LINDAT/CLARIN Repository

Michal Josífko

ÚFAL Seminar

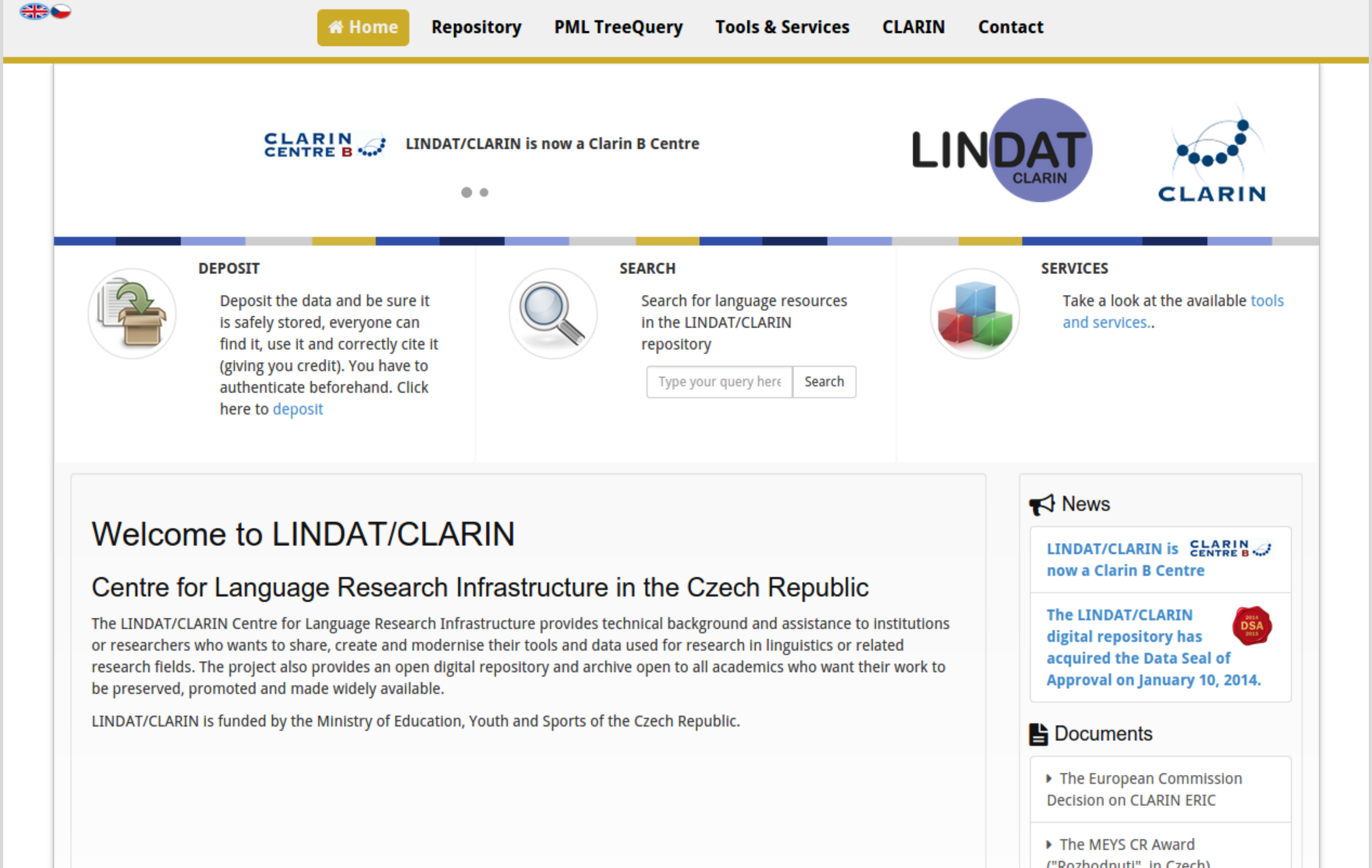
Sedlec-Prčice, September 15 -16, 2014

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



LINDAT/CLARIN is now a Clarin B Centre

DEPOSIT

Deposit the data and be sure it is safely stored, everyone can find it, use it and correctly cite it (giving you credit). You have to authenticate beforehand. Click here to deposit

SEARCH

Search for language resources in the LINDAT/CLARIN repository

SERVICES

Take a look at the available tools and services..

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 wants 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.

