National Czech Programme | eosc

Základy správy výzkumných dat

Jan Vališ

Ústav analytické chemie, Fakulta chemicko-inženýrská, VŠCHT Praha

& Akademické služby, NTK

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

Registrační číslo IPs EOSC-CZ
CZ.02.01.01/00/22_004/0007682

Program

- Představení
- Slovní zásoba
- Motivace
- Životní cyklus dat
- Nástroje pro tvorbu DMP
- DMP
- Dotazy

Výzkumná data – teoreticky

- Informace shromážděné, nebo generované během výzkumu
- Primární data
  - Přímo z daného výzkumu
  - Kvantitativní (hmotnost) vs. Kvalitativní (rozhovor)
  - Experimentální (pH) vs. Observační (migrace ptactva)
- Sekundární data
  - Z jiného výzkumu pro jiný účel

Výzkumná data – prakticky

- Tabulky
- Texty
- Obrázky
- Zvukové záznamy
- Diagnózy
- Statistiky

Metadata - teoreticky

- Deskriptivní
  - název
  - autor
  - datum
  - klíčová slova
  - jazyky
- Strukturální
  - použité dělení na složky/soubory
- Administrativní
  - formát souborů
  - vlastník
  - licence
  - zabezpečení

Metadata - teoreticky

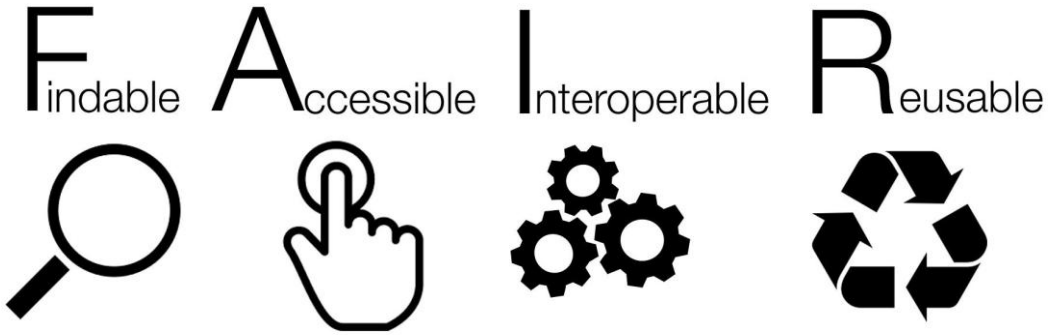
- „Data o datech“
- Kde, kdo, co a jak
- Strojově čitelná → vyhledávání
- Metadatová schémata
  - univerzální (Dublin Core, DataCite)
  - oborově specifická (Open Geospatial Consortium Metadata Standard)

- DataCite
  - Povinná:
    ID, *Creator, Title, Publisher, Year, Type*
  - Doporučená:
    Subject, Contributor, Date, Relataed ID, Description, Geo Location
  - Volitelná:
    Language, Alternative ID, Size, Format, Version, Rights, Funfing, Related Item

