

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology



Czech Academic and Research
Discovery Services
CARDS

Metadata pro výzkumná data a DataCite schéma

Petra Černohlávková , Hana Vyčítalová 

Úvod do metadat II: metadata v kontextu NRP

15. 6. 2023



Spolufinancováno
Evropskou unií

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



Metadatová schémata/standards

Metadata - data o datech

Schémata

- popisují celkovou strukturu metadat;
- stanovují a definují datové prvky (jména autorů, název, rok vydání) a pravidla pro použití těchto prvků (volitelnost, opakovatelnost...);
- jsou v nich vyjádřené vztahy mezi prvky metadat.

Schémata se mohou stát standardem – obecně uznávaným schématem.

- Standard má zlepšit interoperabilitu a omezit zbytečné odchylky ve schématech.
- Zajišťuje společné chápání významu nebo sémantiky dat/údajů o popisovaných objektech, tj. má zajistit správné použití a interpretaci dat mezi jejich vlastníky, systémy a uživateli. To zahrnuje dohodu o jazyce, pravopisu, formátu data atd.

Metadata Standards Catalog - katalog metadatových schémat – univerzálních i oborových

FAIRsharing – zdroj informací o metadatových schématech, repozitářích a dalších nástrojích (politiky pro dlouhodobé uchovávání dat, sdílení dat, ontologie, slovníky...)

Poznatky ze školení Metadata - úvod

MARC21 (případně ve formě MARCXML)

- standard pro reprezentaci a komunikaci bibliografických informací a souvisejících informací ve strojově čitelné podobě
- komplexní, systematický přesný zápis x složitý, příliš detailní

Dublin Core

- standard pro popis různých zdrojů (textové dokumenty, data, další digitální objekty)
- nejrozšířenější, jednoduchý x nedostatečný popis

METS (Metadata Encoding & Transmission Standard)

- standard pro popis složených/složitých objektů („kontejner“), vyjádřený pomocí jazyka XML
- složitá logická struktura

XML

- formát, jazyk, nástroj pro zpracování dokumentů v XML

RDF

- standard a model pro strojově čitelný popis zdrojů na webu

A pokračujeme - další často užívaná schémata

Univerzální

- Dublin Core
- DataCite
- DCAT (zejména pro otevřená data z oblasti veřejné správy)
- DDI
- další

Oborová

- Digital Curation Centre
(přehled oborových standardů)
- Metadata Standards Catalogue
(přehled oborových standardů)



Because good research needs good data



Search by Discipline



Social Science & Humanities



Physical Science



General Research Data



Earth Science



Biology

- Registrační agentura DOI
- Univerzální metadatové schéma - popis pro jednotlivé datové sady nebo další výstupy z výzkumu (nemělo by se jednat o klasické publikované články, spíše např. zprávy, analýzy, metodiky aj.)
- Přenos metadat o zdroji, který získal DOI
- **DataCite Metadata Working Group**
- Zajištění vývoj a pravidelné aktualizace
- **Aktuální verze 4.4 - 20 základních prvků**
- **Draft verze 4.5 a avízo „major change“ 5.0 (2023)**
- Existuje mapování na DublinCore a mapování na DCAT AP



Aktuální verze 4.4 - 20 základních prvků, ve verzi 4.5 avizován nový 21. prvek (Distribution)

Hierarchické schéma (properties – subproperties), určené povinnosti a opakovatelnosti prvků

- mandatory (M) - musí být uvedeny; povinné pro získání identifikátoru DOI u DataCite
- recommended (R) – nejsou povinné ale doporučené kvůli zajištění interoperability
- optional (O) – volitelné

Obsahuje **seznamy povolených hodnot pro některé entity**

- typ dokumentů (resource type general)
- role přispěvatele (contributor type)
- typ časových údajů (data type) – data available, date issued, date created...
- typ identifikátoru souvisejícího zdroje (related identifier type)
- typ vztahu se souvisejícím zdrojem (relation type) – is version of, is published in
- typ popisu (description type)

Aplikace a zobrazení: XML a JSON

Dokumentace <https://schema.datacite.org/>

Příklad komplexního záznamu datasetu v XML (verze 4.4)

Záznam komplexního záznamu SW v XML (verze 4.4)

Struktura zápisu prvku tvůrce/creator v XML (4.4) a JSON (4.3)

```
<creators>
  <creator>
    <creatorName nameType="Personal">Miller,
    Elizabeth</creatorName>
    <givenName>Elizabeth</givenName>
    <familyName>Miller</familyName>
    <nameIdentifier schemeURI="https://orcid.org/"
    nameIdentifierScheme="ORCID">0000-0001-5000-
    0007</nameIdentifier>
    <affiliation>DataCite</affiliation>
  </creator>
</creators>
```