News

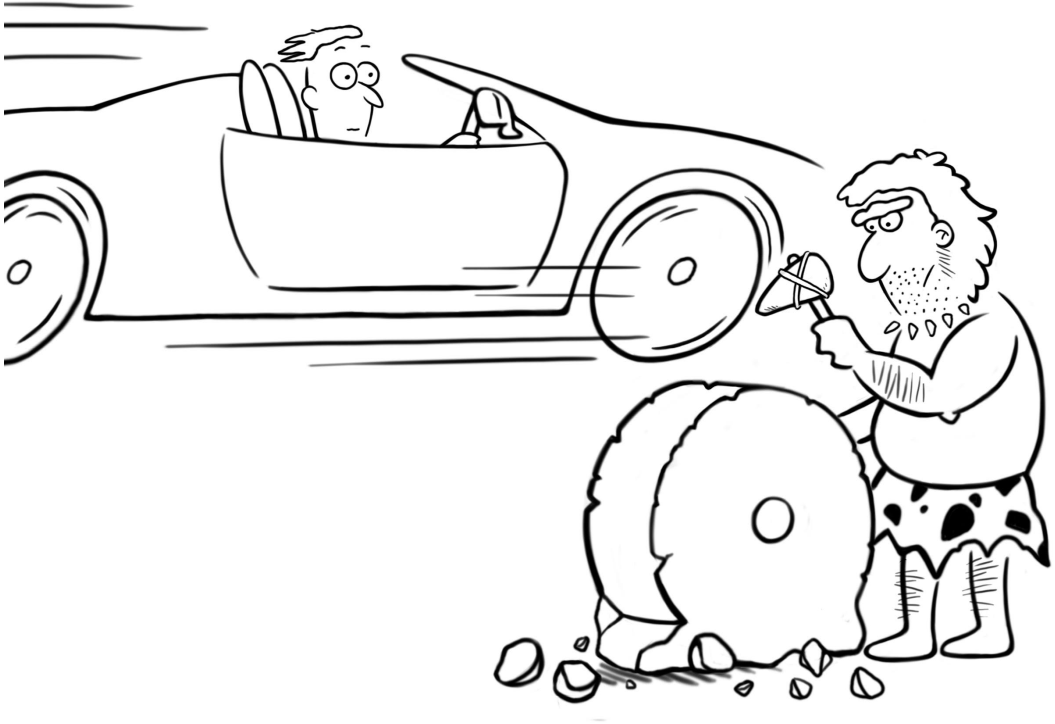
LINDAT/CLARIN is now a Clarin B Centre

The LINDAT/CLARIN digital repository has acquired the Data Seal of Approval on January 10, 2014.

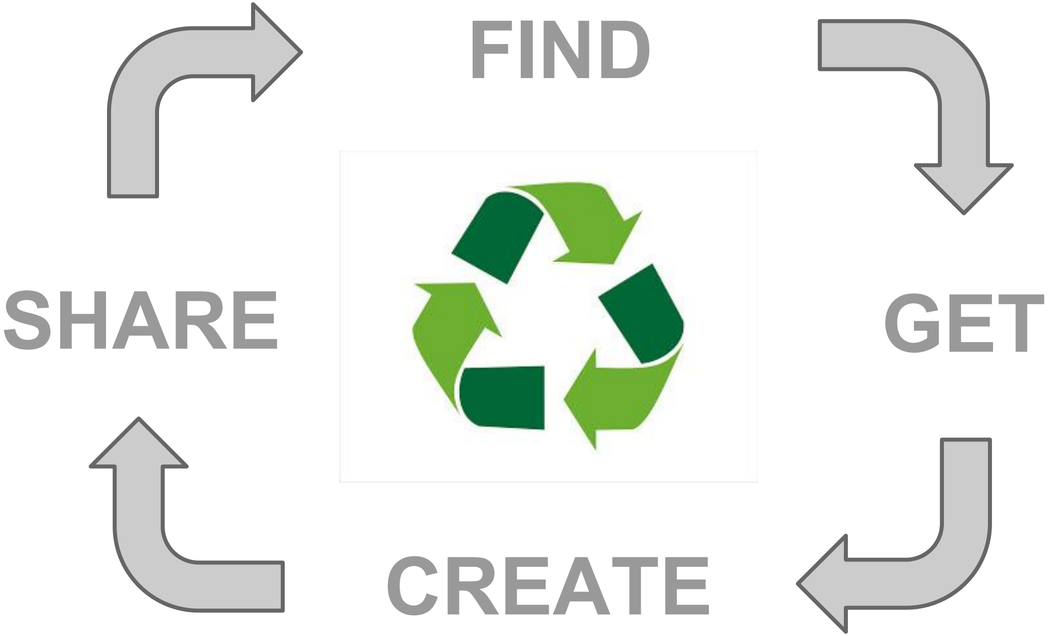
Documents

- The European Commission Decision on CLARIN ERIC
- The MEYS CR Award ("Rozhodnuti", in Czech)

Motivation: don’t reinvent the wheel



...recycle!