podrobněji: DataCite Metadata Schema Documentation

National Czech Programme | eosc

FAIR principy

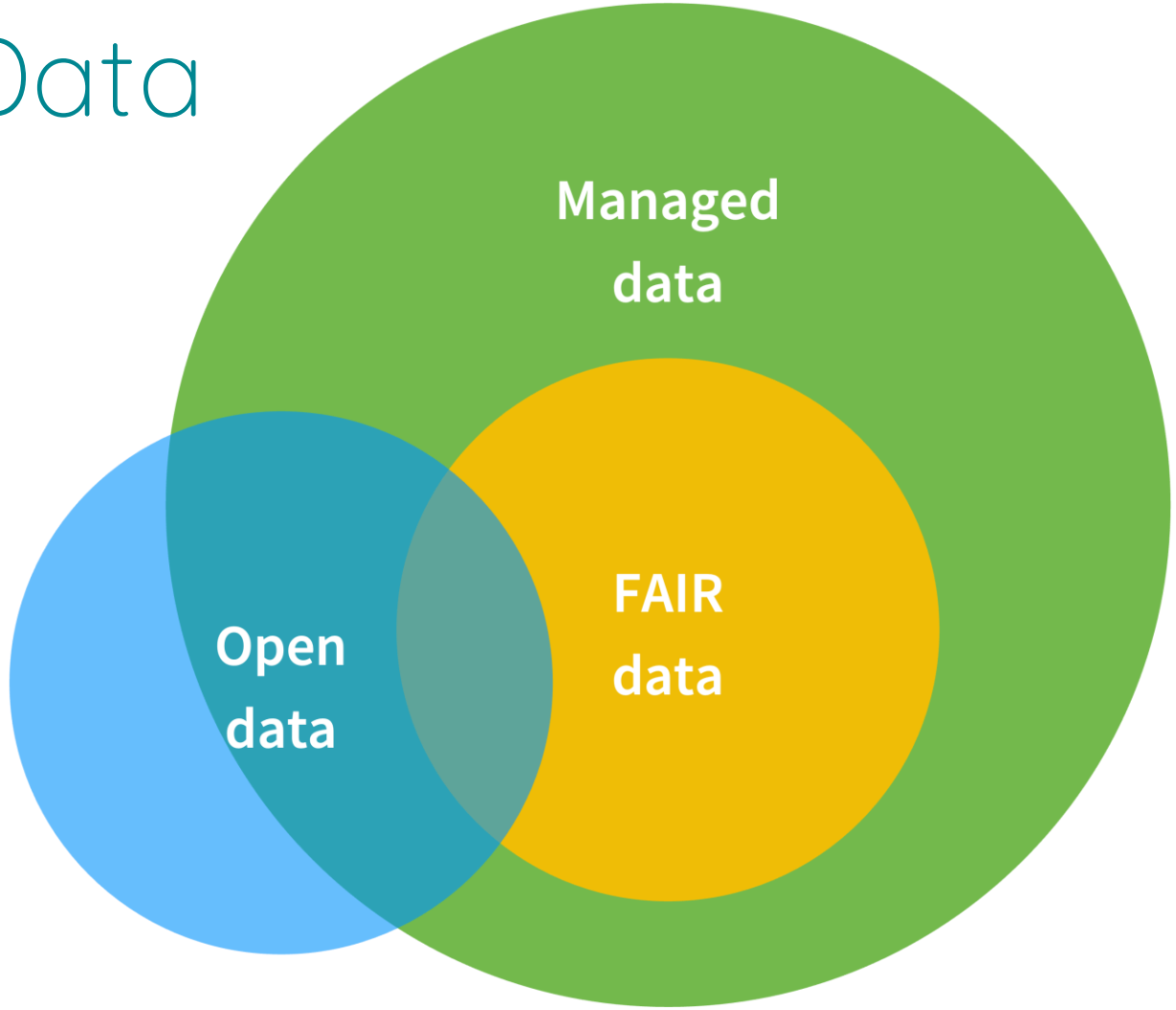


zdroj: SangyaPundir, FAIR data principles, CC BY-SA 4.0

- Findable
  - Metadata
  - Persistentní identifikátory (DOI, ORCiD ID)
  - Registrace a indexace v prohledatelném zdroji (repozitáře)
- Accessible
  - Dostupnost metadat
  - Řízení přístupu k datům
  - Metadata dostupná i když data ne
- Interoperable
  - Široce rozšířený a přístupný jazyk
  - Preferované formáty
  - Slovníky a ontologie
- Reusable
  - Bohatý popis (Read Me File)
  - Licence
  - Oborové/komunitní standardy

podrobněji: GoFAIR

FAIR vs. Open Data



zdroj: Jan Vališ, FAIR vs. Open Data (Venn diagram), CC BY

Co nezveřejňovat?

- Osobní
- Citlivá
- Poškozující legitimní zájem
  - ochrana duševního vlastnictví
  - obchodní tajemství

A co pak s tím?

- (pseudo)anonymizace dat
- Informovaný souhlas
- Embargo, omezení přístupu
- Publikace metadat bez vlastních dat

Motivace pro RDM

- Nutné zlo – plnění požadavků poskytovatelů podpory
- Nastavení procesů vedoucích k:
  - zabezpečení dat před poškozením/ztrátou a
  - větší efektivitě při výzkumu
- Dobře spravovaná data vedou k větší:
  - větší ochotě data sdílet,
  - větší ochotě ostatních Vaše data používat a
  - zvýšení důvěryhodnosti závěrů postavených na datech

Životní cyklus výzkumných dat

DMP – Co to je a k čemu (ne)slouží?

- = Data Managment Plan = Plán správy (výzkumných) dat
- Plán všeho, co se týká dat:
  - finance
  - dokumentace
  - sběr, zpracování, archivace, publikace ev. znovuvyužití
  - úložiště, zálohy
  - řízení přístupu
  - právní a etická problematika
  - nároky poskytovatele podpory, nakladatelství a zaměstnavatele