<https://github.com/datacite/schema/blob/master/source/meta/kernel-4.4/example/datacite-example-full-v4.xml>

```
"creators": [
  {
    "nameType": "Personal",
    "name": "Miller, Elizabeth",
    "givenName": "Elizabeth",
    "familyName": "Miller",
    "nameIdentifiers": [
      {
        "nameIdentifier": "https://orcid.org/0000-0001-5000-0007",
        "schemeUri": "https://orcid.org",
        "nameIdentifierScheme": "ORCID"
      }
    ],
    "affiliation": [
      {
        "name": "DataCite",
        "affiliationIdentifier": "https://ror.org/04wxnsj81",
        "affiliationIdentifierScheme": "ROR"
      }
    ]
  }
],
```

Schéma DataCite – povinné prvky

Identifikátor (Identifier) – DOI

- typ identifikátoru (IdentifierType)

Tvůrce (Creator)

- autor publikace nebo výzkumník, který se hlavně o data zasloužil
- může být osoba nebo organizace
- jméno, příjmení, typ jména, identifikátor (např. ORCID, ISNI), typ identifikátoru, afiliace...

Název (Title)

- typ názvu, pokud tento prvek chybí, automaticky se předpokládá, že se jedná o hlavní název (main title)

Nakladatel (Publisher)

- není pouze nakladatel v pravém slova smyslu, ale i instituce, která disponuje právy k danému zdroji objektům nebo zdroje vydává, archivuje atd. Může se jednat i o repozitář, kde je příslušný zdroj uložen.
- může být pouze jeden

Rok vydání (PublicationYear) – rok zveřejnění dat (kdy zdroj byl nebo bude zveřejněn)

Typ zdroje (ResourceType) – seznam povolených hodnot resourceTypeGeneral, může být doplněno volným textem

Schéma DataCite – doporučené prvky

Předmět (Subject) – může být použito klasifikační schéma nebo volná klíčová slova

Příspěvatel (Contributor) – další osoby, které se podílely na vzniku dat či publikace

- stejné jako tvůrce/creator, ale obsahuje navíc roli příspěvatele (povinné)
- může být osoba i instituce

Časové údaje (Date) – slouží pro zápis dalších dat (jiných než datum zveřejnění)

- seznam povolených hodnot (accepted, created, updated...)

Identifikátor souvisejícího zdroje (RelatedIdentifier)

- související zdroj – článek publikovaný na základě dat, časopis, ve kterém vyšel článek atd.
- DOI, ISBN, Handle atd.

Popis (Description)

- abstrakt, obsah, technické informace...

Geolokace (GeoLocation) – identifikace lokace spojené se zdrojem

- zeměpisná šířka a délka, jméno lokace, popis lokace

Schéma DataCite – volitelné prvky

Jazyk (Language)

- doporučené kódy jazyků ISO 639-1

Další identifikátory (AlternateIdentifier)

- často pro lokální či interní identifikátory popisovaného objektu

Velikost (Size)

- počet stránek, objem v bytech, trvání v hodinách...

Formát (Format)

- Doporučeno použití přípon souboru (file extension) nebo MIME type

Verze (Version)

- číslované menší změny, v případě větších změn se doporučuje popsat zdroj jako novou jednotku

Práva (Rights)

- jakékoliv typy licencí: Creative Commons, Apache License...
- doporučeno uvést identifikaci podle <https://spdx.org/licenses/>

Údaje o financování (FundingReference)

- poskytovatel podpory, identifikace projektu (ID/číslo, název)

Související zdroj (RelatedItem)

- tvůrce, název, rok vydání, nakladatel, stránkování, typ vztahu popisovaného a souvisejícího zdroje...

Další příklady záznamů datasetů

Caltech Data

<https://doi.org/10.22002/8r8hp-8jt82>

TU Wien Research Data

<https://doi.org/10.48436/nr0v5-a9g32>

TU Graz Repository

<https://doi.org/10.3217/mgd4n-gq267>

Dryad

<https://doi.org/10.5061/dryad.wpzgmsbr8>

NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology



**Czech Academic and Research
Discovery Services
CARDS**

Děkujeme za pozornost!

petra.cernohlavkova@techlib.cz
hana.vycitalova@techlib.cz

