Pilotní řešení řízených slovníků/taxonomií pro dokumentový repozitář – dokumentace

**Seznam taxonomií je na GitHubu: https://github.com/Narodni-repozitar/nr-taxonomies/tree/taxonomies-for-nr-docs/data/excel.**

Obsah

[AccessRights 2](#_Toc103073137)

[ContributorType 3](#_Toc103073138)

[ResourceType 4](#_Toc103073139)

[ResourceType\_relatedItem 5](#_Toc103073140)

[Languages 5](#_Toc103073141)

[Countries 5](#_Toc103073142)

[Institutions 5](#_Toc103073143)

[Funders 6](#_Toc103073144)

[itemRelationType 7](#_Toc103073145)

[Licenses 8](#_Toc103073146)

[subjectCategories 9](#_Toc103073147)

[Přehled struktury jednotlivých taxonomií 9](#_Toc103073148)

**Obecné informace**

Slovníky pro dokumentový NR byly původně připraveny v roce 2020. V trochu pozměněné podobě byly použity pro pilotní datový repozitář v roce 2021. Na konci roku 2021 vyšla nová verze Invenia 3.5, která výrazně ovlivnila i podobu slovníků. Tato verze byla vývojáři NR aplikována na začátku roku 2022. Slovníky bylo třeba upravit, a to z těchto důvodů:

1. Bylo třeba, aby stejné položky v různých slovnících byly stejného typu.
2. Slovníky v rámci Invenia 3 nedovolují hierarchické uspořádání (kromě afiliací).

První problém byl odstraněn sjednocením názvů záhlaví, jejich přejmenováním. Při úpravách názvů záhlaví jsme se snažili vycházet z modelu pro slovníky v Inveniu 3. Některé položky (sloupce) byly vypuštěny jako zbytečné.

Druhý problém byl řešen na koncepční schůzce s vývojáři a bylo rozhodnuto, že pro slovníky si vytvoříme vlastní model, aby hierarchické uspořádání bylo možné vždy, když je potřeba.

Diskutován byl také problém slugů, zda je vhodnější, aby zůstaly jako písmenné stringy nebo by měly být vyjádřené ve formě id kódu. Zatím zůstávají stringy, které vychází buď z názvu (title\_cs) s pomlčkami nebo se jedná o převzaté numerické, alfabetické nebo alfanumerické kódy z původních třídění a řízených slovníků (identifikátory jednotlivých položek).

Při definování slovníku byla zohledňována existence relevantních (mezinárodních) slovníků vhodných pro repozitáře. Vycházelo se z praxe NUŠL, pokynů OpenAIRE, slovníků od COAR, pokynů od DataCite (agentury přidělující DOI převážně datům a šedé literatuře) a dalších.

## AccessRights

Úrovně přístupu k plným textům (dokumentům, souborům) pro koncového uživatele, nerozlišuje se kde, tj. ať už přímo v prohlíženém zdroji nebo v jiném, z kterého byl záznam sklizen. Jedná se o aplikaci slovníku COAR (Coalition for OA repositories), tedy slovníku vytvořeného přímo pro potřeby repozitářů. Pro účely Národního repozitáře byl tento slovník rozšířen o české překlady. Zároveň byl v roce 2020 do COAR zaslán požadavek o publikování překladu do češtiny, zatím tento požadavek nebyl zpracován.

Stejný slovník využívá i OpenAIRE, mělo by tedy být usnadněné mapování z NR do OpenAIRE. Tato úprava slovníku obsahuje český a anglický termín a URI COAR slovníku.

