Data citations at LINDAT/CLARIAH-CZ

Make data easy to cite.

How to do it?

How to find those citations?

Pavel Straňák
Charles University

SSHOC Workshop: Data Citation in Practice

Date: June 15 2021
Time: 13:30–15:00 CEST
Venue: online

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Stable, unchanging data with a stable reference

- solution: Repository with a PID
- our solution: lindat.cz/repository
  - https://lindat.cz/faq-repository#what-is-deposit-procedure
  - Fill a form, drag&drop data, submit 🥳
  - data gets a PID that makes a URL
  - URL always leads to the metadata, information about data
  - Stable: once published it stays that way (except typos)
  - Versioned: changes → new version, new PID and new citation

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

The repository

- we use CLARIN-DSpace
  - https://github.com/ufal/clarin-dspace/
  - DSpace is very popular “out-of-the-box” repository system
  - set-up and extended to serve data nicely
- deployed at about 10 CLARIN centres + a few more places
- aimed at ease of use and maintainance
- user perspective: provide safety, visibility and promote citations

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Item submission

Basic Info (✓) — 2. Who's involved — Describe (✓) — Upload (✓) — License (✓) — Note (✓) — Review (✓) — 8. Complete

People and organizations involved

Authors

Last name, *e.g. Smith*

First name(s) + "Jr", *e.g. Donald Jr*

Add

Enter the names of the authors of this item. Start typing the author's last name and use autocomplete form that will appear if applicable. End your input by pressing ESC if you don't want to use the preselected value.

Publisher

Add

Enter the name of the publisher of the previously issued instance of this item, or your home institution. Start typing the publisher and use autocomplete form that will appear if applicable. End your input by pressing ESC if you don't want to

This project is funde

Item submission

Basic Info ✓ — Who's involved ✓ — Describe ✓ — 4. Upload — 5. License — 6. Note — 7. Review — 8. Complete

Upload File(s)

File

Choose File | No file chosen

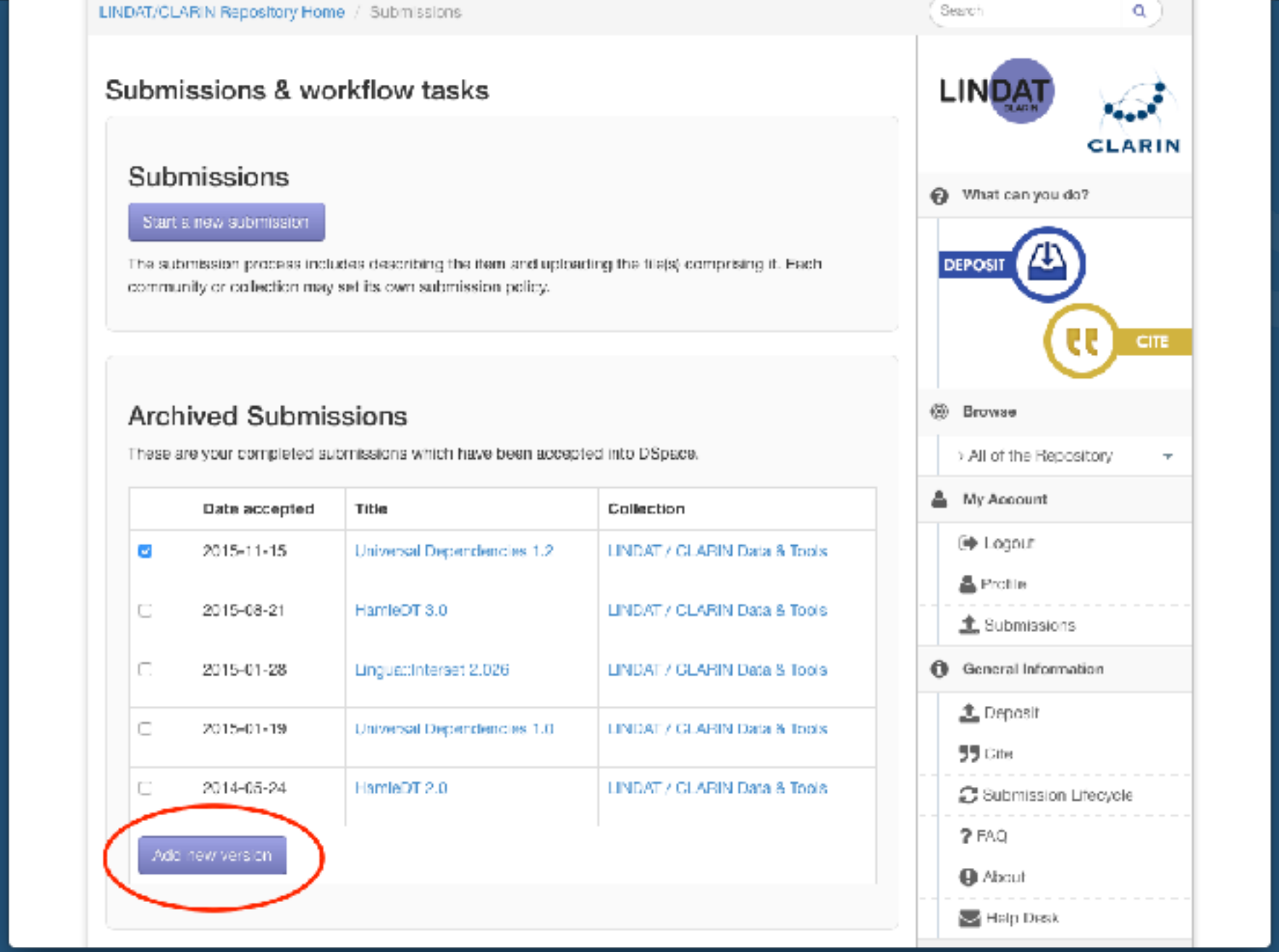
Please enter the full path of the file on your computer corresponding to your item. If you click "Browse...", a new window will allow you to select the file from your computer.