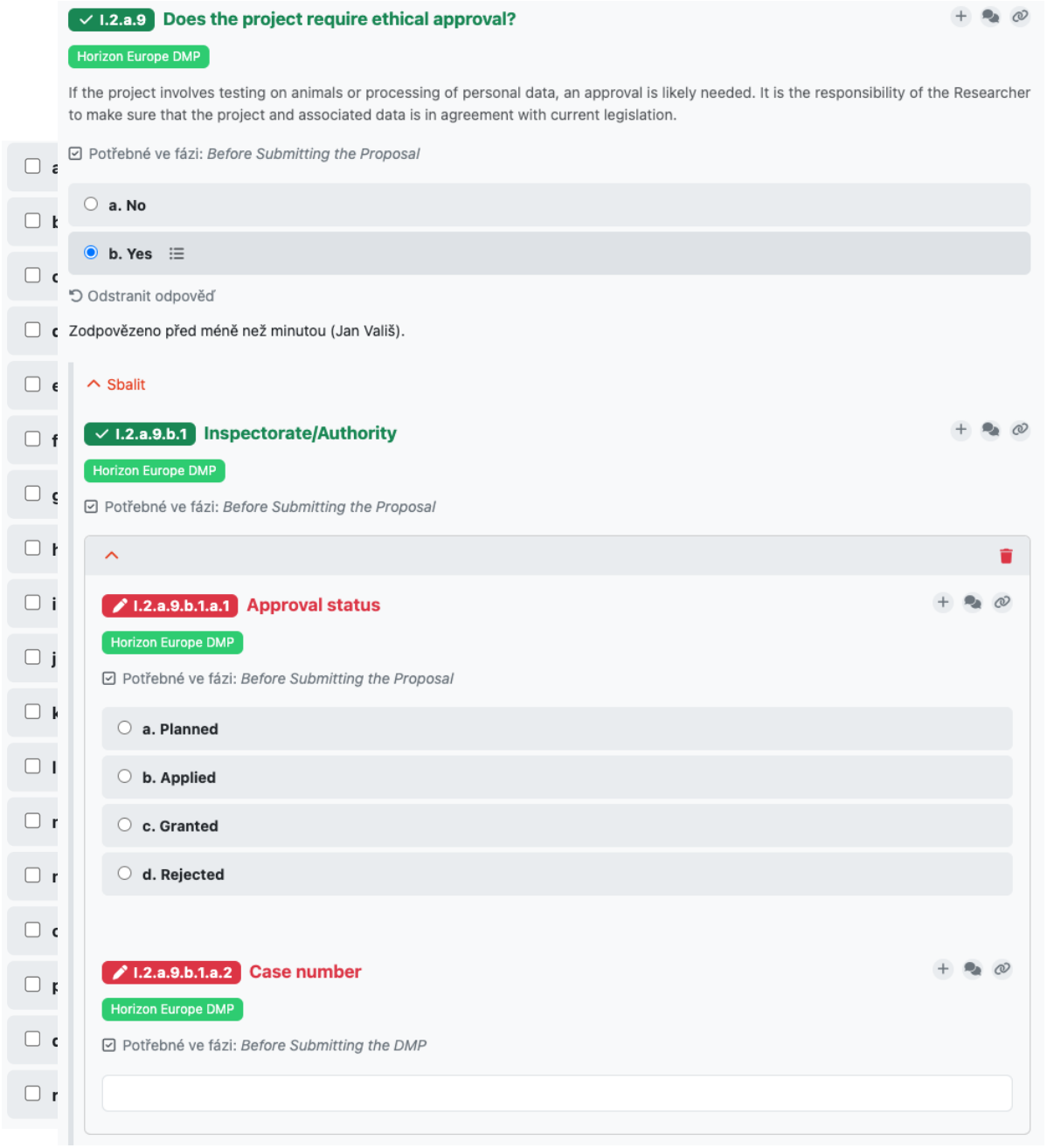
Nástroje pro tvorbu DMP

- Data Stewardship Wizard
- DMP Online
- Argos

National Czech Programme | eosc

DMP – před projektem

- Primárně administrativní informace
  - zapojené osoby (role, ORCiD ID)
  - poskytnutá podpora
  - politika RDM
- K zamyšlení
  - HW, SW a personální zajištění
  - rozpočet
  - souhlas etické komise
  - dohoda v rámci konsorcia



