# Sdílení výzkumných dat:
# As open as possible, as closed as necessary

Open science: od dat k publikaci

CENTRUM PRO PODPORU
OPEN SCIENCE
Univerzita Karlova

Dagmar Hanzlíková & Milan Janíček
Projekt RKV II
3. 2. 2021, ve 14:15
Centrum pro podporu open science
openscience@cuni.cz



Tato prezentace byla podpořena projektem OP VVV Rozvoj kapacit pro výzkum a vývoj UK II, CZ.02.2.69/0.0/0.0/18_054/0015222

UNIVERZITA KARLOVA

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

HR EXCELLENCE IN RESEARCH

# FAIR data

# FAIR Data

- Findable
- Accessible
- Interoperable
- Reusable

# FAIR Principles

**To be Findable:**

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

**TO BE ACCESSIBLE:**

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

**TO BE INTEROPERABLE:**

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

**TO BE RE-USABLE:**

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

https://www.go-fair.org/fair-principles/

# FAIR Data - Findable

- nalezitelná data
  - možnost najít data
  - jak lidé tak počítače by měli být schopni data najít

  - významnou roli v identifikování správného datasetu hrají perzistentní identifikátory
    - např. DOI, nicméně mohou existovat i specifické oborové identifikátory
  - metadata (data popisující dataset) jsou důležitá
    - Pro tvorbu metadatových popisů existují správné postupy. Zeptejte se svých knihovníků;-)

# Perzistentní identifikátory

- "kód" jednoznačně identifikující nějaký objekt, osobu, …
- Za PID (a jeho trvalost) je obvykle zodpovědná nějaká organizace
- Na základě PID je možné získat další informace (metadata) a reprezentaci objektu
- DOI (= digital object identifier) je perzistentní identifikátor (článku, datové sady..) https://doi.org/10.5281/zenodo.4263217
- ORCID je perzistentní identifikátor (osoby) https://orcid.org/0000-0002-8271-3674
- URL není perzistentní identifikátor

# Metadata

- data popisující jiná data
- například název datové sady, autor, klíčová slova, popis vlastností..
  - představte si popis knihy v katalogu knihovny
- podle metadat se vyhledává: špatná metadata => nenalezitelná data!
- při popisu je vhodné používat řízené slovníky
  - záleží na oborových zvyklostech – např. MeSH v medicíně
- popisujte data takovými termíny, které byste zadali do vyhledávače kdybyste je sami chtěli najít

# FAIR Data - Accessible

- přístupná data
  - meta(data) by měla být získatelná pomocí svého identifkátoru
    - za použití standardního komunikačního protokolu
    - autentizace a autorizace jsou možné
      - ne VŠECHNO musí být otevřeně dostupné
  - metadata by měla být dostupná i kdyby už samotná data nebyla

# FAIR Data - Interoperable

- interoperabilní data
  - mělo by být možné kombinovat data s dalšími daty a využívat další nástroje
  - formát by měl být otevřený a interoperabilní pro různé nástroje
  - data i metadata by měla být interoperabilní
    - např při tvorbě metadat k datasetům je vhodné využít (předpřipravené) slovníky (pokud je to možné)

# FAIR Data - Reusable

- znovuvyužitelná data
  - optimalizujte data pro znovuvyužití
    - dostatečně popište data – umožníte tím jak replikaci výsledků tak nové využití dat v jiném kontextu
  - použijte licenci – specifikujte jakým způsobem a za jakých podmínek mohou být data znovu využita

# Jak si usnadnit splnění FAIR principů?

- občas pomůže repozitář kam se data ukládají
- například Zenodo popisuje jako pomáhá naplňovat FAIR principy
  https://about.zenodo.org/principles/

To be Findable:

- **F1**: (meta)data are assigned a globally unique and persistent identifier
  - A DOI is issued to every published record on Zenodo.
- **F2**: data are described with rich metadata (defined by R1 below)
  - Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichements.
- **F3**: metadata clearly and explicitly include the identifier of the data it describes
  - The DOI is a top-level and a mandatory field in the metadata of each record.
- **F4**: (meta)data are registered or indexed in a searchable resource
  - Metadata of each record is indexed and searchable directly in Zenodo's search engine immediately after publishing.
  - Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

# Sdílení dat

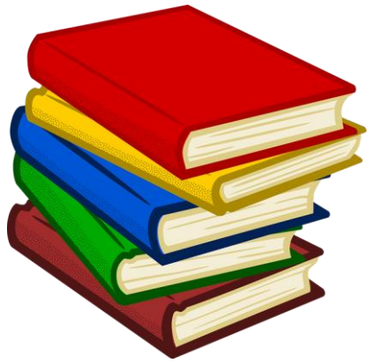
# Otevřená výzkumná data (ORD)

