

Ukládání a zpřístupnění digitálních dat

Ondřej Fúsik

Ústav lingvistiky, Filosofická fakulta University Karlovy
Oddělení rukopisů a starých tisků, Národní knihovna České republiky
ondrej.fusik@ff.cuni.cz / ondrej.fusik@nkp.cz



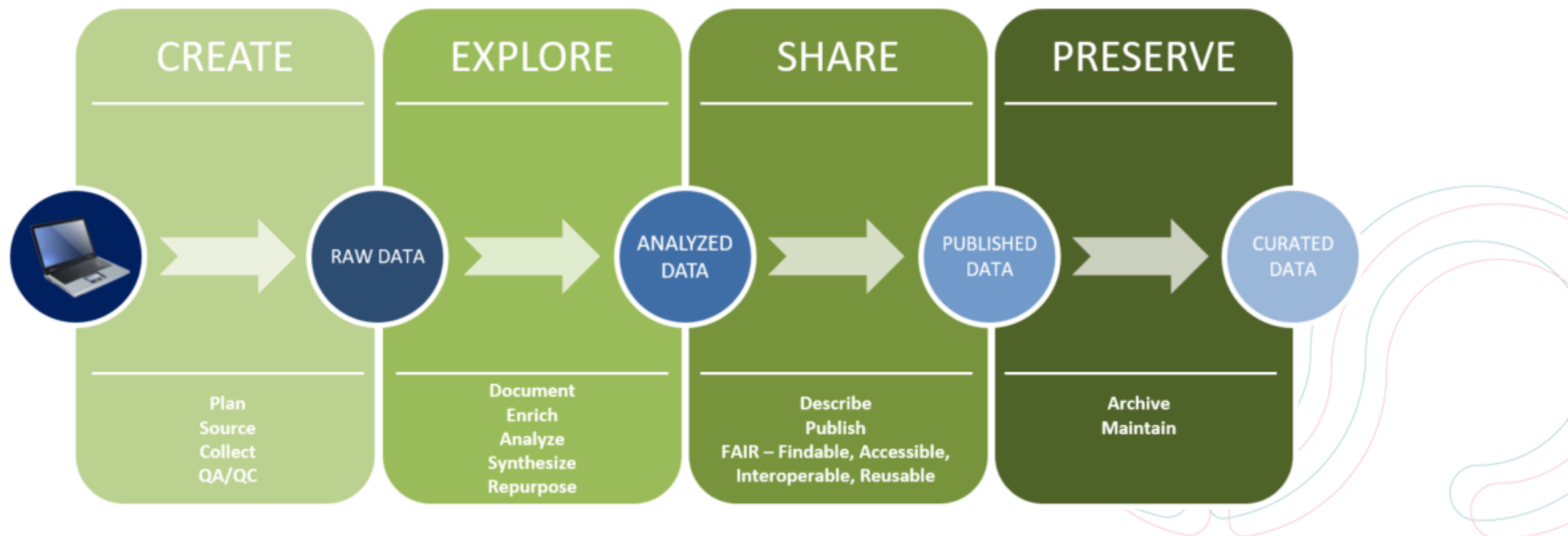
Spolufinancováno
Evropskou unií



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Data management plan

source: <https://rto.asu.edu/research-data-management/data-management-planning/>



Kde hledat pomoc? (1/3)

příklady

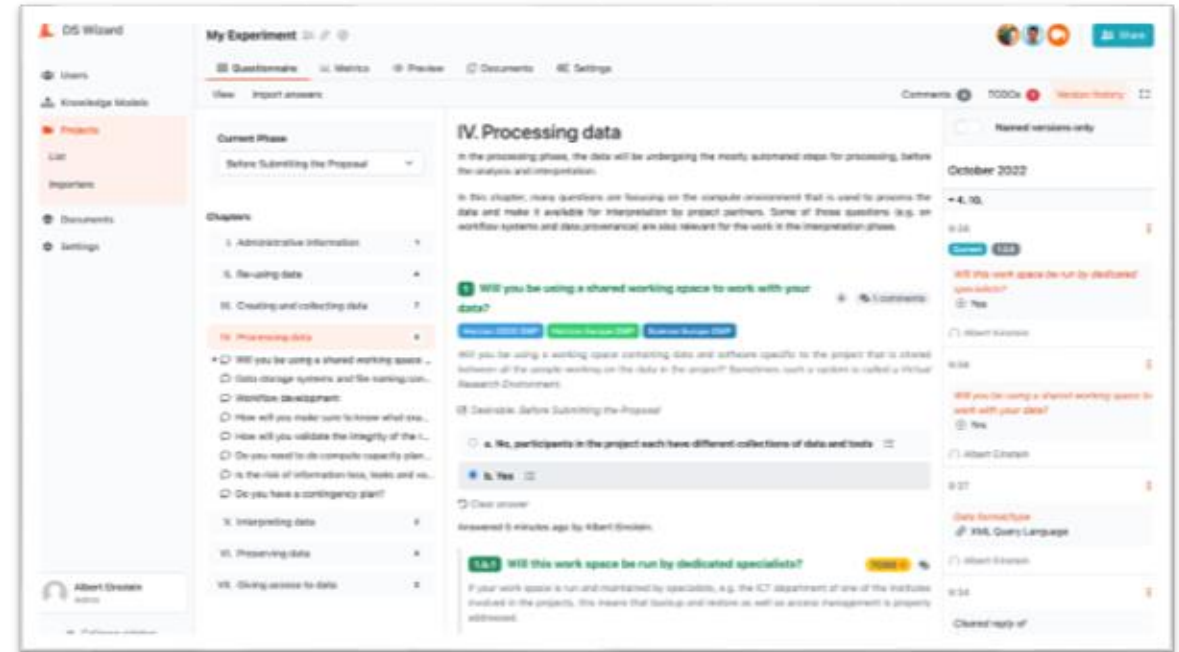


Kde hledat pomoc? (2/3)

příklady

Data steward Wizard

- (source: <https://ds-wizard.org/>)
- DS Wizard je nástroj pro tvorbu a správu Data Management Planu (DMP).
- Umožňuje strukturovaně popsat celý životní cyklus výzkumných dat, nastavovat workflow, standardy, odpovědnosti a způsoby archivace.
- Je určen pro vědecké týmy, grantové projekty a instituce vyžadující kvalitní správu dat dle **FAIR principů**.