National Czech Programme | eosc

DMP – znovuvyužívání dat

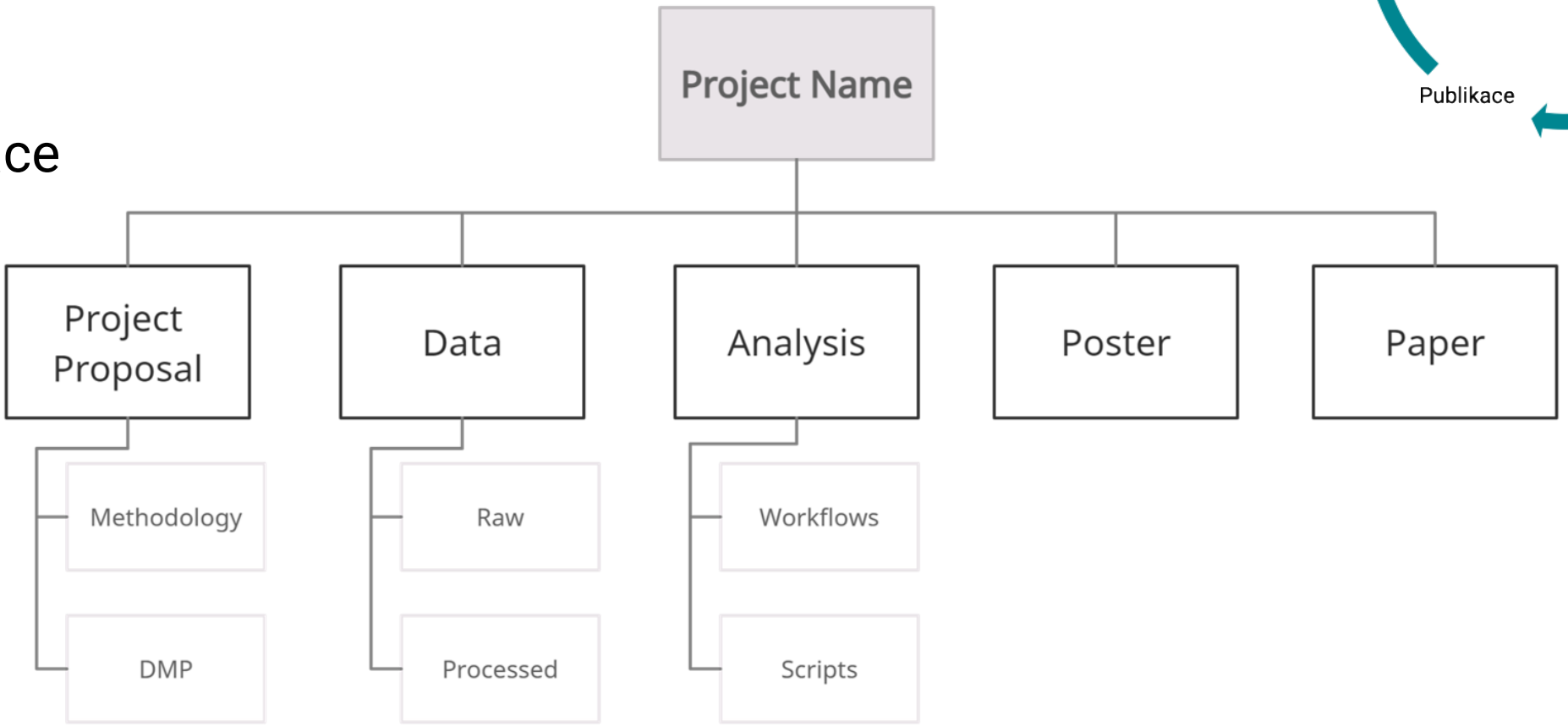
- Benefity
  - úspora času a/nebo financí
  - přístup k jinak těžko dostupným datům
  - větší soubor dat
- Úskalí
  - odlišná metodologie sběru dat
  - různá kvalita dat
  - nekompatibilní a/nebo proprietární formáty
  - etické otázky
- Řešení
  - Implementace FAIR principů
    - komunitní standardy pro RDM
    - preferované formáty
    - permisivní licence
  - Čištění a zpracování dat
  - Harmonizace dat
    - jednotná ontologie
    - transformace a normalizace dat

DMP – znovuvyužívání dat

- Dobrá praxe
  - využití perzistentní identifikátor (DOI)
  - právně a eticky nesporná data
  - znovuvyužití již etablovaných postupů
  - hledání dat i mimo vlastní obor
- Špatná praxe
  - předpoklad správnosti referenčních dat
  - užití bez citace/v rozporu s licencí
- Referenční datasety
  - důvěryhodná ověřená data
  - pro srovnání, validaci atd.
- Nereferenční datasety
  - experimentální, nebo observační data
  - různorodá kvalita