When uploading language resources, please try to use one of the recommended formats mentioned in LRT Standards

Uploading **files larger than** requires special handling. Please contact Help Desk about how to upload these files. Thank you for your understanding.

Drop file(s) here.

This project is funde



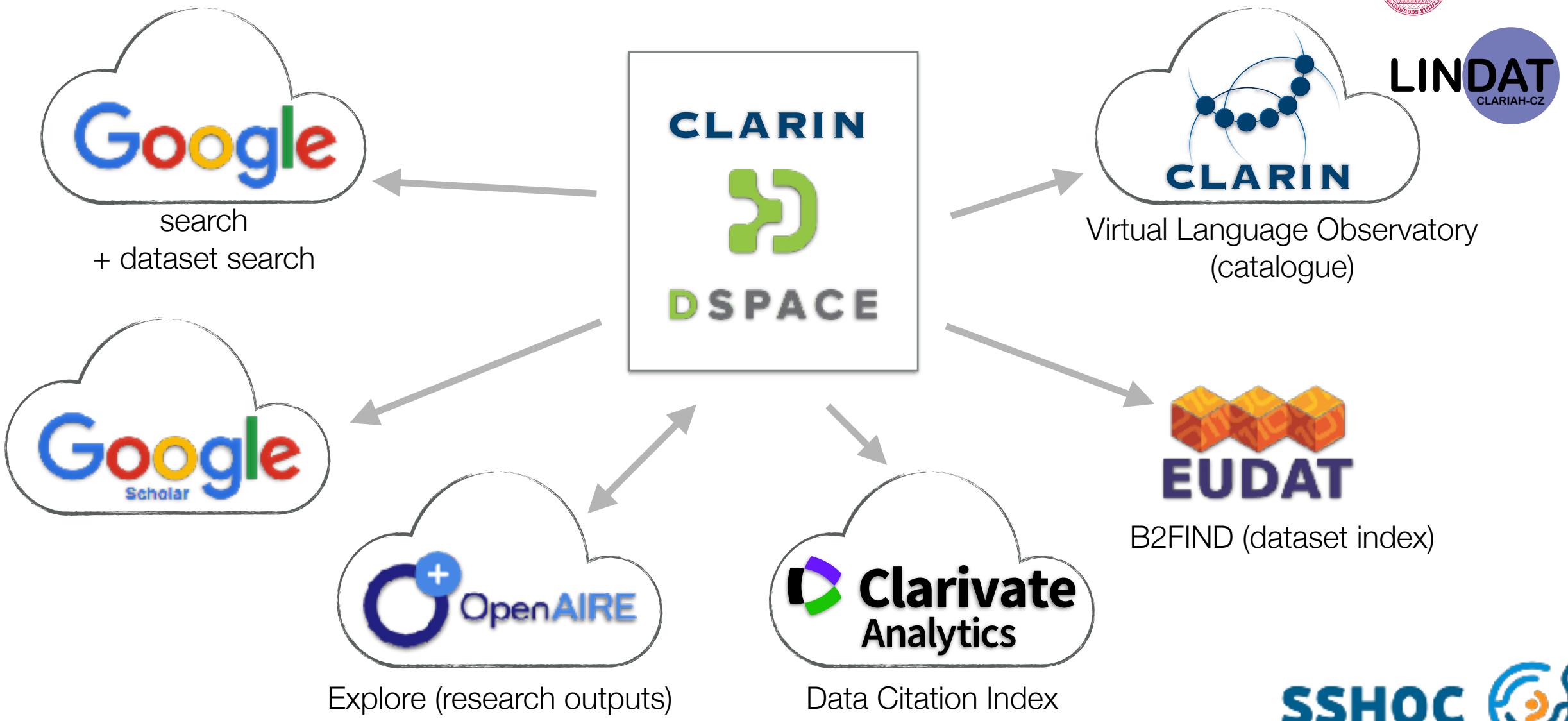
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. XXXXXX