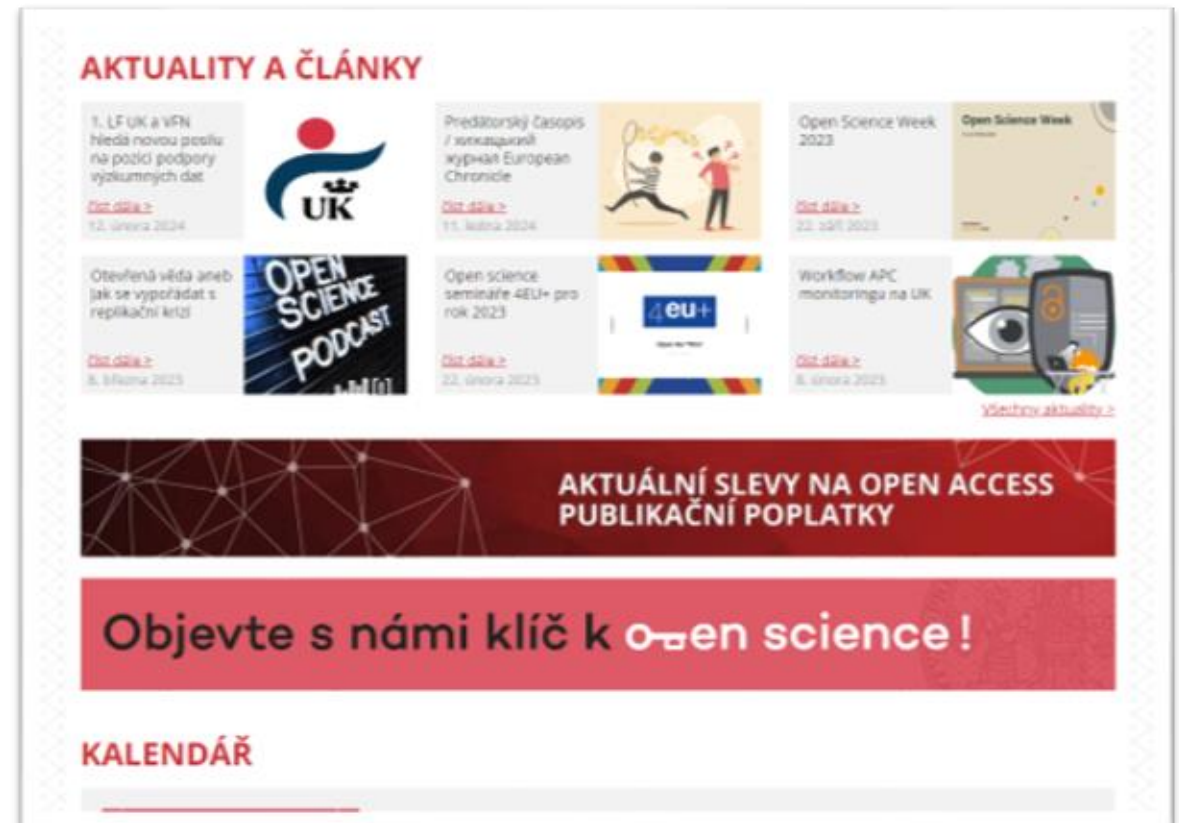


Kde hledat pomoc? (3/3)

příklady

Open Science Centrum

- (source: <https://openscience.cuni.cz/>)
- Open Science Centrum Univerzity Karlovy poskytuje podporu pro otevřenou vědu, správu výzkumných dat, open access publikování a implementaci **FAIR principů**.
- Nabízí školení, konzultace, metodiky a nástroje, které pomáhají výzkumníkům efektivně plánovat, sdílet a dlouhodobě uchovávat vědecké výstupy.

