LINDAT/CLARIN Repository

Michal Josífko

ÚFAL Seminar
Sedlec-Prčice, September 15-16, 2014
Outline

1. Basic information
2. Motivation
3. Key components
4. Current status
5. What next?
6. Problems
Basic information

- CLARIN = Common Language Resources and Technology Infrastructure
- Objective: preserve linguistic data and tools and make them accessible
- Cooperation: AT, BG, CZ, DE, DK, EE, NL, PO + DLU
- LINDAT/CLARIN = CLARIN Centre in Czech Republic
- Under the Programme of Large Infrastructures (2010+)
- Partners: UK, ZČU, ÚJČ, MU, MŠMT
- Now in the process of prolonging
Welcome to LINDAT/CLARIN

Centre for Language Research Infrastructure in the Czech Republic

The LINDAT/CLARIN Centre for Language Research Infrastructure provides technical background and assistance to institutions or researchers who want to share, create and modernise their tools and data used for research in linguistics or related research fields. The project also provides an open digital repository and archive open to all academics who want their work to be preserved, promoted and made widely available.

LINDAT/CLARIN is funded by the Ministry of Education, Youth and Sports of the Czech Republic.
Motivation: don’t reinvent the wheel
...recycle!
Key components

- Find
  - Repository (WWW, OAI-PMH, PID, interoperability)
- Get
  - Repository (AAI, licenses, services)
- Create
  - Know-how, data, tools and services
- Share
  - Repository (infrastructure, PID, backup)
Infrastructure

- 24/7 uptime, 100% redundancy (mirroring)
- Virtualization using Proxmox platform
- 10+ TB RAID6 disk array
- Online replicas over iSCSI, failover
- Backups: locally, to CESNET, to CINES (France)
- Monitoring using Nagios, uptimerobot.com, custom scripts, Piwik implementation for access statistics coming soon
- Data i CT se ukládají společně na iSCSI pole
- Přečte jen jeden APL servery a oběma CT
- Linhot 2 zálohuj
- Pacemaker vyhodnocuje dostupnost zdrojů a řídí přesunu

Výbavy:
- Power
- HVAC v CT (vrtulník, klima)
- Disk
- HVAC APL Racks (HV/SW)
Repository

- repository of linguistic data and tools
- fork of open source project DSpace (v1.8.2)
- enhanced user interface
- enhanced authentication (Shibboleth)
- more export formats for metadata (OAI-PMH)
- persistent identifiers (Handle, also for services)
- open repository - also for linguistic data outside of project CLARIN
- data in repository, tools source code also on github
“There ought to be only one grand dépôt of art in the world, to which the artist might repair with his works, and on presenting them receive what he required...”
—Ludwig van Beethoven, 1801

<table>
<thead>
<tr>
<th>Author</th>
<th>Subject</th>
<th>Language (ISO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hajíč, Jan (21)</td>
<td>corpus (17)</td>
<td>ces (65)</td>
</tr>
<tr>
<td>Bojar, Ondřej (17)</td>
<td>machine translation (10)</td>
<td>eng (28)</td>
</tr>
<tr>
<td>Pajas, Petr (11)</td>
<td>treebank (10)</td>
<td>slk (9)</td>
</tr>
<tr>
<td>Straňák, Pavel (11)</td>
<td>Czech (9)</td>
<td>hin (6)</td>
</tr>
<tr>
<td>Štěpánek, Jan (10)</td>
<td>parallel corpus (9)</td>
<td>deu (5)</td>
</tr>
</tbody>
</table>

... View More

What’s New

AKCES S (CzeSL-SGT)

Authors:
Šebesta, Karel; Bedřichová, Zuzanna; Šormová, Kateřina; Šindlová, Barbora; Hrdlička, Milan; Hrdličková, Tereza; Hama, Jiří; Petkevič, Vladimír; jelinek, Tomas; Skůdalová, Sofie; Poláčková, Marta; Janeš, Petr; Zumdělková, Kateřina; Skoumalová, Hana; Sládek, Šimon; Pierscieniuk, Piotr; Toufarová, Dagmar; Richter, Michal; Strako, Milan; Rosen, Alexander

Description:
Essays written by non-native learners of Czech, a part of AKCES/CLAC — Czech Language Acquisition Corpora. CzeSL-SGT stands for Czech as a Second Language with Spelling, Grammar and Tags. Extends the “foreign” (ciz) part...
OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- standard pro publishing metadata
- basic format of metadata: CMDI
- metadata are converted to different formats (DC, Metashare, ORE, ...)
- metadata are harvested by meta-search engines (VLO, ...)

Current status

- CLARIN Centre level B
- Data Seal of Approval 2014-2015
- 102 items with data, ~100 GB of data
- ~1000 records imported from central CLARIN database
- 12 on-line services publicly available
- integrated in European infrastructures (Weblicht, Federated Content Search, VLO)
- work on compatibility with OpenAire (HORIZONT 2020)
- implementing license selection tool
What next?

- more satisfied users
- more intuitive user interface of the repository
- enhanced workflow for item deposition (support for preallocation of PID, hassle-free upload of large files,...)
- more data
- more interoperable services
- CLARIN Centre level A certification
Problems

- unique identification of authors, objects, users (Researcher ID, PIDs, AAI)
- global authorization
- evolution of data and tools (PID, versioning, actual “runnability” of the stored tools from a long term perspective)
- legal issues (licensing, user consent with licences)
- maintaining the quality of records
Thank you for your attention

http://lindat.mff.cuni.cz