Dokumentace COAR slovníku: <http://vocabularies.coar-repositories.org/documentation/access_rights/>

Slovník v RDF z github: <https://github.com/coar-repositories/vocabularies/tree/master/access_rights>

Hodnoty:

* open access
  + cz: otevřený přístup
  + Volně dostupný dokument.
* embargoed access
  + cz: odložené zpřístupnění
  + Odložený přístup definovaný datem, po jehož uplynutí se soubory zveřejní. Většinou se týká akademických článků, kde vydavatel umožňuje autorovi volně zpřístupnit článek až po určité době. Po vypršení stanoveného času se práva změní na open access.
* restricted access
  + cz: omezený přístup
  + Dokument je přístupný pouze vybrané skupině uživatelů (např. zaměstnancům a studentům univerzity) nebo za určitých podmínek (např. registrace do systému); většinou se jedná o omezený přístup pro zaměstnance konkrétní instituce; přístup může být udělen po přihlášení, registraci nebo žádosti přes email.
* metadata only access
  + cz: pouze metadata
  + Neexistuje elektronický soubor nebo není veřejně dostupný. V podstatě pokud nelze použít nějakou možnost výše, tak bude užita tato možnost.
  + Bude se týkat převážně tištěných dokumentů nebo dokumentů vyloučených ze zveřejnění (např. z důvodu patentového řízení nebo ochrany osobních údajů).

## ContributorType

Jedná se o slovník hodnot, kterých může nabývat role přispěvatele.

Náš slovník je kombinací výběru položek z těchto schémat/seznamů: DataCite, MARC21 a EVSKP-MS. Výběr probíhal na základě praxe s existujícími daty v NUŠL a užití v jiných repozitářích.

**DataCite** posloužil jako základ a většina termínů byla převzata. Vyřazeny byly takové, které prozatím nenajdou v repozitáři využití. Kompletní DataCite slovník převzal OpenAIRE. Pro předávání dat do OpenAIRE budou hodnoty převzaté z jiných systémů nahrazeny hodnotou „Other“, která je součástí slovníku DataCite. Definice hodnot v DataCite jsou dostupné v Apendixu 1 <https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf>, nemají však trvalé identifikátory jako slovníky COAR.

**Label kód v MARC**

ContactPerson

Data Collector

DataCurator

DataManager

Distributor dst

Editor edt

Producer pro

ProjectLeader rth

ProjectManager

ProjectMember rtm

Researcher res

ResearchGroup

RightsHolder asg

Supervisor

Z **EVŠKP-M** byl převzat termín advisor a referee (vedoucí a oponent práce) viz <http://www.evskp.cz/standardy/evskp/#_Toc203923576>. Definice Supervisora v DataCite by nezahrnula i roli vedoucího práce, přestože se tento termín často užívá, a roli oponenta neobsahuje vůbec. Z těchto důvodů byly obě hodnoty převzaty z evškp-ms.

Referee opn

Advisor ths

**MARC 21** obsahuje velmi obsáhlý seznam kódů rolí tvůrců, který by jako celek pro záběr Národního repozitáře pro výstupy VaVaI neměl smysl, jako např. iluminátor rukopisů, dřevorytec apod. Byly tedy vybrány pouze hodnoty považované za relevantní. V katalozích (a jejich OAI-PMH) jsou pak zaznamenány třímístné písmenné kódy. Kódy a definice jednotlivých rolí v MARC 21 viz <https://text.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/informativni-materialy/kody-roli>

ill ilustrátor

cur kurátor výstavy

mod moderátor

trl překladatel

pht fotograf

rev recenzent

clb spolupracovník

art umělec

ive dotazovaný

ivr dotazovatel

orm organizátor akce

spk spíkr

pan panelista

pbl nakladatel

pfr proofreader

own owner

fmo former owner

rsp respondent

## ResourceType

Slovník resourceType (typ dokumentu) byl vytvořený na míru českému prostředí a Národnímu repozitáři. Při jeho tvorbě byly vzaty v potaz Definice druhů výsledků pro IS VaVaI (RIV), slovník Resource Types od COAR (<https://vocabularies.coar-repositories.org/resource_types/3.0/>; použitá též v OpenAIRE), typologie DataCite Resource Type General, typologie Národního úložiště šedé literatury a obdobné typologie užité v jiných repozitářích (ASEP, Zenodo aj.). Cílem bylo dosáhnout co nejvyššího souladu mezi používanými typologiemi, proto slovník obsahuje rovněž mapování na hodnoty slovníku COAR a DataCite. To by následně mělo usnadnit mapování a výměnu dat u datových i dokumentových repozitářů.

