Sharing data in repositories
Class #7, March 29 2022
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

Metadata

<table>
<thead>
<tr>
<th>Code</th>
<th>Date</th>
<th>Type de document</th>
<th>Nb de f.</th>
<th>Langue</th>
<th>Auteur du document</th>
<th>Lieu</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMA.8.7.1</td>
<td>10.03.1908</td>
<td>Lettre manuscirote</td>
<td>1</td>
<td>allemand</td>
<td>Hartmann, Erich</td>
<td>Bautzen</td>
</tr>
<tr>
<td>AMA.8.12.34</td>
<td>24.11.1913</td>
<td>Lettre dactylographiée</td>
<td>1</td>
<td>allemand</td>
<td>Böhlme, Erich</td>
<td>Berlin</td>
</tr>
<tr>
<td>AMA.8.13.40</td>
<td>26.04.1914</td>
<td>Carte manuscirote</td>
<td>1</td>
<td>allemand</td>
<td>Irmer, Hermann</td>
<td>Harkov</td>
</tr>
<tr>
<td>AMA.8.13.27</td>
<td>01.04.1914</td>
<td>Lettre manuscirote</td>
<td>1</td>
<td>allemand</td>
<td>Leskien, August</td>
<td>Leipzig</td>
</tr>
<tr>
<td>AMA.8.13.31</td>
<td>06.04.1914</td>
<td>Lettre manuscirote</td>
<td>1</td>
<td>anglais</td>
<td>Minrs, Ellis H.</td>
<td>Cambridge</td>
</tr>
<tr>
<td>AMA.8.12.12</td>
<td>04.04.1913</td>
<td>Lettre dactylographiée</td>
<td>1</td>
<td>anglais</td>
<td>Miller, Arthur William Kaye</td>
<td>Londres</td>
</tr>
<tr>
<td>AMA.8.12.25</td>
<td>29.07.1913</td>
<td>Lettre manuscirote et dactylographiée</td>
<td>1</td>
<td>anglais</td>
<td>Miller, Arthur William Kaye</td>
<td>Londres</td>
</tr>
<tr>
<td>AMA.8.15.4</td>
<td>06.05.1919</td>
<td>Lettre manuscirote</td>
<td>2</td>
<td>bulgare</td>
<td>Stolov, Nikola</td>
<td>indéterminé</td>
</tr>
<tr>
<td>AMA.8.13.37</td>
<td>16.04.1914</td>
<td>Lettre manuscirote</td>
<td>1</td>
<td>français</td>
<td>Demidov, Elim Pavlovich</td>
<td>Athènes</td>
</tr>
</tbody>
</table>
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, now non-linguistic content 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.

▶ show everyone

This item contains 1 file (21.87 GB).

**Academic Use**
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):
Ševčíková, Magda; Žabokrtský, Zdeněk; Straková, Jana and Straka, Milan

This item contains 1 file (10.48 MB).

Publicly Available
Linguistic (language) data

= resources for Linguistics and Natural Language Processing

treebanks
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.

show everyone

This item contains 3 files (689.16 MB).

Publicly Available

url
Linguistic (language) tools

- Natural Language Processing tools
  - Machine translation, UDPipe (see Lec. #5), NameTag, …
- 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. #5)
Linguistic (language) tools

- Can be downloaded
- Users install at their computing environment (PC, notebook, OS, …)
- Users run 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](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`
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).

Publicly Available

url
Two ways how to get into LINDAT repository

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

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

   ![Search Bar]

   - 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
  (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2017-09-01)
- Author(s): Krž, 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: c0ae10f0c49ff159db68e59d37525e4

Download file  Preview

File Preview

* data
  * sentences
    * prfl
      * 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
Repository items :: Data types

Corpus

Lexical conceptual

Language description

Technology / Tool / Service

Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Lexical Conceptual Resource" includes lexica, ontologies, dictionaries, word lists etc. "Language Description" covers language models and grammars. "Technology / Tool / Service" is used for tools, systems, system components etc.

Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2020)
Author(s): Hajic, Jan; et al.
► show everyone
This item contains 1 file (2.6 GB).
Publicly Available

MorphoDiTa: Morphological Dictionary and Tagger
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14)

EngVallex - English Valency Lexicon 2.0
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2021-02-25)
Author(s): Cinková, Silvie; Fučíková, Eva; Šindlerová, Jana and Hajič, Jan
This item contains 1 file (9.52 MB).
Publicly Available
User activities

- Searching the repository
- Downloading data
- Storing data
Searching the repository :: [http://lindat.cz](http://lindat.cz)

**Catalogue**  **Corpora**  **Treebanks**  **CDK**  **Bibliography**

```
treebank
```
e.g. [corpus](#) or [lexicon](#) or [editor](#)
Search results and faceted search
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](url)
User activities

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

= uploading datasets = creating new items in the repository

Catalogue

LINDAT/CLARIAH-CZ is a Czech centre for data providing certified storage and natural language processing services

Who can download? Who can deposit? Publication schedule

Cite  Deposit
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
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)
    - 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

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

Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)

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

Hajič, Jan; et al., 2020, Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), 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/11234/1-3185.

This resource is also integrated in following services:

- TEITOK
- TEITOK@LINDAT

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

Corpora

Below is the list of corpora in the TEITOK/Context hybrid set-up, hosted at the UFAL institute. To get a repository. For corpora that have previous in TEITOK, you can click on the version number to see a

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Latest Token size</th>
<th>Corpus Type</th>
<th>Corpus Status</th>
<th>Corpus Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDT-C</td>
<td>1.0</td>
<td>4M Treebank</td>
<td>stable</td>
<td>Czech</td>
</tr>
<tr>
<td>ParCzech</td>
<td>27M</td>
<td>Parliamentary</td>
<td>stable</td>
<td>Czech</td>
</tr>
<tr>
<td>Skipt 2015</td>
<td>400k</td>
<td>Learner Corpus</td>
<td>live</td>
<td>Czech</td>
</tr>
<tr>
<td>Universal</td>
<td>26M</td>
<td>Treebank</td>
<td>stable</td>
<td>Many</td>
</tr>
</tbody>
</table>


- **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

HW #4 assignment

1. Explore LINDAT repository https://lindat.cz/repository
   a. Select a data set that interests you the most
   b. Explain why you are interested in this data set and what you would use it for
   c. Optional: If you have enough space on your hard drive/in your cloud storage, download it and examine it carefully. Provide more details about the data set.

2. Train the LINDAT submission procedure
   a. Select any data set you want to share
   b. Fill a submission form out, see

   Download a template from https://ufal.mff.cuni.cz/courses/npfl134/credit), add your answers and send the file to hladka@ufal.mff.cuni.cz by Apr 5 2022.
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