

Metadata a jejich správa

Petra Černohlávková  <https://orcid.org/0000-0001-8683-6156>,

Ondřej Zamazal  <https://orcid.org/0000-0002-7442-9016>

Správa výzkumných dat a práce s repozitáři, MUNI

3.3.2026



Spolufinancováno
Evropskou unií

Národní technická knihovna (NTK)

NTK
50°6'14,083"N, 14°23'26,365"E
Národní technická knihovna
National Library of Technology

- Praha, Dejvice
- Historie sahá až do roku 1718
- Poskytuje širokou škálu služeb na národní úrovni (např. CzechELib nebo Národní centrum pro PID) tak i klasické knihovnické služby pro studenty ze STEM oborů
- Je součástí realizátorů iniciativy EOSC v ČR



Jaká metadata dokážeme z produktu
vyčíst?

Jak by vypadal jejich metadatový
zápis?



FAIR: Findable

- Nalezitelná data
- Cílem je zabezpečit možnost **najít data**
- Data by měli být schopni najít jak lidé tak počítače
- Významnou roli hrají **perzistentní identifikátory**
- **Metadata** datové sady jsou základem pro nalezení v rámci různých vyhledávacích služeb

DOI:

DOI 10.5281/zenodo.3978090

Keyword(s):

pine wood nematode

Bursaphelenchus xylophilus

quarantine pest

survey

Finland

Subject(s):

Bursaphelenchus 

surveillance 

data 

Related identifiers:

Supplement to

[10.3897/neobiota.58.38313](https://doi.org/10.3897/neobiota.58.38313) (Journal article)

[10.5281/zenodo.3842358](https://doi.org/10.5281/zenodo.3842358) (Software)

FAIR: Accessible

- Přístupná data
- Meta(data) by měla být získatelná pomocí svého identifikátoru
- Za použití standardního komunikačního protokolu, který může použít kdokoliv
- Autentizace a autorizace jsou možné
- Nemusí být nutně „online“ protokol
- Ne všechno musí být volně dostupné (!)



FAIR: Interoperable

= možnost *integrovat* data s jinými datasety a zajištění *kompatibility* s nástroji pro jejich zpracování

- používání **standardizovaných výrazů** pro popis a **metadata**
- oborové standardy
- otevřené formáty, např.:
 - Obrazová data: TIFF, PNG, GIF, BMP, SVG
 - Tabulky : CSV, TXT, XLSX
 - Texty : XML, PDF/A, HTML, JSON, TXT, RTF
 - Kontejnery : TAR, GZIP, ZIP
 - Databáze: XML, CSV, JSON, POR
 - Souřadnice : SHP, DBF, GeoTIFF, NetCDF
 - Video : MPEG, AVI, MXF, MKV
 - Zvuk : WAVE, AVI, MP3, MXF, FLAC
 - Statistické soubory : DTA, POR, SAS, SAV

FAIR: Reusable

= optimalizace dat pro co nejvyšší možnost *opětovného využití*

- důkladný popis dat – **dokumentace**
 - popisná **metadata**
 - provenance & kontext
- použití **licencí**

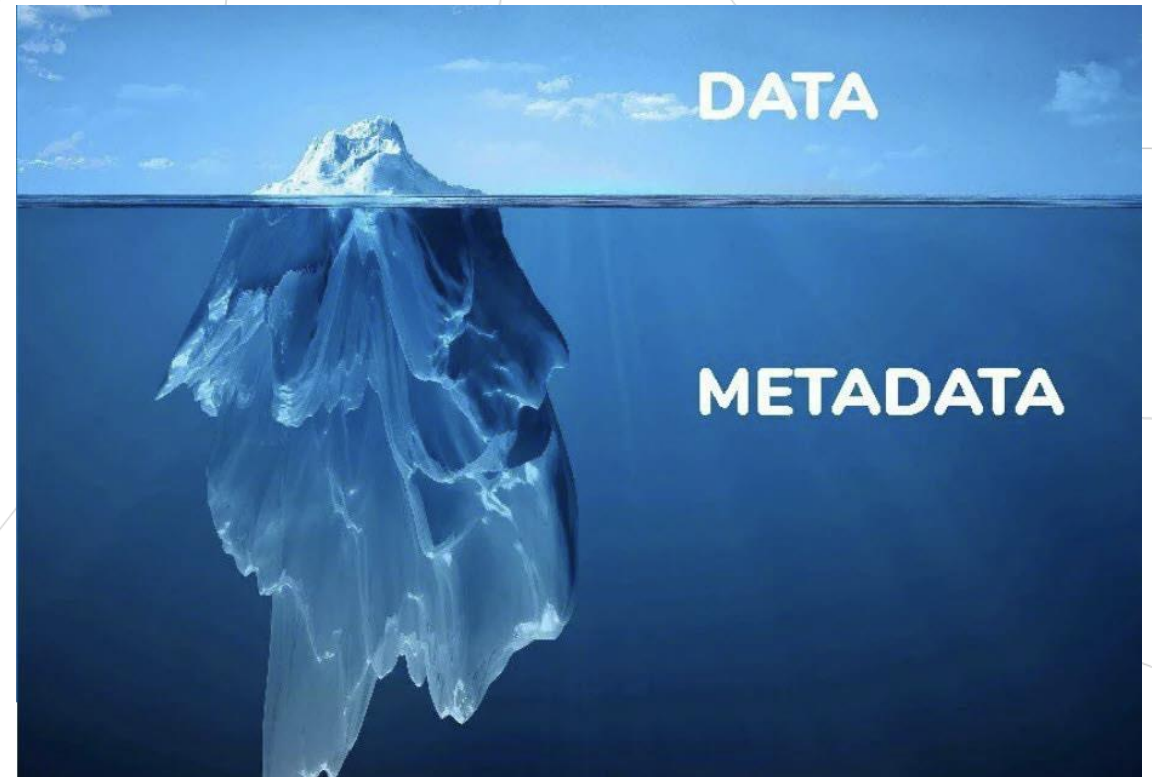
Teplota

- 36,6
- 36,1
- 37,8
- 41,5



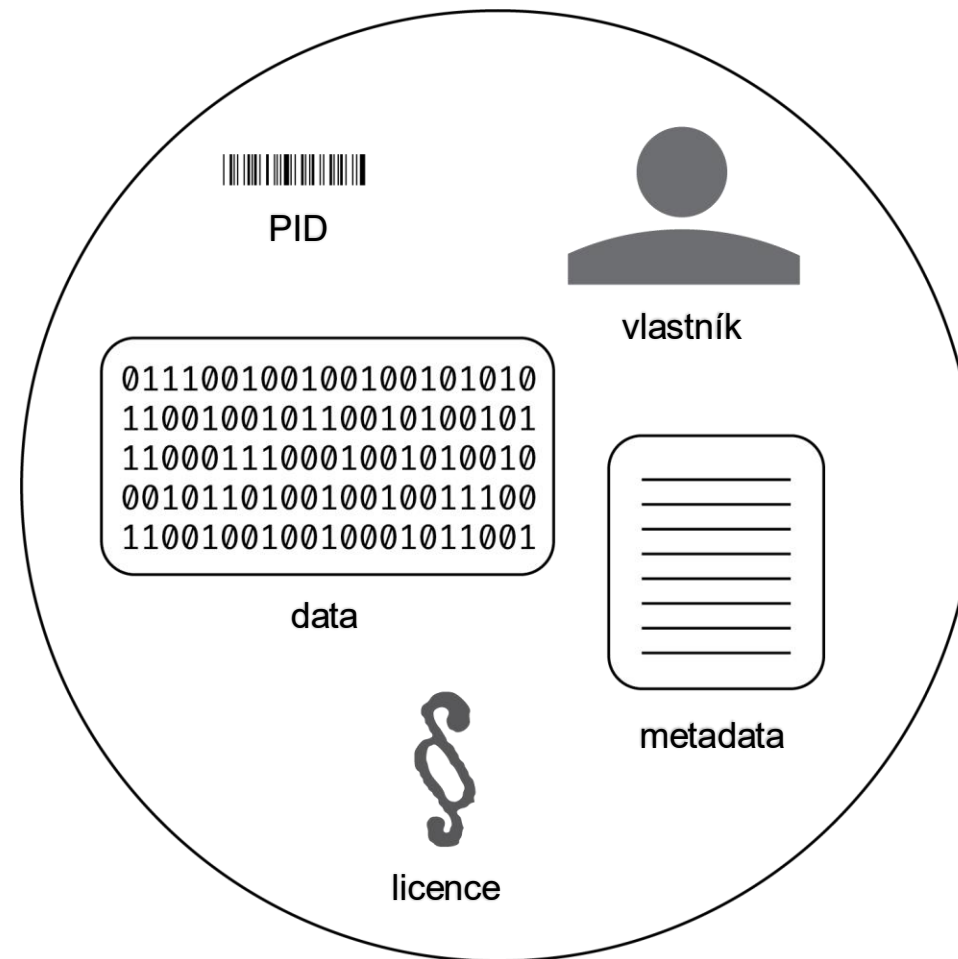
Metadata

- Klasická definice: metadata jsou "data o datech"
- Metadata = informace o vzniku položky, názvu, tématu, vlastnostech, ceně...
- Čím standardizovanější je popis metadat, tím lepší jsou výsledky vyhledávání, filtrační operace atd.
- Metadata by měla být dostupná, i kdyby už samotná data nebyla
- Pokud by už data nebyla dostupná, identifikátor by měl vést na „tombstone“ - náhrobek