Slovník má dvě úrovně a obsahuje termíny v českém a anglickém jazyce, v případě potřeby také alternativní označení typů dokumentů (jako např. knihy a monografie) obsahuje též mapování na hodnoty slovníku COAR Resource Types (včetně trvalých identifikátorů <https://vocabularies.coar-repositories.org/resource_types/>) a na hodnoty slovníku DataCite Resource Type General.

Pro označení typu souvisejícího zdroje v metadatech popisovaného zdroje byl zkušebně vytvořen nový slovník resourceType\_relatedItem, který obsahuje i speciální typy, které určitě v počátcích Národní repozitář neplánuje evidovat.

## ResourceType\_relatedItem

Slovník obsahuje celý slovník ResourceType a další speciální typy zdrojů, jejichž přijímání do Národního repozitáře se zatím neočekává, ale bude pravděpodobné, že jejich metadata budou připojována k přijímaným dokumentům coby související zdroje přes pole relatedItem. Strukturou se od ResourceType nijak neliší. Jedná se o zkušební řešení za účelem otestování technických možností/zjednodušení technického řešení různého použití.

## Languages

Slovník jazyků vychází ze standardu ISO 639-2 B (určen pro bibliografický popis. K jednomu záznamu lze přiřadit více jazyků. Seznam viz <https://cs.wikipedia.org/wiki/Seznam_k%C3%B3d%C5%AF_ISO_639-2>

Slovník obsahuje dvoumístné kódy (ve slugu), název jazyku v češtině a angličtině a třímístné kódy jazyku v angličtině a v originálním jazyce.

## Countries

Slovník zemí (countries) vychází ze standardu ISO 3166 pro země, obsahuje dvoumístné kódy zemí (ve slugu), název země v českém a anglickém jazyce a třímístné kódy zemí. Seznam viz https://en.wikipedia.org/wiki/List\_of\_ISO\_3166\_country\_codes.

## Institutions

Slovník institucí evidovaných v Národním repozitáři – jedná se převážně o výzkumné organizace, neziskové organizace, muzea, galerie, knihovny apod. Vycházelo se ze seznamu partnerských institucí Národního úložiště šedé literatury, slovník tedy obsahuje zejména instituce, poskytující data do NUŠL, včetně zaniklých ústavů AV ČR. Přidány byly instituce, které jsou členy EduID a dá se od nich očekávat zájem o ukládání výzkumných dat. Slovník byl vytvořen pro české prostředí, tvoří ho převážně české instituce. Z tohoto slovníku se budou čerpat data různá pole jako je provider, degree grantor, publisher atd. Dá se předpokládat, že tento slovník bude dále rozšiřován dle potřeb. Poskytovatelé dotací jsou v samostatné taxonomii.

Ve většině případů má záznam instituce pouze jednu úroveň. U vysokých škol a několika dalších institucí obsahuje i podřízené organizační jednotky jako fakulty a katedry.

Obsahuje následující údaje o instituci:

* Názvové údaje:
  + Název instituce, v českém a anglickém jazyce (title\_cs a title\_en).
  + Zkratky názvů institucí (acronym).
  + Další známé názvy institucí i organizačních jednotek (nonpreferredLabels v čj a aj).
* URL odkaz na webové stránky instituce.
* Identifikace instituce:
  + ROR,
  + identifikátor RID z registru univerzit,
  + DOI z Crossref,
  + IČO instituce (bude sloužit jako její identifikátor).
* Typ instituce (tags).
  + Vychází z původního třídění institucí v NUŠL.
  + Současné hodnoty, kterých může toto pole nabývat: archiv, AV ČR, galerie, knihovna, muzeum, nezisková organizace, soukromá VŠ, státní správa, státní VŠ, veřejná VŠ, VVI, výzkumný ústav, osobní archiv, soukromý sektor, ostatní.
  + Nejedná se o povinné pole, může být vybráno více hodnot, mohou být instituce bez typu; může sloužit pro následné filtrování nebo zobrazování dat.
* Pole pro účely technického řešení:
  + „context“ pro odlišení vysokých škol coby degreeGrantor (DegreeGrantor pro VŠ, ostatní default),
  + „nameType“ s hodnotou organizational pro rozlišení institucí a osob coby autorů v záznamech, při importech.

