Sharing data in repositories
Class #14, May 16 2023

Barbora Hladká hladka@ufal.mff.cuni.cz

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

Data

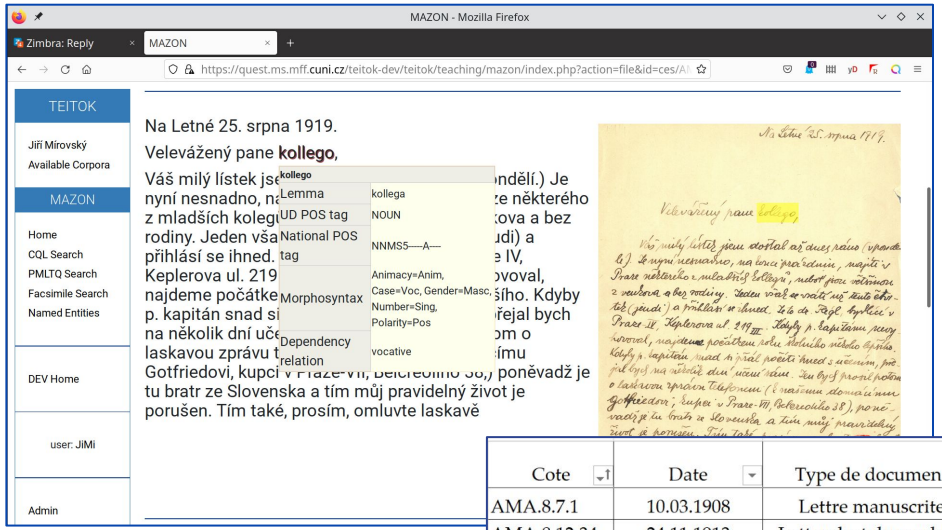
= information in digital form for computer processing

- text, audio, video, image, software

Metadata

= data about data

Data



Metadata

Cote	Date	Type de document	Nb de f.	Langue	Auteur du document	Lieu
AMA.8.7.1	10.03.1908	Lettre manuscrite	1	allemand	Hartmann, Erich	Bautzen
AMA.8.12.34	24.11.1913	Lettre dactylographiée	1	allemand	Böhme, Erich	Berlin
AMA.8.13.40	26.04.1914	Carte manuscrite	1	allemand	Irmer, Hermann	Harkov
AMA.8.13.27	01.04.1914	Lettre manuscrite	1	allemand	Leskien, August	Leipzig
AMA.8.13.31	06.04.1914	Lettre manuscrite	1	anglais	Minns, Ellis H.	Cambridge
AMA.8.12.12	04.04.1913	Lettre dactylographiée	1	anglais	Miller, Arthur William Kaye	Londres
AMA.8.12.25	29.07.1913	Lettre manuscrite et dactylographiée	1	anglais	Miller, Arthur William Kaye	Londres
AMA.8.15.4	06.05.1919	Lettre manuscrite	2	bulgare	Stoilov, Nikola	indéterminé
AMA.8.13.37	16.04.1914	Lettre manuscrite	1	français	Demidov, Elim Pavlovich	Athènes