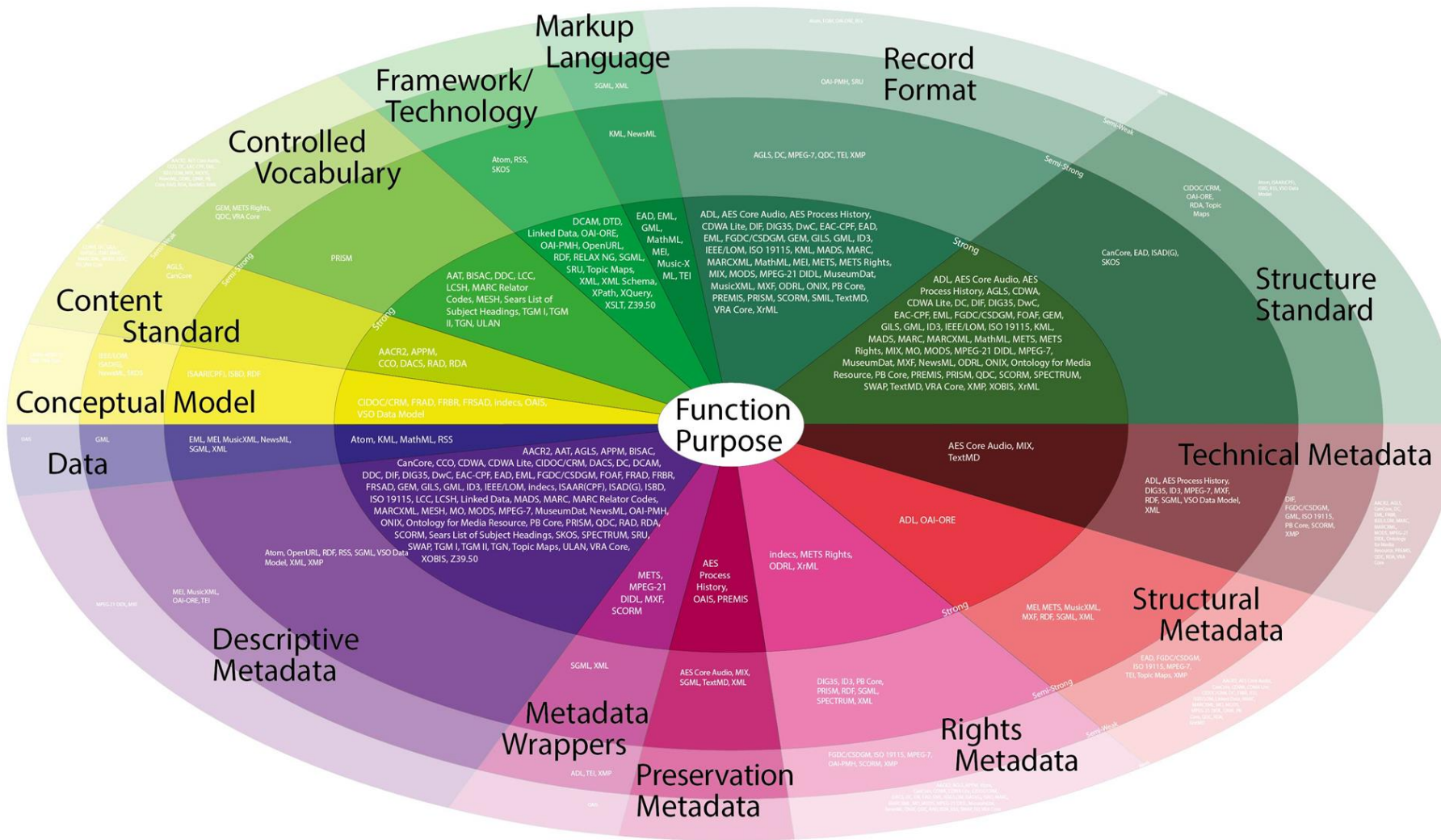


Metadata

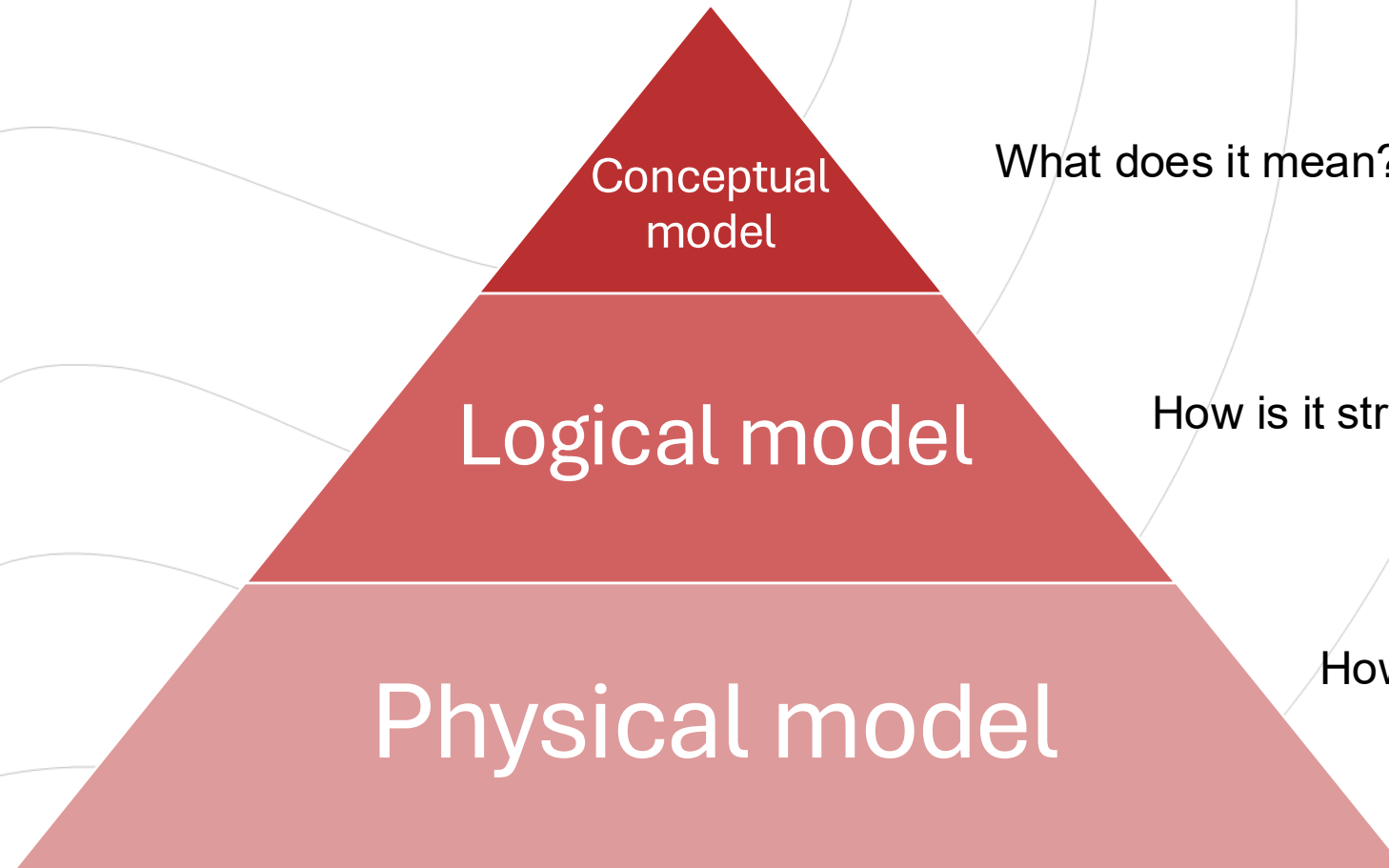
- Strojově čitelná
- Podle metadat záznamu se vyhledává
- Špatná metadata => nenalezitelná data!
- Řízené slovníky
- Užívání PID v metadatach (autor, organizace, ontologie...)
- Popisujte data takovými termíny, které byste zadali do vyhledávače, kdybyste je sami chtěli najít



Nejsou metadata jako metadata



(Meta)data modelling



Conceptual
model

What does it mean?

UML diagram, classes,
properties, vocabularies

Logical model

How is it structured?

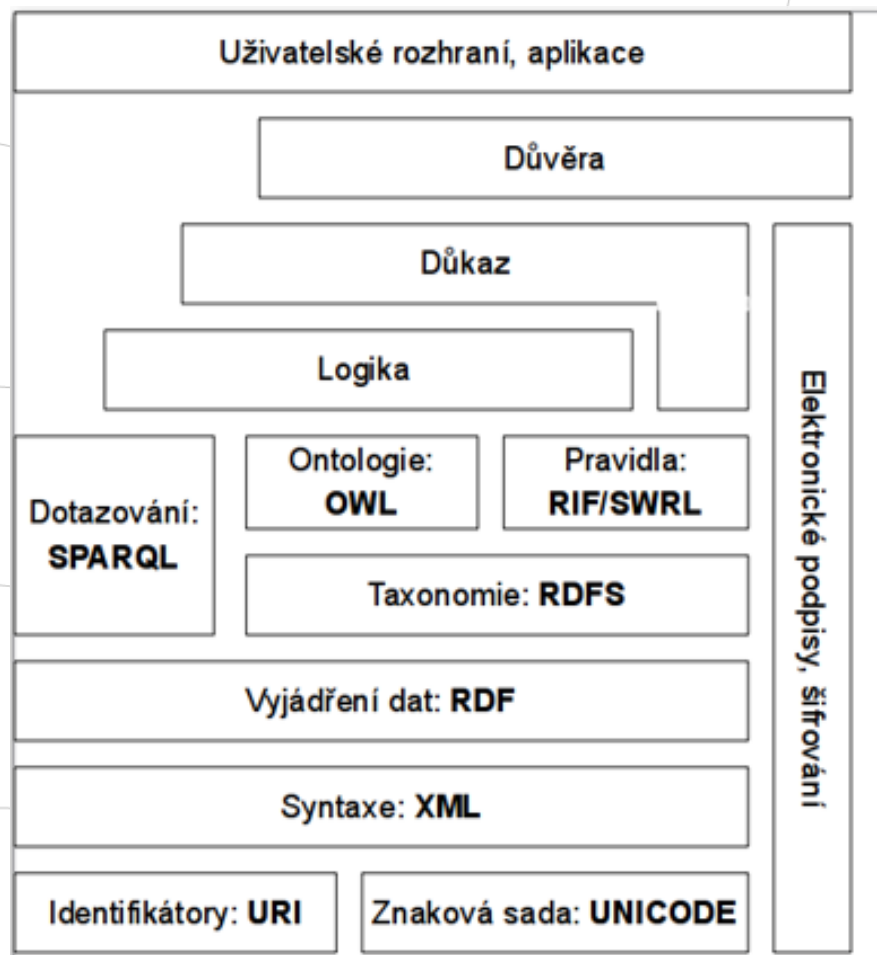
XML Schema,
RDF/OWL, JSON
schema, logical ER

Physical model

How is it implemented?