= *data, která jsou volně dostupná online komukoli a mohou být dále využívána, upravována a sdílena za jakýmkoli účelem.*

**As open as possible,**
**as closed as necessary.**

# Proč sdílet?

- Podmínky poskytovatelů financí
- Evropská směrnice 2019/1024 o otevřených datech a opakovaném použití informací veřejného sektoru
- Datové politiky časopisů



# DATA POLICY

At *Cognitive Linguistics*, we strongly believe that research data should be made widely available to the research community in order to demonstrate the robustness and validity of the research presented in our journal, to encourage replication of published results, and to provide the community with opportunities to learn. We believe that such transparency improves the quality of science and benefits not only the wider research community but the researchers as well by increasing their impact and enhancing their citation rates.

For all manuscripts submitted as of July 2020, *Cognitive Linguistics* requires, as a condition for publication, that all data (and related metadata) and any code supporting the results presented in the paper should be made publicly available, at the latest at the time of acceptance. Exemption may be granted by the Editor-in-Chief, for example, in case of sensitive data. Research data should be FAIR (Findable, Accessible, Interoperable, Reusable), should be deposited in an appropriate open repository and needs to be assigned a persistent identifier and an appropriate license specifying the conditions for reuse.

## Policy summary for authors

| Required | Optional |
| --- | --- |
| • Sharing of research data via repositories | • Data citation |
| • Data availability statements | • Prepare and share Data Management Plans |

# Proč sdílet?

## Benefity pro vědce

- Zajištění robustního výzkumu
- Vylepšení reputace a větší informační dopad
- Vyšší citovanost (data & publikace)
- Kombinací dat k novým poznatkům

## Benefity pro společnost

- Efektivní využití zdrojů
- Urychlení výzkumného procesu
- Podpora občanské vědy
- Omezení podvodů ve vědě

# Jaká data sdílet?

*dobrou praxí je zpřístupnit všechna data, která jsou potřeba k replikaci vaší výzkumné práce*

## Na co si dát pozor

- Etické a právní otázky
- Kdo rozhoduje o zveřejnění dat
- S kým data sdílet

# Jak sdílet?

**Způsoby sdílení dat**

- Supplementary materials
- Datový repozitář
- Datový časopis (data journal)

# Supplementary Materials

- Doplňující materiály k publikovanému článku
- Data publikuje vydavatel
- Může být omezená velikost či typ dat
- Data nelze samostatně citovat
- Práva mohou přejít na vydavatele!

# Datový repozitář

- Oborový repozitář
  - Data zaměřená na konkrétní obor
  - Registr datových repozitářů re3data.org
- Institucionální repozitář
  - Spravuje instituce
  - Např. ASEP
- Obecný repozitář
  - Bez ohledu na oborové zaměření
  - Např. Zenodo, Figshare, Dryad

Filter

**Subjects**

Humanities and Social Sciences (4)
- Humanities (4)
  - Linguistics (4)

Engineering Sciences (3)
- Computer Science, Electrical and System Engineering (3)
  - Systems Engineering (1)
    - Human Factors, Ergonomics, Human-Machine Systems (1)
  - Computer Science (3)
    - Artificial Intelligence, Image and Language Processing (2)

**Content Types**

**Countries**

**API**

**Certificates**

**Data access**

**Data access restrictions**

**Database access**

**Database access restrictions**

**Data licenses**

**Data upload**

**Data upload restrictions**

**Enhanced publication**

**Institution responsibility type**

**Institution type**

**Keywords**

**Metadata standards**

czech linguistics

Search

Toogle short help

← Previous | 1 | Next →

Sort by

Found 4 result(s)

## Czech National Corpus

CNK

| | |
|---|---|
| Subject(s) | Humanities and Social Sciences · Humanities · Linguistics |
| Content type(s) | Databases · Audiovisual data · Standard office documents |
| Country | European Union · Czech Republic |

The aim of the project is systematic mapping of Czech and other languages in comparison with Czech. CNC corpora are accessible to everybody interested in studying the language after free registration.

