Jihočeská univerzita v Českých Budějovicích
Přírodovědecká fakulta

Otevřená data v oblasti vědeckého výzkumu

Diplomová práce

Bc. Ondřej Doktor, DiS.

Školitel: PhDr. Miloš Prokýšek, Ph.D.

České Budějovice 2020

Otevřená data v oblasti vědeckého výzkumu [Open data in scientific research. Mgr. Thesis, In Czech.] – 128 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Abstract

This thesis focuses on the topic of open data in academic and research sectors and the openness, sustainability, and reusability of data which are being processed in diverse scientific communities. Since this topic is very broad, the thesis narrows its focus down to the author's home institution, the Faculty of Science, University of South Bohemia, Czech Republic, and its data sources in respect to the FAIR data principles.

The main questions being discussed are: Is it possible to convert the current data foundation of research groups of the Faculty of Science of the University of South Bohemia into a FAIR form and how demanding would this transformation be? What data foundations now exist at the institution and in what form? What needs to be done to consider these data foundations as FAIR? What prevents the transformation of these data foundations into FAIR form? How is it possible to carry out this transformation in practice?

Prohlášení

Prohlašuji, že svoji diplomovou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 7. 12. 2020

Bc. Ondřej Doktor, DiS.

Poděkování

Rád bych na tomto místě poděkoval zejména PhDr. Miloši Prokýškovi, Ph.D., svému školiteli, za veškerou odbornou podporu a pomoc, dále Ing. Rudolfovi Vohnoutovi, Ph.D., vedoucímu Ústavu aplikované informatiky PřF JU a delegátovi reprezentujícího JU v konsorciu ELIXIR_CZ, za zprostředkování cenných podkladů a kontaktů, doc. RNDr. Milanovi Předotovi, Ph.D., proděkanovi PřF JU pro vědu a výzkum, za záštitu a odbornou podporu při zpracování dotazníkového šetření na instituci, a také všem vědeckým pracovníkům, respondentům tohoto šetření, za jejich vstřícný, kolegiální a veskrze pozitivní přístup a za jejich otevřenost novým myšlenkám. V neposlední řadě děkuji též své rodině a svým nejbližším za veškerou jejich podporu během celého studia.

Obsah

1 Úvod

Tato diplomová práce se zabývá problematikou otevřených dat (OpenData) v akademické a vědecko-výzkumné sféře. Především pak otevřeností, udržitelností a znovupoužitelností zpracovávaných dat v různorodých vědeckých komunitách.

Jak vyplývá z některých studií, publikovaná data a výsledky vědeckého výzkumu svou formou a kvalitou často neumožňují reprodukování procesu jejich získání, nebo jejich využití pro další zpracování [1]. Spolu s neustále rostoucím objemem zpracovávaných dat [2] a nástupu výzkumných metod, spočívajících ve velkoobjemovém zpracování dat napříč mnoha heterogenními zdroji, se tyto problémy stále intenzivněji dostávají na povrch a sílí vůle k jejich nápravě [3]. Navíc lze s vysokou mírou jistoty předpokládat, že financování budoucích výzkumných projektů bude podmíněno existencí dobrého data management plánu. Tyto tendence se již staly realitou například v evropském prostředí při financování vývoje infrastrukturních vědeckých nástrojů – viz stanovisko konsorcia ELIXIR z roku 2017 [4]. Je tedy v zájmu výzkumných skupin tento problém řešit, a to nejen z důvodu financování, ale i budoucího vědeckého pokroku.

Je také třeba zdůraznit, že principy otevřených dat se nutně uplatní také v případech, kdy nedochází ke zpřístupnění dat široké veřejnosti, ale např. jen malé vědecké komunitě. I v takovém případě je nutné zajistit, aby byla data interoperabilní, přičemž zásady podmiňující interoperabilitu jsou s těmi, které podmiňují otevřenost dat v principu shodné. Proto považujme pro účely této práce za „otevřená data" taková data, která je možné sdílet bez ohledu na rámec, ve kterém se tak děje.

Jelikož problematika otevřených dat či FAIR dat je velmi široká, zaměřuje se tato práce specificky na konkrétní akademickou instituci, Přírodovědeckou fakultu Jihočeské univerzity v Českých Budějovicích. V praktickém části pak na její vybrané výzkumné skupiny a zvolený reprezentativní datový zdroj nálezových dat.

V kontextu výše uvedených předpokladů a formulovaného problému, lze hlavní výzkumný problém práce transformovat do otázky:

Lze převést současnou datovou základnu výzkumných skupin PřF JU do FAIR podoby a jak náročná by tato transformace byla?

Tuto základní otázku je možné dále rozvinout do systému dílčích otázek, a to: Jaké datové zdroje nyní na instituci existují a v jaké podobě? Co je potřeba udělat, aby bylo možné datové

zdroje považovat za FAIR? Co brání transformaci datových zdrojů do FAIR? Jakým způsobem je možné transformaci prakticky provést?

Zodpovězení položených výzkumných otázek je možné dosáhnout realizací soustavy dílčích cílů práce a souvisejících úkolů:

1. Cíl: Identifikovat možné přístupy pro sdílení dat ve zkoumané oblasti
    a) Provést rešerši přístupů a metodik
    b) Porovnat a sestavit doporučení
2. Cíl: Analyzovat stávající institučí stav
    a) Vybrat a aplikovat vhodnou metodiku hodnocení
    b) Provést analýzu stávající datové základny
    c) Identifikovat potenciální překážky bránící zlepšení
3. Cíl: Navrhnout technické řešení (proof-of-concept)
    a) Definovat funkční a nefunkční požadavky
    b) Vyřešit správný způsob strojově čitelného zápisu sémantiky dat
    c) Implementovat a nasadit řešení
4. Cíl: Demonstrovat aplikaci metodiky a technického řešení na reprezentativní datové sadě
    a) Získat a analyzovat datovou sadu
    b) Aplikovat změny dle sestavených doporučení a metodiky
    c) Nasadit do navrženého technické řešení
    d) Certifikovat výsledek a navržené řešení pomocí relevantních validátorů

2 Teoretická část

2.1 Open a FAIR data

Z obecného pohledu je základem otevřených dat jejich strojová zpracovatelnost, tedy podpora pro tzv. *machine-to-machine* komunikaci (M2M). V prostředí obchodu a e-governmentu je tato vlastnost v současnosti zajišťována technologiemi obecně sdruženými pod pojmy *Linked-Data*, příp. *Linked Open Data* (dále v textu také jako „LOD"), jejímž smyslem je přidat nad hypertext, který tvoří strukturu webových stránek, další sémantickou vrstvu, která dodá informacím strojově čitelný význam.[5]

Do oblasti vědeckého výzkumu se obdobné principy začaly promítat přirozeně také, avšak s nižší razancí. Je vhodné zmínit zejména doporučení OECD z roku 2007 [8]. Ucelené a systematické uchopení problému se však začalo dít až poměrně nedávno. Diskuse ohledně aplikace těchto doporučení na tzv. Lorentzské konferenci v roce 2014 dala vzniknout propracované a lépe definované sadě doporučení, později (2016) formalizované s pomocí vědecké koalice FORCE11 jako **FAIR Data Principles** [9], které se nyní dostává značné popularity a uznání v široké vědecké komunitě [3]. FAIR je zkratkou pro čtyři skupiny obecných kritérií, která musí data a nástroje splnit, aby byla považována za tzv. „férová" a tedy kvalitní. Musí být dohledatelná (**F**indable), přístupná (**A**ccessible), vhodná ke spolupráci (**I**nteroperable) a znovupoužitelná (**R**eusable).

Podstata a náplň těchto kritérií je samozřejmě mnohem propracovanější a je podrobněji popsána dále v textu práce. FAIR data a proces implementace tzv. „férovosti" dat jsou pojmy v současnosti pro vědeckou komunitu natolik zásadní [3], že je nutné je alespoň velmi stručně popsat již v úvodu.

2.2 Metodiky hodnocení pro analýzu stávajícího stavu

Jako základní metodiku pro hodnocení otevřenosti obecně jakýchkoliv dat, je možné použít **5-star Open Data**, kterou navrhl Tim Berners-Lee, mj. tvůrce WWW a ředitel konsorcia W3C. Metodika je velmi jednoduchá a vychází z pojmů propojených a sémantického webu, které Berners-Lee nastínil ve svém původním článku z roku 2006 [6]. Metodika klasifikuje datové zdroje do pěti skupin, které pro názornost označuje počtem „hvězdiček" [7]. Tyto skupiny přehledně shrnuje Tabulka 1.

Prakticky je možné si toto představit na příkladu s tabulkou otevírací doby obchodu: Mějme na webových stránkách svého podniku (★) klasickou HTML tabulku s otevírací dobou

(★★ a ★★★), ideálně na samostatné podstránce (★★★★). Tato je významově srozumitelná pro člověka – návštěvníka naší webové stránky, není však strojově čitelná. Pokud tuto tabulku sémanticky popíšeme dle vhodného standardu (★★★★★), může informaci o otevírací době zpracovat webový vyhledávač a zobrazit ji uživatelům rovnou ve výsledcích vyhledávání, např. i při vyhledávání v mapě apod. To přinese užitek jak našim návštěvníkům, tak nám samotným v podobě rozšíření okruhu potenciálních zákazníků.