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

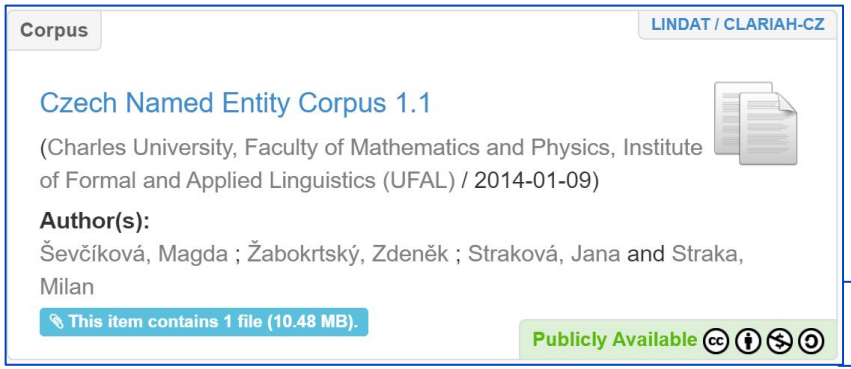
corpora

url

Linguistic (language) data

= resources for Linguistics and Natural Language Processing

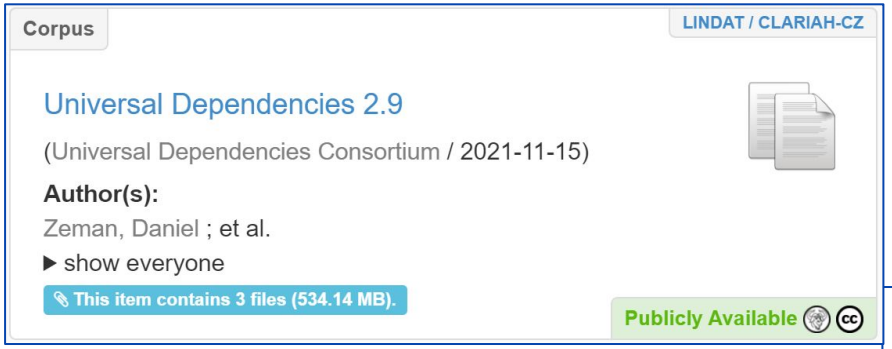
annotated corpora

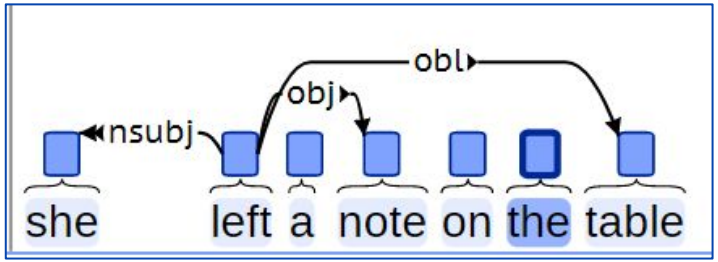


Linguistic (language) data

= resources for Linguistics and Natural Language Processing

treebanks



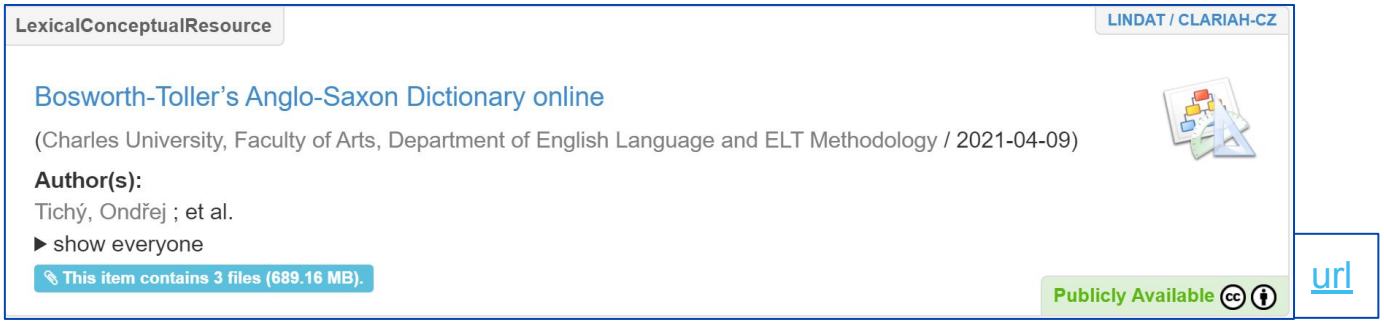


see Lec. #6

Linguistic (language) data

= resources for Linguistics and Natural Language Processing

lexicons



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
```

data: `'data=@sentences.txt'`

output format: `'output=vertical'`

service: `http://lindat.mff.cuni.cz/services/nametag/api/recognize`