## LINDAT/CLARIN repository

| | |
|---|---|
| Subject(s) | Humanities · Linguistics · Artificial Intelligence, Image and Language Processing · Humanities and Social Sciences · Computer Science · Computer Science, Electrical and System Engineering · Engineering Sciences |
| Content type(s) | Standard office documents · Audiovisual data · Plain text · Structured text · Archived data · Source code |
| Country | Czech Republic · European Union |

LINDAT/CLARIN is designed as a Czech “node” of Clarin ERIC (Common Language Resources and Technology Infrastructure). It also supports the goals of the META-NET language technology network. Both networks aim at collection, annotation, development and free sharing of language data and basic technologies between institutions and individuals both in science and in all types of research. The Clarin ERIC infrastructural project is more focused on humanities, while META-NET aims at

zenodo

Search

Upload Communities

Log in Sign up

Featured

Recent up

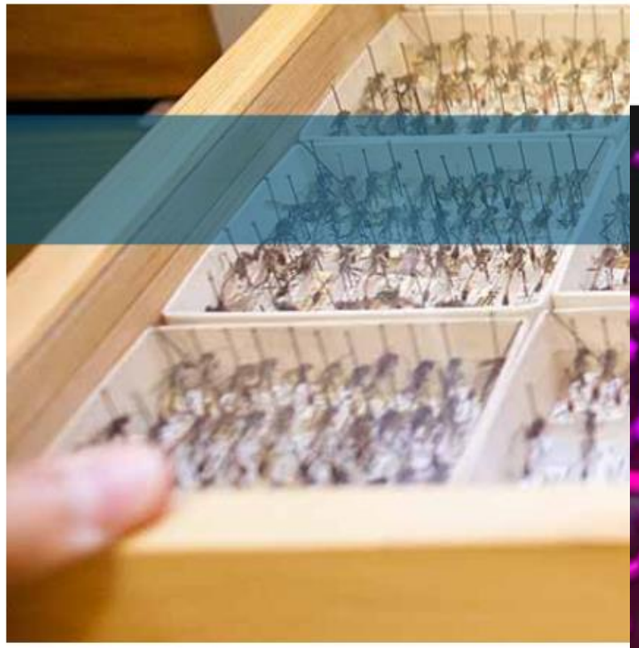
January 24, 2021 (v46)

A large-scale C
international co

Banda, Juan M.;
Elena; Chowell, G

Version 46 of the da
acquired from the T

DRYAD



Dryad is a community-ov
Learn more about our organiza

figshare

Browse Search on figshare... Log in Sign up

Search

Explore Data | About | Help | Login

store, share, discover **research**

get more citations for all of the outputs of your academic research
over 30,000 citations of fig**share** content to date

ALSO FOR INSTITUTIONS & PUBLISHERS

*"fig**share** wants to open scientific data to the world"* WIRED

*The background figure: Comparative model of novel coronavirus 2019-nCoV… by Christian Gruber in Virology*

# Data journal

*= recenzovaný časopis, který publikuje příspěvky popisující datové sady uložené v repozitáři*

- Vědecky cenné datové soubory
- Data uložená v repozitáři
- Možnost snadného citování a vykazování
- Otázka kvality recenzního řízení
- Čistě datové časopisy nebo kombinované
- Např. *Scientific Data*, *Earth System Science Data*, *Journal of Open Archaeology Data*

# ORD: Jak sdílet?

- Opatřit data vhodnou (veřejnou) licencí, např. Creative Commons nebo Open Data Commons
- Přidělit datům trvalý identifikátor
- Doporučený formát citace
- Zároveň s daty sdílet dokumentaci

# Multi-Dimensional Analysis of Czech

Version 1.0

Cvrček, Václav, 2018, "Multi-Dimensional Analysis of Czech", https://doi.org/10.18710/QAJKZW, DataverseNO, V1, UNF:6:5rqhrfGF8iJspOAQER3OCA== [fileUNF]

Cite Dataset

Learn about Data Citation Standards.

**Dataset Metrics**

39 Downloads

**Description**

Original data for a general-purpose multi-dimensional analysis model of register variation in Czech.

This post contains a CSV data set of 137 linguistic features measured on 3428 Czech text chunks, and an R script which performs a factor analysis on this data set. The results of this factor analysis were used as a basis for an 8-dimensional model of register variation in Czech (see Related Publications), following the methodology introduced by Douglas Biber (see e.g. his 1988 seminal work Variation Across Speech and Writing for details on the methodology, or his 2014 article "Using multi-dimensional analysis to explore cross-linguistic universals of register variation" for a review of MDA results across a variety of languages).

The data is derived from the Koditex corpus , which aims to be as diversified as possible, covering various forms of spoken and written (both print and on-line) Czech. In compiling this corpus, the purpose was to provide a solid empirical basis for a comprehensive general-purpose model of register variation in Czech.

Apart from this data set and related publications, additional resources pertaining to the project are available via the czcorpus/mda GitHub repository.