FIND

GET

CREATE

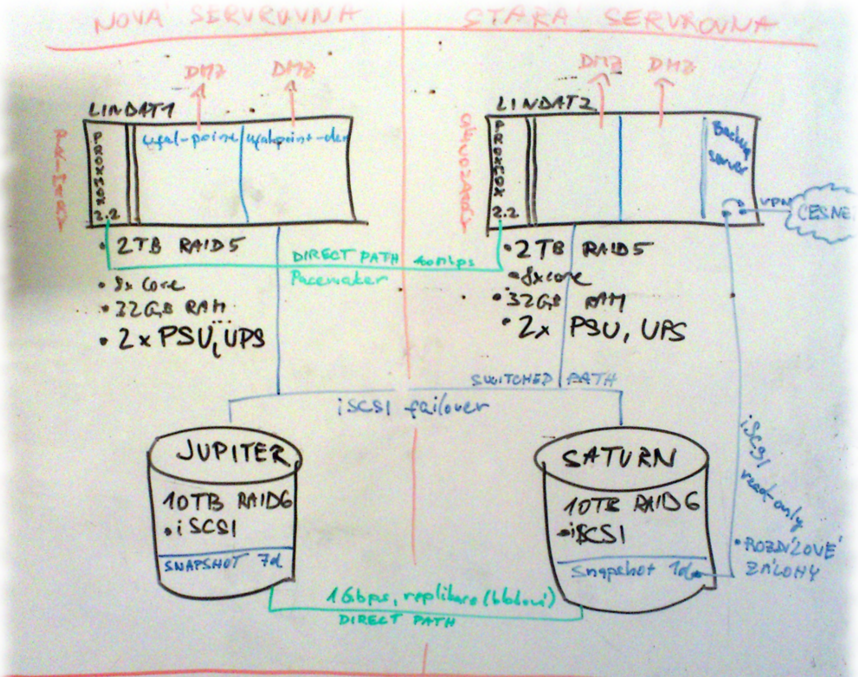
SHARE

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 skripts, Piwik implementation for access statistics coming soon