SQL table, Turtle, XML
file, JSON(-LD), YAML

Sémantický web a Linked Data



SW: Rozšíření tradičního webu o sémantiku (2001)

- Jazyky pro zachycení sémantiky: OWL, RDFS, SKOS
- Jazyk pro reprezentaci dat: RDF

LD: principy Linked Data pro strukturovaná data na webu (2006)

- Používání URI
- Dereferencovatelnost URI
- Na URI dostupná užitečná informace
- Propojitelnost URI

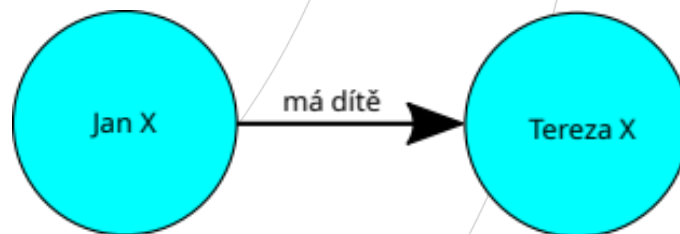
Znalostní graf (80. léta, 2012), příklady:

- Google Knowledge Graph
- DBpedia
- Wikidata
- Microsoft Academic Graph (do r. 2021)

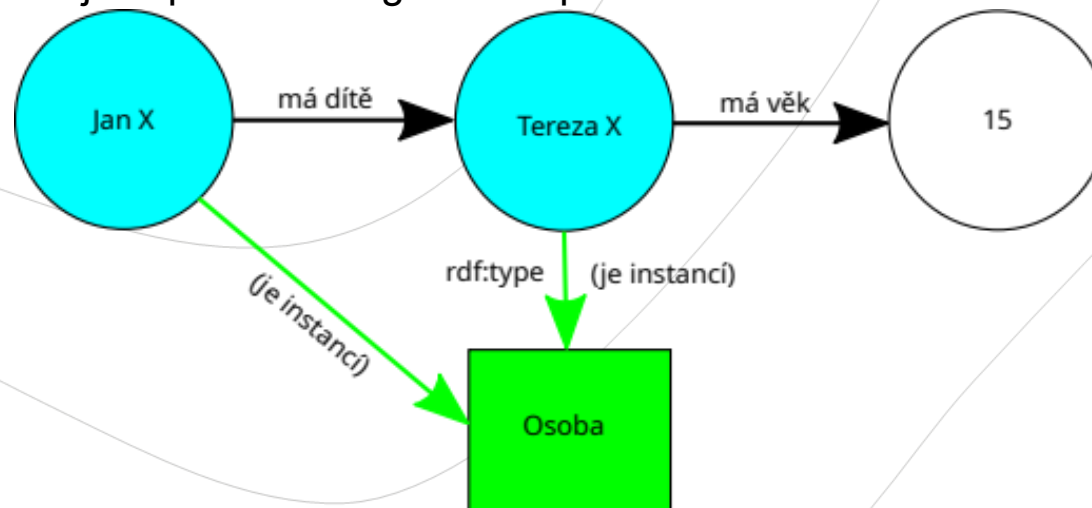
Reprezentace dat v RDF

RDF (Resource Description Framework)

- Jazyk pro reprezentování tvrzení (faktů)
- Tvrzení se skládají z trojic: subjekt-predikát-objekt



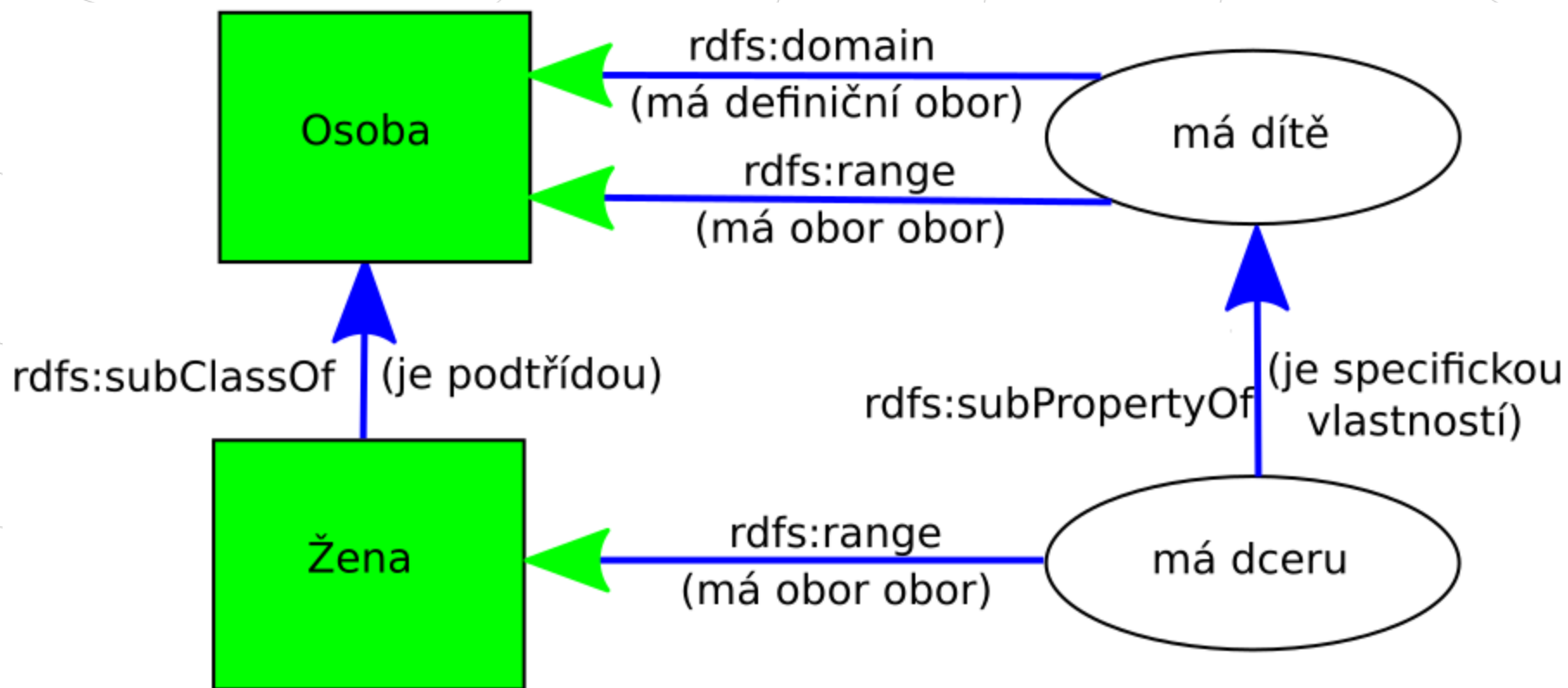
- Datová sada RDF je množina trojic s přirozenou grafovou podobou



Reprezentace (méně expresivní) ontologie

RDFS (Resource Description Framework)

- Jazyk pro zachycení významu entit
- Pomocí definování **tříd**, **vlastností**, **taxonomii tříd**, **taxonomii vlastností** a **vymezení vlastností**



OWL (Web Ontology Language)

- Expresivnější jazyk pro zachycení významu entit
- Založený na deskripční logice
 - Ontologii lze chápat jako logickou teorii obsahující množinu subsumpčních a ekvivalenčních logických axiomů
 - Subsumpce dvou pojmenovaných tříd
 - Žena $\sqsubseteq$ Osoba
 - Subsumpce dvou vlastností
 - mítDceru $\sqsubseteq$ mítDítě
 - Ekvivalence pojmenované třídy a složeného konceptového výrazu
 - BakalářskýStudent $\equiv$ Student $\sqcap \exists$ studujeObor.BakalářskýObor

- Řízený slovník = množina referenčních termínů (pojmu) pro propojené informační systémy
- Různé varianty: hesláře, tezaury, taxonomie, slovníky, ontologie
- SKOS: univerzální jazyk pro reprezentaci řízených slovníků s výhodami linked data
- skos: <<http://www.w3.org/2004/02/skos/core#>>
- SKOS hlavní prvky:
 - skos:Concept, skos:ConceptScheme
 - skos:prefLabel, skos:altLabel, skos:hiddenLabel
 - skos:notation, skos:note, skos:inScheme
 - skos:broader, skos:narrower
 - skos:broadMatch, skos:narrowMatch, skos:exactMatch, skos:closeMatch

