Sharing data in repositories

Class #14, May 16 2023

Barbora Hladká    hladka@ufal.mff.cuni.cz
Data
= information in digital form for computer processing
  ▪ text, audio, video, image, software

Metadata
= data about data
Data Analytics for Students of Social Studies and Humanities  
https://ufal.mff.cuni.cz/courses/npfl134
Data repository

= a digital infrastructure to share data
  - i.e. to preserve data and help others to find them
  - for many subjects, check e.g.,
    - https://fairsharing.org/
    - https://www.re3data.org/
LINDAT repository :: https://lindat.cz

initially LINguistic DATa only, non-linguistic content now as well
Linguistic (language) data
= resources for Linguistics and Natural Language Processing

corpora

SYN v9: large corpus of written Czech
(Charles University, Faculty of Arts, Institute of the Czech National Corpus / 2021-12-05)

Author(s):
Křen, Michal ; et al.

This item contains 1 file (21.87 GB).
Linguistic (language) data

= resources for Linguistics and Natural Language Processing

annotated corpora

Czech Named Entity Corpus 1.1
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-01-09)

Author(s):
Ševčiková, Magda; Žabokrtský, Zdeněk; Straková, Jana and Straka, Milan

This item contains 1 file (10.48 MB).

Publicly Available

url
Linguistic (language) data

= resources for Linguistics and Natural Language Processing

treebanks

Universal Dependencies 2.9
(Universal Dependencies Consortium / 2021-11-15)

Author(s):
Zeman, Daniel; et al.

This item contains 3 files (534.14 MB).

Publicly Available

see Lec. #6
Linguistic (language) data

= resources for Linguistics and Natural Language Processing

lexicons

Bosworth-Toller’s Anglo-Saxon Dictionary online
(Charles University, Faculty of Arts, Department of English Language and ELT Methodology / 2021-04-09)

Author(s):
Tichý, Ondřej ; et al.

This item contains 3 files (689.16 MB).

Publicly Available

url
Linguistic (language) tools

- Natural Language Processing tools
  - Machine translation, UDPipe, NameTag (see Lec. #7), …
- Lexicons
  - Morphology - MorphoDiTa (basic morphological analysis and synthesis)
  - Syntax - Valency lexicons for Czech and English
  - Semantics - SynSemClass (synonym verbal lexicon alias event-type ontology)
- Search services
  - many corpora, all lexicons – KonText, PML-TQ, TEITOK (see Lec. #7)
Linguistic (language) tools

- Can be downloaded
- Users install them at their computing environment (PC, notebook, OS, …)
- Users run them on their data locally (or any other computing facility, cloud, etc.)
Language tools vs. applications

- Web applications accessed through web browsers
- User opens service’s web application page, enters data or uploads file
- Run link at https://lindat.cz/services
  - e.g., NameTag

For twenty-two years, Bohumil Veselý filmed important personalities, both well-known and lesser-known, on the courtyard balcony of his flat at 38 Školská Street in Prague’s New Town.
Language tools vs. services :: https://lindat.cz/services

- Tools running @LINDAT in the background, 24/7
- Each service has its application
- Users can use it remotely, by calling scripts, or from their own code
- Means: Application Programming Interface (API)
  = a software that allows two applications to talk to each other (send request, return response), e.g., each time you check the weather on your phone, you’re using an API

- **NameTag**: curl -F 'data=@sentences.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize service  
  data  output format
Non-linguistic content :: Video

Bohumil Veselý’s Gallery (url)

Tennis player Karel Koželuh teaching at his tennis school in Horní Liboc (url)
Non-linguistic content :: Img

Danish Fungi 2020
(IEEE/CVF / 2022-01-01)

Author(s):
Picek, Lukáš; et al.
▶ show everyone

This item contains 2 files (138.95 GB).
Two ways how to get into LINDAT repository

1. [http://lindat.cz](http://lindat.cz)

   ![Search interface](image)

   **Catalogue**  **Corpora**  **Treebanks**  **ČDK**  **Bibliography**

   ![Search box](image)

   Search by type of data

   e.g. corpus or lexicon or editor

2. [http://lindat.cz/repository](http://lindat.cz/repository)
Repository items :: Basic view

- title: Czech Legal Text Treebank 2.0
- publisher: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2017-09-01
- author(s): Kríž, Vincent and Hladká, Barbora
- type: Publicly Available
- licenses: This item contains 1 file (6.67 MB)
- how many files to download: 1
Repository items :: Detailed view
Repository items :: Files to download

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Name: clt_2.0.zip
Size: 6.67 MB
Format: application/zip
Description: CLT 2.0
MD5: c0af10f6c44ff159db68e59d37825e4

Download file  Preview

File Preview
- data
- sentences
- xml
  - document_02_004.w 201 kB
  - document_02_006.a 771 kB
  - document_01_004.m 479 kB
  - document_01_003.w 143 kB
  - document_01_005.a 766 kB
  - document_01_001.w 356 kB
  - document_01_002.m 646 kB

Data Analytics for Students of Social Studies and Humanities  https://ufal.mff.cuni.cz/courses/npfl134
Repository items :: Data types

- Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)
  - Author(s): Hajič, Jan; et al.