★	Data jsou zpřístupněná na webu (v libovolném formátu)
★★	Data mají strukturovaný formát (např. Excel místo naskenované tabulky)
★★★	Formát dat je neproprietární a otevřený (např. CSV navíc k Excelu)
★★★★	Všechny prvky jsou přímo adresovatelné pomocí URI (deep linking)
★★★★★	Data jsou napojena na jiná data, jsou tak zasazena do kontextu (dodržení ontologie, schématu a odkazem na něj)

Tabulka 1: Kategorie zdrojů dle metodiky 5-star Open Data

Výše uvedený příklad ilustruje, proč jsou LOD nyní poměrně široce implementována v obchodním sektoru, o čemž se čtenář může snadno přesvědčit při svém příštím googlování – princip je srozumitelný, realizace není nákladná a instantně přináší konkurenční výhodu.

Další metodikou jsou již zmíněné **FAIR Data Principles**, které jsou oproti LOD již mnohem propracovanější [9]. Podrobně se jimi zabývá kapitola 2.3 této práce. Z hlediska aktuálně diskutovaného cíle, výběru metodiky hodnocení stávající datové základny, jsou relevantní zejména související publikace, které se určování míry „férovosti" zabývají. Protože FAIR není návod, ale sada technologicky neutrálních principů, není snadné vytvořit „checklist" k odškrtnutí jednotlivých požadavků. V souvislosti s FAIR se proto používají tzv. „indikátory vyspělosti" (Maturity Indicators) definované zpočátku (2018) pro manuální vyhodnocování [10], pro něž se aktuálně (2019) vyvíjí automatizovaný framework [11]. Vývojem indikátorů vyspělosti se nyní intenzivně zabývá pracovní skupina FAIR Metrics Group, výstupy její práce jsou průběžně dostupné na GitHubu na adrese http://fairmetrics.org. Je třeba upozornit, že seznam a podoba indikátorů vyspělosti FAIR záměrně není definitivní a je třeba počítat, že se bude vyvíjet a reagovat na faktické a technologické změny v oblasti sdílení vědeckých dat, aby si její výstupy zachovaly vypovídací schopnost [11].

Oba popsané přístupy LOD a FAIR sdílí mnoho společných znaků. FAIR je obecně použitelnější, zejména díky své technologické neutralitě a myšlence, že kvalitní data nutně nemusí být otevřená pod svobodnou licencí, nicméně vyhodnocování jeho indikátorů vyspělosti je oproti LOD složitější [13]. Velmi zajímavou meta-metodiku, která spojuje výhody obou zmíněných, vyvinula australská Organizace vědeckého a průmyslového výzkumu Commonwealthu (CSIRO) v rámci projektu OzNome [14], která se zabývá především zpracováním big data ze senzorických sítí a mnoha heterogenních zdrojů. **CSIRO 5-star Data Rating** je 5-ti hvězdičkový systém hodnocení, který integruje přístupy více metodik, přičemž výsledkem je číselné skóre, a to jak celkové, tak v pěti sledovaných kategoriích:

- Findable
- Accessible
- Interoperable
- Reusable
- Trusted

Zjevná podobnost s FAIR Data Principles není náhodná – principy FAIR uznává metodika CSIRO 5-star Data Rating jako obecně validní a tvoří jejich nepřímou nadmnožinu. Nadto přidává posouzení důvěryhodnosti zdroje (Trusted), v rámci kterého sleduje i to, jak jsou data spravována, jak často jsou aktualizována a zda jsou sledovány statistiky používání.

Podrobný popis prvků metodiky CSIRO 5-star Data Rating a jejich souvztažnost k výše uvedeným FAIR Data Principles a 5-star Open Data popisuje dokumentace [15]. Níže je uveden základní přehled sledovaných kvalit, ze kterých vychází otázky, které metodika klade, se stručným komentářem (názvy ponechány v originále, komentáře přeloženy autorem, zdroj: [15]). Zkoumá se, zda data jsou:

published — chápána jako přístupná i jiným uživatelům, než je jejich tvůrce či vlastník

hosted — dostupná na webu

curated — doplněna závazkem, že data budou k dispozici v dlouhodobém horizontu

updated / maintained — tvoří výstup pravidelného sběru dat, s jasnými opatřeními pro jejich údržbu a harmonogramem jejich aktualizace

licensed — jasně licencována tak, že podmínky pro další užití jsou jednoznačně vyjádřeny

citeable — označena pomocí standardních, persistentních a veřejných identifikátorů

described — popsána a označena metadaty, která jsou vyjádřena formálně s pomocí standardu

findable — indexována ve známém systému, lhostejno zda obecném nebo komunitně zaměřeném

loadable — reprezentována pomocí běžného nebo standardního formátu (např. DOC nebo PDF se nepočítají)

useable — strukturována pomocí veřejně dostupného a v komunitě známého schématu nebo datového modelu

comprehensible — podpořena jednoznačnými definicemi obsahu všech prvků, a to pomocí odkazů na veřejně dostupné slovníky či ontologie

connected / linked — provázána s jinými daty pomocí externích identifikátorů (např. URI) a tudíž umožňující strojové procházená (tzv. crawling)

assessable — doplněna přímo nebo odkazem o hodnocení kvality a popis původu a pracovních postupů, který vedly k pořízení těchto dat

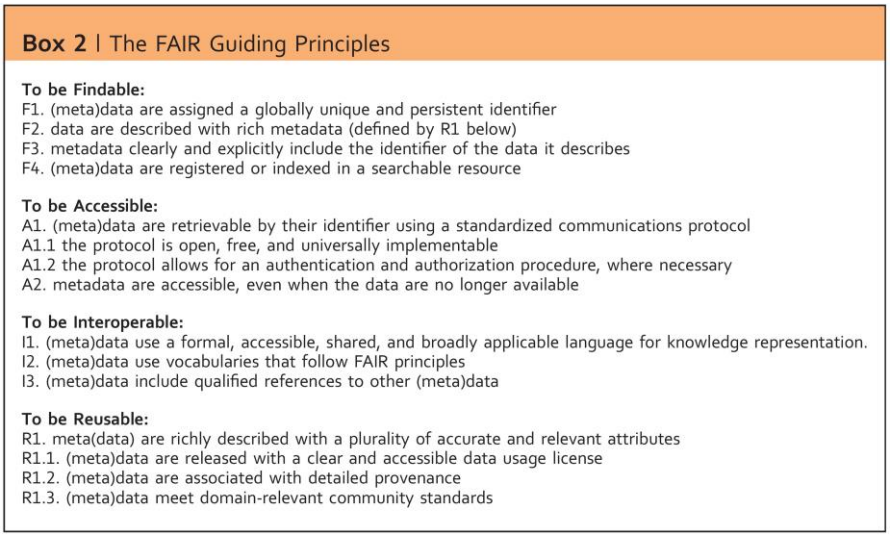
trusted — doplněna přímo nebo odkazem o informace o tom, jak jsou data používána, kým a jak často

2.3 Správné postupy pro sdílení dat – FAIR Data Principles

V době psaní této práce, tedy na přelomu let 2019 a 2020, se vzhledem k doporučení evropských institucí [3] a nadnárodních odborných skupin vědecké komunity [4] jeví jako správná cesta k dosažení požadované kvality datových sad aplikace již v úvodu zmíněných **FAIR Data Principles** [9].

Jak již bylo zmíněno, FAIR je zkratkou pro čtyři skupiny kritérií či principů, které obecně říkají, že data musí být dohledatelná (**F**indable), přístupná (**A**ccessible), vhodná ke spolupráci (**I**nteroperable) a znovupoužitelná (**R**eusable). Tato základní kritéria jsou dále rozpracována na 15 konkrétních principů, které mají vést k naplnění těchto kritérií. Jejich přehled viz Obrázek 1.

Společným jmenovatelem všech principů je zajištění podpory pro strojové zpracování. Je však důležité uvést, že samotný FAIR nestanovuje konkrétní řešení nebo technické postupy k naplnění těchto principů, aby byly tyto stále relevantní s ohledem na budoucí technologický vývoj. Vzhledem k obecnosti těchto principů je proto vždy nutné jejich význam vůči konkrétním aplikacím vykládat.



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Obrázek 1: Přehled jednotlivých principů FAIR, zdroj: [9]

Pochopení pojmů používaných FAIR principy a jejich souvztažnosti je důležitý předpoklad k realizaci praktické části této práce, proto jsou v následujících kapitolách jednotlivá kritéria a principy stručně popsány (názvy jednotlivých principů přeloženy z originálu [9] autorem). Jejich pořadí přitom není náhodné a reprezentuje logickou souslednost kroků k dosažení optimálního stavu datové základny.

2.3.1 Principy dohledatelnosti (Findable)

Smyslem principů této skupiny je zajistit první krok – aby se o existenci dat mohl vůbec někdo dozvědět. A to jak lidé, tak především stroje.