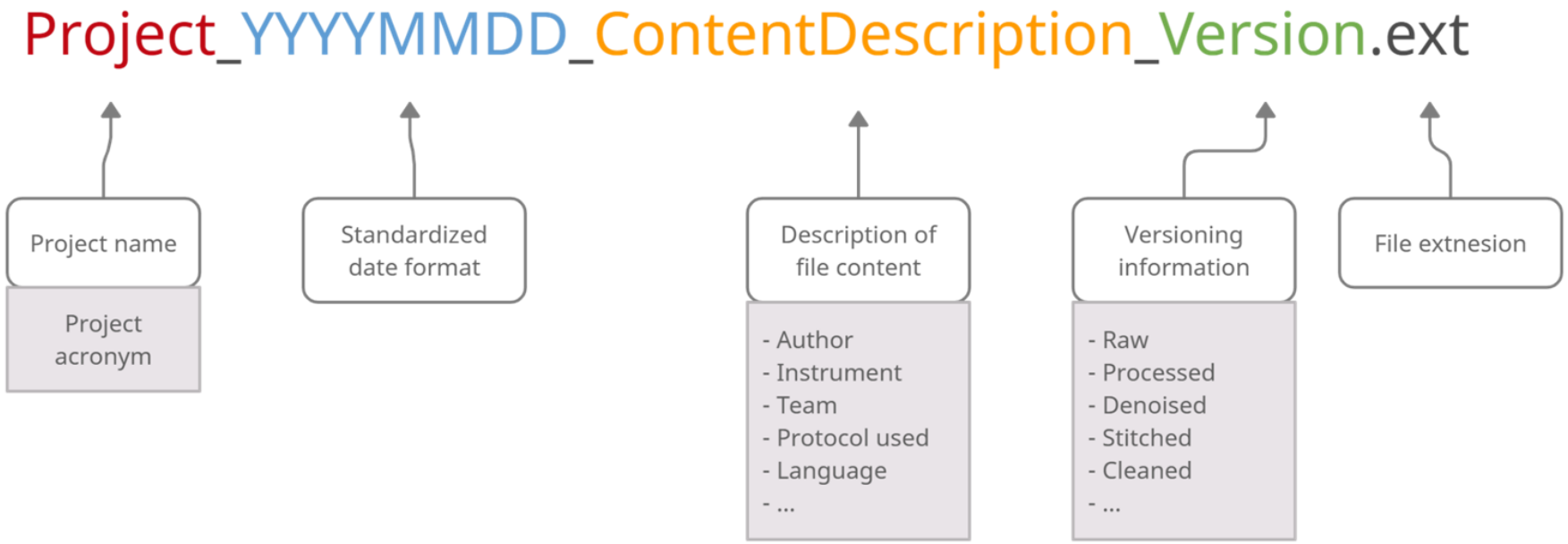
DMP –tvorba a sběr dat

- Organizace



DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
  - bez speciálních znaků
  - bez diakritiky
  - konzistentní



National Czech Programme | eosc

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
  - metadatové konvence (Dublin Core, DataCite …)
  - ontologie a kontrolované slovníky
  - na úrovni souborů i celého datasetu
  - použití elektronických laboratorních deníků (ELN)

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
- Formáty souborů
  - preferované (vhodné pro archivaci, otevřené, dobře dokumentované, strojově i lidsky čitelné, bezztrátová komprese, nezávislé na specifickém SW)
  - TIFF místo BMP; CSV místo XLSX; TXT nebo PDF/A místo DOCX
  - možno publikovat stejný soubor v preferovaném i nepreferovaném, ale populárním formátu

National Czech Programme | eosc

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
- Formáty souborů
- Instrumentální data (teplota, tlak, pH)
  - Popis metody a vybavení znám?

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
- Formáty souborů
- Instrumentální data (teplota, tlak, pH)
- Neinstrumentální data (zdravotní záznamy, dotazníky, matematický model)

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
- Formáty souborů
- Instrumentální data (teplota, tlak, pH)
- Neinstrumentální data (zdravotní záznamy, dotazníky, matematický model)
- Citlivá data

DMP –tvorba a sběr dat

- Organizace
- Pojmenování souborů
- Metadata
- Formáty souborů
- Instrumentální data (teplota, tlak, pH)
- Neinstrumentální data (zdravotní záznamy, dotazníky, matematický model)
- Citlivá data
- Osobní data

National Czech Programme | eosc

DMP – zpracování a analýza dat

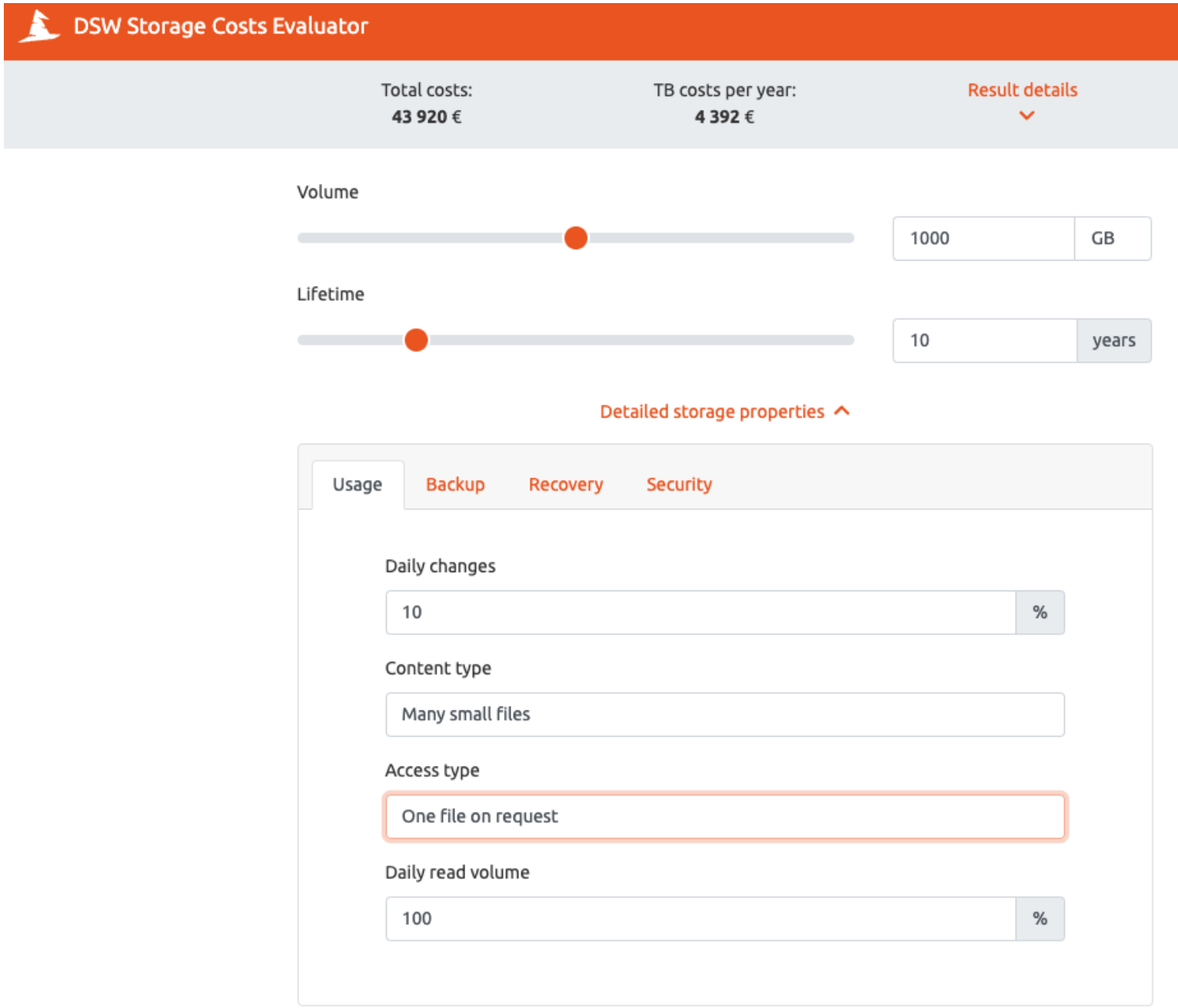
- Využití algoritmických přístupů (pipelines)
- Práce vždy s kopií dat – originál nezměněný
- Minimalizace práce s citlivými a/nebo osobními údaji