- EngVallex - English Valency Lexicon 2.0
  - Author(s): Cinková, Šilvie; Fučíková, Eva; Šíndlerová, Jana and Hajič, Jan

- NameTag 2 Models (2021-09-16)
  - Author(s): Straková, Jana and Straka, Milan

- MorphoDiTa: Morphological Dictionary and Tagger (2014-02-14)
User activities

- Searching the repository
- Downloading data
- Storing data
Searching the repository :: http://lindat.cz

treebank

e.g. corpus or lexicon or editor
Search results and faceted search

Limit your search

Author
Subject
Rights
Language (ISO)
Type
Contain Files
Community

Showing 1 through 10 out of 113 results

Czech Legal Text Treebank 2.0
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)) (2017-09-01)
Author(s):
Kolář, Vincenc and Hájková, Barbora
Publicly Available

Prague Czech-English Dependency Treebank 2.0
Coref
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)) (2018-03-30)
Author(s):
Nedoluzhko, Anna; Nováček, Michal; Cimolová, Silvia; Mikulovič, Marek and Mikolová, Jana
Restricted Use

Data Analytics for Students of Social Studies and Humanities  https://ufal.mff.cuni.cz/courses/npfl134
User activities

- Searching the repository
- **Downloading data**
- Storing data
Downloading data

- **Login is not required** when downloading data the license of which allows free sharing, i.e. Creative Commons licenses and Open Access licenses.
Downloading data

- **Login is required** when downloading data and tools with the required which require a license agreement to be signed.
Login

- Find an academic institution where you have an account
- If it does not exist (= you’re a “homeless” researcher), create a CLARIN account at [url](https://ufal.mff.cuni.cz/courses/npfl134)
User activities

- Searching the repository
- Downloading data
- Storing data
Storing data :: Upload and don’t worry

= uploading datasets = creating new items in the repository
Why to create a new item

Your work is visible
Why to create a new item

Your work is visible :: Direct data citation, integration with Google scholar

enTenTen

Please use the following text to cite this item or export to a predefined format:

Masaryk University, NLP Centre, 2011, enTenTen. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8.

bibtex

@misc{11858/00-097C-0000-0001-CCDF-8,
title = {{enTenTen}},
url = {http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8},
note = {{{LINDAT}/CLARIAH}-cz digital library at the Institute of Formal and Applied Linguistics {{(ÚFAL)}}, Faculty of Mathematics and Physics, Charles University},
copyright = {{(NLP Centre Web Corpus License)}},
year = {2012} }
Why to create a new item

Your work is **permanently visible**

- Persistent identifier PID (= stable reference)
- PID makes a URL that always leads to the metadata
  - it always works, even when moving data
  - included in the citation
Why to create a new item

Your work can **make other people happy**

- Licencing
- LINDAT team supports open data sharing while respecting the licenses of your data
  - Open Access in 5 minutes ([url](https://ufal.mff.cuni.cz/courses/npfl134))
    - free and open online access to academic information, such as publications and data
Why to create a new item

Your work is **safe**

- on/off-site backup
Storing data

= uploading datasets = creating new items in the repository, [help](https://ufal.mff.cuni.cz/courses/npfl134)

1. Login
2. Submissions, Start
3. Select the data type
4. Fill in the metadata
   = describe a new item
5. Upload the files
6. Select a license
7. Fill in a note for the editors
8. Check the added information
9. Submit the item to the editors for review
Storing data :: Versions

- Prefer latest, preserve all
- Changes → new version, new PID, new citation
Storing data :: Connection to services

- Greater user comfort

https://lindat.cz/services/teitok/


- **Findable**  
The first step in (re)using data is to find them, incl. metadata
- **Accessible**  
Once we find the data we need to know how to access them
- **Interoperable**  
The data can be exchanged and used across different applications and systems
- **Reusable**  
The data are well documented and curated
LINDAT Helpdesk

Harvesting metadata

- extract metadata from various sources
- **OAI-PMH** = a protocol for metadata harvesting that specifies how repositories can expose their metadata for other to harvest
CLARIN Virtual Language Observatory

- **CLARIN** is a digital infrastructure offering data, tools and services to support research based on language resources. It operates through the national centers all over Europe (LINDAT is one of them)
- Virtual Language Observatory (VLO) [https://vlo.clarin.eu](https://vlo.clarin.eu)
  - online overview of the data that is available at a variety of computing centres