Poznámka k organizačním jednotkám (2 a nižší level – fakulta, katedra/ústav, tzv. children) – obsahují pouze tyto hodnoty:

* Názvové údaje:
  + Název instituce, v českém a anglickém jazyce (title\_cs a title\_en).
  + Zkratky názvů institucí (acronym).
  + Další známé názvy institucí i organizačních jednotek (nonpreferredLabels v čj a aj).
* Identifikace instituce:
  + ROR,
  + identifikátor RID z registru univerzit,
  + DOI z Crossref,
  + IČO instituce (bude sloužit jako její identifikátor).
* Pole pro účely technického řešení
  + „context“ pro odlišení vysokých škol coby degreeGrantor (DegreeGrantor pro VŠ, ostatní default),
  + „nameType“ s hodnotou organizational pro rozlišení institucí a osob coby autorů v záznamech, při importech.

## Funders

Slovník funders (poskytovatelé financí) je vytvořený na míru českému prostředí a Národnímu repozitáři. Hodnoty slovníku jsou převzaty zejména z Informačního systému pro výzkum, vývoj a inovace (IS VaVaI), doplněni jsou někteří nejčastější poskytovatelé ze zahraničí. Do budoucna se počítá s komplexnějším řešením vytěžování databází o financování jako je např. IS VaVaI a OpenAIRE, aby se eliminovalo riziko překlepů a duplicitní vyplňování všech údajů o financování VaVaI.

Obsahuje následující informace o poskytovatelích:

* název poskytovatele v češtině a angličtině, popřípadě dalších jazycích,
* předchozí názvy nebo nepreferovaná zkratka názvu instituce (označení záhlaví jako nonpreferredLabels),
* acronym (nejznámější česká zkratka názvu instituce),
* relatedURI (identifikátory ROR a DOI z Crossref),

## itemRelationType

Slovník popisuje vztah popisovaného zdroje k připojené jednotce, např. kapitola odkazuje na knihu. Do slovníku byly převzaty hodnoty z DataCite (https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel\_v4.4.pdf.

Slovník obsahuje název vztahu v češtině a angličtině. Pokud se jedná o párovou vazbu (př. Je verzí/Má verzi), je uvedena rovněž párová vazba. Do budoucna považujeme za užitečné připojit do taxonomie (nebo mimo ni) i nápovědu pro jednotlivé vazby (příklady jejich užití) a zobrazovat je uživateli, pokud to bude technicky možné, v rozhraní repozitáře ve vkládacím formuláři.

## Licenses

Tento slovník obsahuje licence, které je možné dílům udělit. Zatím se jedná pouze o Creative Commons licence, známé pod zkratkou CC licence. Slovník bude nutné do budoucna rozšířit o další typy licencí. Vycházeli jsme z modelu Zenoda a NUŠL, kde jsou přidělovány rovněž CC licence. Slovník CC licencí je zatím pojatý hierarchicky, aby bylo možné při vyhledávání v repozitáři podle licencí filtrovat výsledky. Na první úrovni je název sady licencí (zatím tedy pouze CC). Na druhé úrovni je verze licencí (pro CC 1 – 4). Na třetí úrovni se pak nachází konkrétní licence.