F1 — (Meta)datům jsou přiřazeny globálně-unikátní a persistentní identifikátory

Cílem tohoto principu je především zcela eliminovat nejednoznačnosti, které by mohli nastat při práci s datovou základnou. Mít možnost jasně a spolehlivě odkazovat na (meta)datové elementy je přirozeně nutnou podmínkou pro jakoukoliv další činnost. Toho by mělo být dosaženo přiřazením identifikátoru *každému* elementu dat a metadat – tedy zajistit jednoznačnou identifikaci celku i všech jeho atomických součástí. Konkrétní podobu takových (i třeba složených) identifikátorů tento princip nestanovuje, klade na ně však jisté obecné požadavky.

Prvním požadavkem je, aby byl identifikátor globálně-unikátní. Jinými slovy je nutné zajistit, aby zvolený identifikátor nemohl být použit znovu, a to nejen námi, ale i kýmkoliv jiným. Toho lze dosáhnout použitím vhodného algoritmu pro generování identifikátorů, registrací identifikátoru v externí službě, nebo kombinací obojího.

Druhým požadavkem je, aby byl identifikátor persistentní, tedy aby se neměnil v čase. Zajištění tohoto požadavku bývá u odkazů webových protokolů problematické, zejména proto, že udržovat tyto odkazy aktivní stojí čas a peníze. Z toho důvodu je vhodné využít existující a důvěryhodnou směrovací/registrovou službu.

F2 — Data jsou popsána s pomocí bohatých metadat

Metadata jsou doprovodné informace k datům samotným, která dokumentují zejména postup, který vedl k jejich vzniku a další související okolnosti. Smyslem principu F2 je umožnit data nalézt i bez znalosti jejich identifikátorů, což půjde tím lépe, čím více budou metadata tzv. „bohatá" – termín *bohatá metadata* (jinak též rich metadata) zavádí, a jeho smysl popisuje, princip R1.

F3 — Metadata jasně a výslovně zahrnují identifikátor dat, která popisují

Tento zdánlivě primitivní princip představuje požadavek na propojení dat a metadat, která je popisují. Je explicitně zmíněn proto, že jeho naplnění nemusí být automatické a samozřejmé, protože data a metadata jsou obvykle reprezentována oddělenými entitami nebo soubory. Z toho důvodu je nutné, aby metadata obsahovala odkaz na globálně unikátní, persistentní identifikátor s pomocí kterého jsou data označena, viz princip F1.

F4 — (Meta)data jsou zaregistrovaná nebo indexovaná v prohledatelném zdroji

Tento princip vychází z premisy, že data a metadata, jakkoliv kvalitní a dobře strukturovaná, nemají hodnotu, pokud je nelze objevit. Cílem je zajistit, aby (meta)data mohl autonomně objevit v prostředí sítě internet prohledávací stroj – tzv. crawler. Crawlery, nebo též „pavouky" či „roboty", běžně používají webové vyhledávající služby typu Google, Seznam.cz apod. Pomyslná armáda těchto strojů neustále prohledává web, traverzuje přes jednotlivé hypertextové odkazy na webových stránkách a tvoří komplikovaný vyhledávací index, který je možné využít ke hledání odpovědí na neméně komplikované dotazy.

Webové vyhledávače fungují na bázi přirozeného jazyka, a proto se nedají použít pro datové vyhledávání. Za tímto účelem v době psaní této práce, tedy v letech 2019-20, teprve vznikají obecné datové vyhledávače, které budou na stejném principu prohledávat datové zdroje a poskytovat odpovědi na datově orientované dotazy [16]. První základní funkční implementaci nabízí Google se službou s názvem Data Search, která byla uvedena do testovacího provozu ve druhé polovině roku 2018 [17]. Tato služba však zatím funguje pouze na úrovni celých datasetů a rozsah zpracovávaných dat je značně omezený [18].

Ultimativním způsobem naplnění tohoto principu je tedy poskytnout veřejně přístupný datový bod (viz např. snahy o tzv. FAIR data point [19]), který bude bez dalšího automatizovaně prohledatelný crawlery datových vyhledávačů. Popisovaný princip lze v současnosti naplnit prakticky pouze manuální registrací dat, nebo integrací datového zdroje do indexovaných služeb třetích stran, které však bývají oborově úzce zaměřené (např. GBIF.org).

2.3.2 Principy přístupnosti (Accesible)

Principy skupiny přístupnosti formalizují druhý krok – způsob jakým mají být data a metadata dostupná, poté co je člověk nebo stroj objevil. Je vhodné připomenout, že tyto principy nejsou v konfliktu s požadavky na případnou autentizaci a autorizaci, ale naopak navádí k jejich podpoře. Zajímavé však je, že FAIR principy explicitně nezmiňují požadavek na šifrování, ani od protokolů nevyžadují zajištění integrity nebo důvěrnosti přenášených zpráv.

A1 — (Meta)data lze získat na základě jejich identifikátoru s pomocí standardizovaného komunikačního protokolu

Účelem tohoto principu je v přeneseném slova smyslu povinnost používat univerzální dorozumívací jazyk. Ten je definován obecně přijímaným standardem. Rozdíl mezi nestandardním a standardním protokolem tkví především v tom, že nestandardním protokolem obvykle může komunikovat jen konkrétní software nebo zařízení, a tudíž není univerzálně použitelný. Tento princip se z praktických důvodu rozpadá na dva sub-principy, které popisují následující dvě části.

A1.1 — Protokol je otevřený, svobodný a univerzálně implementovatelný

Tento princip v zásadě pouze stanoví, aby se získávání dat dělo pouze prostřednictvím otevřených protokolů. Otevřenost v tomto kontextu znamená, že kdokoliv může vytvořit software, který s daným protokolem bude pracovat, protože je k němu dostupná dokumentace, není zatížen patenty nebo jinými právními či licenčními omezeními a za jeho použití se neplatí.

A1.2 — Protokol umožňuje autentizaci a autorizaci, pokud jsou potřeba

V určitých případech může být nutné přístup k datům omezit pro konkrétní okruh osob. To může být nutné např. v situaci, kdy data obsahují citlivé údaje, tvoří určité tajemství nebo se je vlastník ze své vlastní vůle rozhodne dát k dispozici pouze za poplatek apod. Z hlediska tohoto principu je rovněž konformní podmínit přístup k datům pomocí uživatelského účtu vytvořeného na žádost, aby bylo možná přístupová oprávnění řídit na individuální bázi. Jak

již bylo zmíněno, FAIR není s takovými požadavky v konfliktu. Je však nutné, aby byla bez omezení veřejně dostupná alespoň metadata, viz následující princip A2. Z toho také plyne, že omezení přístupu také nesmí kolidovat s principem F1, který implikuje že aspoň identifikátory dat a metadat musí být veřejně známé.

Aby bylo možné toto zajistit, je nutné použít takový komunikační protokol, který obsahuje mechanismy řízení přístupu a oprávnění. Opět je nutné zvolit zavedený standard tak, aby podmínkám pro přístup k datům mohl porozumět stroj.

A2 — Metadata jsou dostupná i když data již nejsou k dispozici

Tento princip vychází z premisy, že udržovat zdroje dostupné online stojí čas a peníze, a proto v dlouhodobém horizontu může logicky dojít k jejich nedostupnosti. Pokud se tak stane, má to za následek rozpad vazeb, protože odkazy, přes které byly data a metadata dostupné již nefungují, což představuje velký problém pro zajištění integrity propojených dat a mimo jiné i citační provázanosti.

Tento princip rovněž předpokládá, že náklady spojené s udržováním dostupnosti metadat budou výrazně nižší, než náklady nutné pro zajištěni dostupnosti dat samotných, a to zejména kvůli tomu, že data mohou mít formu velmi velkých souborů, v řádech gigabytů i terabytů dat.

Smyslem tohoto principu je tedy z výše uvedených důvodů zajistit dostupnost alespoň základních metadat i poté, co úložiště dat takříkajíc „umře", nebo data přestanou být dostupná kvůli změně přístupových oprávnění. Vedlejším důsledkem tohoto principu je také to, že data a metadata by měla být jasně oddělena, aby s nimi bylo možné samostatně nakládat alespoň v rozsahu nutném pro naplnění tohoto principu.

V souvislosti s tím je doporučeno, aby byla formalizována a zpřístupněna politika persistence dat (persistence policy), ve která bude definováno, zda a jaký závazek poskytovatel dat má z hlediska zajištění jejich dostupnosti v dlouhodobém horizontu.

2.3.3 Principy interoperability (Interoperable)

Smyslem principů této skupiny je zajistit možnost vzájemné spolupráce výpočetních systémů bez nutnosti manuálních zásahů, nebo tyto alespoň minimalizovat. Naplnění těchto principů by mělo být dosaženo zejména s pomocí oboustranného provázání dat a zejména metadat strojově čitelnými křížovými odkazy.

Tato skupina principů se zaměřuje nejen na syntaktické aspekty takového provázání, ale též klade důraz na jejich korektní pochopení z hlediska sémantiky. Pracuje proto mimo jiné s pojmem **ontologie**, který je před popisem jednotlivých principů vhodné nejprve zavést, cit.:

„Cílem ontologie je definovat společné, jednotné chápání určité třídy pojmů. Ontologie by ve výsledku měla podporovat porozumění mezi lidmi (vědečtí pracovníci), komunikaci mezi počítačovými systémy či usnadnění návrhu znalostně-orientovaných aplikací. Ontologií dokážeme znázornit komplexnost vztahů mezi znalostmi." [20]

I1 — (Meta)data používají formální, přístupný, sdílený a široce použitelný jazyk pro reprezentaci znalostí

Hlavním tématem tohoto principu je umožnění a usnadnění korektní interpretace dat napříč různými zdroji.