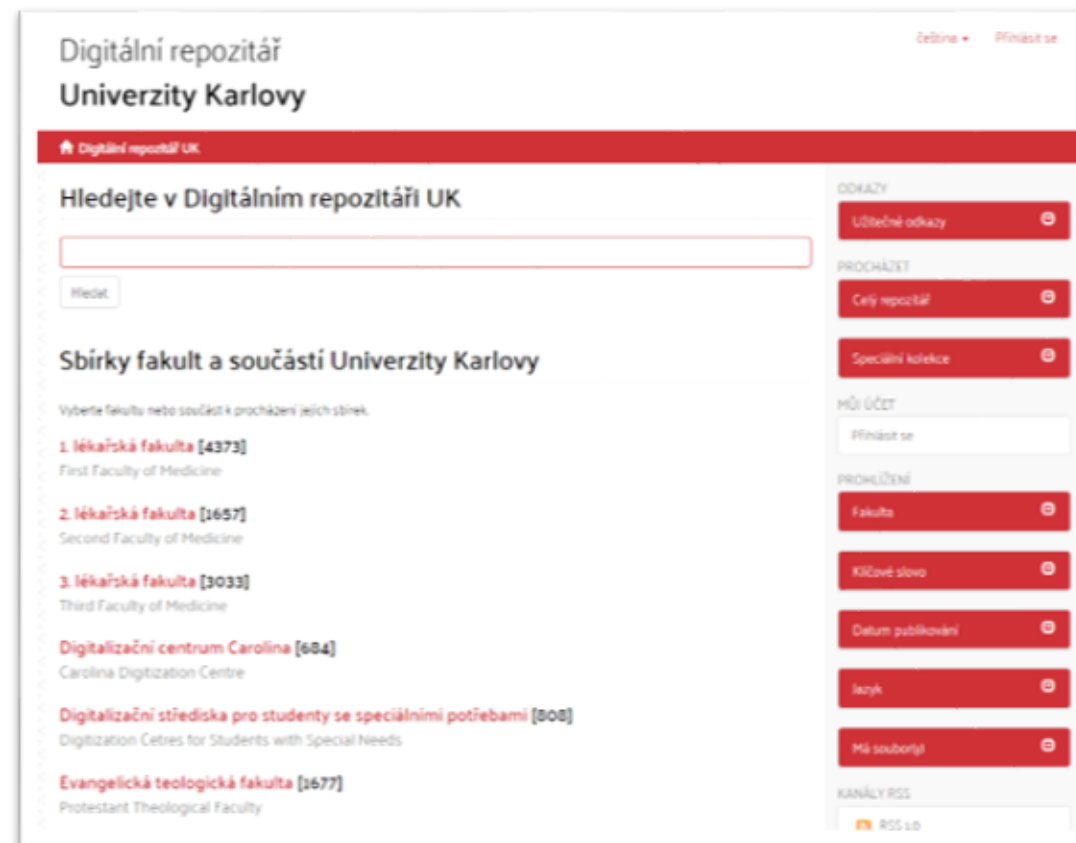


Repositáře (1/6)

příklady

Digitální repozitář Univerzity Karlovy

- **PRO**
 - Oficiální, institucionálně zaručený archiv – garantovaná dlouhodobá dostupnost.
 - Povinné uložení závěrečných prací – jednotné místo pro Bc., Mgr., PhD práce.
 - Stabilní infrastruktura – DSpace, standardní metadatové formáty.
 - Možnost embarga – autor může určit podmínky zveřejnění.
 - Automatické indexování v knihovních systémech UK a často i ve vyhledávacích (Google Scholar).
 - Podpora pro publikace akademiků UK – lze ukládat postprinty, preprinty apod.
- **PROTI**
 - Nevytváří vlastní DOI – pokud chceš DOI, musí ho poskytnout vydavatel nebo jiný repozitář (např. Zenodo).
 - Není určen pro velké vědecké datasety nebo software – omezené možnosti verzování a kapacity.
 - Méně uživatelsky přívětivé rozhraní ve srovnání se Zenodo / Figshare / Lindat.
 - Slabší podpora pro mezinárodní viditelnost dat – není to globální repozitář.
 - Primárně pro závěrečné práce, nikoli pro komplexní týmové projekty nebo korpusy.

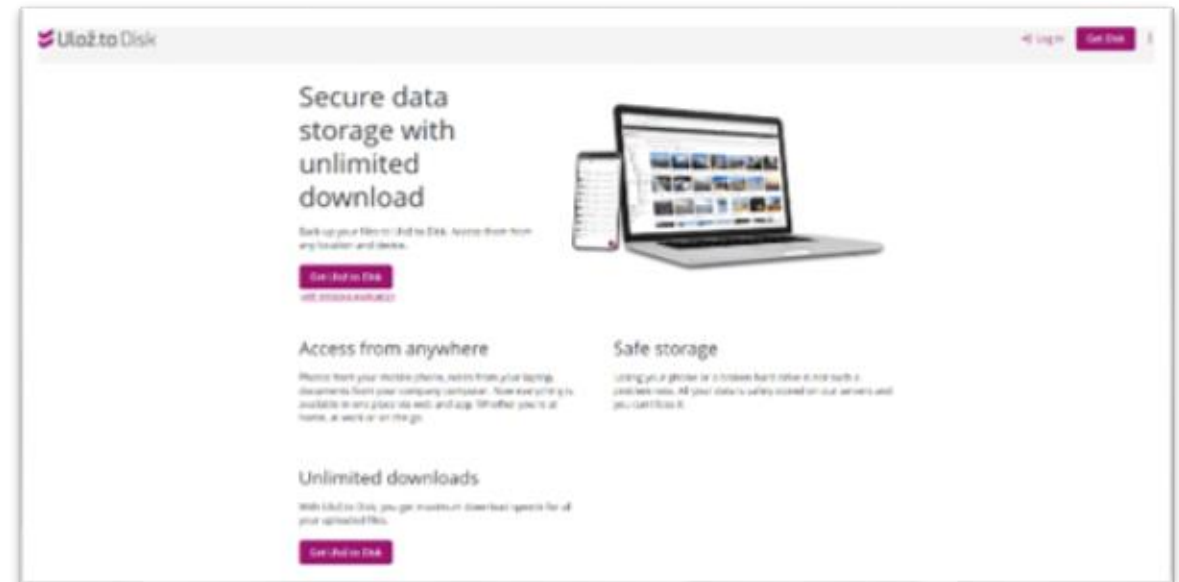


Repositáře (2/6)

příklady

Ulož.to

- **Pro**
 - jednoduché nahrávání i sdílení
 - velké limity pro nahrávání
 - rychlý přenos souborů
 - populární v ČR → snadné sdílení mezi laiky
- **Proti**
 - nevědecké, žádná archivní garance
 - právní rizika (často pirátské sdílení)
 - neexistence PID
 - soubory mohou být v budoucnu nedostupné