The Linked Open Data Cloud

Legend

Cross Domain

Geography

Government

Life Sciences

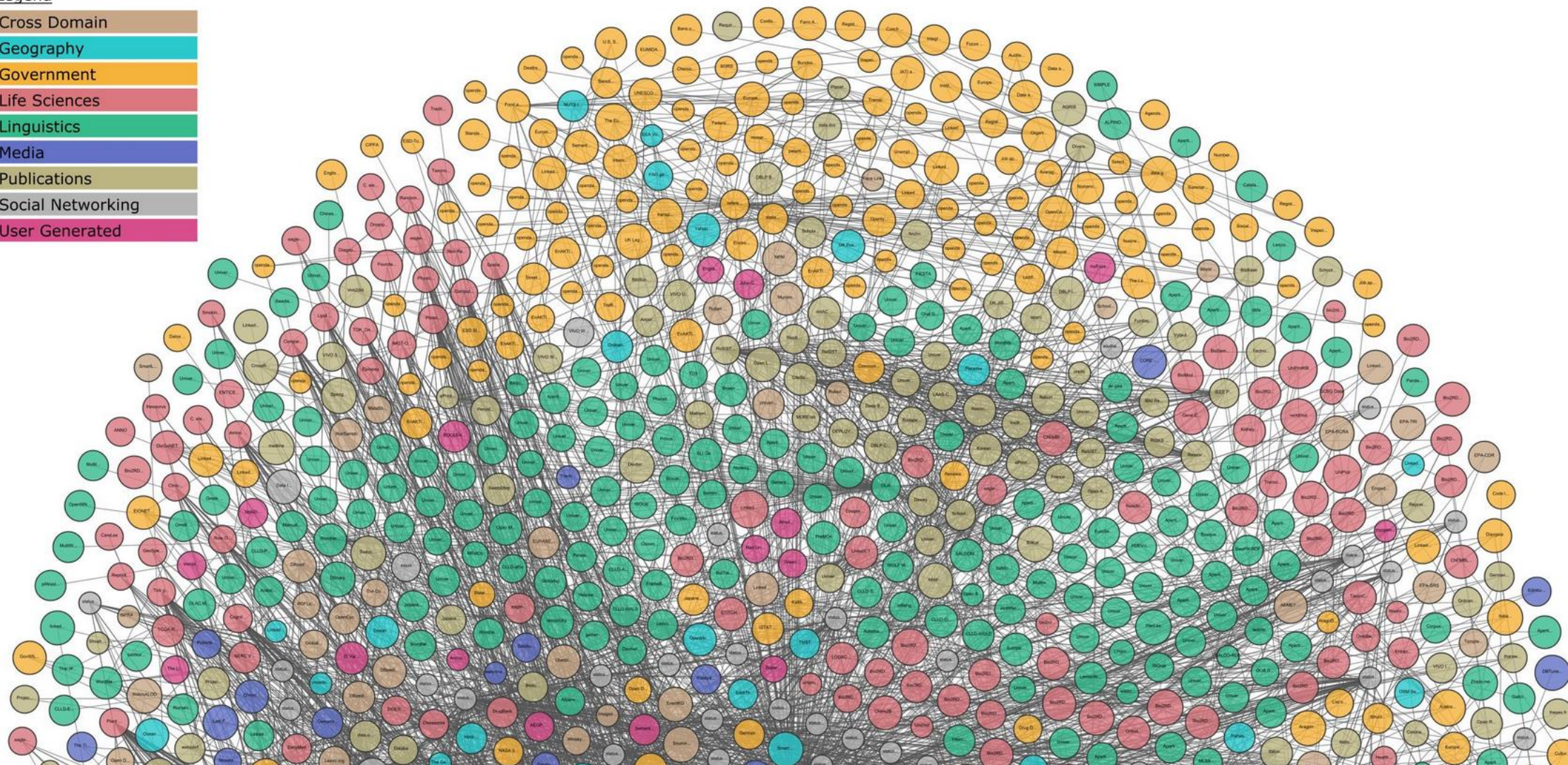
Linguistics

Media

Publications

Social Networking

User Generated



Řízený slovník a SKOS (příklad)

PREFEROVANÝ TERMÍN	Knihovnictví 
NADŘÁZENÝ KONCEPT	Média a komunikace
V JINÝCH JAZYCÍCH	Library science
URI	https://vocabs.ccmm.cz/registry/codelist/SubjectCategory/50000/50800 
STÁHNOUT TENTO KONCEPT:	RDF/XML TURTLE JSON-LD

```
<https://vocabs.ccmm.cz/registry/codelist/SubjectCategory/50000/50800>
```

```
  skos:prefLabel "Media and communications"@en,  
  "Média a komunikace"@cs ;
```

```
  a skos:Concept ;
```

```
  skos:narrower
```

```
<https://vocabs.ccmm.cz/registry/codelist/SubjectCategory/50000/50800/50804> .
```

```
<https://vocabs.ccmm.cz/registry/codelist/SubjectCategory/50000/50800/50804>
```

```
  skos:prefLabel "Knihovnictví"@cs, "Library  
  science"@en ;
```

```
  skos:broader
```

```
<https://vocabs.ccmm.cz/registry/codelist/SubjectCategory/50000/50800> ;
```

```
  a skos:Concept .
```

Řízené slovníky, ontologie - zdroje

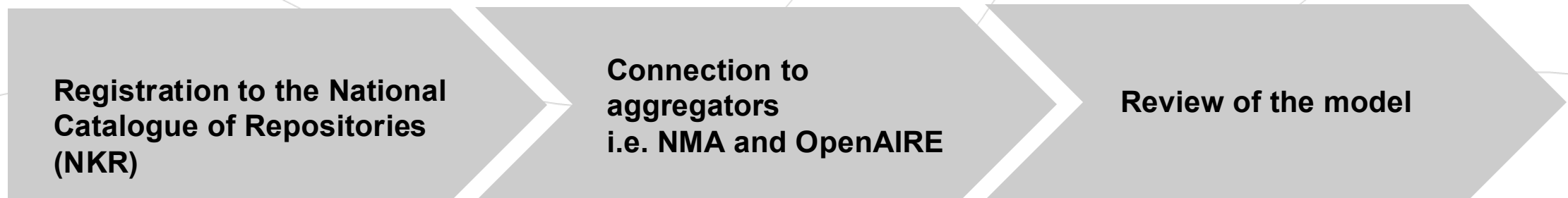
- **ISO** – jazyky, země, čas, datum
- **COAR** – řízené slovníky pro repozitáře
- **FairSharing** – oborové ontologie, řízené slovníky
- **EU Vocabularies** – evropské slovníky v různých jazycích a překladech – užíváno především státní správou
- **OSF** – Otevřené formální normy; státní správa a W3C asi

Roadmap of metadata

About to build repository:



Next:



Czech Core Metadata Model for Research Data (CCMM)

CCMM - aplikační profil pro česká výzkumná data

- V rámci českého výzkumného prostředí se pro strukturování metadat používá CCMM model jako základní model pro popis výzkumných dat
- CCMM = Czech Core Metadata Model for Research Data
- CCMM je vytvořené jako aplikační profil, kde se přepoužívají existující entity (třídy a vlastnosti) s možností určení kardinalit vlastností a určením řízených slovníků pro použití
- CCMM schéma vychází ze známých metadatových modelů
 - **DCAT** (Data Catalog Vocabulary) - schéma pro popis datových sad a katalogů
 - **DataCite** - schéma zaměřené pro identifikaci výzkumných dat a spojené s přidělováním trvalých identifikátorů DOI
 - **DublinCore** (DCMI) - univerzální soubor jednoduchých metadatových prvků



Dále zahrnuje existující doménové slovníky sémantického webu

- FOAF
 - slovník pro popis lidí, jejich zájmů, aktivit a vztahy k jiným lidem (např. foaf:Agent, foaf:Person, foaf:name, foaf:knows)
- ISA Programme Location Core Vocabulary
 - základní slovník pro popis míst
- GeoSPARQL
 - slovník pro reprezentaci geografických dat (např. geo:Feature, geo:Geometry) a geografické rozšíření jazyka SPARQL (např. geof:distance)
- vcard
 - slovník pro kontaktní údaje
- SPDX 2.3
 - Slovník pro popis SW (např. kontrolní součet, použité algoritmy)
- SKOS
 - Slovník pro reprezentaci schémat typu tezaury, číselníky, řízené slovníky
- OWL-Time
 - Slovník pro reprezentaci časových aspektů entit