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

Corpus

LINDAT / CLARIAH-CZ

Danish Fungi 2020

(IEEE/CVF / 2022-01-01)

Author(s):

Picek, Lukáš ; et al.

▶ show everyone

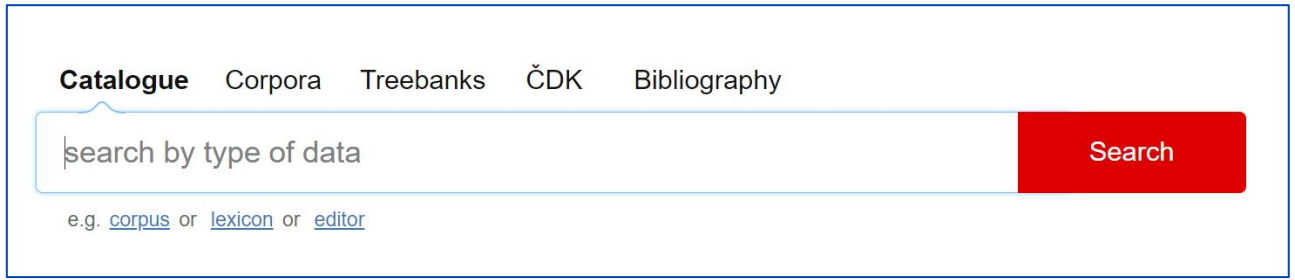
This item contains 2 files (138.95 GB).

Publicly Available

url

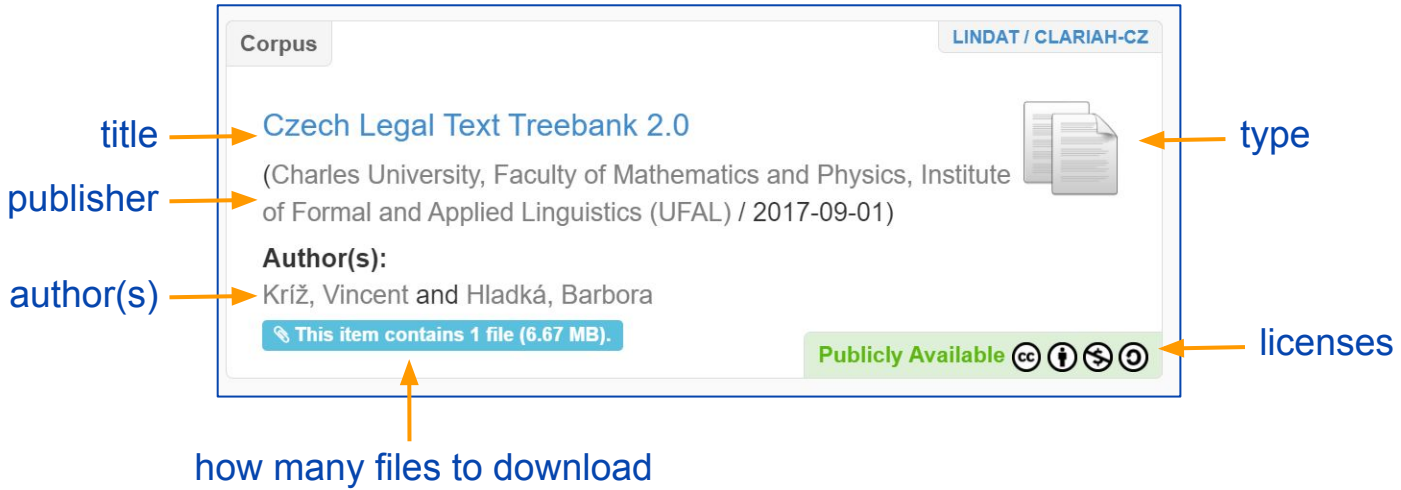
Two ways how to get into LINDAT repository

1. http://lindat.cz

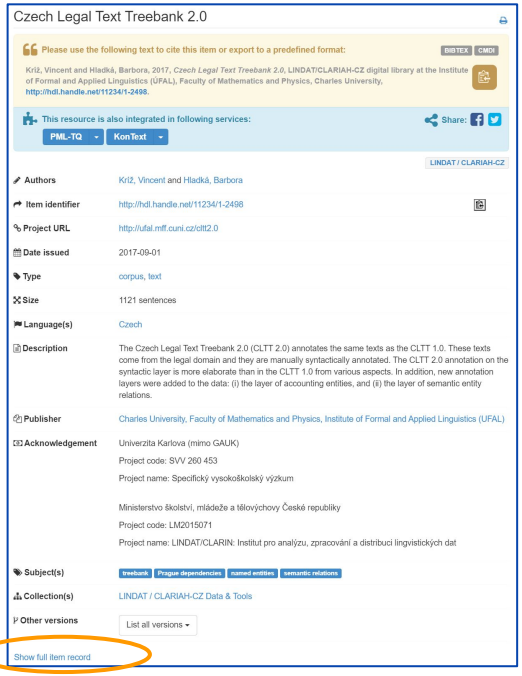


2. http://lindat.cz/repository

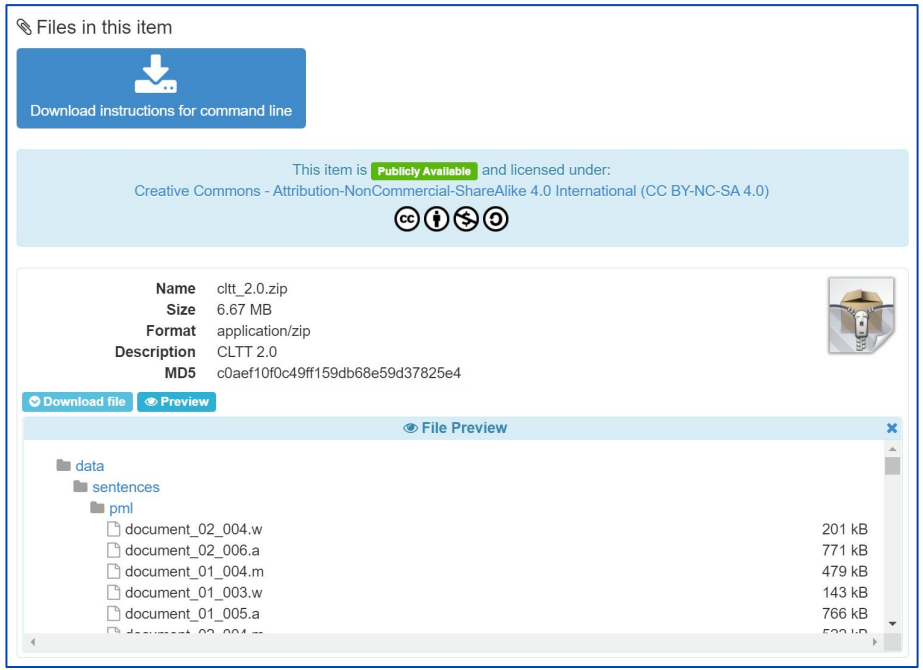
Repository items :: Basic view



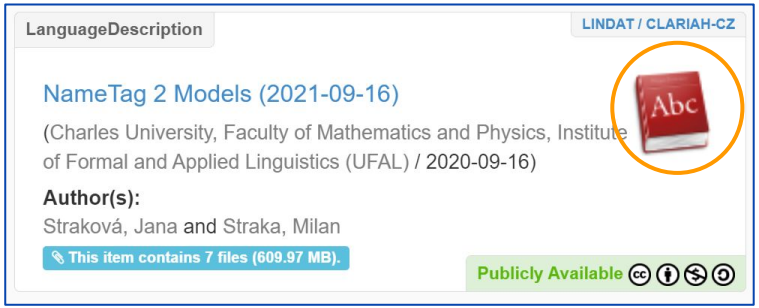
Repository items :: Detailed view

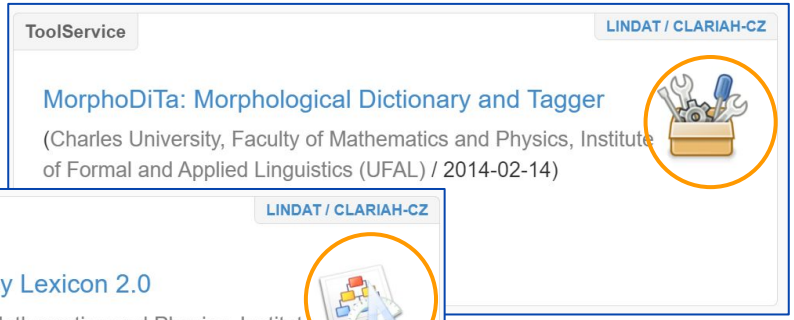


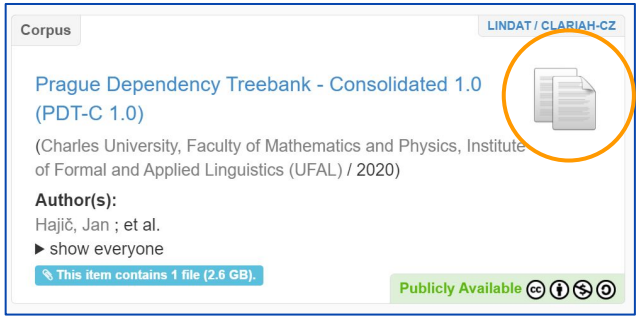
Repository items :: Files to download