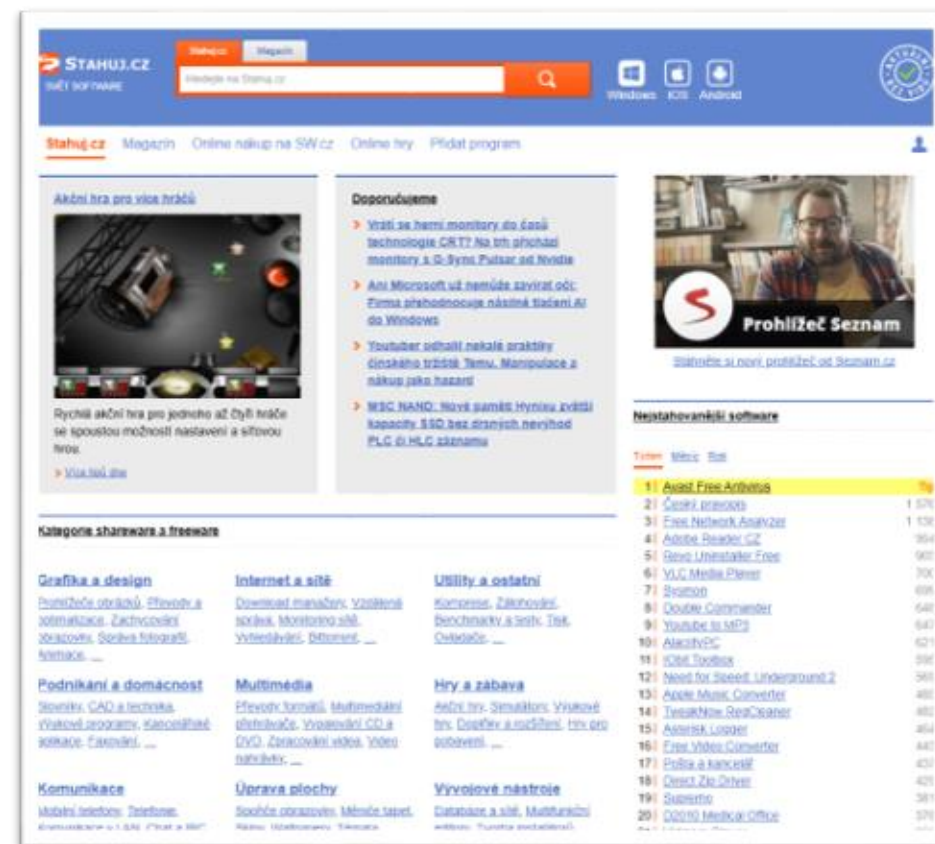
Repositáře (3/6)

příklady

Stahuj.cz

Zaměření: stahování programů a utilit (software katalog).

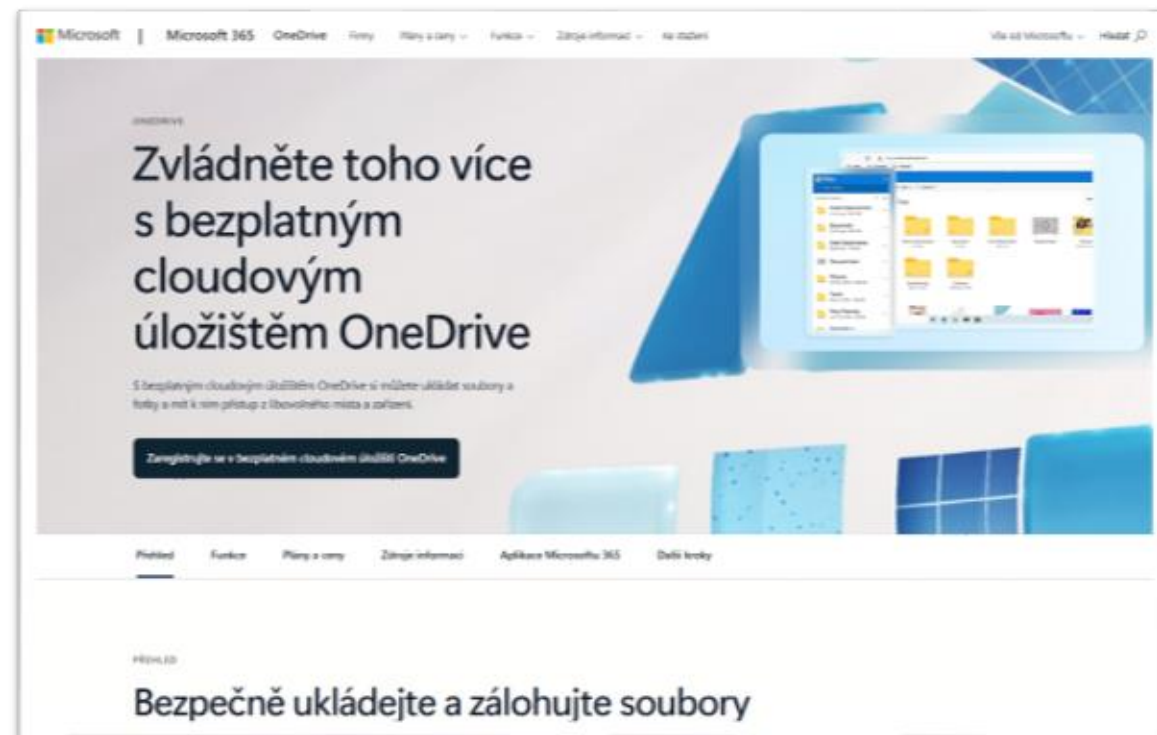
- **Pro**
 - velký katalog freewaru/sharewarudůvěra díky značce Seznam.cz
 - rychlé stahování
 - jednoduché pro běžné uživatele
- **Proti**
 - není to repositář pro vlastní obsah (nelze nahrát vlastní data)
 - extrémně nevhodné pro vědu nebo archivaci
 - katalog bývá zastaralý u méně udržovaných programů
 - nulová kontrola integrity dat z pohledu vědecké komunity