Metadata identification

Data identification

Distribution
info

Data quality

Descriptive
properties

Agent
relationships

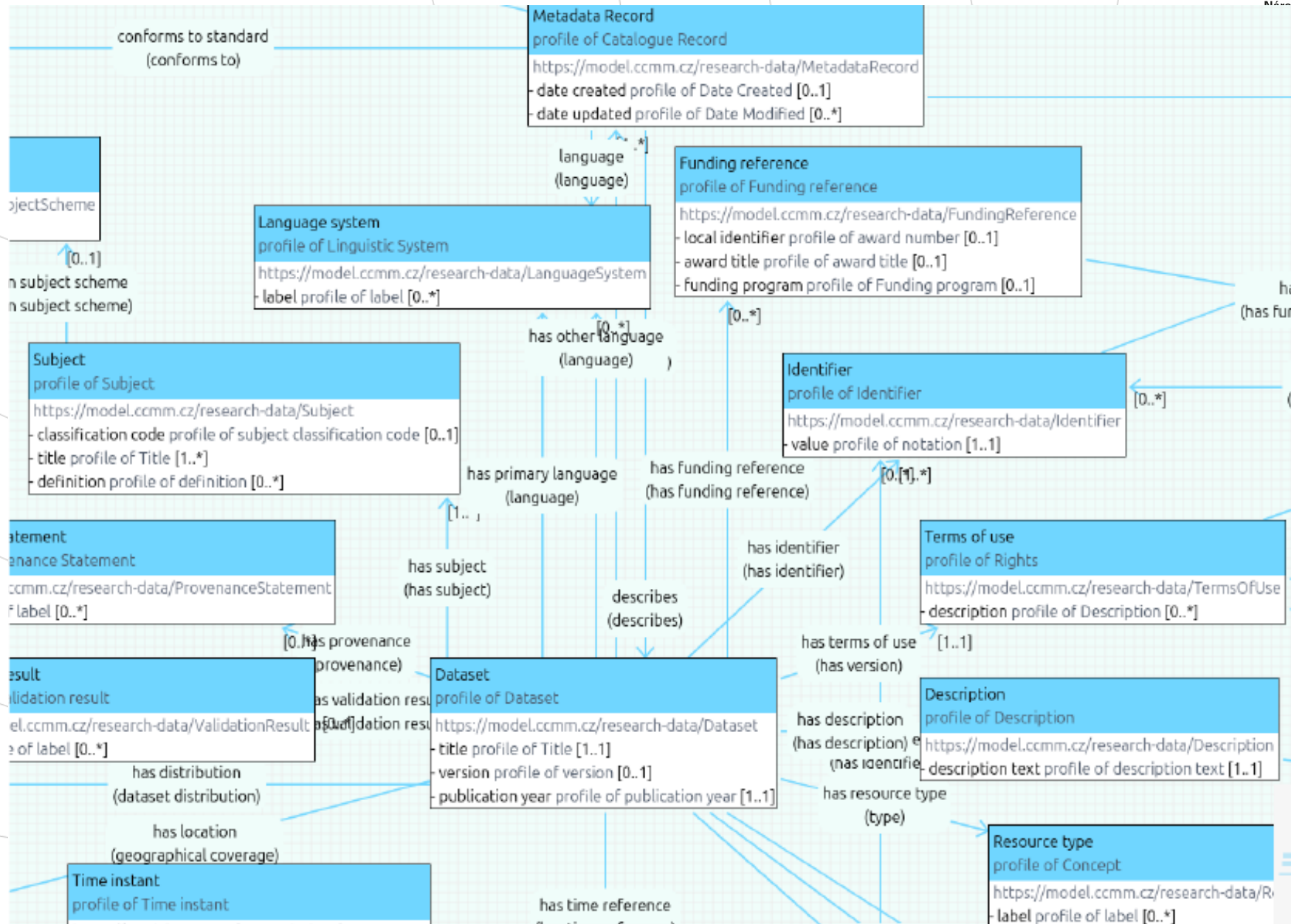
Administrative
properties

Distribution

Validation

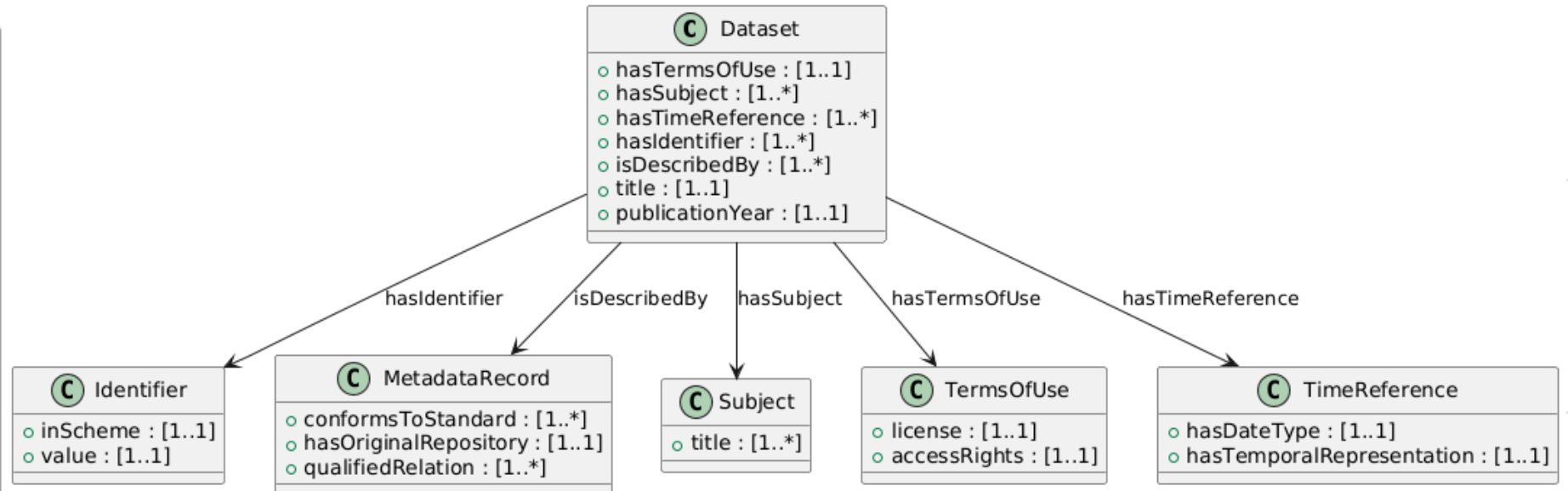
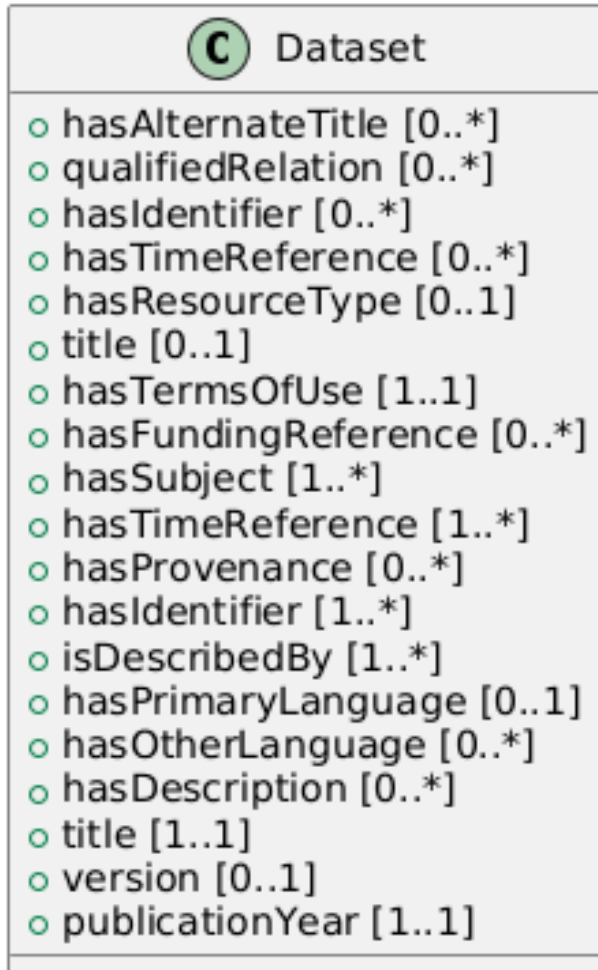
Provenance

CCMM – konceptuální model



Profiled class IRI(s)	foaf:Agent
IRI	https://model.ccmm.cz/research-data/Agent
Label	Agent
Definition	Any entity carrying out actions with respect to the entities Catalogue and the Catalogued Resources.
Hierarchy	▶ profiles class profile Agent (dcat-ap:Agent) Definition: <i>Any entity carrying out actions with respect to the entities Catalogue and the Catalogued Resources.</i> ▶ profiles class profile Agent (dcat-dap:Agent) ▶ profiles class Agent (foaf:Agent) Definition: <i>An agent (eg. person, group, software or physical artifact).</i>

CCMM – hlavní třída Dataset a její povinné části



Metadata podle CCMM (RDF v Turtle vs. XML)

```
v:qualifiedRelation [  
    dcterms:relation <https://orcid.org/0000-0003-  
0328-0725> ;  
    dcat:hadRole  
<https://vocabs.ccmm.cz/registry/codelist/AgentRole/  
Creator> ;  
];
```

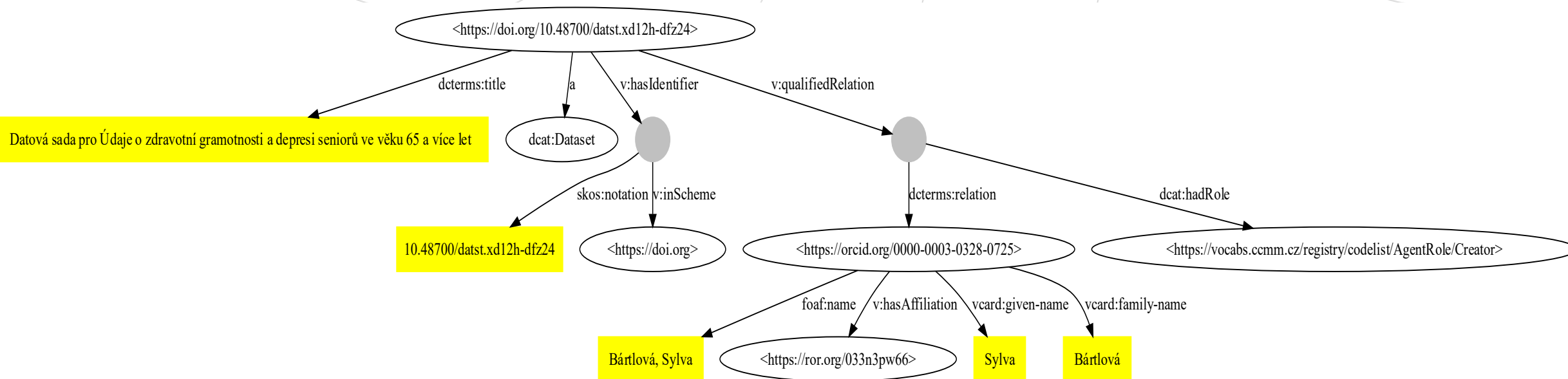
```
<https://orcid.org/0000-0003-0328-0725>  
    vcard:family-name "Bártlová" ;  
    vcard:given-name "Sylva" ;  
    foaf:name "Bártlová, Sylva" ;  
    v:hasAffiliation <https://ror.org/033n3pw66> .
```

```
<qualified_relation>  
  <relation>  
    <person>  
      <name>Bártlová, Sylva</name>  
      <given_name>Sylva</given_name>  
      <family_name>Bártlová</family_name>  
    </person>  
  </relation>  
  <role>  
    <iri>https://vocabs.ccmm.cz/registry/codelist/Agent  
Role/Creator</iri>  
  </role>
```

Metadata podle CCMM (lidé ad.)

<https://doi.org/10.48700/datst.xd12h-dfz24>

<https://orcid.org/0000-0003-0328-0725>

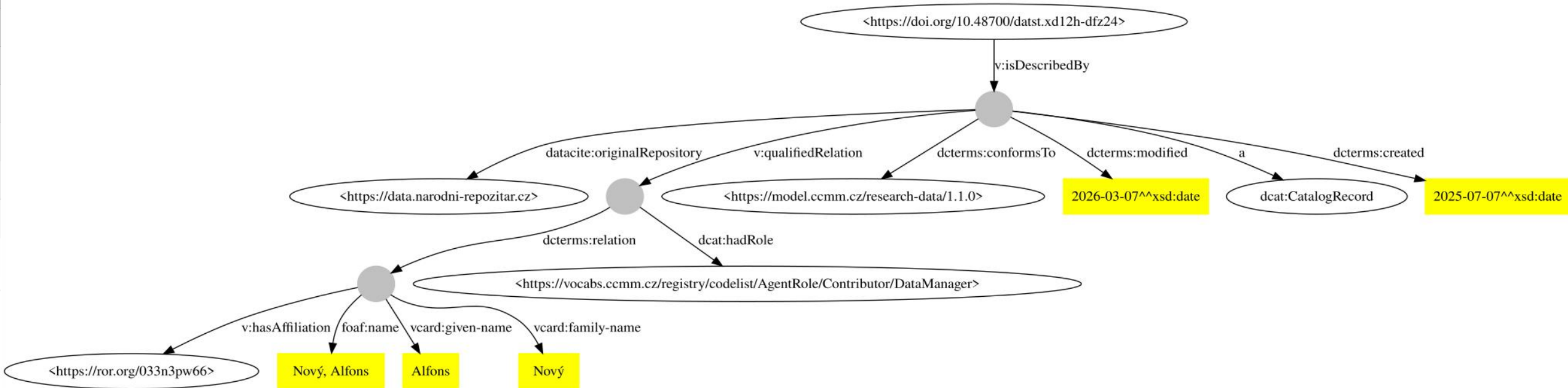


<https://vocabs.ccmm.cz/registry/codelist/AgentRole/Creator>

Metadata podle CCMM (data o metadatach)