(2018-10-12)

**Subject** Arts and Humanities

**Keyword** multi-dimensional analysis, register variation, factor analysis, corpus, Czech

**Related Publication** Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. Corpus Linguistics and Linguistic Theory. doi: 10.1515/cllt-2018-0020

**Related Publication** Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. Corpus Linguistics and Linguistic Theory. doi: 10.1515/cllt-2018-0020

Files | Metadata | Terms | Versions

Export Metadata

## Citation Metadata

| | |
|---|---|
| **Dataset Persistent ID** | doi:10.18710/QAJKZW |
| **Publication Date** | 2018-10-30 |
| **Title** | Multi-Dimensional Analysis of Czech |
| **Author** | Cvrček, Václav (Czech National Corpus) - ORCID: 0000-0003-3977-2393 |
| **Contact** | Use email button above to contact.<br>Lukeš, David (Czech National Corpus) |
| **Description** | Original data for a general-purpose multi-dimensional analysis model of register variation in Czech.<br><br>This post contains a CSV data set of 137 linguistic features measured on 3428 Czech text chunks, and an R script which performs a factor analysis on this data set. The results of this factor analysis were used as a basis for an 8-dimensional model of register variation in Czech (see Related Publications), following the methodology introduced by Douglas Biber (see e.g. his 1988 seminal work Variation Across Speech and Writing for details on the methodology, or his 2014 article "Using multi-dimensional analysis to explore cross-linguistic universals of register variation" for a review of MDA results across a variety of languages).<br><br>The data is derived from the Koditex corpus , which aims to be as diversified as possible, covering various forms of spoken and written (both print and on-line) Czech. In compiling this corpus, the purpose was to provide a solid empirical basis for a comprehensive general-purpose model of register variation in Czech. |

# Chtěl bych data sdílet ale...

- …obávám se, že má data někdo mylně interpretuje
  - *Sdílení dokumentace*
- ...obávám se, že má data někdo odcizí
  - *Licence a doporučený formát citace*
- ...obávám se, že má data využije někdo dřív, než je stačím vytěžit sám
  - *Sdílení metadat, časové embargo, předregistrace*

# Chtěl bych data sdílet ale...

- ...mého výzkumu se účastnili lidé a nejsem si jistý, jestli mohu takto nasbíraná data zveřejnit
  - *Informovaný souhlas, anonymizace*
- ...má data obsahují osobní údaje
  - *Odstranění údajů, anonymizace, informovaný souhlas, pověřenec GDPR*
- ...součástí mého datasetu jsou data třetích stran a nejsem si jistý, jestli je mohu sdílet
  - *Licenční podmínky, souhlas autora*

# Prostor pro diskuzi

# Co nás dnes ještě čeká?

| | | |
|---|---|---|
| EU quō vādis? Evropské vědecké infrastruktury a open science | 15:15 – 16:15 | Ing. Milan Janíček |
| Přestávka 16:15 – 16:30 | | |
| Služby e-infrastruktury CESNET | 16:30 – 17:00 | RNDr. David Antoš, Ph.D. |
| Podpora na UK: Co nám kdy univerzita dala? | 17:00 – 18:00 | Centrum pro podporu Open Science |

# Užitečné odkazy

- Go FAIR Initiative: FAIR Principles
- Registr datových repozitářů re3data.org
- Obecné repozitáře
  - Zenodo
  - Figshare
  - Dryad
- Center for Open Science: Preregistration
- Creative Commons Česká republika

# Použité obrázky


- [15] OpenClipart-Vectors. Bag money wealth revenue finance. *Pixabay* [online]. Dostupné z https://pixabay.com/images/id-147782/ Podléhá licenci Pixabay.
- [15] OpenClipart-Vectors. Book books library books reading. *Pixabay* [online]. Dostupné z https://pixabay.com/images/id-2022464/ Podléhá licenci Pixabay.
- [15] OpenClipart-Vectors. European Union Europe flag EU. *Pixabay* [online]. Dostupné z https://pixabay.com/images/id-155207/ Podléhá licenci Pixabay.
- [16] Screenshot https://www.nature.com/nature-research/editorial-policies/reporting-standards, 29. 1. 2021
- [16] Screenshot https://journals.plos.org/plosone/s/data-availability, 29. 1. 2021
- [16] Screenshot https://www.degruyter.com/supplemental/journals/cogl/cogl-overview.xml/DataPolicy.pdf, 29. 1. 2021
- [22] Screenshot https://www.re3data.org/search?query=czech%20linguistics 25. 1. 2021
- [23] Screenshot https://zenodo.org/ 25. 1. 2021
- [23] Screenshot https://datadryad.org/stash 25. 1. 2021
- [23] Screenshot https://figshare.com/ 25. 1. 2021
- [26-27] Screenshot https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QAJKZW, 29. 1. 2021

# Děkujeme za pozornost.

Centrum pro podporu open science
Ústřední knihovna UK
openscience@cuni.cz