Repositáře (4/6)

příklady

- **OneDrive / Google Drive / Dropbox:**
- **Zaměření:** cloudové úložiště a sdílení mezi zařízeními.
- **Pro**
 - spolehlivé synchronizace
 - velmi robustní desktopové aplikace
 - dobré sdílení složek v týmech
 - kvalitní integrace API
- **Proti**
 - nevědecké, bez DOI
 - placené tarify dražší než konkurencemenší bezplatný prostor
 - vhodné spíše pro provoz, ne archivační nebo publikační účely



Repositáře (5/6)

příklady

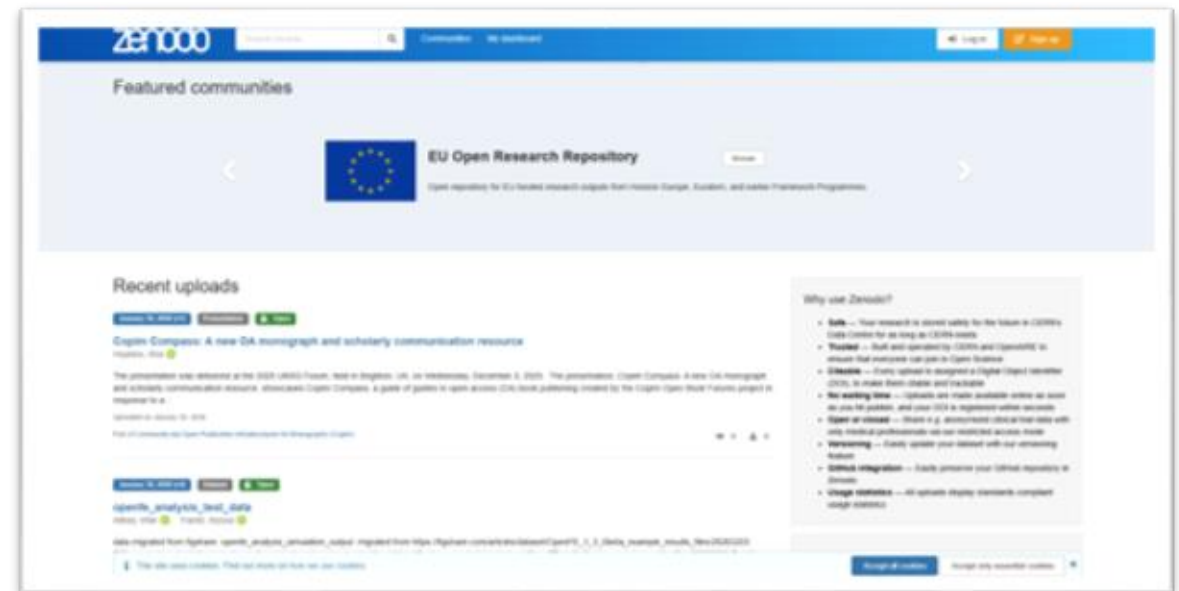
- **Zenodo**

- **Charakteristika**

- Mezinárodní otevřený repositář provozovaný CERNem a OpenAIRE
- Vytváří PID pro každý záznam
- Splňuje požadavky FAIR (Findable, Accessible, Interoperable, Reusable)
- Dlouhodobá archivace dat, publikací a software

- **Co lze ukládat**

- vědecké články
- datasets, korpusy, TEI-XML
- výzkumné prezentace, postery
- obrázky, multimédia
- software (vč. GitHub integrace)

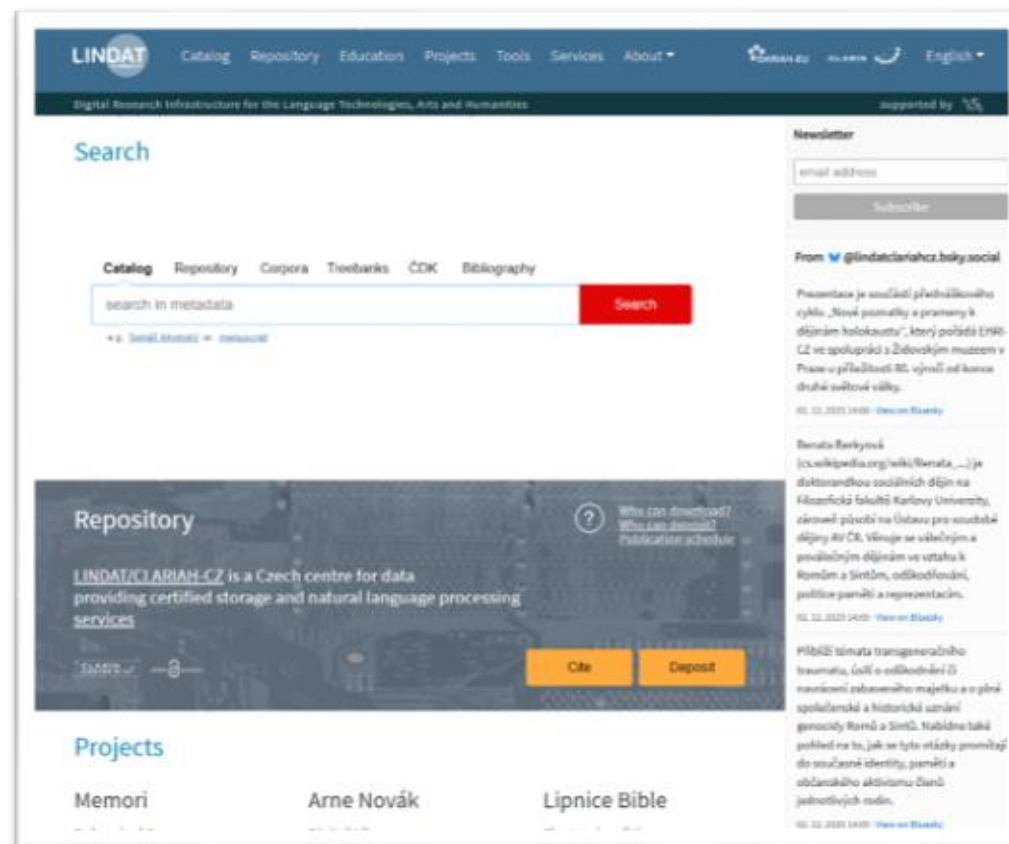


Repositáře (6/6)