<https://doi.org/10.48700/datst.xd12h-dfz24>

<https://ror.org/033n3pw66>



<https://vocabs.ccmm.cz/registry/codelist/AgentRole/Contributor/DataManager>

CCMM a používané řízené slovníky

Název slovníku / Odkaz	Užití slovníku	Hlavní zdroj	Spravuje
Agent Role	určení role	DataCite	NTK
Alternate Title	určení typu dalšího názvu datasetu	DataCite	NTK
Description Type	určení typu dalšího názvu datasetu	DataCite	NTK
Location Relation	určení vztahu mezi datasetem a geografickou lokací		NTK
Relation Type	určení vztahu mezi datasetem a dalším souvisejícím objektem	DataCite	NTK
Subject Category	určení oborové kategorie, do které dataset obsahově spadá	OECD FORD Frascati	NTK
Time Reference	určení typu datace	DataCite	NTK
Resource types	určení typu digitálního objektu	COAR	COAR
Language	určení jazyku digitálního objektu	Evropská unie	Evropská unie
Format	určení formátu souboru	Evropská unie	Evropská unie
Media type	určení typu média, dříve známé jako MIME	IANA	IANA
Access rights	určení typu přístupu k datasetu	COAR	COAR

Metadatová podpora při NTK

- Konzultace pro práci s metadaty výzkumných dat a jejich interoperabilitou
- <https://model.ccmm.cz/research-data/en/>
- <https://github.com/techlib/CCMM/>



Základní model CCMM ▾ Metadatové profily ▾ O nás ▾ Novinky

CCMM (Czech Core Metadata Model for Research Data) je základní metadatový model pro popis výzkumných dat v prostředí České republiky. Cílem CCMM je stanovit obecné prvky pro popis datových sad, které budou použitelné v různých datových repozitářích, a zajistit interoperabilitu základních metadat napříč obory.

Model CCMM vytváří a udržuje metadatové pracoviště Národní technické knihovny v rámci projektu CARDS a poskytuje také podporu k práci s metadaty výzkumných dat.

Rychlé odkazy

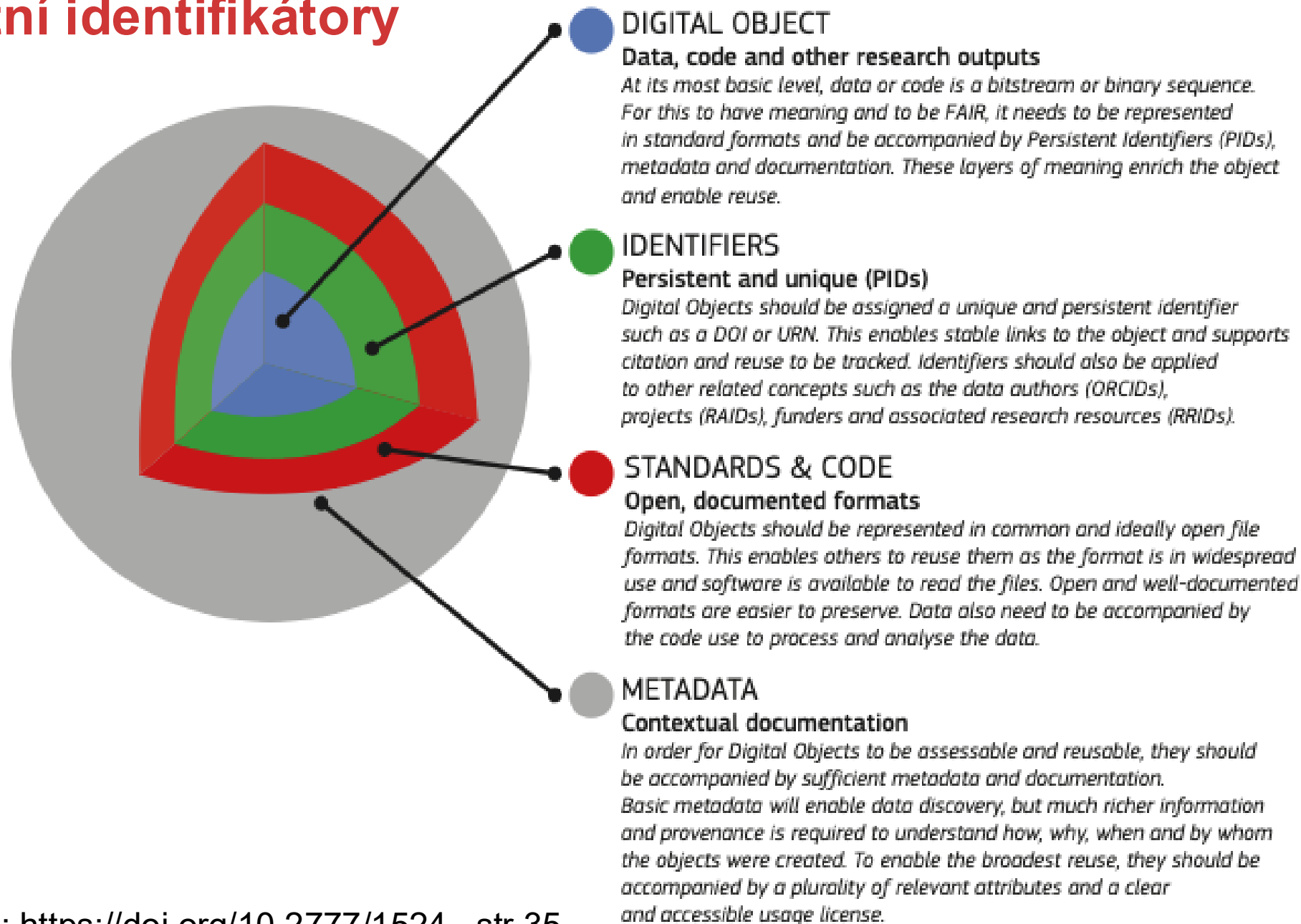
Základní model
CCMM

Řízené slovníky

Tvorba a správa
metadatových
profilů

FAQ

Perzistentní identifikátory



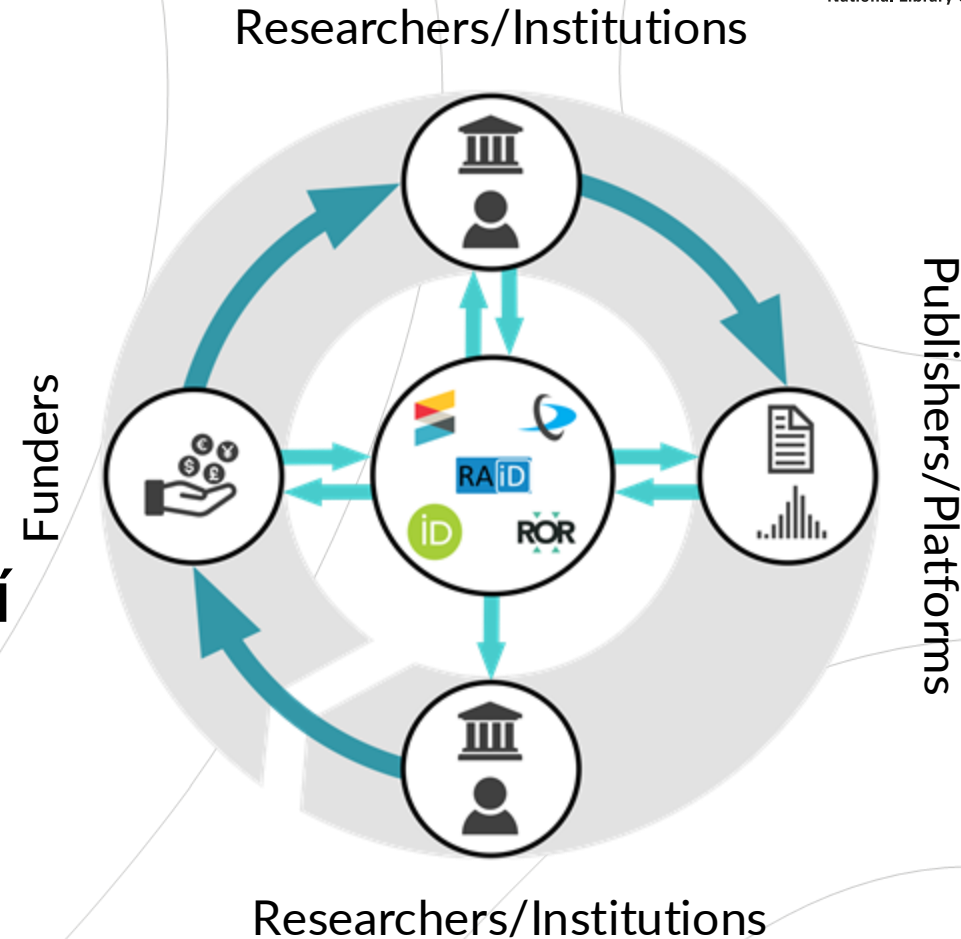
„A PID is a **digital** identifier that is **globally unique**, **persistent**, **machine resolvable**, has an associated **metadata** schema, **identifies** an entity, and is frequently used to disambiguate between entities.

PIDs are **long-lasting**, **managed**, and **registered unique** digital references (often in the form of a URL) to an object that can be represented or described online.”

Proč jsou PID důležité?

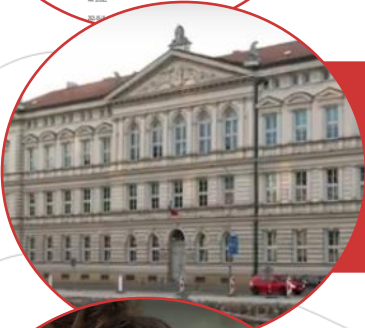
PIDS...

- ...jsou unikátní
- ...se nemění v čase
- ...jsou vzájemně propojené
- ...jsou strojově čitelné
- ...šetří čas při ukládání i vyhledávání informací
- ...jsou součástí řady doporučení a politik jakou FAIR principy





Digitální objekty: články, datasets, software, audio-vizuální záznamy aj.



Abstraktní objekty: organizace, projekty, ocenění, konference aj.



Fyzické objekty: osoby, vzorky, instrumenty aj.

Existuje mnoho typů PID...