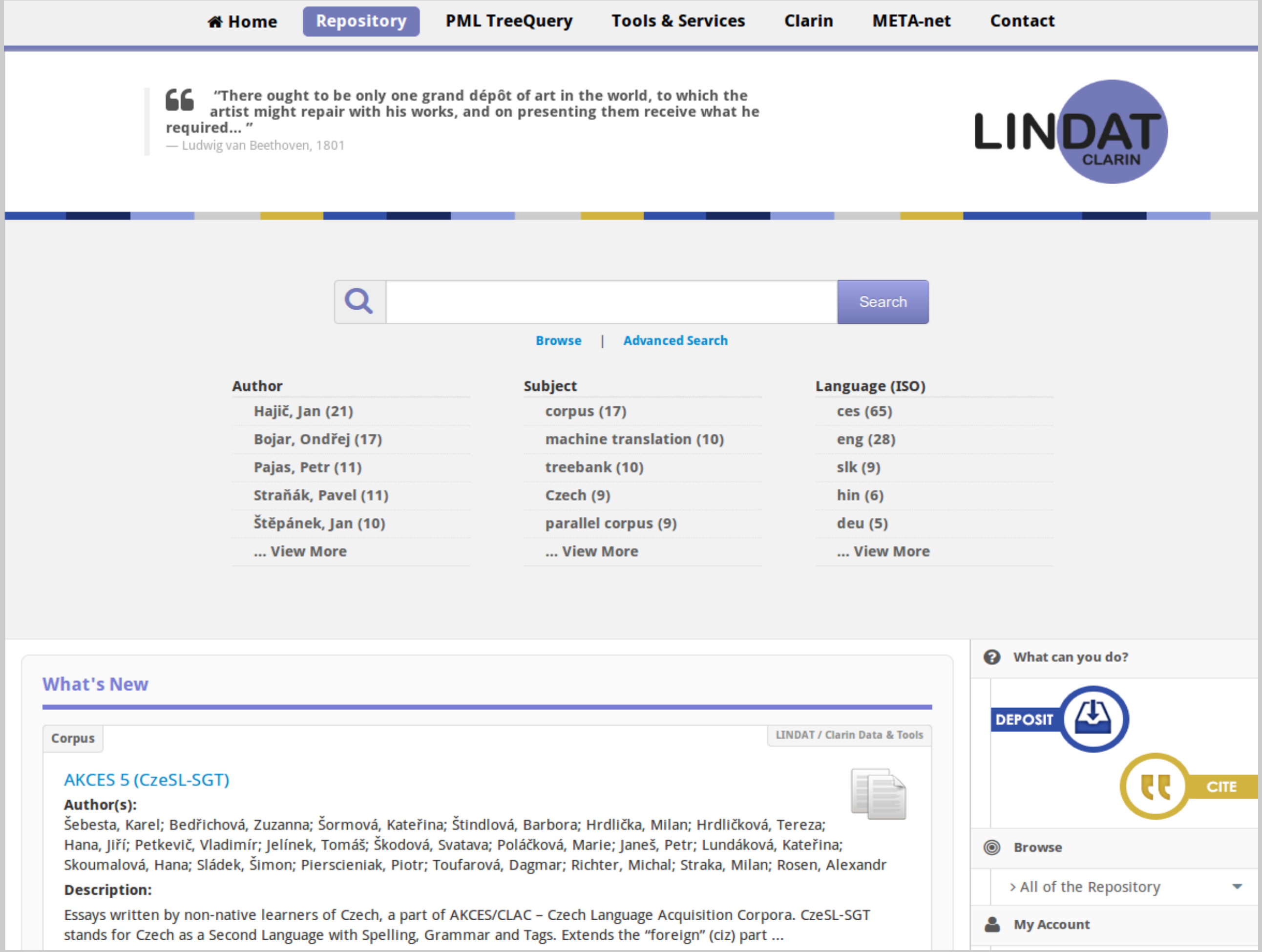


- Data i CT se ukládají společně na iSCSI pole
- Pracuje jen jeden APL server s oběma CT
- Lindat 2 zálohuje
- Pacemaker vyhodnocuje dostupnost zdrojů a řídí přesuny

VÝPADKY:
- POWER ✓
- DISK ✓
- SÍŤ
- HAVÁRIE v CT (virtuální systém)
- HAVÁRIE APL serveru (HW/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

Author	Subject	Language (ISO)
Hajič, Jan (21)	corpus (17)	ces (65)
Bojar, Ondřej (17)	machine translation (10)	eng (28)
Pajas, Petr (11)	treebank (10)	slk (9)
Straňák, Pavel (11)	Czech (9)	hin (6)
Štěpánek, Jan (10)	parallel corpus (9)	deu (5)
… View More	… View More	… View More