Repository items :: Data types



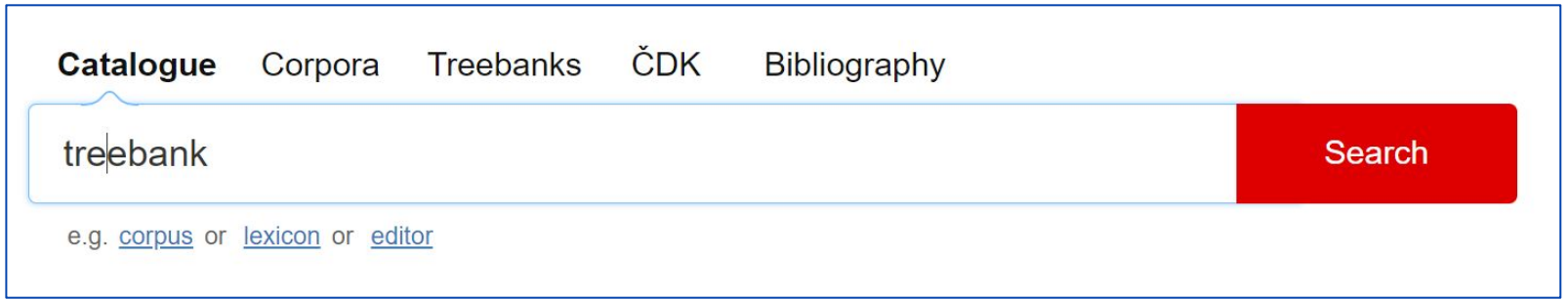




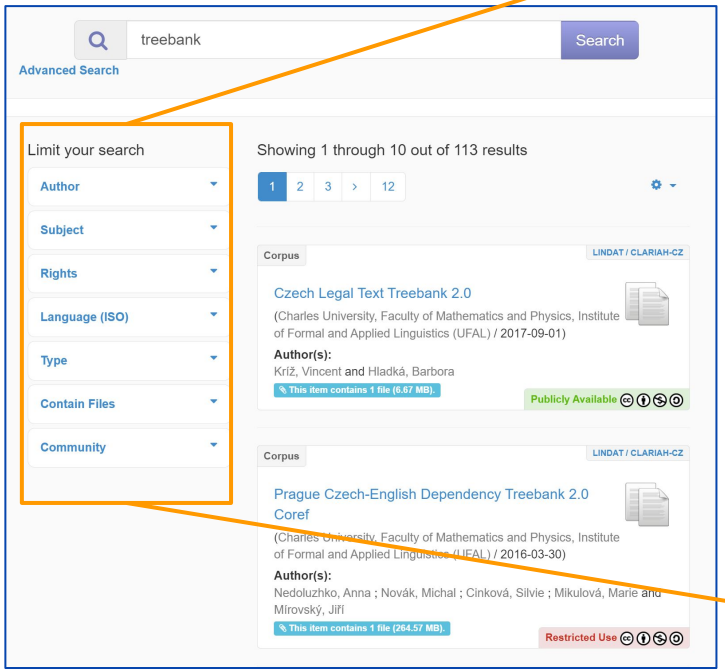
User activities

- Searching the repository
- Downloading data
- Storing data

Searching the repository :: http://lindat.cz



Search results and faceted search

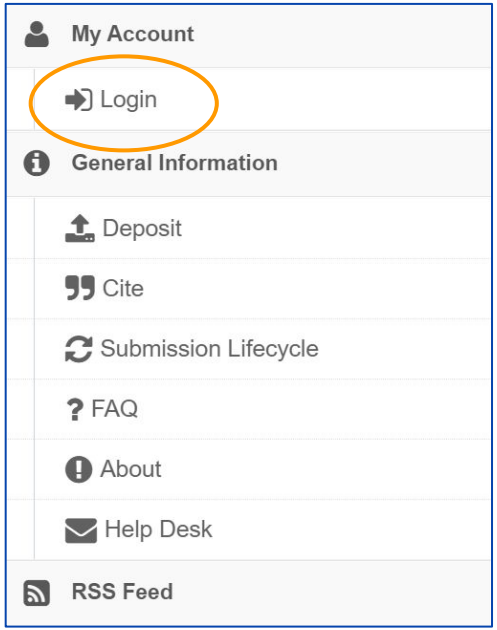


User activities

- Searching the repository
- Downloading data
- Storing data

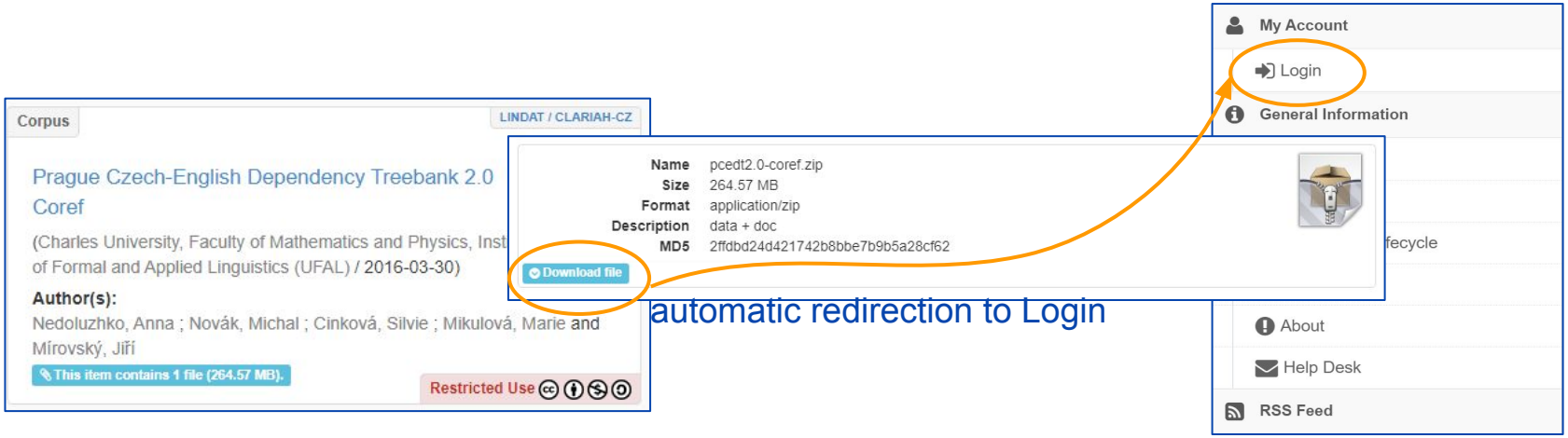
Downloading data

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



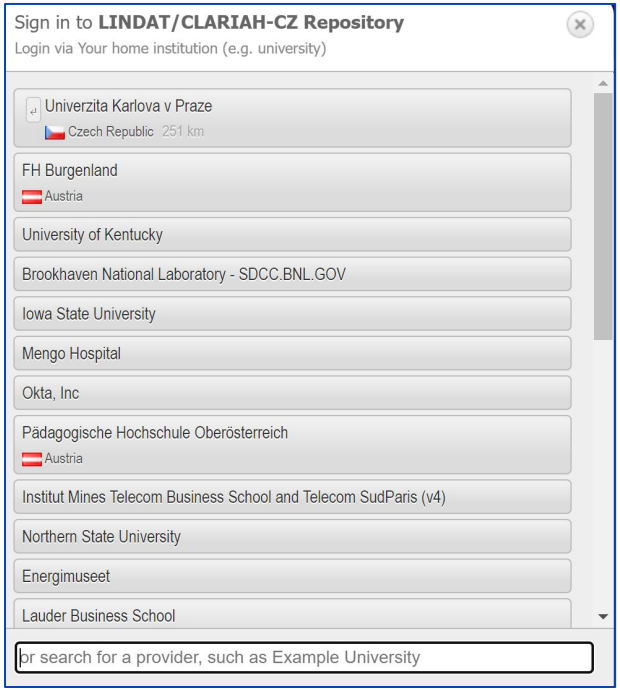
Downloading data

- 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



User activities

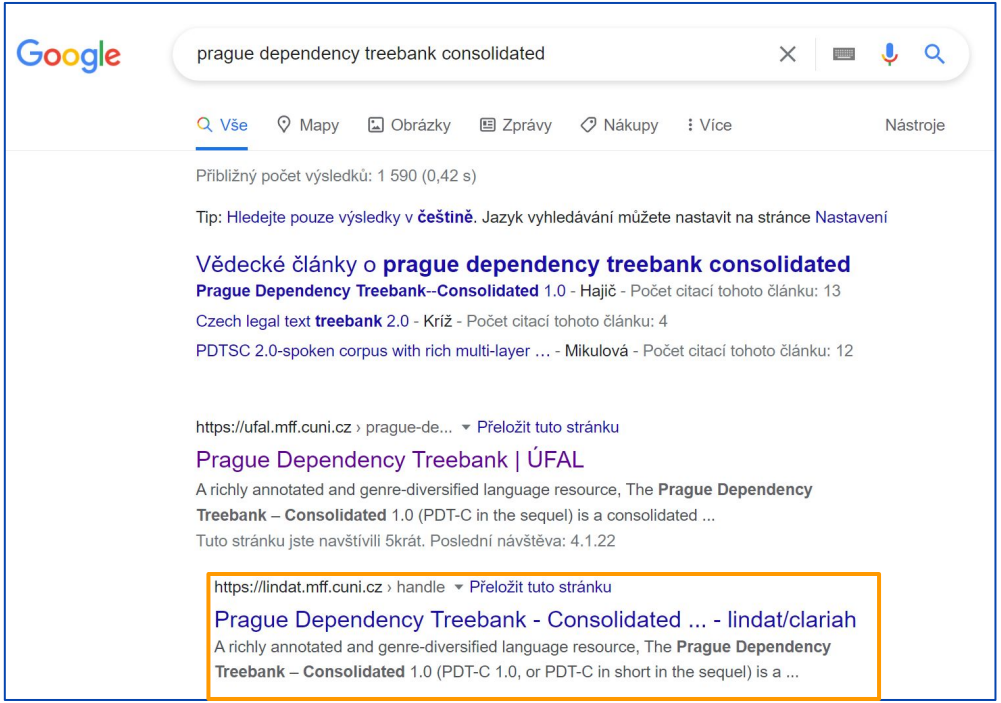
- Searching the repository
- Downloading data
- Storing data