Pro výsledky/výstupy
výzkumu



Handles



Pro výzkumníky a
další přispěvatele



Pro výzkumné
organizace



Doporučené PID v NDI

NTK

50°6'14,083"N, 14°23'26,365"E
Národní technická knihovna
National Library of Technology

For **objects / outputs**



For **researchers** and
other contributors



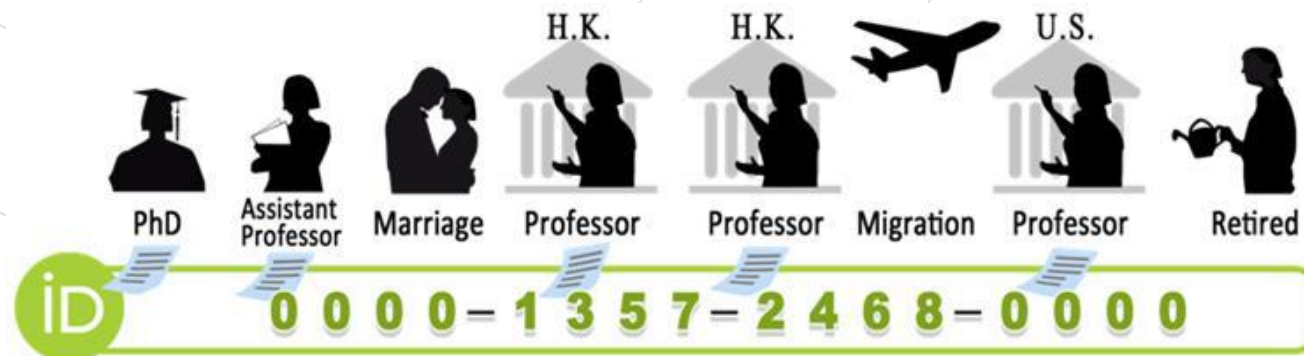
For **research
organizations**





ORCID iD – základní principy

- ✓ Máte vždy **plnou kontrolu** nad svým profilem
- ✓ Můžete **nastavovat viditelnost** pro každou položku zvlášť
- ✓ Vy udělujete a spravujete **oprávnění organizacím a jednotlivcům práva na zápis i čtení údajů** na Vašem profilu
- ✓ **ORCID Privacy Policy:** <https://info.orcid.org/cs/privacy-policy/>



ORCID iD – základní principy

Setting the visibility of each item



Everyone



Trusted parties



Only me

Trust
markers



✓ Employment (2)

National Library of Technology: Prague, CZ

2024-04 to present | ORCID Consortium Coordinator
Employment

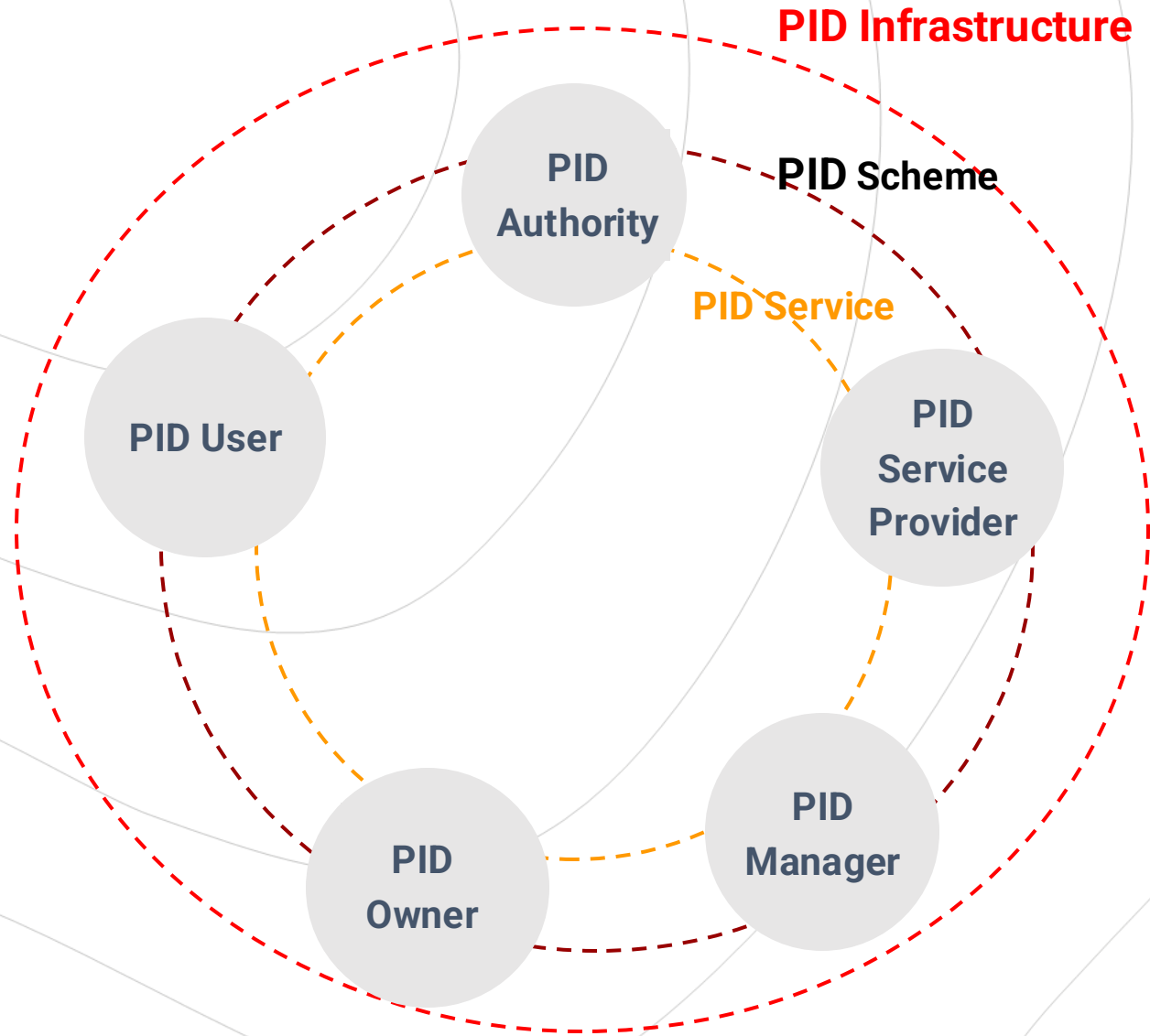
Source: ✓ Czech National Library of Technology via ORCID Member Portal

Biology Centre of the Czech Academy of Sciences: České Budějovice, CZ

2016-06-01 to 2024-02-29
Employment

Source: 👤 Hana Šilha Machová

Části PID





ve spolupráci s
Národní
technickou
knihovnou



Administrativní, metodická a finanční podpora
pro zavádění perzistentních identifikátorů (PIDs)
kompatibilních s evropskými a mezinárodními
standards

- **Národní centrum ISSN**
- **Národní centrum ORCID**
- **Národní centrum DOI**
- Činnost financována z projektu **CARDS** (2023-28)
- Součástí Centra repozitářů a správy metadat v NTK

Centrální registr určený pro evidenci a správu údajů o českých repozitářích výzkumných dat.

Cílem je poskytovat důvěryhodné, strojově čitelné a dlouhodobě udržitelné informace o tom, jaké repozitáře v České republice existují, kdo je provozuje, jaké metadatové standardy a ontologie používají a zajišťovat napojení na externí služby.



Práce s metadaty dle rolí

Výzkumník



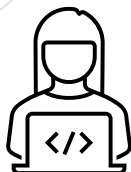
- **Hlavní cíl:** rychle najít, vytvořit a použít data pro vlastní vědecký výzkum s pouze nezbytnými metadaty, která umožní splnění tohoto cíle.
- **Primárním nástrojem** jsou metadata exportovaná/generovaná z přístrojů a základní popisná v readme.

Data steward



- **Hlavní cíl:** zajistit, aby data a metadata byla správná, konzistentní, dobře popsána v rámci specifické komunity, které je součástí (obor, projekt, instituce) a řídila se zavedenými standardy oboru, projektu/grantu a instituce.
- **Primárním nástrojem** je metodická podpora skupině výzkumníků/instituci, jak si procesy správně optimalizovat, aby fairifikace nepředstavovala další zátěž a data z výzkumu se neztrácela.

Kurátor repozitáře



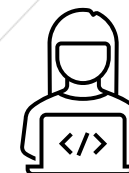
- **Hlavní cíl:** zajistit, aby data byla dostatečně a správně popsána bohatými metadaty, aby mohla být data bezpečně uložena, dlouhodobě dostupná, vyhledatelná pro budoucí uživatele.
- **Primárním nástrojem** je implementovaný metadatový model s řízenými slovníky, napojením na externí zdroje (PID, dtb) a nastavenými validacemi