In order ti cite it, you have to find it first

- Provide metadata in many ways, as the services require them
- Google + Scholar: rules + dc metadata in the html header
- Google Dataset Search: metadata in JSON+LD
- CLARIN VLO: CMDI metadata over OAI:PMH
- OpenAIRE: DataCite metadata over OAI:PMH
- Data Citation Index: Dublin Core metadata
- Etc.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

In order ti cite it, you have to find it first



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

RobeCzech Base

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

Straka, Milan; Náplava, Jakub; Straková, Jana and Samuel, David, 2021, *RobeCzech Base*, 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/11234/1-3691.

Share:

LINDAT / CLARIAH-CZ

Authors	Straka, Milan ; Náplava, Jakub ; Straková, Jana and Samuel, David
Item identifier	http://hdl.handle.net/11234/1-3691
Date issued	2021-05-25

authors =
DC.creator + DC.creator …

LINDAT CLARIAH-CZ CLARIN

Login

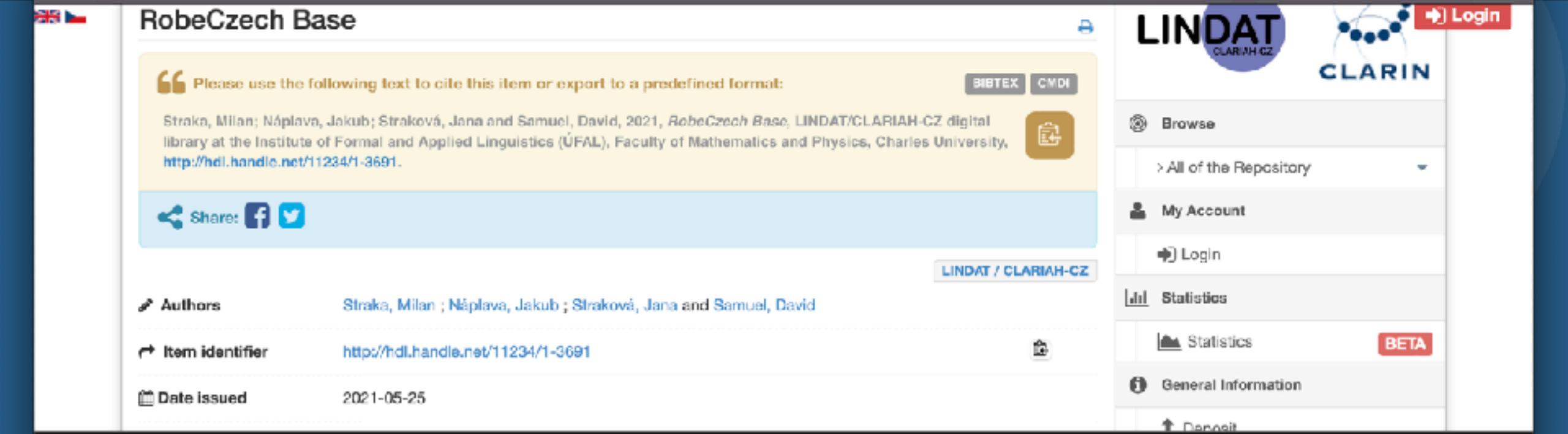
Browse
> All of the Repository
My Account
Login
Statistics
Statistics BETA
General Information
Deposit