DMP – Úschova dat

- Zabezpečení
  - průběžné zálohování (on-campus, off-campus)
  - hesla, šifrování a školení pracovníků
  - omezení přístupu
- Různá stádia
  - V průběhu projektu – ukládání (např. lokálně, na serveru)
  - Po zpracování – publikace (např. repozitář)
  - Po skončení projektu – archivace (lokláně/repozitář/cold-storage?)
- Finanční aspekt

National Czech Programme | eosc

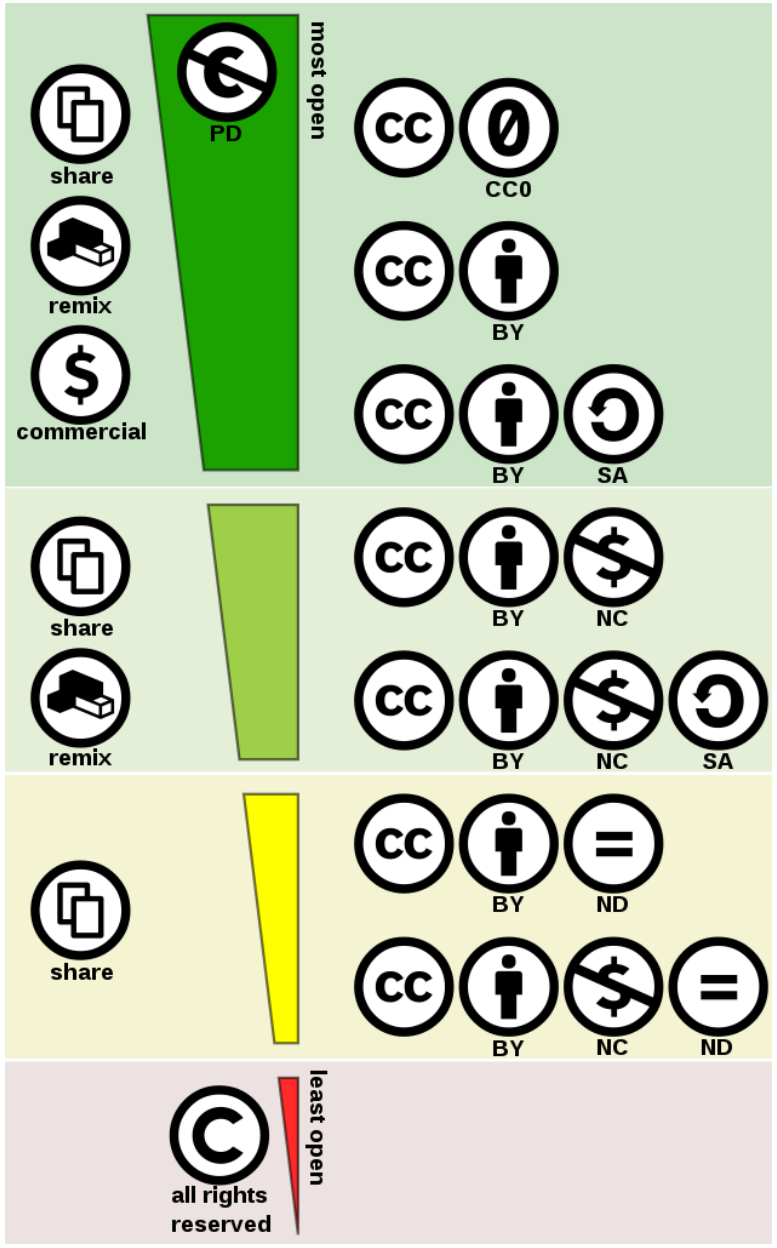
DMP – Úschova



National Czech Programme | eosc

Publikace dat

- Licence
  - Creative Commons (CC)
  - GNU General Public License (GPL)
  - MIT
  - Apache