příklady

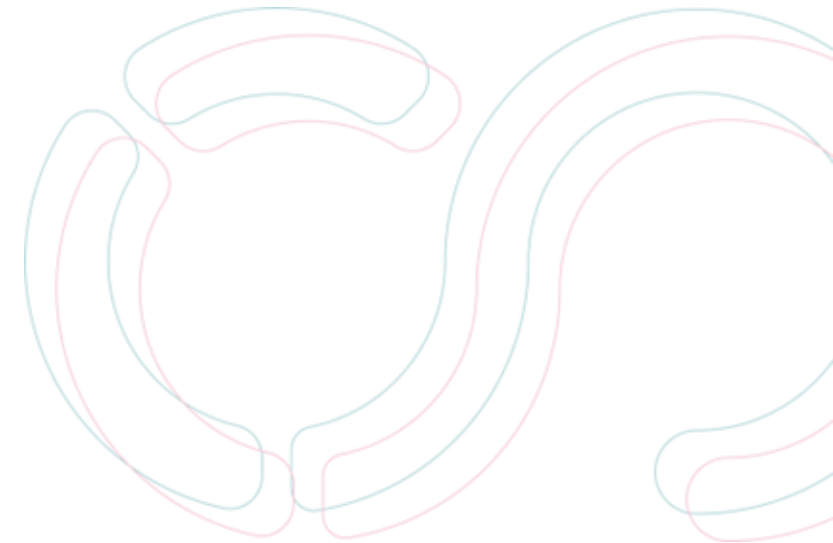
• Lindat/Clariah-CZ

- **PRO**
 - Oficiální národní infrastruktura — vysoká stabilita a institucionální garance (MFF UK + CLARIAH).
 - Automatické přidělování PID — ideální pro publikaci datasetů, korpusů a nástrojů.
 - FAIR principy — splňuje požadavky evropských grantů.
 - Dlouhodobá archivace — data jsou garantovaně uchovávána.
 - Vhodné pro jazyková data, korpusy, DH projekty — zaměření, které běžné repositáře nepokrývají.
 - Licenční podpora — možnost uložení i právně citlivějších datasetů (např. s omezeným přístupem).
 - Napojení na evropské infrastruktury CLARIN ERIC a DARIAH-ERIC — vyšší viditelnost.
- **PROTI**
 - Užší profilace — není univerzální repositář pro všechny disciplíny (primárně lingvistika, NLP, DH).
 - Složitější metadatová struktura — vyžaduje přesné popisy a validaci.
 - Technické rozhraní může být náročnější než Zenodo / Figshare (méně “user-friendly” pro začátečníky).
 - Není primárně určen pro publikace — lepší pro data než pro články či prezentace.
 - Vyžaduje povědomí o licencích (zejména u jazykových dat, osobních údajů, práv třetích stran).



Case study: Cíl

- **Ukážeme celý proces:**
 - přihlášení do systému
 - zahájení nového záznamu
 - výběr typu dat
 - vyplnění metadat
 - nahrání souborů
 - licence a právní náležitosti
 - schválení a vydání záznamu (vč. PID)



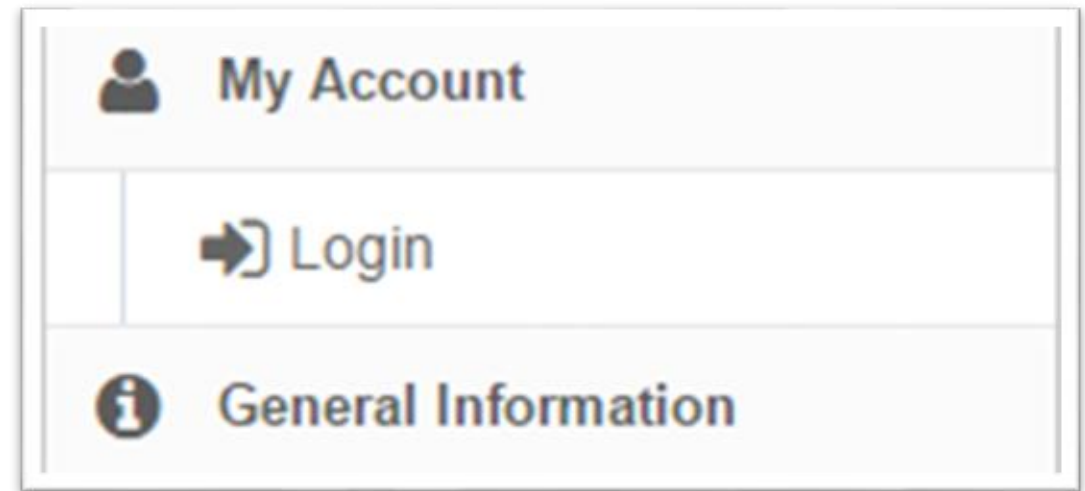
Case study: Proč nahrát data právě do LINDATu?