Elements Console Sources Network Timelines Storage Graphics Layers Audit

1-3691 Response

Google metadata – SEO

```
<title xmlns:i18n="http://apache.org/cocoon/i18n/2.1">RobeCzech Base</title>
<link rel="schema.DCTERMS" href="https://purl.org/dc/terms/"/>
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/"/>
<meta name="DC.creator" content="Straka, Milan"/>
<meta name="DC.creator" content="Náplava, Jakub"/>
<meta name="DC.creator" content="Straková, Jana"/>
<meta name="DC.creator" content="Samuel, David"/>
<meta name="DCTERMS.dateAccepted" content="2021-05-25T08:15:54Z" scheme="DCTERMS.W3CDTF"/>
<meta name="DCTERMS.available" content="2021-05-25T08:15:54Z" scheme="DCTERMS.W3CDTF"/>
<meta name="DCTERMS.issued" content="2021-05-25" scheme="DCTERMS.W3CDTF"/>
<meta name="DC.identifier" content="http://hdl.handle.net/11234/1-3691" xml:lang="*" scheme="DCTERMS.URI"/>
<meta name="DC.description" content="RobeCzech is a monolingual RoBERTa language representation model trained on Czech data. RoBERTa is a robustly optimized Transformer-based pretraining approach. We show that RobeCzech considerably outperforms equally-sized multilingual and Czech-trained contextualized language representation models, surpasses current state of the art in all five evaluated NLP tasks and reaches state-of-theart results in four of them. The RobeCzech model is released publicly at https://hdl.handle.net/11234/1-3691 and https://huggingface.co/ufal/robeczech-base, both for PyTorch and TensorFlow." xml:lang="en_US"/>
<meta name="DC.language" content="ces" xml:lang="en_US" scheme="DCTERMS.RFC1766"/>
<meta name="DC.publisher" content="Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL)" xml:lang="en_US"/>
<meta name="DC.rights" content="Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)" xml:lang="*"/>
<meta name="DC.rights" content="http://creativecommons.org/licenses/by-nc-sa/4.0/" xml:lang="*" scheme="DCTERMS.URI"/>
<meta name="DC.subject" content="Czech" xml:lang="en_US"/>
<meta name="DC.subject" content="BERT" xml:lang="en_US"/>
<meta name="DC.subject" content="RoBERTa" xml:lang="en_US"/>
<meta name="DC.title" content="RobeCzech Base" xml:lang="en_US"/>
<meta name="DC.type" content="languageDescription" xml:lang="en_US"/>
<meta name="DC.rights" content="PUB" xml:lang="*"/>
```