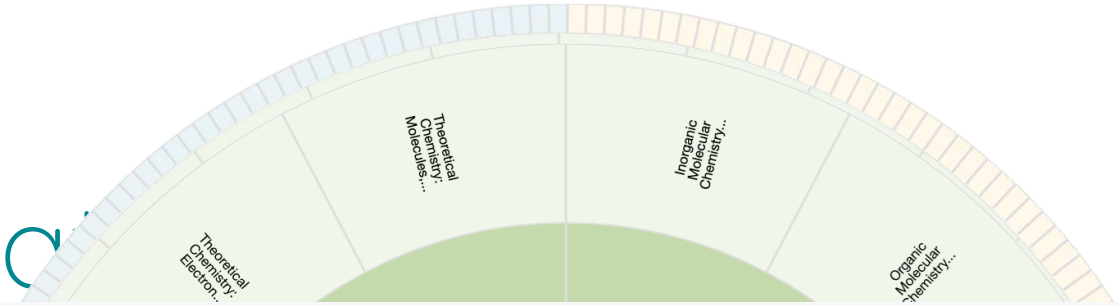


Publikace dat

- Licence
- Volba repozitáře
  - Instituciuonální/všeobecný/oborový
  - Řízení přístupu (otevřený, embargo, omezený)
  - Perzistentní identifikátory (ORCiD, DOI …)
  - Metadata
  - Licence
  - Cena

National Czech Programme | eosc

Publikace da



re3data.org

Search | Browse | Suggest | Resources | Contact

Filter

Reset all

Subjects

- Humanities and Social Sciences (1)
- Life Sciences (3)
  - Biology (3)
    - Basic Research in Biology and Medicine (2)
      - Biochemistry (1)
      - Biophysics (1)
  - Agriculture, Forestry and Veterinary Medicine (1)
- Natural Sciences (10)
  - Chemistry (10)
    - Molecular Chemistry (10)
      - Inorganic Molecular Chemistry - Synthesis and Characterisation (3)
      - **Organic Molecular Chemistry - Synthesis and Characterisation (10)**
    - Chemical Solid State and Surface Research (3)
      - Solid State and Surface Chemistry, Material Synthesis (2)
      - Physical Chemistry of Solids and Surfaces, Material Characterisation (2)
    - Physical Chemistry (2)
      - Physical Chemistry of Molecules, Liquids and Interfaces, Biophysical Chemistry (2)
    - Analytical Chemistry (2)
      - Analytical Chemistry (2)
    - Biological Chemistry and Food Chemistry (2)
      - Biological and Biomimetic Chemistry (1)
      - Food Chemistry (1)
    - Polymer Research (2)
      - Preparatory and Physical Chemistry of Polymers (1)
      - Experimental and Theoretical Physics of Polymers (1)
      - Polymer Materials (1)
    - Theoretical Chemistry (2)
      - Theoretical Chemistry: Electron Structure, Dynamics, Simulation (1)
  - Physics (2)
    - Condensed Matter Physics (1)
    - Statistical Physics, Nonlinear Dynamics, Complex Systems, Soft and Fluid Matter, Biological Physics (1)
    - Particles, Nuclei and Fields (1)
- Engineering Sciences (1)
  - Materials Science and Engineering (1)
    - Materials Science (1)

Countries

AID systems

API

Search... | Search

Toogle short help

← Previous | 1 | Next → | Sort by

Found 10 result(s)

Store.Synchrotron Data Store
myTARDIS Diffraction Image Repository

Subject(s)	Life Sciences, Biology, Basic Research in Biology and Medicine, Biophysics, Natural Sciences, Chemistry, Molecular Chemistry, Organic Molecular Chemistry - Synthesis and Characterisation, Physics, Condensed Matter Physics, Statistical Physics, Nonlinear Dynamics, Complex Systems, Soft and Fluid Matter, Biological Physics, Particles, Nuclei and Fields
Repository type(s)	institutional, disciplinary
Provider type(s)	dataProvider, serviceProvider
Country	Australia

<<<!!!<<< The repository is offline >>>!!!>>> Store.Synchrotron is a fully functional, cloud computing based solution to raw X-ray data archival and dissemination at the Australian Synchrotron, largest stand-alone piece of scientific infrastructure in the southern hemisphere. Store.Synchrotron represents the logical extension of a long-standing effort in the macromolecular crystallography community to ensure that satisfactory evidence is provided to support the interpretation of structural experiments.