Storing data :: Upload and don't worry

= uploading datasets = creating new items in the repository

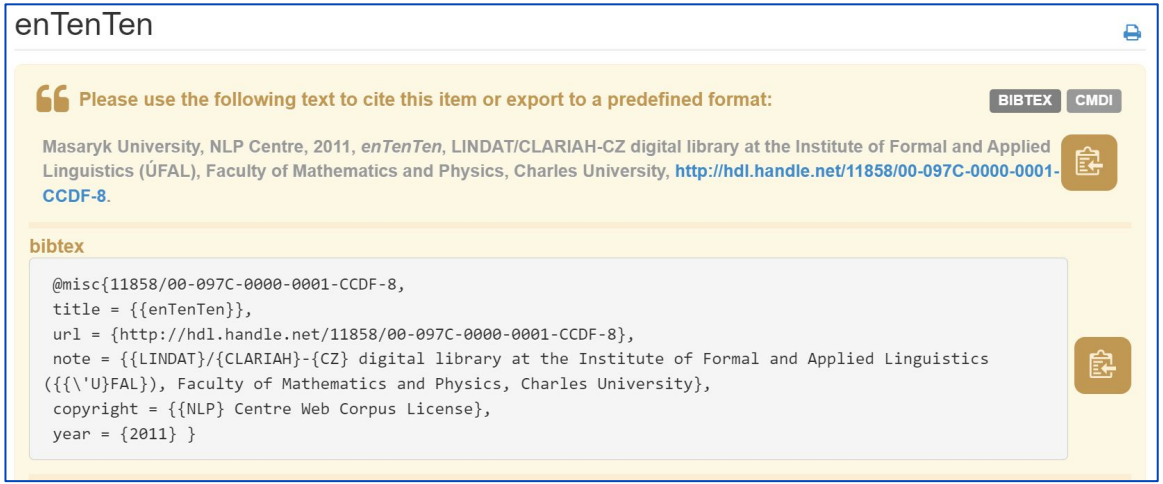
Why to create a new item

Your work is visible



Why to create a new item

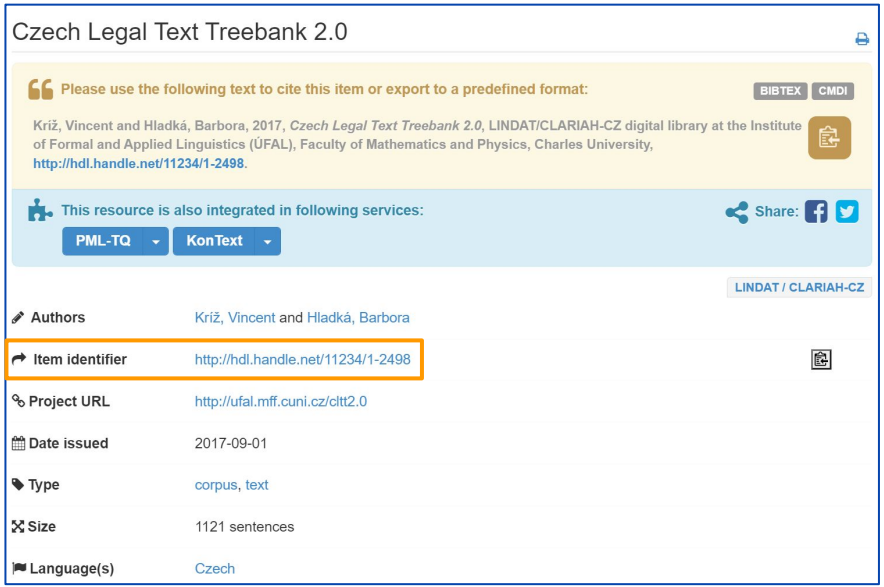
Your work is visible :: Direct data citation, integration with Google scholar



Why to create a new item

Your work is <u>permanently</u> 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

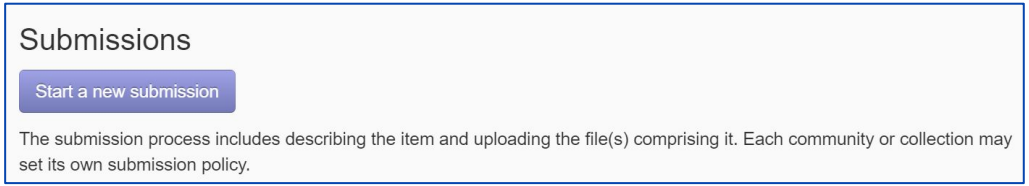
Your work is safe

- on/off-site backup

Storing data

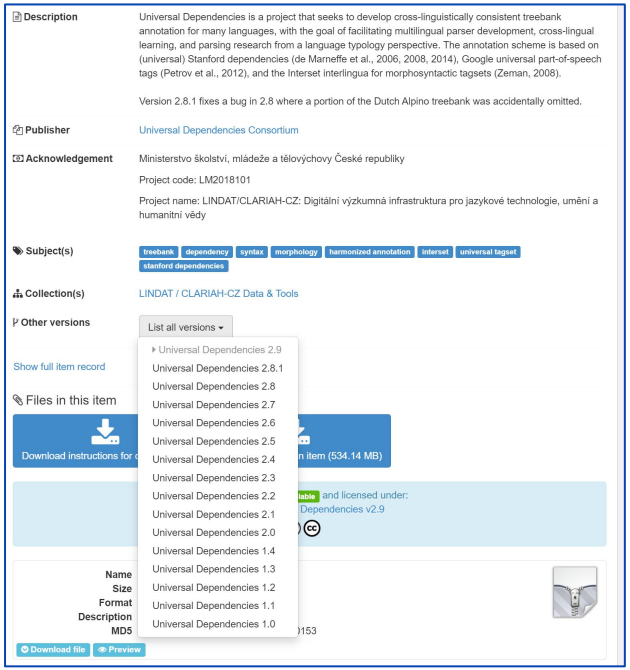
= uploading datasets = creating new items in the repository, help

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



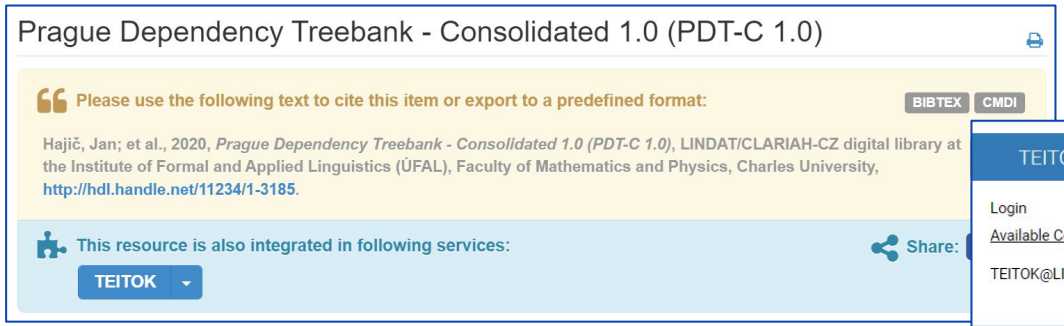
Storing data :: Versions