RobeCzech Base
Please use the following text to cite this item or export to a predefined format:
BIBTEX
CMDI
Straka, Milan; Náplava, Jakub; Straková, Jana and Samuel, David, 2021, RobeCzech Base, 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/11234/1-3691.
Share:
LINDAT / CLARIAH-CZ
Authors
Straka, Milan ; Náplava, Jakub ; Straková, Jana and Samuel, David
Item identifier
http://hdl.handle.net/11234/1-3691
Date issued
2021-05-25
LINDAT
CLARIAH-CZ
CLARIN
Login
Browse
> All of the Repository
My Account
Login
Statistics
Statistics
BETA
General Information
Deposit
Elements
Console
Sources
Network
Timelines
Storage
Graphics
Layers
Audit
1-3691
Response
<meta content="Straka, Milan" name="citation_author"/>
<meta content="Náplava, Jakub" name="citation_author"/>
<meta content="Straková, Jana" name="citation_author"/>
<meta content="Samuel, David" name="citation_author"/>
<meta content="https://lindat.mff.cuni.cz/repository/xmlui/bitstream/11234/1-3691/2/robeczech-base-tf.zip" name="citation_pdf_url"/>
<meta content="2021-05-25" name="citation_date"/>
<meta content="https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3691" name="citation_abstract_html_url"/>
<script type="application/ld+json">
{"@context":"https://schema.org","@type":"Dataset","identifier":"http://hdl.handle.net/11234/1-3691","license":"http://creativecommons.org/licenses/by-nc-sa/4.0/","creator":
[{"@type":"Person","name":"Straka, Milan","familyName":"Straka","givenName":"Milan"},{"@type":"Person","name":"Náplava, Jakub","familyName":"Náplava","givenName":"Jakub"},
{"@type":"Person","name":"Straková, Jana","familyName":"Straková","givenName":"Jana"},{"@type":"Person","name":"Samuel, David","familyName":"Samuel","givenName":"David"}],"keywords":
["Czech","BERT","RoBERTa"],"name":"RobeCzech Base","description":"RobeCzech is a monolingual RoBERTa language representation model trained on Czech data. RoBERTa is a robustly optimized
Transformer-based pretraining approach. We show that RobeCzech considerably outperforms equally-sized multilingual and Czech-trained contextualized language representation models, surpasses
current state of the art in all five evaluated NLP tasks and reaches state-of-theart results in four of them. The RobeCzech model is released publicly at https://hdl.handle.net/11234/1-3691 and
https://huggingface.co/ufal/robeczech-base, both for PyTorch and TensorFlow.","url":"http://hdl.handle.net/11234/1-3691"}
</script>
Google Dataset Search metadata
<!--[if lt IE 7]> <body class="ie6"> <![endif]-->
<!--[if IE 7]> <body id="lindat-repository" class="ie7"> <![endif]-->
<!--[if IE 8]> <body id="lindat-repository" class="ie8"> <![endif]-->
<!--[if IE 9]> <body id="lindat-repository" class="ie9"> <![endif]-->
<!--[if (gt IE 9)|!(IE)]><!-->
<body id="lindat-repository">
<!--<![endif]-->
<nav class="lindat-header lindat-common" role="navigation" data-version="2.4.0" data-build="4804a6cc9b428607c4040cbebbed83884739b89c">

Cite data properly and easily

- Force-11 Data Citation Principles
  - Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
- Formated citation in simple text: copy&paste for Word
- BibTeX record for those who typeset papers in LaTeX

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Etalon 1.0

for word processors

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

Skoumalová, Hana, 2021, *Etalon 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/11234/1-3698.

for LaTex users

bibtex

```
@misc{11234/1-3698,
 title = {Etalon 1.0},
 author = {Skoumalov{\'a}, Hana},
 url = {http://hdl.handle.net/11234/1-3698},
 note = {{LINDAT}/{CLARIAH}-{CZ} digital library at the Institute of Formal and Applied Linguist
ics ({{\'U}FAL}), Faculty of Mathematics and Physics, Charles University},
 copyright = {Creative Commons - Attribution-{NonCommercial}-{ShareAlike} 4.0 International ({CC
} {BY}-{NC}-{SA} 4.0)},
 year = {2021} }
```

Share:

LINDAT / CLARIAH-CZ

Authors	Skoumalová, Hana
Item identifier	http://hdl.handle.net/11234/1-3698

How well does it work?

- We look at Google Scholar and Publish or Perish services
- We count citations, remove pre-prints, …
- leave only “real publications”
- only by authors not employed on the project

year	citations
2016	32
2017	72
2018	99
2019	133
2020	120

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Room for improvement

- DOI services for Handles:
  - support for Altmetrics, CrossRef and other services
  - easier finding of citations
  - easier formating of citations
- Add support for Citeproc JSON format
  - used by Zotero, Mendeley and others
  - use it with JS to offer more citation style

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Thank you!

Questions, please!

LINDAT CLARIAH-CZ

SSHOC social sciences & humanities open cloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782