Teprve nejnižší úroveň nese následující informace:

* název česky a anglicky,
* odkaz do prohlášení k dané licenci CC, které je povinné uvádět, ve struktuře <https://creativecommons.org/licenses/by-nc/2.5/>,
* kvůli technickému řešení slovník obsahuje také odkaz do obrázku, který by měl následně v záznamech představovat ikonu, reprezentující informaci o licenci.

Více o verzích licencí dostupné z <https://wiki.creativecommons.org/wiki/License_Versions>

Celé detailní znění konkrétní licence lze získat přes odkaz s legalcode, který má následující strukturu - <https://creativecommons.org/licenses/by-nd-nc/1.0/legalcode>

**Poznámka pro užití**

* Udělení licence je nevratné.
* Každé dílo by mělo být publikováno pouze pod jedním typem licencí CC.
* U licence 1.0 jsou u poslední možnosti prohozena písmena oproti jiným verzím. Tj. ve verzi 1.0 je BY-ND-NC, ve zbylých verzích je to BY-NC-ND.
* Verze 3 vyšla v národních mutacích, lze ji použít obecně pro mezinárodní organizace, více lze dohledat na wiki creativecommons; v tomto slovníku se prozatím pracuje jen s českou národní verzí, s možností pozdějšího rozšíření, pokud to bude potřeba.

## subjectCategories

Taxonomie byla vytvořena a zařazena pro potřebu oborové klasifikace zejména datových sad. Byla přejata klasifikace Struktura oborů OECD (Frascati manual) jako celek. Jeho základ (první tři úrovně) jsou užívány také jako číselník oborů v IS VaVai. Zvažována byla ještě klasifikace, užívána pro systém re3data, vytvořena původně v Deutsche Forschungsgemeinschaft (https://www.dfg.de/en/dfg\_profile/statutory\_bodies/review\_boards/subject\_areas/index.jsp), nakonec byla ale vybrána obecnější klasifikace FRASCATI, důvodem bylo především užití v IS VaVaI.

Slovník obsahuje název předmětové kategorie v angličtině a jeho překlad do češtiny, provedený pracovníky NTK, částečně převzatý z <https://www.tacr.cz/wp-content/uploads/documents/2019/09/27/1569592365_Struktura_oboru_FORD.pdf>.

Třídění a označování datových sad a případně dalších zdrojů pomůže k lepší práci s nimi v rámci systému (poskytování specifických oborových popisů) a k lepší a přesnější diseminaci (předávání relevantních záznamů do oborových repozitářů).

K jednomu zdroji/záznamu může být přiřazeno více oborových kategorií.

## Přehled struktury jednotlivých taxonomií

itemRelationType (slug, title\_cs, title\_en, pair, hint)

accessRights (slug, title\_cs, title\_en, relatedURI\_COAR)

languages (slug, title\_cs, title\_en, alpha3CodeENG, alpha3CodeNative)

countries (slug, title\_cs, title\_en, alpha3Code)

licenses (slug, title\_cs, title\_en, icon, relatedURI\_URL)

contributor type (slug, title\_cs, title\_en, marcCode, dataCiteCode)

institutions (slug, title\_cs, title\_en, contexts\_0, tags\_0, tags\_1, acronym, RID, ICO, relatedURI\_ROR, relatedURI\_CrossrefFunderID, relatedURI\_URL, nameType, nonpreferredLabels\_cs\_0 až 5, nonpreferredLabels\_en\_0)

resourceType (slug, title\_cs, title\_en, nonpreferredLables\_cs\_0, onpreferredLables\_en\_0, coarType, relatedURI\_COAR, dataCiteType)

resourceType\_relatedItem (slug, title\_cs, title\_en, nonpreferredLables\_cs\_0, onpreferredLables\_en\_0, coarType, relatedURI\_COAR, dataCiteType)

funders (slug, title\_cs, title\_en, nonpreferredLabels\_en\_0, title\_fr, title\_sk, title\_pl, title\_de, acronym, relatedURI\_CrossrefFunderID, relatedURI\_ROR)

….subjectCategories (slug, title\_cs, title\_en)