What's New

Corpus

LINDAT / Clarin Data & Tools

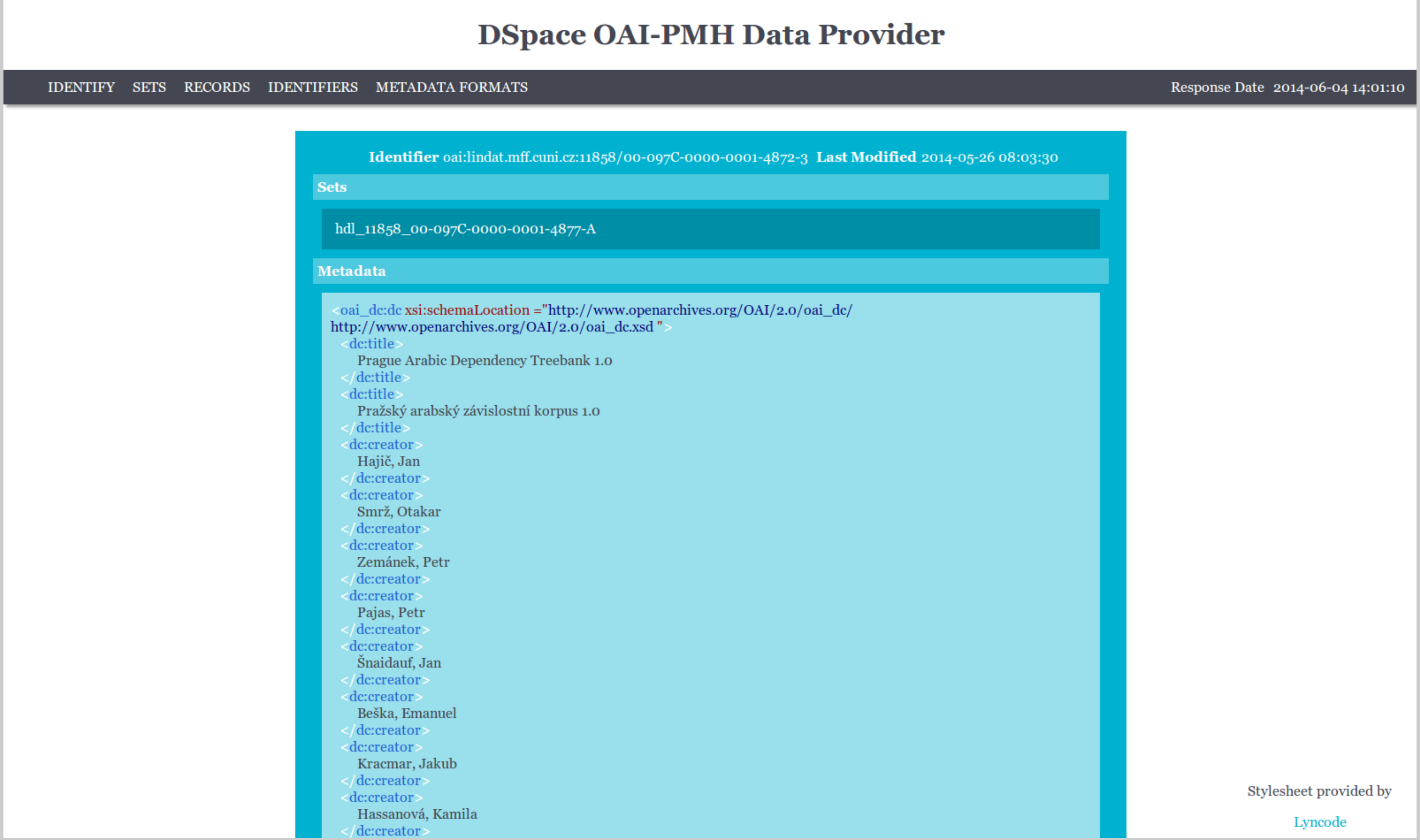
AKCES 5 (CzeSL-SGT)

Author(s):
Šebesta, Karel; Bedřichová, Zuzanna; Šormová, Kateřina; Štindlová, Barbora; Hrdlička, Milan; Hrdličková, Tereza; Hana, Jiří; Petkevič, Vladimír; Jelínek, Tomáš; Škodová, Svatava; Poláčková, Marie; Janeš, Petr; Lundáková, Kateřina; Skoumalová, Hana; Sládek, Šimon; Pierscieniak, Piotr; Toufarová, Dagmar; Richter, Michal; Straka, Milan; Rosen, Alexandr

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, ...)



DSpace OAI-PMH Data Provider

Response Date 2014-06-04 14:01:10

Identifier oai:lindat.mff.cuni.cz:11858/00-097C-0000-0001-4872-3 **Last Modified** 2014-05-26 08:03:30

Sets

hdl_11858_00-097C-0000-0001-4877-A

Metadata

```
<oai_dc:dc xsi:schemaLocation ="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd ">
  <dc:title>
     Prague Arabic Dependency Treebank 1.0
  </dc:title>
  <dc:title>
     Pražský arabský závislostní korpus 1.0
  </dc:title>
  <dc:creator>
     Hajič, Jan
  </dc:creator>
  <dc:creator>
     Smrž, Otakar
  </dc:creator>
  <dc:creator>
     Zemánek, Petr
  </dc:creator>
  <dc:creator>
     Pajas, Petr
  </dc:creator>
  <dc:creator>
     Šnaidauf, Jan
  </dc:creator>
  <dc:creator>
     Beška, Emanuel
  </dc:creator>
  <dc:creator>
     Kracmar, Jakub
  </dc:creator>
  <dc:creator>
     Hassanová, Kamila
  </dc:creator>
```

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