Nanomaterial Registry
NanomaterialRegistry

Subject(s)	Natural Sciences, Chemistry, Molecular Chemistry, Organic Molecular Chemistry - Synthesis and Characterisation, Physics, Engineering Sciences, Materials Science and Engineering, Materials Science
Repository type(s)	disciplinary
Provider type(s)	dataProvider
Country	United States

<<<!!!<<< This repository is no longer available. >>>!!!>>>

ioChem-BD

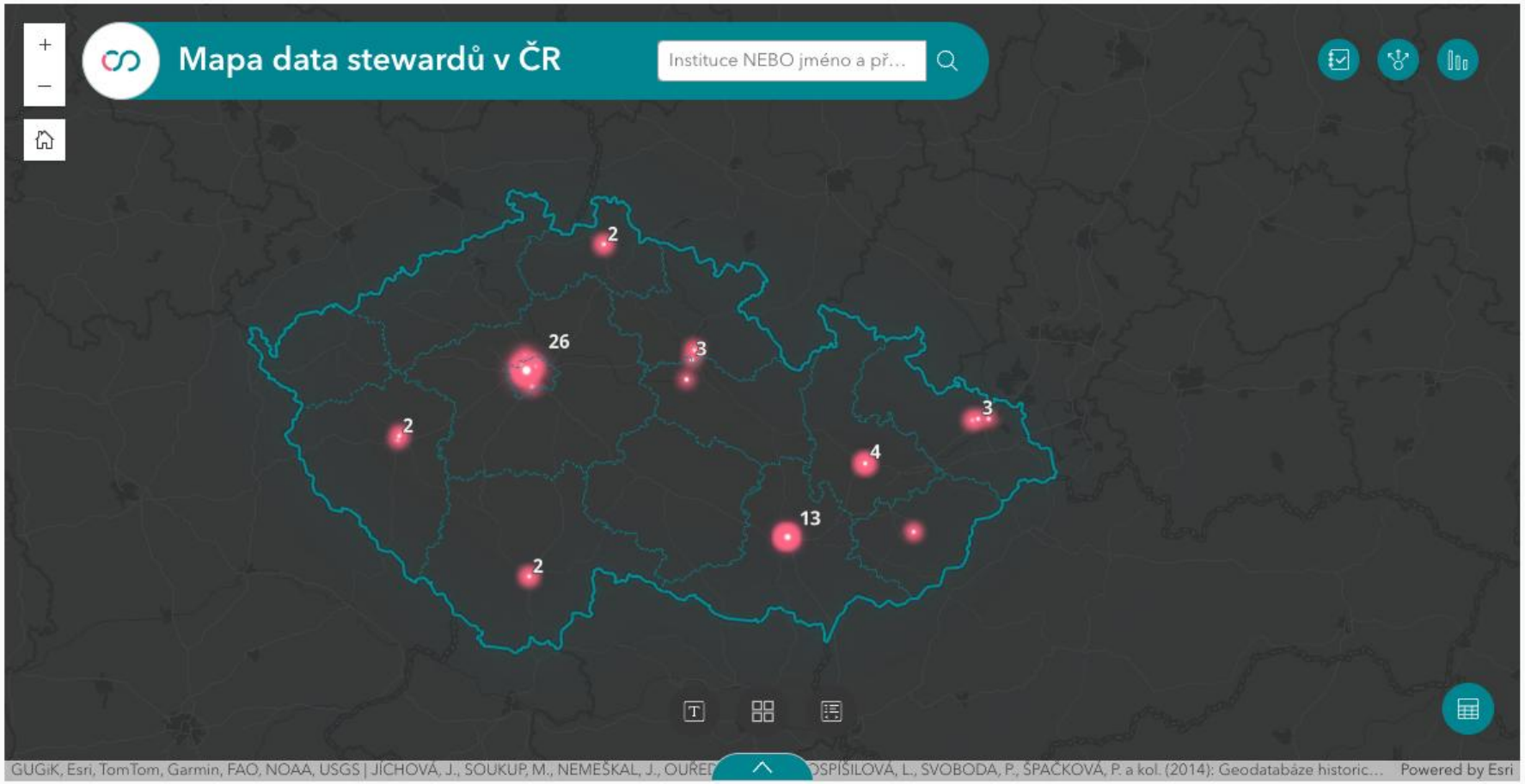
Kde se dozvědět více?

- Samostudium
  - EOSC CZ – školení (pro všechny)
  - NTK – Research Data Managment Guide (pro všechny, zejména STEM)
  - RDMkit (zejména biologické vědy)
  - OpenAIRE (pro všechny)
- U Vašich institucionálních
  - Data Stewardů
  - Knihovníků
- Komunita
  - Manuál začínajícího data stewarda

více informací: eosc.cz/komunity/data-stewardi/komunita-data-stewardu

Komu mapa pomůže?

- **Vědcům:** Snadnější nalezení podpory při správě dat.
- **Institucím:** Možnost prezentovat své odborníky a ukázat, jak podporují otevřenou vědu.
- **Data stewardům:** Získání uznání, inspirace a příležitostí ke spolupráci.



Co si odnést

- RDM a tvorba DMP
  - nejsou samoúčelná administrativa
  - pomáhají plánovat a dokumentovat výzkum
  - podporují dobré praktiky vč. sdílení
- Sdílení dat
  - má pozitiva pro komunitu i pro Vás
  - není vždy vhodné/možné, ale i tak existují řešení
- Nebojte se obrátit na svou instituci a/nebo na komunitu

National Czech Programme | eosc

Dotazy?

jan.valis@techlib.cz

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

cesnet

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM

Registrační číslo IPs EOSC-CZ
CZ.02.01.01/00/22_004/0007682