- **Hlavní důvody:**

- **Zdarma a bezpečné uložení** s dlouhodobou archivací
- **Licenční podpora** – pomohou ti s výběrem i s omezenými daty
- **Data jsou viditelná** v Google Scholar, DataCite, OLAC, arXiv apod.
- **Data jsou snadno citovatelná** (PID/handle, hotové citace v BibTeXu, RIS atd.)
- **Další služby** (např. fulltext či stromové vyhledávání u některých korpusů)

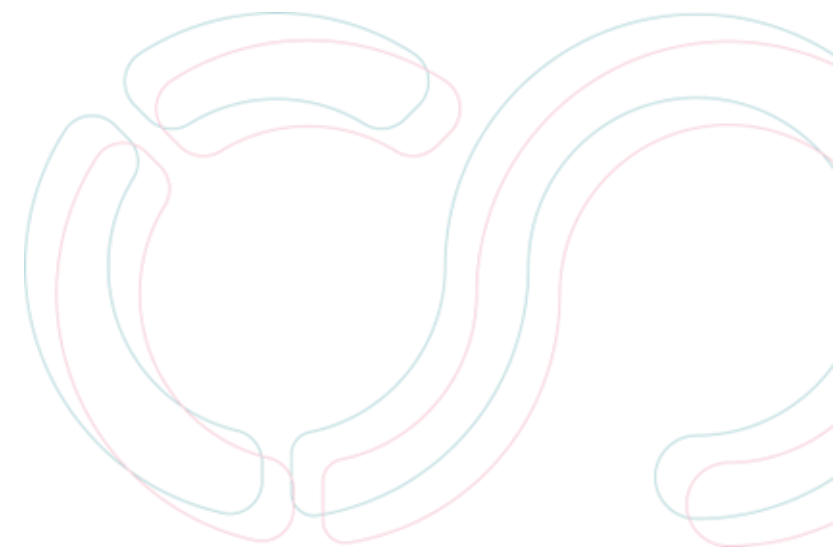
Jak probíhá ukládání dat: principy

- **Každé nahrané data:**
 - prochází **kurátorským procesem** → kontrola formátu, konzistence a kvality
 - musí mít **přiřazena metadata**
 - musí mít **jasně vyřešené licence a práva**
 - dostanou **persistentní identifikátor (PID)**
 - musí být ověřen uživatel (přihlášení skrze federaci eduID / CLARIN)
- **Smysl:**
 - garantovat dlouhodobou kvalitu a legální bezpečí.



Case study: Nahráváme do Lindatu

- **Krok 1: Přihlášení**
 - Přihlášení přes eduID, Charles University, nebo CLARIN účet.
- **Krok 2: Zahájení nového záznamu**
 - V menu „Uložení dat“ → „Start new submission“.
- **Krok 3: Výběr typu dat**
 - Korpus / lexical resource / language description / tool / service.
- **Krok 4: Vyplnění metadat**
 - Osoby, projekt, financování, popisy, klíčová slova, jazyk atd.
- **Krok 5: Nahrání souborů**
- **Krok 6: Licence**
 - Výběr licence, případně dodání vlastní licence.
- **Krok 7: Review a publikace**
 - Kurátor schválí → přidělí PID → záznam je zveřejněn.



Metadata – proč jsou kritická a kde vznikají problémy

- **Co jsou metadata**

- Strukturované informace popisující datovou položku (autoři, jazyk, licence, žánr, velikost, klíčová slova, popisy, projekty, financování).

- **Nejčastější problémy s metadaty**

- **Neúplná metadata** → položka je nevyhledatelná / nepoužitelná.
- **Nekonzistentní metadata** (např. různé formáty jmen, rozdílné názvy souborů, chybné datumy).
- **Špatně vyplněná práva a licence** → potenciální právní rizika.
- **Nevhodná klíčová slova** → data nejsou v mezinárodních indexech dohledatelná.
- **Chybějící jazykové kódy, formátové specifikace** → LINDAT nemůže správně klasifikovat datové sady.

- **Doporučení pro kvalitní metadata**

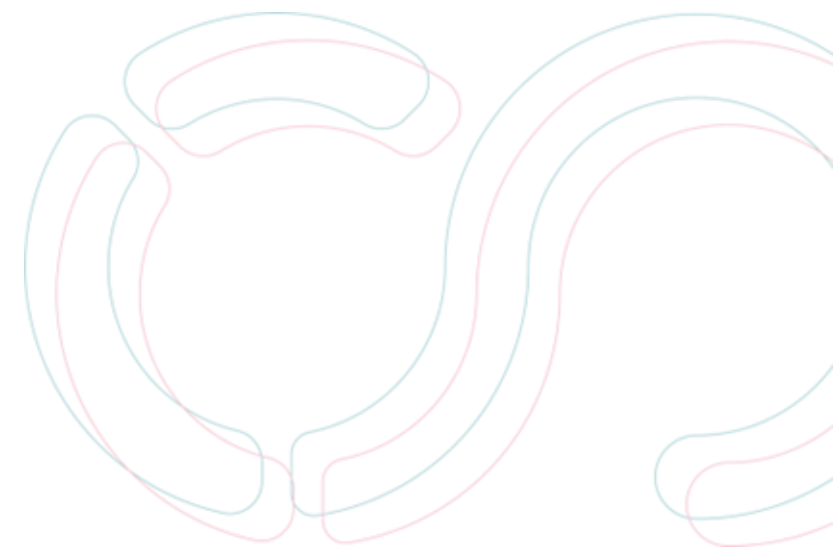
- Použít **standardizované formáty** (např. ORCID, ISO jazykové kódy, DataCite schema).
- Vždy vyplnit **úplné popisy**, ne minimální.
- Použít **konzistentní terminologii** a klíčová slova z odborných slovníků (např. OLAC, CLARIN VLO).
- Přesně uvést **projekty, granty a financování**, pokud existují (kvůli auditovatelnosti).
- Dokonale vyplnit pole o **původu dat**, způsobu vytvoření, metodologii a verzi.