Produkuje data



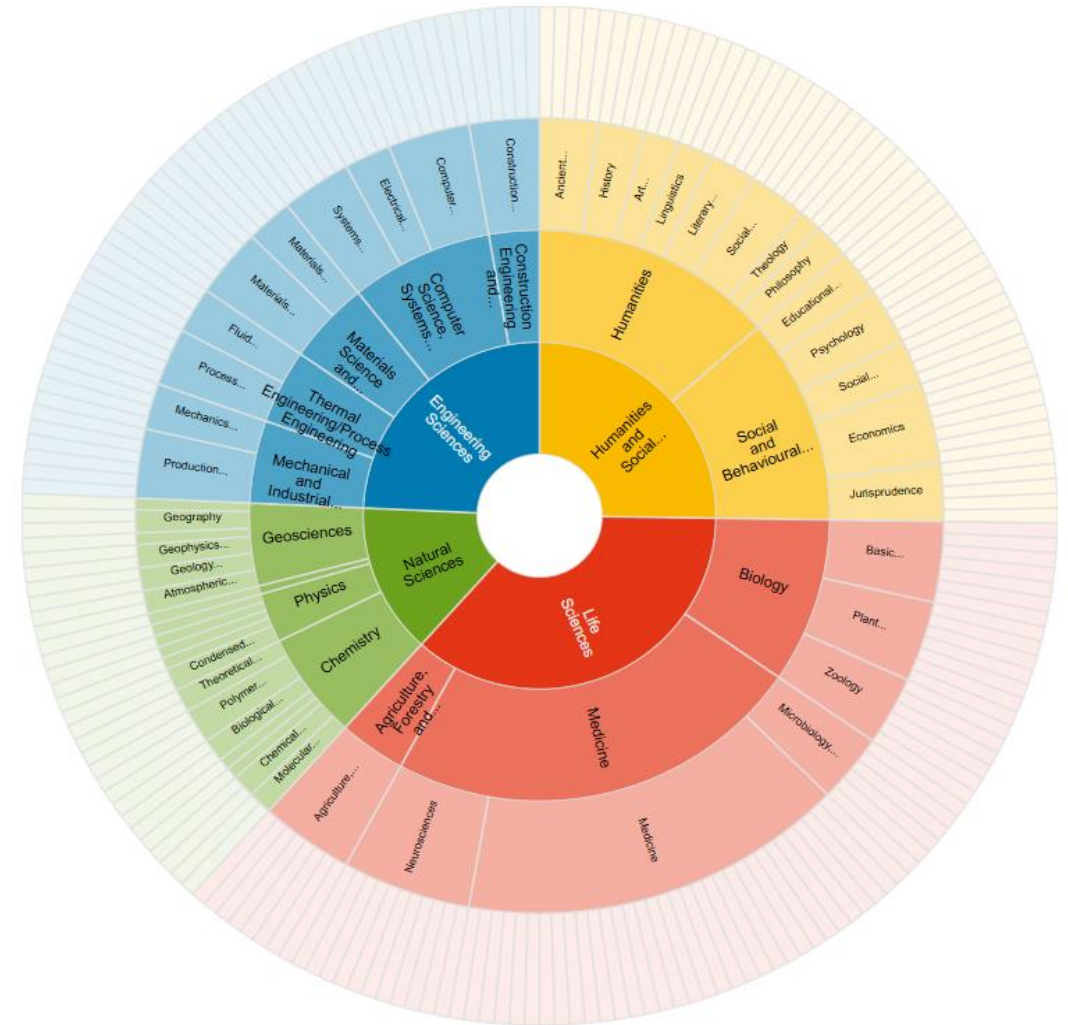
Metodická podpora,
kontrola pravidel



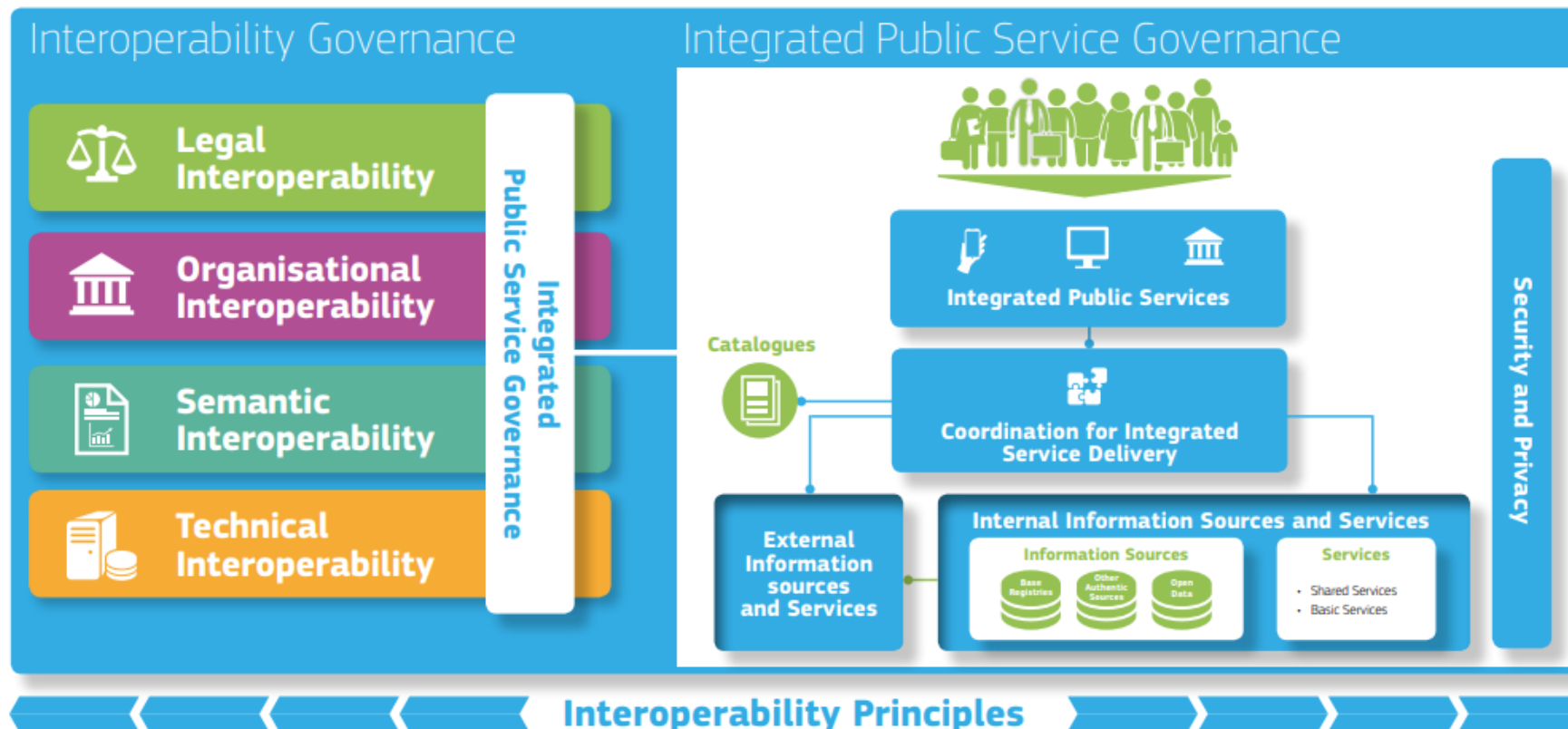
Revize požadavku,
hlídání interoperability

Keep in mind

- Kdo je vaše cílová skupina/komunita
- Jaké disciplíny a datové typy pokrývá
- Identifikuj její potřeby pro práci s daty a jejich popisem
- Identifikuj, co svět nabízí
- Vyber, co je pro komunitu nejoptimálnější a zároveň naplňuje principy interoperability a FAIR



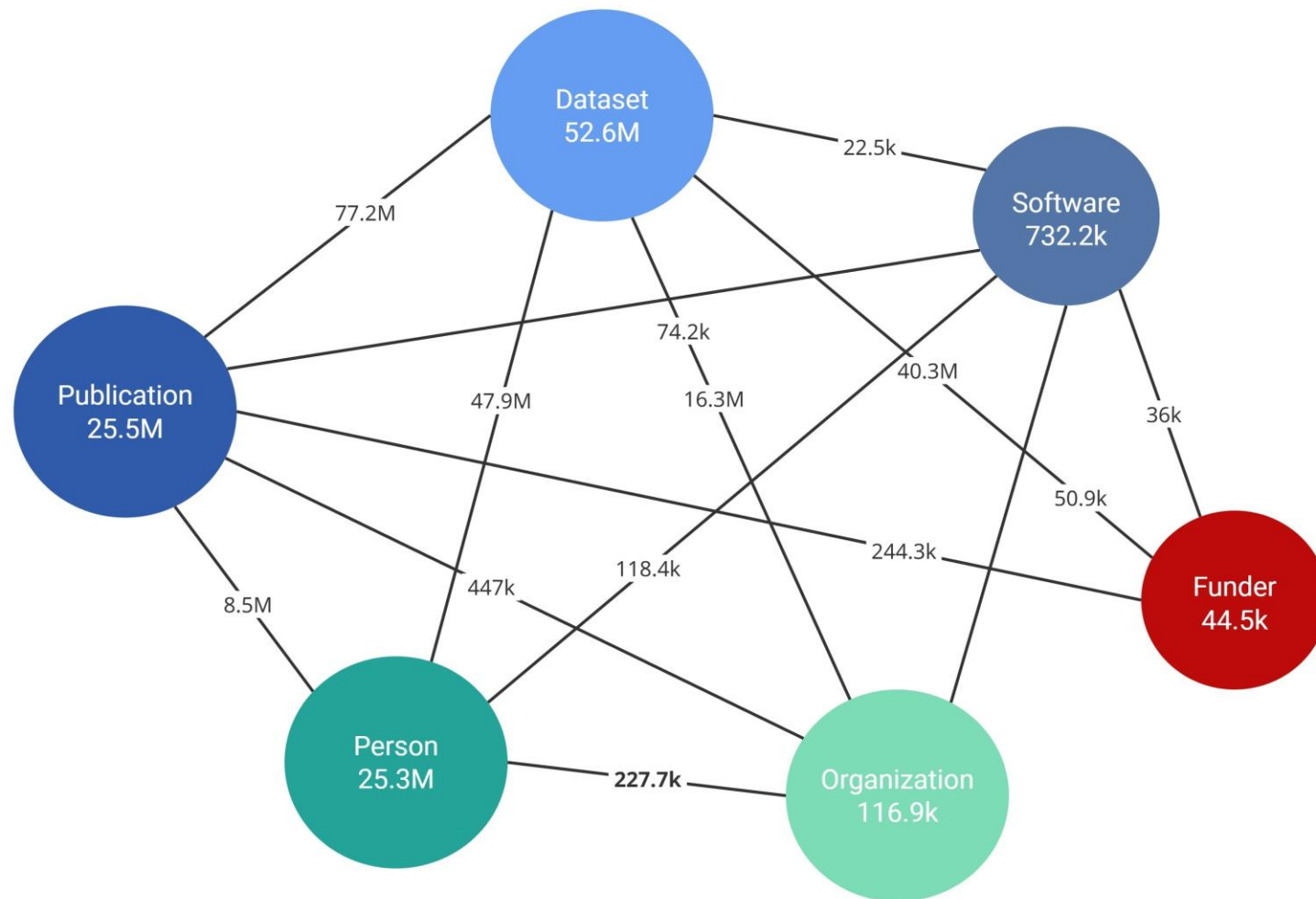
EIF Conceptual Model



Děkujeme za pozornost

Research is a graph...

**...connected through the
power of metadata and PID**



web: ccmm.cz; metadata@techlib.cz

web: identifikatory.cz; identifikatory@techlib.cz
nkr@techlib.cz