Pokud data nepoužívají stejný jazyk, je nutné pro jejich vzájemnou integraci použít mapování. Tento proces dělá člověk při zpracování heterogenních dat intuitivně ve své mysli, definici neznámých termínů si dohledá v přirozeném jazyce a přiřadí známému pojmu překlad do dosud neznámého vyjádření. Protože však stroje takové míry abstrakce obvykle nejsou schopny, je třeba vazby mezi daty, pojmy a jejich definicemi formalizovat ve standardním strojově zpracovatelném formátu, a eliminovat tak nutnost nestandardního ad-hoc mapování a překladů. Takto formalizované vazby poté tvoří sémantický model, jehož existence je spolu s využitím běžně užívaného názvosloví a ontologie prostředkem pro naplnění tohoto principu.

I2 — (Meta)data používají názvosloví, které se řídí principy FAIR

Tento princip dále rozvíjí předchozí princip I1 v tom smyslu, že názvosloví používané k popisu dat a metadat musí být dokumentované, přičemž forma této dokumentace musí sama o sobě splňovat principy FAIR – tedy být dostupná komukoliv, kdo s daty pracuje ve strojově zpracované podobě pomocí otevřeného protokolu na základě globálně-unikátních a persistentních identifikátorů.

I3 — (Meta)data obsahují kvalifikované odkazy na jiná (meta)data

Cílem, který tento princip sleduje, je provázat data a zejména metadata v co největší míře pomocí smysluplných kvalifikovaných křížových odkazů. U kvalifikovaných odkazů je pojmenován vztah mezi odkazující a odkazovanou entitou (např. „A je součástí B" / „B má součást A"), čímž se liší od prosté reference (např. „A, viz také B"), jejíž smysl nemusí být bez dalšího kontextu zřejmý. Provázání může být přítomno jak v rámci našich (meta)dat, tak mezi našimi a externími (meta)daty. Provázání kvalifikovanými odkazy pak může vyjadřovat, že např. jedna datová sada je součástí jiné, nebo je na jiné založena nebo je jinou doplněna. Poslední jmenovanou možnost lze s výhodou využít také pro naplnění souvisejícího principu I1, a to vhodným propojením dat a metadat na vhodné související externě spravované

ontologické definice. Žádoucím vedlejším efektem je, že při korektním použití kvalifikovaných odkazů je automaticky zajištěno řádné ocitování odkazovaných dat a metadat.

V této souvislosti je vhodné zmínit, že FAIR specifikuje tento princip s vědomím nutnosti vynaložit určité úsilí pro jeho naplnění, a proto též stanovuje, že energie vynaložená za účelem zavedení výše popsaných kvalifikovaných odkazů má být v zájmu zachování principu hospodárnosti přiměřená míře odpovídající přínosu a důležitosti, které takové propojení představuje.

2.3.4 Principy znovupoužitelnosti (Reusable)

Dosažení znuvupoužitelnosti je konečným cílem FAIRu jako celku, přičemž by se tak mělo dít pomocí naplnění principů této skupiny.

R1 — (Meta)data jsou bohatě popsána pomocí velkorysé množiny přesných a relevantních atributů

V kontextu FAIR a této práce je obecně možné metadata rozlišit z kvalitativního hlediska na základní, specializovaná a tzv. bohatá:

Základní metadata obvykle tvoří název, kdo, kdy a proč data pořídil, z čeho vznikla, kdo k nim přispěl, v jakém jsou jazyce či formátu, kdo k nim má jaká licenční práva apod. Lze říct, že primárním účelem základních metadat je zasazení dat do kontextu okolností jejich vzniku.

Specializovaná metadata jsou již specifická konkrétním typům dat nebo oboru či účelu, za kterým byla data pořízena. Specializovaná metadata tedy přináší užitek zejména vlastníkům dat a zpracovatelům dat ze stejného oborového nebo zájmového okruhu.

O *bohatých metadatech* (tzv. rich metadata) hovoříme obvykle tehdy, pokud jsou kromě metadat přímo užitečných vlastníkovi uchovány i ostatní informace, které byly při pořízení těchto dat nashromážděny. Informace tohoto charakteru typicky vznikají při činnosti různých přístrojů. Jednoduchým příkladem jsou EXIF a IPTC metadata, které k obrazovým datům automaticky připojují digitální fotoaparáty – tato metadata obvykle obsahují informace o datu a času pořízení, rychlosti závěrky, velikosti clony, typu fotoaparátu, vlastnostech objektivu, ale například také GPS souřadnice nebo okolní teplotu. Tato metadata nemusí být nutně užitečná pro autora snímku, kterého třeba zajímá pouze snímek samotný, ale v budoucnu mohou být hodnotná pro někoho jiného – třeba i pro výzkum, který leží zcela mimo obor či zájem původního autora a který tedy jejich hodnotu ani nemůže dostatečně posoudit.

Smyslem principu R1 je umožnit člověku nebo stroji rozhodnout, zda jsou pro něj data, která metadata popisují, užitečná. K tomu mají dopomoci právě bohatá metadata. Naplnění

principu R1 tak prakticky znamená, že pokud již nějaká metadata vzniknou, nebudou aktivně zahozena, ale naopak budou uchována pro další i zcela nečekané využití kdykoliv v budoucnu – viz velkorysost zmiňovaná v názvu principu (v originálu *„...plurality of...”* [9]).

R1.1. (Meta)data jsou uvolněna pod jasnou a přístupnou licencí

U dat, která se rozhodneme umístit do nějakého datového zdroje, musíme definovat podmínky, za kterých je možné je dále využít. Tyto podmínky jsou formalizovány pomocí licence, což je dokument, kterým autor vyjadřuje svoji vůli, jaké užití jeho díla je přípustné a jaké naopak ne.

Licenci si buď můžeme napsat sami, nebo zvolit nějakou standardně používanou. Vlastní licence, vyjádřené pouze slovně, představují veliký problém, protože je musí analyzovat člověk (obvykle navíc právník), který poté rozhodne, zda data lze v konkrétním případě užít či nikoliv. U licencí sepsaných diletantským způsobem navíc hrozí zvýšené riziko vzniku právních vad, což má zásadní negativní vliv na znovupoužitelnost předmětného díla a jde tak přímo proti smyslu FAIR.

Mnohem lepší je adoptovat některou z obecně známých a používaných licencí, u které navíc postačí indikovat její název s příp. parametry. Pokud tak navíc učiníme pomocí strojově čitelného odkazu, umožníme datovým vyhledávačům jasně rozpoznat, co dovolujeme s našimi daty dělat a co ne.

Typickým zástupcem používaným pro vědecká data jsou licence z rodiny Creative Commons (mají několik variant, ze kterých můžeme volit), dále jsou často používané GNU GPL, MIT, BSD, Apache – ty se velmi dobře hodí pro skripty nebo počítačový kód. Často se lze setkat s označením Public Domain, které značí volné dílo (např. v USA jsou v tomto režimu všechny výstupy vládních organizací, vč. např. NASA) - tento stav je však vhodné formalizovat označením s pomocí licence CC0.

R1.2. (Meta)data jsou spojena s detailní dokumentací původu

Dokumentace původu (angl. *provenance*) popisuje okolnosti, za kterých data a metadata vznikla nebo byla pořízena. Základní dokumentaci tvoří informace, kdo data pořídil, jak při tom postupoval, a z jakých podkladů vycházel.

Obvykle jsou tyto informace obsaženy v laboratorních protokolech, na základě, kterých by mělo být možné daný pokus opakovat, příp. zhodnotit, zda jsou data dostatečně vypovídající. Součástí těchto informací by měla být i metodika a příp. také citační odkaz na jiné (cizí nebo dřívější) práce či datové sady, které byly využity. Dokumentace o kvalitě a

původu bývá vedena neformálně, existují však standardy pro její formalizaci, např. pomocí známé ontologie PROV-O, jejichž zápis je strojově čitelný.

R1.3. (Meta)data odpovídají relevantním komunitním standardům

Tento princip lze v zásadě vyložit tak, že pokud v dané vědecké komunitě (oblasti, doméně) již existují zavedené postupy či standardy určující způsob organizace informací, je vhodné je dodržovat. Tyto standardy a postupy obvykle definují určitou minimální množinu informací, kterou by měl daný datový záznam obsahovat, bez které by jinak byl nespolehlivý či přímo nepoužitelný. Nejde tedy o technickou formu či způsob vyjádření – ty řeší zejména principy skupiny I.

Je nutné podotknout, že způsob naplnění tohoto principu zcela závisí na tom, v jaké vědecké oblasti se pohybujeme, a tedy i v jakém rozsahu a formě pro tuto oblast relevantní standardy a postupy existují (příp. zda vůbec). Z toho důvodu v této věci nelze vydat konkrétnější doporučení, ani naplnění daného principu automatizovaně testovat.