Formáty dat – rizika, limity a doporučení (1/2)

- **Typické problémy s formáty**

- **Obskurní nebo proprietární formáty**
(např. staré formáty MS Office, proprietární binární soubory).
- **Data bez popisu formátu** – uživatel neví, jak je otevřít.
- **Neexistující nebo nestandardní kódování**
(častý problém u jazykových dat).
- **Chybějící deklarace schématu** (např. u XML, JSON).
- **Nekompletní data** – některé nástroje

vyžadují doprovodné soubory (např. číselníky, tagsety, lexikální zdroje).



Formáty dat – rizika, limity a doporučení (2/2)

• Co LINDAT doporučuje

- Vždy přiložit **původní formát**, i když je exotický (kvůli autenticity a archivaci).
- Současně přiložit také **archivní / standardní formát**, např.:
 - dokumenty → PDF/A
 - textová data → UTF-8 plaintext, XML, JSON
 - audio → WAV (bez komprese)
 - video → MP4/H.264
 - korpusy → TEI-XML nebo CoNLL-U
- Uložit **specifikaci formátu**, případně README o struktuře dat.
- Uložit **verzi nástroje**, který byl použit pro vytvoření dat.

• Doporučení pro dlouhodobou použitelnost

- Volit **otevřené formáty s existující podporou konverze**.
- Vyhnout se závislosti na jedné proprietární aplikaci.
- Interně používat **kódování UTF-8** jako standard.
- U binárních dat přidat **checksumy** (LINDAT s nimi pracuje při archivaci).

Licencování – rizika, povinnosti a workflow

Problémy s licencováním:

- **Autoři nerozumí licencím** → volí nesprávný typ (např. příliš restriktivní nebo naopak příliš otevřený).
- **Data obsahují osobní údaje** → LINDAT **nemůže** provést GDPR kontrolu.
- **Nejasná autorská práva** → autor neví, zda může data sdílet (typické u multimodálních nebo převzatých dat).
- **Střet licencí** u kombinovaných datasetů (např. různé části dat mají jiné licence).



Důležité upozornění pro uživatele:

- **Uživatel se nejprve musí licenčně zavázat vůči LINDATu**, že má právo data distribuovat.
- LINDAT **nenese odpovědnost za GDPR, osobní údaje ani autorské právo** – odpovědnost je vždy na tom, kdo data ukládá.
- Pokud existují omezení (např. neveřejné části datasetu), musí být specifikována už v metadatech.

Licencování – rizika, povinnosti a workflow

- **Doporučené licenční workflow:**

1. **Posoudit, zda má uživatel právo data sdílet.**

2. **Zkontrolovat přítomnost osobních údajů.**

> pokud ano → anonymizace, pseudonymizace nebo restricted access.

3. Vybrat **odpovídající otevřenou licenci** (např. CC-BY, CC-BY-SA, CC0).

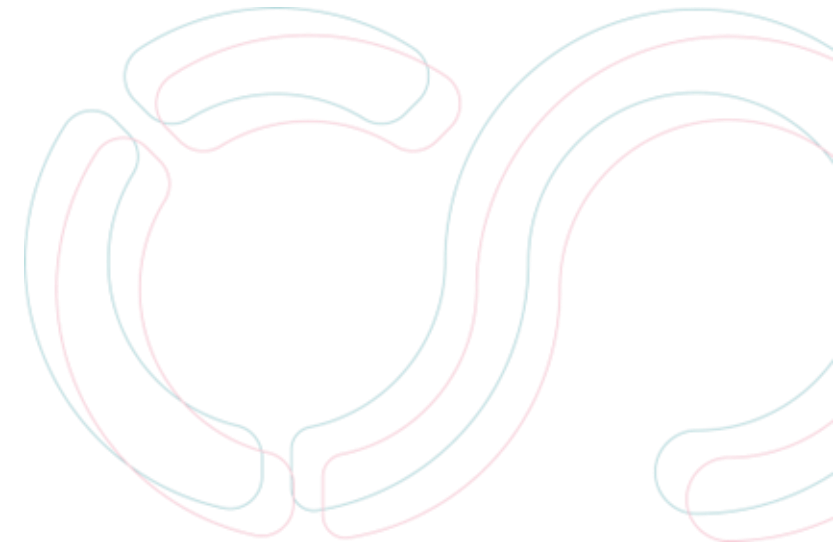
4. Pokud má dataset vlastní smluvní podmínky → přiložit je jako samostatný soubor.

5. Pokud žádná licence nevyhovuje → napsat na LINDAT Helpdesk → přidají vlastní licenci.

6. Po výběru licence kliknout na **“Accept distribution license agreement”**.

Licencování – rizika, povinnosti a workflow

- **Kdy použít otevřené licence (výhoda)**
 - Chcete maximální citovanost.
 - Chcete komunity (NLP, DH) umožnit práci s daty.
 - Data neobsahují citlivé informace.
 - Data byla vytvořena z veřejných zdrojů.
- **Kdy použít omezené licence**
 - Data obsahují osobní údaje nebo transkripty lidských mluvčích.
 - Data nejsou zcela vaše (část je převzatá).
 - Grant vyžaduje omezený přístup.
 - Chcete kontrolovat znovupoužití nebo přístup.



Děkuji za pozornost.

Budu rád za vaše otázky, připomínky a sdílení zkušeností s ukládáním dat v repozitářích.

Email:

ondrej.fusik@ff.cuni.cz;

ondrej.fusik@nkp.cz



Spolufinancováno
Evropskou unií



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY