Editor
  - Personendatenrepositorium
  - Virtuelles Skriptorium

- Developer portal
  - Chili Projects
  - Confluence
  - Jenkins
  - Etherpad
  - E-Mail-Liste
GeoBrowser

- Developed within Europeana /DARIAH-DE(?)
- Spatio-temporal visualization
- KML as input format
- Accepts external data
- Integrated with “data sheet”

http://geobrowser.de.dariah.eu/
# DataSheet editor

![DataSheet Editor](http://geobrowser.de.dariah.eu/edit/)

The DataSheet editor is a tool for managing and visualizing geospatial data. It allows users to add geocoordinates for places, select places from the DARIAH Getty Thesaurus Service, and view raw data. The raw CSV data of this sheet is stored in the DARIAH Storage Service and has the ID [263801](http://geobrowser.de.dariah.eu/edit/).

### Place Selection

The coordinates automatically added are for the most probable place, you may adjust the places below. The coordinates let you lookup and finetune the coordinates or set unrecognized places.

### Wien:

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Description</th>
<th>Longitude</th>
<th>Latitude</th>
<th>TimeStamp</th>
<th>TimeSpan:begin</th>
<th>TimeSpan:end</th>
<th>GettyID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wien</td>
<td></td>
<td></td>
<td>16.3867</td>
<td>48.2167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Map Selection

![Map Selection](http://geobrowser.de.dariah.eu/edit/)

The datasheet will be checked for errors and completeness and could be viewed with the DARIAH Geo-Browser.
PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies also, Athena Parthenos, the Greek goddess of wisdom, inspiration and civilization.

- Horizon 2020 thematic cluster of European research infrastructures and integrating initiatives
- Started in May 2015
- Coordinated by Franco Niccolucci (PIN, Italy)
- ACDH-OEAW involved in WPs 2, 3, 4, 5, 6, 8 (~ 3 FTEs) leading the task Resource Discovery Tools
- Built around DARIAH and CLARIN
- [http://parthenos-project.eu/](http://parthenos-project.eu/)
PARTHENOS works on …

» the implementation of common AAA (authentication, authorization, access)
» the implementation of data curation policies
» IPR management (also addressing sensitive data and privacy issues)
» standardization and interoperability
» common tools for data-oriented services such as resource discovery, search services, quality assessment of metadata, annotation of sources
» guidelines, standards, methods, services and tools to be used by its partners and by all the research community.
Architecture

Proposed technologies as the backbone of the architecture

- **D-Net**
  *infrastructure, enables a run-time environment where services can be dynamically deployed, shared and combined to form applications*

- **gCube** – Hybrid Data Infrastructure

- **X3ML** – Mapping tool
  - currently mainly XML to RDF (CIDOC-CRM)
Current view of the overall architecture

- Knowledge aggregation
- Index 1
- Index 2
- ... 

Directory service of providers/people

- Minimal identity metadata, part-of, agency about:
  - Datasets, metadata sets, software, services, mappings, and more

Keep identities of the who, what in the infrastructure

Mappings, Resources integration

Joint resource registry

Aggregation mechanism

Common design requirements

Resources: institutional research environments

Content cloud
## D-Net

### Data Provision Area
- OAI-ORE Publisher Service
- SRW/CQL Publisher Service
- OAI-PMH Publisher Service
- Generic portal service
- User Service
- Collection Service
- Recommendation Service

### Enabling Area
- Information Service
- Authorization Authentication Service
- Orchestration Service
- ResultSet Service
- ChronJobs Service

### Data Curation and Enrichment Area
- Deduplication Service
- Citation Identification Service
- Classification Service
- User Feedback Service
- User Behavior Analysis Service
- Text Similarity Service
- Record Tagging Service
- Metadata Editor Service

### Data Storage and Indexing Area
- Object Store Service
- Full-Text Index Service
- Metadata Store Service
- Graph Storage Service
- Database Service

### Data Conversion Area
- Feature Extraction Service
- Metadata (un)packaging Service
- Metadata Transformation Service
- Metadata Cleaner Service

### Data Mediation Area
- FTP Import Service
- OAI-PMH Harvester Service
- Data Sources Man Service
- Data Source Validator Service

### External Data Sources
- File Systems
- Repositories
- CRIS systems
- Archives
Research
On-demand Virtual Research Environments

HDI
Hybrid Data Infrastructure

Apps
Application Bundles: rich array of facilities off-the-shelf

gcube-system.org
Where do we want to go?

- **Vertical integration - VREs**
  - Harmonized portfolio of high-quality services supporting the whole research process
  - Bring together Resources & Services (→ LR Switchboard)

- **Horizontal integration**
  - broader coverage (of resources & services)
  - From different domains

- **Innovation vs. Reliability**

- **Safe home for the data**

- **LR -> DH**

- **Semantics! (Named entities – Who did what when)**
Thank you!

Questions?
=> Discussion
Discussion: Methodical issues / Best practices

Software development, project management, provision of services

- **Reuse**
  - “Deeper levels”
  - Functionality (libraries, modules)
  - but also UI-design (→ CLARIN-PLUS D3.1 Human Interface Guidelines)

- **Modular setup**

- **Open source – public repos**

- **Integrate with the infrastructure**
  - Outsource hardware - Push to computing centres
  - Adhere to contracts/protocols (PID, CMD, OAI-PMH, FCS)
  - Monitoring
Provision of services

Innovation vs. reliability/stability

▸ innovation: develop new experimental tools – may fail by definition
▸ but innovation needs to build on reliable/stable environment
▸ Therefore development must be (also) conservative:
  ◆ striving to achieve industrial level quality of service: apps that just work

quality of

▸ hardware (processing/storage capacity, throughput, uptime),
▸ software (performance, correctness / no bugs),
▸ docs/descriptions
▸ data (curation, provenance)
▸ ergonomy (as people are used to from commercial applications)
▸ “human factor” (response time, expertise, knowledge exchange)
▸ reliable technical infrastructure / seamless technology
How can technology support research?

▸ enable **collaboration**:
  ▸ Federated Identity (use all services with your institutional account)
  ▸ Virtual Research Environments

▸ enable **sharing**
  ▸ Open Access - make (raw) data available
  ▸ Open Source - make source code available (under permissive licenses)
    (https://github.com/acdh-oeaw)

▸ exploit sharing / **reuse**
  ▸ loads of resources freely available (data, applications) “everything is there, just use it”
  ▸ but (/ on the other hand) always need for customization
Discussion: ??