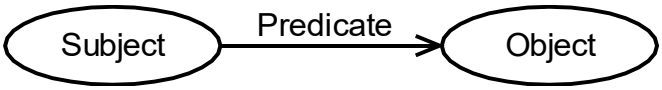
2.4 Technické aspekty FAIR a prostředky strojově čitelného vyjádření sémantiky dat

Jak již bylo zmíněno, principy FAIR popsané v předchozí kapitole jsou technologicky neutrální. Pro jejich praktickou implementaci je však zapotřebí technologicky-specifické metodiky. Pro tento úkol lze doporučit především sadu doporučených postupů Data on the Web Best Practices W3C Recommendation [21], která umožní uchopit všechny dosud diskutované problémy z praktického pohledu a je s principy FAIR v souladu.

Z množství doporučovaných postupů jsou v kontextu této práce zajímavé především ty, které se zabývají způsoby strojově čitelného vyjádření sémantiky dat. K tomuto účelu se standardně používá již zavedený a otevřený (a tedy splňující FAIR princip I1) Resource Description Framework (RDF), který umožňuje vyjádření jak v lidsky-, tak ve strojově-čitelné formě [22]. Jelikož jde o zásadní téma z pohledu FAIR principů jsou v této kapitole stručně popsány jeho základní stavební prvky doplněné názornými příklady.

2.4.1 Resource Description Framework (RDF)

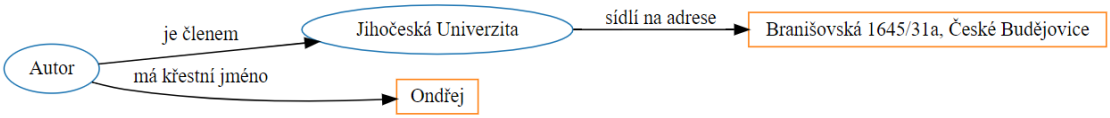
RDF používá k popisu sémantiky model tvrzení (RDF statement), který tvoří sémantická trojice *subjekt-predikát-objekt* (nazývaná triple, RDF triple), viz Obrázek 2. Množina těchto trojic pak tvoří sémantický graf.



Obrázek 2: Grafické znázornění modelu RDF trojice/tvrzení, zdroj: [22]

Model trojice vychází z poznatků lingvistiky a představuje formalizaci vyjadřování přirozeného jazyka – např. věta *„Autor této práce má křestní jméno 'Ondřej'.“*, kde je subjektem autor této práce a predikátem „má křestní jméno“. Objektem přitom může být nejen literál (v příkladu ‚Ondřej‘), ale také jiný subjekt, díky čemuž je možné vytvářet složitější konstrukce – např. *„Autor této práce je členem Jihočeské univerzity. Jihočeská univerzita sídlí na adrese 'Branišovská 1645/31a, České Budějovice'.“* Množina trojic, tedy sémantický model, má charakter orientovaného grafu (RDF graph).

Grafické, a tedy i lidsky čitelné znázornění uvedeného příkladu je naznačeno na Obrázku 3:



Obrázek 3: Grafické znázornění příkladu

Pro strojově čitelný zápis takového grafu je však nutné nahradit volné konstrukce subjektů a predikátů a specifikovat je pomocí identifikátorů. RDF používá k tomuto účelu IRI (Internationalized resource identifier, neplést s IDN[1]). IRI je ve většině případů zobecněním URI[2] (Uniform Resource Identifier), který definovaným způsobem umožňuje zápis všech Unicode znaků. URI je inherentně globálně unikátní v tom smyslu, že jeden URI identifikuje právě jeden konkrétní subjekt, jeden subjekt však muže být identifikován více URI (např. URL

[1] Internationalized domain names, způsob zápisu znaků národních abeced v doménových jménech, viz RFC 3490. V normalizované podobě RDF se naopak tyto prvky vyskytovat nesmí [22].

[2] Z definice: Každý absolutní URI a URL je IRI, ale ne všechny IRI jsou URI. Případy, kdy IRI není URI jsou pro kontext této práce irelevantní a pro zjednodušení je tedy pomineme s tím, že zájemce o hlubší vhled lze odkázat přímo na standard RFC 3987.

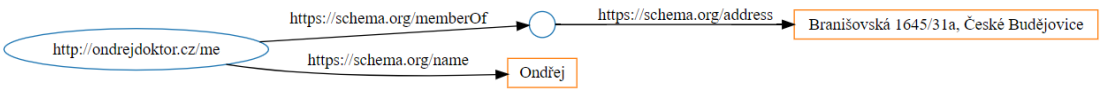
adresou a DOI identifikátorem). Připomeňme, že URI nemusí být jen webová adresa (přesněji zdroj identifikovaný pomocí schématu HTTP), i když v daném kontextu tomu zpravidla tak bývá.

Pamatováno je i na situaci, kdy subjekt žádný identifikátor nemá, příp. jej neznáme – v sémantickém grafu tak představuje tzv. prázdný uzel (blank node). Z technického pohledu je pak v této situaci nutné prázdným uzlům i tak nějaký identifikátor přiřadit, aby bylo možné sémantický graf vůbec zapsat. Konkrétní řešení pak záleží na technické implementaci systému pro ukládání dat (data storage engine), také ve vazbě na syntaxi vyjádření RDF, kterou používá. Běžná řešení spočívají buď v generování umělých identifikátorů, které jsou platné pouze lokálně v rámci grafu, nebo se přistupuje k tzv. skolemizaci IRI [23] pokud je vyžadováno zachování určitého jmenného prostoru.

Obě řešení samozřejmě porušují požadavek FAIR principu F1, proto je nelze použít k identifikaci subjektů, které představují instanci entit datového modelu. Přípustné je však užití v případech, kdy je to technicky vyžadováno strukturální povahou datového typu popisovaného elementu (meta)dat – např. GPS souřadnice chápané jako anonymní subjekt s predikáty „má zeměpisnou šířku“ a „má zeměpisnou délku“. Je také nutné zdůraznit, že zcela nepřípustné je užití generovaných identifikátorů pro predikáty, a to z prostého důvodu, že predikáty jinak než s pomocí IRI technicky zapsat nelze.

Naopak z hlediska popisu sémantiky není existence více identifikátorů na překážku, a není v konfliktu ani s FAIR principy (pokud o existenci takových identifikátorů víme, můžeme naopak tuto skutečnost vyjádřit vhodným predikátem typu „je stejné jako“).

Na základě této znalosti můžeme navázat na výše uvedený příklad: Předpokládejme, že autor chce tyto informace o sobě ve strojově čitelné podobě publikovat v rámci svého webu a pro subjekt „sebe“ vytvoří IRI http://ondrejdoktor.cz/me, a dále předpokládejme predikáty „je členem“ = https://schema.org/memberOf a „sídlí na adrese“ = https://schema.org/address (jak k nim dojít bude vysvětleno v další podkapitole). Kromě grafického znázornění RDF grafu je nyní již možné tvrzení z příkladu serializovat do prakticky použitelné podoby, např. formátu N-Triples – ten je již technicky bez problémů možné publikovat na webu (např. tak že server při HTTP požadavku na adresu http://ondrejdoktor.cz/me vrátí tento obsah s hlavičkou *Content-type: application/n-triples*). Obrázek 4 zachycuje výsledek.




```
<http://ondrejdoktor.cz/me> <https://schema.org/name> "Ondřej"
<http://ondrejdoktor.cz/me> <https://schema.org/memberOf> _:b0
_:b0 <https://schema.org/address> "Branišovská 1645/31a, České Budějovice"
```

Obrázek 4: Příklad po doplnění IRI – schéma a zápis N-Triples

Ačkoliv je příklad již technicky publikovatelný, stále trpí zásadním nedostatkem – stále není jasné o čem RDF věty hovoří. Přeloženo zpět do přirozeného jazyka celé vyjádření v podstatě znamená pouze, že *„Něco se jmenuje 'Ondřej' a toto něco je členem něčeho jiného co sídlí na adrese 'Branišovská 1645/31a, České Budějovice', a ani to nemá identifikátor."*

Důsledkem je, že bez předchozí znalosti není pro stroj možné odvodit co (tedy jaký koncept) přítomné subjekty představují a kvůli absenci původní indicie to není možné ani pro člověka. Tento příklad tak názorně demonstruje důsledky, které má nedodržení FAIR principů F1, F2 a R1. Mohli bychom také uvažovat v tom smyslu, že uzlům budeme přidávat další predikáty, třeba „název", „jméno", „příjmení", „IČO", apod. (v rámci principu plurality). To je jistě žádoucí, samo o osobě nám to však hlavní problém nevyřeší. Vyvstávají tedy důležité otázky: Jak určit co jednotlivé subjekty představují? Jaké predikáty bychom subjektům mohli či přímo měli dát? V jakém formátu mají být jejich hodnoty? Příklad tedy stále není kompletní. Odpovědi na tyto otázky nalezneme aplikací vhodné ontologie, tedy výrazového rámce, do kterého sémantické trojice usadíme.

2.4.2 Ontologie v kontextu RDF

Ontologie v kontextu informačních věd a RDF je uspořádanou kolekcí výrazových prostředků, definuje zejména výrazové koncepty (třídy), jejich atributy, podobu těchto atributů a vzájemné vztahy mezi těmito prvky [20]. Nejde tedy o pouhý slovníček pojmů ale komplexní sémantický framework[3].