- Prefer latest, preserve all
- Changes → new version,
  new PID,
  new citation

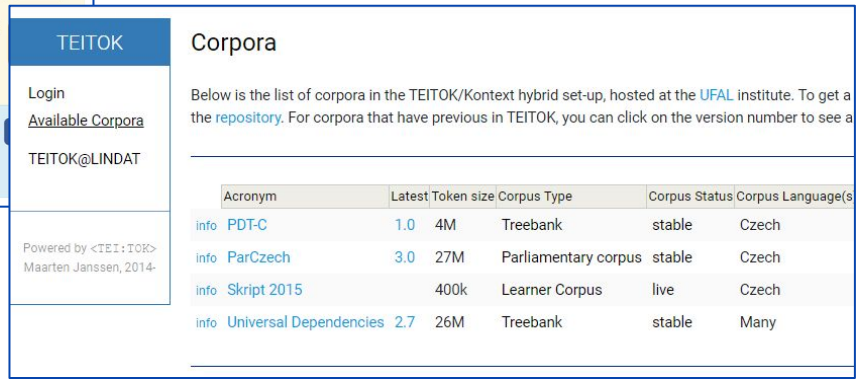


Storing data :: Connection to services

- Greater user comfort



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



http://lindat.mff.cuni.cz/services/teitok/pdtc10/index.php

Sum up :: FAIR data principles https://en.wikipedia.org/wiki/FAIR_data

- 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

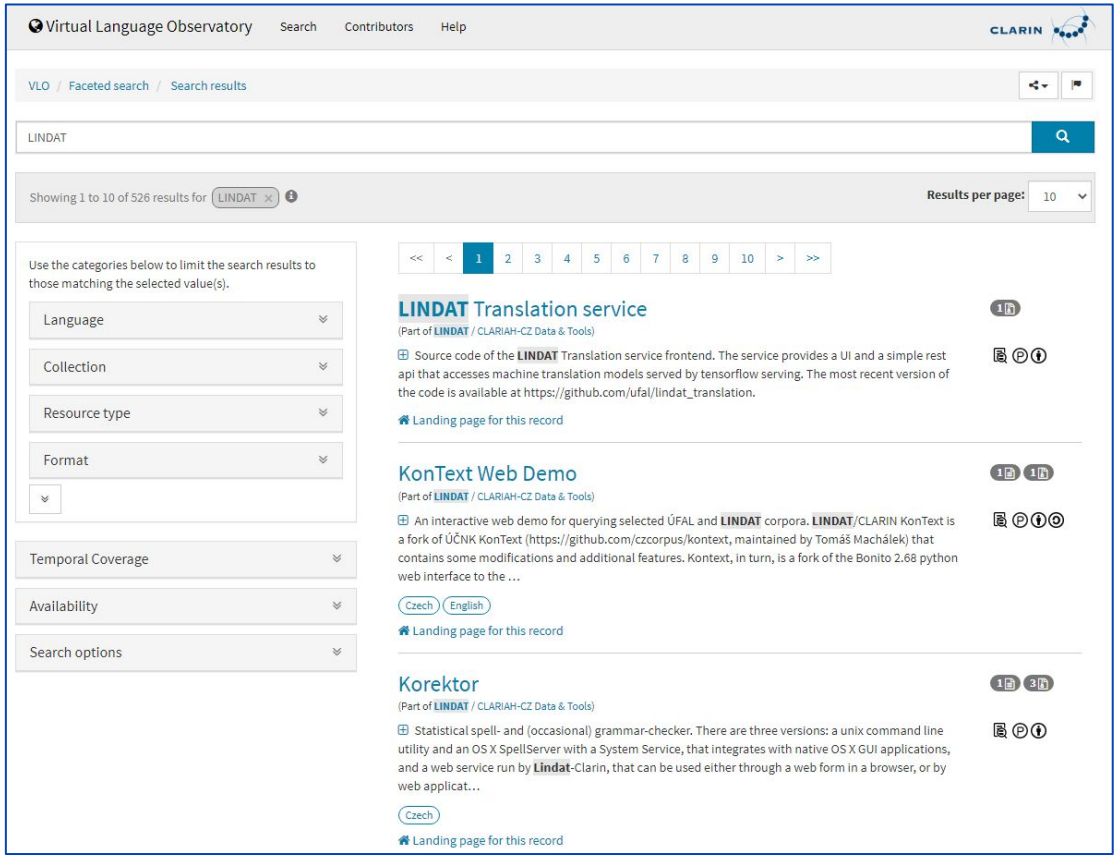
- Frequently Asked Questions https://lindat.cz/faq-repository

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
  - online overview of the data that is available at a variety of computing centres


Virtual Language Observatory
Search
Contributors
Help
CLARIN
VLO / Faceted search / Search results
LINDAT
Showing 1 to 10 of 526 results for LINDAT
Results per page: 10
Use the categories below to limit the search results to those matching the selected value(s).
Language
Collection
Resource type
Format
Temporal Coverage
Availability
Search options
<< < 1 2 3 4 5 6 7 8 9 10 > >>
LINDAT Translation service
(Part of LINDAT / CLARIAH-CZ Data & Tools)
Source code of the LINDAT Translation service frontend. The service provides a UI and a simple rest api that accesses machine translation models served by tensorflow serving. The most recent version of the code is available at https://github.com/ufal/lindat_translation.
Landing page for this record
KonText Web Demo
(Part of LINDAT / CLARIAH-CZ Data & Tools)
An interactive web demo for querying selected ÚFAL and LINDAT corpora. LINDAT/CLARIN KonText is a fork of ÚČNK KonText (https://github.com/czcorpus/kontext, maintained by Tomáš Machálek) that contains some modifications and additional features. Kontext, in turn, is a fork of the Bonito 2.68 python web interface to the …
Czech
English
Landing page for this record
Korektor
(Part of LINDAT / CLARIAH-CZ Data & Tools)
Statistical spell- and (occasional) grammar-checker. There are three versions: a unix command line utility and an OS X SpellServer with a System Service, that integrates with native OS X GUI applications, and a web service run by Lindat-Clarin, that can be used either through a web form in a browser, or by web applicat…
Czech
Landing page for this record