Technicky jsou ontologie formalizovány pomocí Web Ontology Language (OWL) [24]. Stejně jako RDF jde o otevřený standard – je možné využívat existující ontologie z otevřeného ekosystému, nebo vytvářet a publikovat vlastní. Rovněž není definována „jediná správná ontologie" a výběr (jedné nebo více) ontologií pro popis daných (meta)dat v rámci procesu

[3] Samotný výraz *ontologie* lze v širším pojetí přeložit jako „nauka o jsoucnu". Jeho kořeny je proto nutné hledat, podobně jako u RDF vycházejícího z lingvistiky, v humanitních vědách.

FAIRifikace, který bude blíže popsán v následující kapitole, tak zcela závisí na charakteru popisovaných (meta)dat a vědecké oblasti které se týkají. Při volbě ontologie je nutné vzít v úvahu její popularitu, resp. používanost, jelikož je logicky (a v souladu s FAIR principy I2 a R1.3) žádoucí „hovořit" jazykem, kterému rozumí co nejširší publikum. **Volba vhodných ontologií je tedy klíčovým momentem v procesu popisu sémantiky dat.**

Jako vhodné nástroje k tomuto úkolu, v kontextu této práce zaměřené na biologická data, lze zmínit především repozitáře Ontology Lookup Service [25] a Fairsharing.org [26], kde kromě oborově-specifických ontologií najdeme i ontologie obecně použitelné a široce přijímané, zejména pak základní ontologii Schema.org, kterou používají a doporučují mj. velcí IT hráči jako je Google či Microsoft [27].

S těmito znalostmi tak můžeme náš příklad dokončit, třeba právě s využitím ontologie Schema.org a ontologie syntaxe RDF (ano, i struktura RDF je sama popsána vlastní ontologií):

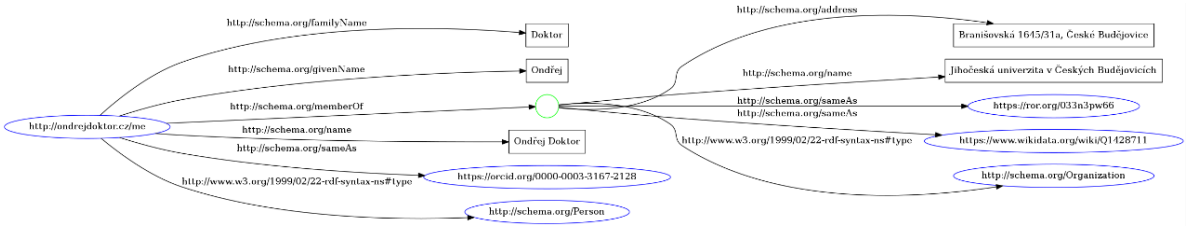
Po prostudování se pro naše subjekty skvěle hodí třídy *Person* (https://schema.org/Person) a *Organization* (https://schema.org/Organization)[4], které k subjektům přiřadíme pomocí predikátu *type* (https://www.w3.org/1999/02/22-rdf-syntax-ns#type).

Ontologie Schema.org sice nestanovuje žádné povinné predikáty pro tyto třídy, nicméně je ustálenou zvyklostí, aby pro instance těchto tříd byl k dispozici alespoň predikát vyjadřující „má jméno" (https://schema.org/name) – zde mj. vidíme praktický význam FAIR principu R1.3. Tento predikát obě zvolené třídy dědí od společné nadtřídy *Thing* (https://schema.org/Thing), která je mimo jiné předkem všech tříd této ontologie. Vhodnými specifickými predikáty jsou pro třídu *Person* atomické součásti jména „má křestní jméno" (http://schema.org/givenName) a „má příjmení" (http://schema.org/familyName), pro třídu *Organization* pak „nachází se na adrese" (http://schema.org/address).

Protože jednou ze základních a klíčových vlastností sémantických dat je možnost jejich propojování, doplníme také odkazy na aliasy subjektů v externích systémech s pomocí predikátu „je stejné jako" (http://schema.org/sameAs). Praktickým důsledkem takového propojení je robustní zasazení našich dat do bohatého kontextu a z toho vyplývající možnost získat o daných subjektech mnohem více informací, než máme sami k dispozici – což je mimo jiné smysl FAIR principu I3, který právě tímto ukázkově naplníme. V našem konkrétním

[4] Výtažek z definice:
Person – „A person (alive, dead, undead, or fictional)." (sic)
Organization – „An organization such as a school, NGO, corporation, club, etc."

příkladě je možné využít u subjektu autora jeho ORCID (https://orcid.org/0000-0003-3167-2128), u subjektu Jihočeské Univerzity identifikátory Research Organzation Registry (https://ror.org/033n3pw66) a na portálu Wikidata (https://www.wikidata.org/wiki/Q1428711)[5]. Výsledek usazení příkladu do rámce vhodné ontologie zachycuje Obrázek 5.

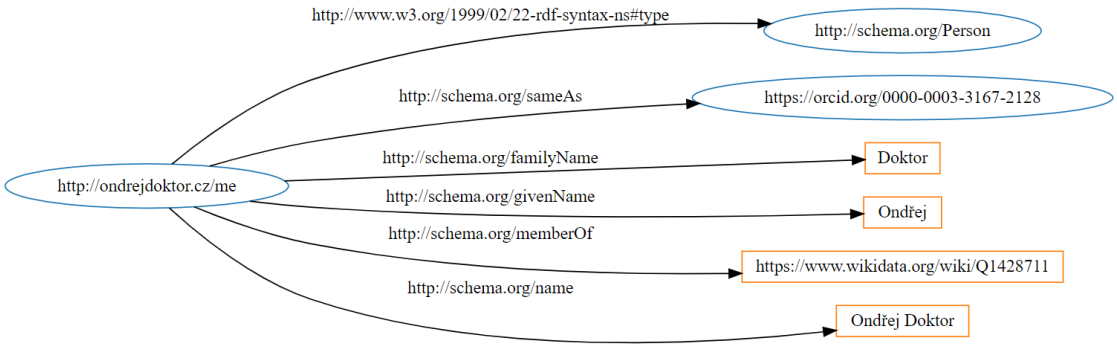



```
<http://ondrejdoktor.cz/me> <http://schema.org/familyName> "Doktor"
<http://ondrejdoktor.cz/me> <http://schema.org/givenName> "Ondřej"
<http://ondrejdoktor.cz/me> <http://schema.org/memberOf> _:b0
<http://ondrejdoktor.cz/me> <http://schema.org/name> "Ondřej Doktor"
<http://ondrejdoktor.cz/me> <http://schema.org/sameAs>
↪<https://orcid.org/0000-0003-3167-2128>
<http://ondrejdoktor.cz/me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
↪<http://schema.org/Person>
_:b0 <http://schema.org/address> "Branišovská 1645/31a, České Budějovice"
_:b0 <http://schema.org/name> "Jihočeská univerzita v Českých Budějovicích"
_:b0 <http://schema.org/sameAs> <https://ror.org/033n3pw66>
_:b0 <http://schema.org/sameAs> <https://www.wikidata.org/wiki/Q1428711>
_:b0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/Organization>
```

Obrázek 5: Příklad zasazený do kontextu ontologie - schéma a RDF N-Triples
(pozn.: znak ↪ značí pokračování předešlého řádku)

Téma propojování je vhodné uzavřít poznámkou, že v případě, kdy pro daný subjekt již existuje záznam v renomovaném externím systému obsahující potřebná data vyjádřené ve vhodné ontologii, můžeme přímo odkázat na jeho IRI namísto toho, abychom vytvářeli vlastní reprezentace nebo pomocné prázdné uzly. Pro názornost můžeme náš příklad takto upravit a prázdný uzel reprezentující Jihočeskou Univerzitu přímo nahradit IRI odpovídajícího subjektu na portálu Wikidata, který obsahuje aktualizované a bohaté informace, a to včetně odkazů na další subjekty reprezentující tuto instituci (propojení je tranzitivní), viz Obrázek 6.

[5] Čtenáře by mohlo nyní napadnout použít URI oficiálního webu instituce https://www.jcu.cz/. To však bohužel není možné protože ten sémantická RFD data v jakémkoliv formátu neobsahuje.




```
<http://ondrejdoktor.cz/me> <http://schema.org/familyName> "Doktor"
<http://ondrejdoktor.cz/me> <http://schema.org/givenName> "Ondřej"
<http://ondrejdoktor.cz/me> <http://schema.org/memberOf>
↳"https://www.wikidata.org/wiki/Q1428711"
<http://ondrejdoktor.cz/me> <http://schema.org/name> "Ondřej Doktor"
<http://ondrejdoktor.cz/me> <http://schema.org/sameAs>
↳<https://orcid.org/0000-0003-3167-2128>
<http://ondrejdoktor.cz/me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
↳<http://schema.org/Person>
```

Obrázek 6: Využití externích IRI - schéma a RDF N-Triples
(pozn.: znak ↳ značí pokračování předešlého řádku)

2.4.3 JSON-LD a další formáty pro reprezentaci RDF dat

Závěrem kapitoly je vhodné alespoň stručně zmínit syntaxi s pomocí které jsou RDF data prakticky reprezentována. Jedná se o nutnou znalost pro realizaci praktické části této práce.

Kromě již zmíněné základní podoby N-Triples [28] a její varianty N-Quads [29] (rozšíření o vyjádřením formátu či jazyka literálů) jsou v praxi používány formáty Turtle [30], N3 [31] a RDF/XML [32], jejichž obecným smyslem je poskytnout jiné vyjádření RDF trojic, které je více syntakticky přívětivé pro člověka (Turtle, N3) nebo stroje (RDF/XML). V důsledku jsou však jiným zápisem téhož.

Moderním a doporučovaným formátem (např. dle [33]) je JSON-LD, který technicky staví na syntaxi univerzálního strukturovaného formátu JSON.

JSON-LD byl vyvinut s cílem maximálně usnadnit dodání sémantické informační vrstvy do existujících strukturovaných dat [34]. Díky tomu, že je po technické stránce zcela totožný s formátem JSON, je automaticky zajištěna kompatibilita s existujícími a rozšířenými knihovnami, nástroji a postupy pro provádění základních úloh, jako je např. uchování, transport, serializace a deserializace.

JSON-LD nad prostými JSON daty zavádí sémantickou vrstvu zejména pomocí konceptu tzv. kontextu, což je prakticky realizováno přidáním atributů se speciálním významem, které dle konvence začínají znakem @ (zavináč), a které umožňují doplnit surová data o informace nutné k jejich interpretaci jako RDF trojice. Pro demonstraci použití základních atributů *@context* a *@type* zde alespoň uveďme minimální příklad na Obrázku 7, který zachycuje původní data v JSON formátu, sémanticky doplněná data pomocí JSON-LD s využitím ontologie Schema.org a výsledná sémantická tvrzení, která JSON-LD reprezentuje (na která se přeloží).

```
// JSON
{
  "mojeAdresa": "http://ondrejdoktor.cz/me",
  "jmeno": "Ondřej Doktor",
  "odkazNaProfil": "https://orcid.org/0000-0003-3167-2128"
}

// JSON-LD
{
  "@context": {
    "mojeAdresa": "@id",
    "jmeno": "http://schema.org/name",
    "odkazNaProfil": {
      "@id": "http://schema.org/sameAs",
      "@type": "@id"
    }
  },
  "@type": "http://schema.org/Person",
  "mojeAdresa": "http://ondrejdoktor.cz/me",
  "jmeno": "Ondřej Doktor",
  "odkazNaProfil": "https://orcid.org/0000-0003-3167-2128"
}


// RDF N-triples
<http://ondrejdoktor.cz/me> <http://schema.org/name> "Ondřej Doktor"
<http://ondrejdoktor.cz/me> <http://schema.org/sameAs>
↪<https://orcid.org/0000-0003-3167-2128>
<http://ondrejdoktor.cz/me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
↪<http://schema.org/Person>
```

Obrázek 7: Minimální příklad postupu transformace prostého JSON objektu na JSON-LD data (pozn.: znak ↪ značí pokračování předešlého řádku)

Povšimněme si zejména, že záměrně svévolně zvolené názvy atributů původního JSON objektu (tzv. terms) byly u JSON-LD v jeho kompaktní formě zachovány, aniž by to mělo negativní vliv na možnost jejich sémanticky správné interpretace na straně jedné (tzv. term-to-IRI expansion), ani na možnost zpětného získání původní podoby dat na straně druhé.

Kromě této kompaktní formy (compacted) má JSON-LD dále podobu expandovanou (expanded), plochou (flattened). V expandované formě je kontext zcela odstraněn, původní atributy jsou nahrazeny normalizovanými IRI a dokument je převeden do lépe předvídatelné uniformní struktury, které je vhodná zejména pro optimální strojové zpracování. Plochá forma vychází z expandované a činí její tvar deterministický zploštěním vnořených struktur.

Uvedené formy jsou vzájemně převoditelné pomocí postupů popsaných ve specifikaci *JSON-LD 1.1 Processing Algorithms and API* [35], díky čemuž je možné postupnými transformacemi tvarovat podobu totožných dat pro různá rozhraní či aplikace. Výsledek provedení expanze příkladu z Obrázku 7 a následné zpětné kompaktace je zachycen na Obrázku 8. U kompaktace bylo pro názornost rovněž provedeno zavedení tzv. prefixů – syntaktického cukru, který umožňuje snadnější zápis IRI pomocí vlastních zástupek.

```
// Výstup provedení expanze:
[
  {
    "@type": [
      "http://schema.org/Person"
    ],
    "http://schema.org/name": [
      {
        "@value": "Ondřej Doktor"
      }
    ],
    "@id": "http://ondrejdoktor.cz/me",
    "http://schema.org/sameAs": [
      {
        "@id": "https://orcid.org/0000-
↪0003-3167-2128"
      }
    ]
  }
]
```

```
// Následná kompaktace s prefixem:

// parametr:
{
  "@context": {
    "sorg": "http://schema.org/",
    "dktr": "http://ondrejdoktor.cz/",
    "orcid": "https://orcid.org/"
  }
}

// výstup:
{
  "@context": {
    "sorg": "http://schema.org/",
    "dktr": "http://ondrejdoktor.cz/",
    "orcid": "https://orcid.org/"
  },
  "@id": "dktr:me",
  "@type": "sorg:Person",
  "sorg:name": "Ondřej Doktor",
  "sorg:sameAs": {
    "@id": "orcid:0000-0003-3167-2128"
  }
}
```

Obrázek 8: Expanze JSON-LD a následná kompaktace s prefixem
(pozn.: znak ↪ značí pokračování předešlého řádku)

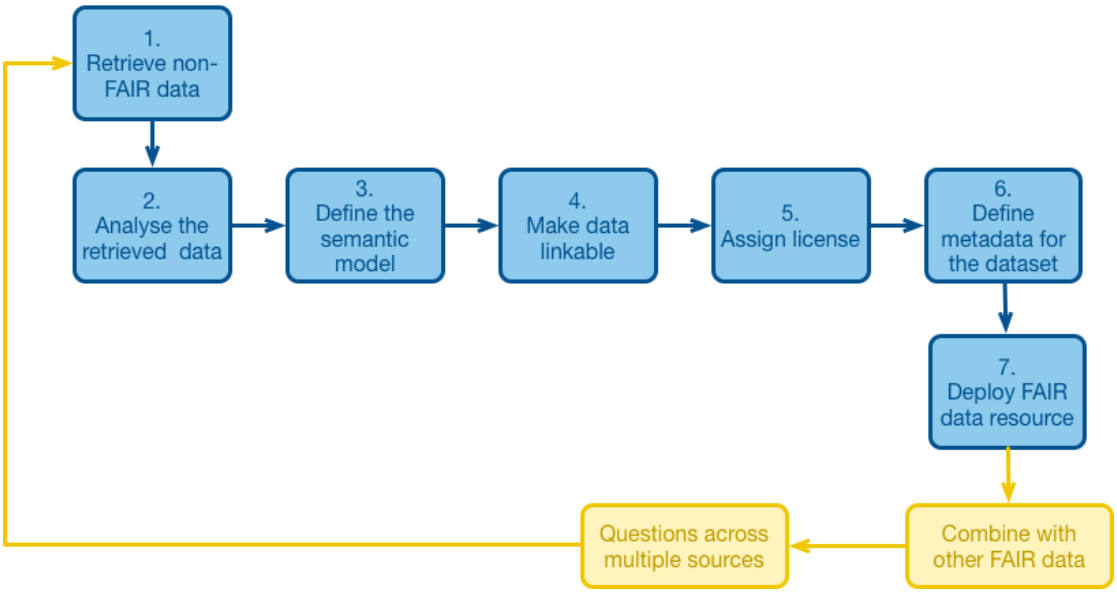
JSON-LD nabízí kromě uvedených základních postupů mnohem více možností pro popis a transformaci strukturovaných dat. Jejich kompletní výčet a popis přesahuje rozsah této práce a pro hlubší studium je vhodnější čtenáře odkázat přímo na specifikace standardu *JSON-LD 1.1* [34].

2.5 Proces převodu dat do FAIR podoby – FAIRifikace

2.5.1 FAIRifikace dle GO FAIR

Pro popis postupu tzv. FAIRifikace, tedy převodu stávajících dat do vhodnější sémanticky definované podoby, si můžeme jako vhodný výchozí bod vypůjčit proces a schéma, které definovala a používá iniciativa *GO FAIR* [36], viz Obrázek 9 níže.

Schéma v obecné rovině popisuje jak produkci (modře, očíslované uzly), tak zpětný proces použití (žlutě) FAIRového zdroje. V interpretaci GO FAIR se proces produkce skládá z několika pro sobě logicky jdoucích kroků, které jsou stručně popsány a diskutovány v následujících podkapitolách:



Obrázek 9: Proces FAIRifikace, zdroj: [36]

Krok 1. – Získat data

Prvním krokem je získat či spíše zvolit data, která se mají optimalizovat. I když může působit banálně takovou věc vůbec zmiňovat, naráží na skutečnost, že **pro veškerou další práci je nezbytně nutná podpora vedení, resp. vlastníků dat**. Ta nemusí být vůbec samozřejmá. Rozhodnutí, jaká data a v jaké míře pro proces FAIRifikace zvolit závisí především na vůli povolaných osob. Ekonomická a pracovně-psychologická rovina implementace FAIR principů je věcí managementu dané instituce a přesahuje rozsah této práce, v této souvislosti je však vhodné zmínit některé skutečnosti, které je vhodné při rozhodování vzít v úvahu [3][9]:

- **Mít otevřená data ≠ poskytovat vše komukoliv zdarma**. FAIR plně podporuje omezení přístupu k datům, dokonce přímo vyžaduje použití technologií, které řízení přístupu na základě autentizace umožňují (princip A1.2) a rovněž podporuje to, aby sami vlastníci dat rozhodli o způsobu jejich licencování (princip R1.1). Pro dobré fungování je však vhodné technicky zajistit, aby identifikátory a metadata byly dostupné, i když samotná data nejsou (principy skupiny F a A).
- **Mít FAIR data znamená především future-proofing v rámci organizace / týmu**. Ať už budou na datové sadě pracovat v přísném utajení pouze dva prověření pracovníci nebo bude bez omezení sdílena s celým světem, je vhodné mít data bezpečně uložena v sémanticky srozumitelné podobě, tedy nikoliv roztroušeně po papírech nebo souborech, jejichž obsahu je schopen porozumět pouze člověk, který je vytvořil (a jehož případných odchod z organizace tak de facto znamená ztrátu těchto informací).
- **Mít FAIR data bude pravděpodobně nutnou podmínkou pro získání dalšího financování v blízké budoucnosti.** Z toho důvodu je vhodné problematiku chápat jako investici.

Krok 2. – Analyzovat získaná data

Dalším krokem je analýza získaných dat, jejímž výsledem je popis datového modelu daného zdroje. Postup se liší v závislosti na charakteru analyzovaných dat. Obecně se identifikují existující entity a jejich vazby (příp. jediná entita, pokud jsou data plochá). V tomto kroku je rovněž vhodné analyzovat obsah a datový typ atributů, identifikovat případné problémy a připravit postupy k jejich korekci, tedy normalizaci dat, např:

- převod hodnot uložených v podobě řetězce na vhodný datový typ, nebo naopak (např. číselné hodnoty)
- odstranění duplicit a nejednoznačností
- normalizace číselníkových hodnot (např. měna, kódy oblastí či zemí, telefonní čísla)
- převod hodnot reprezentující datum a čas do kulturně-invariantním formátu s vyjádřením časové zóny (ISO 8601)
- doplnění mezer a chybějících údajů

V souvislosti s těmito postupy je vhodné seznámit se s pojmem *frictionless data* [37] a pro praktickou realizaci použitelnou sadou nástrojů *Frictionless Toolkit* [38].

Krok 3. – Definovat sémantický model

S datovým modelem v rukou je možné postoupit k dalšímu kroku a přiřadit prvkům datového modelu význam – definovat sémantický model. Zde je zejména možné aplikovat poznatky z kapitoly 2.4 teoretické části práce.

Předně je třeba vyhledat vhodné ontologie, do jejichž kontextu je možné data usadit. K tomuto úkolu je možné doporučit využití již zmíněných služeb Ontology Lookup Service [25] a Fairsharing.org [26]. Samotný sémantický model je pak možné realizovat abstraktním RDF grafem s prázdnými uzly na místě subjektů odpovídajícím třídám datového modelu z předchozího kroku.

Na tomto místě je rovněž vhodné vydělit ze vstupního souboru dat prvky, které mají povahu metadat. Rozdělení zcela závisí na chápání vstupních dat v kontextu vědecké oblasti, ve které se pohybujeme – co jsou data pro jednoho, jsou metadata pro jiného a naopak. Rozdělení je důležité zejména pro naplnění FAIR principu A2, kde je též smysl rozlišování na data a metadata popsán. Obecně dle tohoto principu mají být metadata dostupná i případě kdy data z nějakého důvodu dostupná nejsou (např. kvůli omezení přístupu, byla smazána atp.), což je možné použít jako vhodné vodítko.

Krok 4. – Transformovat data do odkazovatelné podoby

V tomto kroku je aplikován sémantický model z kroku 3 jakožto transformace nad získanými (meta)daty. Jednotlivé záznamy jsou připraveny pro uložení v cílovém datovém zdroji a jsou jim přiřazeny persistentní globálně unikátní identifikátory.

Konkrétní pracovní postup závisí na zvoleném technickém řešení. V tomto bodě je rovněž možné aplikovat transformace pro technickou normalizaci dat, tedy k odstranění případných problémů identifikovaných v kroku 2.

Krok 5. – Přiřadit licenci

Upozornění: *Informace obsažené v této kapitole i kdekoliv jinde v této práci nejsou právním názorem, doporučením ani jinou právní službou a nemohou být pokládány za základ pro jakékoli rozhodnutí nebo jednání. Autor práce nenese žádnou odpovědnost v souvislosti s užitím těchto informací. Pro řešení jakékoliv konkrétní právní otázky se doporučuje vyhledat individuální právní službu.*

Přiřazení licence pro data a metadata (tedy obecně tzv. „autorská díla“) je důležité z hlediska možnosti tyto, jakožto předměty práva duševního vlastnictví následně rozumným způsobem užít dalšími osobami odlišnými od autora, resp. vlastníka autorských práv

majetkových, neboť se v případě její absence uplatní výchozí právní úprava místně příslušného práva. Princip FAIR R1.1, který přiřazení licence upravuje, je obecně třeba chápat tak, že osoby, kterých se to v rámci procesu FAIRifikace týká by měli mít jasno v tom, kdo vykonává k předmětným dílům autorská práva majetková a osobnostní (či jejich zahraniční ekvivalenty), jak chtějí tyto osoby se svými právy naložit a aby svou vůli dali jasným způsobem najevo.

Běžným omylem, který se nevyhýbá také vědecké komunitě, je například představa, že autoři zaměstnaní na základě pracovněprávním vztahu automaticky „mají“ všechna práva k výsledkům vědeckého bádání, které vytvořili v rámci tohoto vztahu pro svého zaměstnavatele, např. univerzitu. Obvykle tomu tak není, přičemž v této situaci v rámci standardní právní úpravy legislativního prostředí České republiky vykonává autorská práva majetková k těmto dílům zaměstnavatel, neboť se obvykle jedná o tzv. dílo zaměstnanecké, tedy dílo vytvořené ke splnění závazku vyplývajícího z pracovního nebo služebního poměru. Určité překvapení z pohledu akademického světa by v této souvislosti mohla působit skutečnost, že fakulty ani katedry obvykle sami o sobě nemají právní subjektivity, tedy nemohou být účastníky právních vztahů, přičemž vykonavatelem práv k dílu zaměstnaneckému je zaměstnavatel, tedy obvykle přímo vysoká škola či univerzita. Praktickým důsledkem je tedy, obdobně jako v kroku 1, nutnost zajistit dostatečnou podporu vedení a respektovat jeho politiku v této oblasti, pokud je formulována.

Právní úprava v oblasti duševního vlastnictví navíc podléhá teritorialitě a je vždy třeba vycházet z podmínek, které jsou relevantní vzhledem k místu vzniku a skutečného užití daného díla, čemuž je v oblasti vědeckého výzkumu vzhledem k časté mezinárodní spolupráci rovněž nutné věnovat zvýšenou pozornost.

Samotná volba licence vzhledem k výše uvedenému by tedy z praktického pohledu měla být racionální a oprávněné osoby by ideálně měli volit některou ze zralých, obecně přijímaných, mezinárodně kompatibilních a také strojově čitelných licencí. Dobrý příkladem jsou licence Creative Commons [39], které nejenže mají uvedené vlastnosti, ale též disponují vysokou flexibilitou a pomocí systému modulárních licenčních prvků umožňují nakombinovat nejběžnější požadavky, kterými oprávněné osoby upravují výkon svých práv, viz přehledová tabulka na Obrázku 10.

Označení	Popis	Zkratka
	Nejširší možné užití díla (i bez uvedení autorství)	0
	Pouze uvedení autora	BY
	Uvedení autora + Žádná odvozená díla	BY-ND
	Uvedení autora - týká se původního díla i jeho modifikací	BY-SA
	Uvedení autora + Pouze nekomerční užití	BY-NC
	Uvedení autora + Pouze nekomerční užití + Žádné modifikace	BY-NC-ND
	Uvedení autora + Nekomerční užití + Zachovejte licenci	BY-NC-SA

Obrázek 10: Přehledová tabulka variant licencí Creative Commons, zdroj: [40]

Krok 6. – Definovat metadata

Jak již bylo uvedeno v předchozích částech této práce, věnují principy FAIR zvláštní pozornost jasnému rozdělení informací na data a metadata a jejich vzájemným vazbám. Smyslem tohoto oddělení je pak zejména snaha zachovat základní informace o datech, pokud již tato nejsou dostupná z technických či organizačních důvodu, zejména pak informaci o tom, že data v určitém bodě vůbec existovala, kde se nacházela, kdo a kdy je pořídil, pod jakou licencí byla dostupná apod.

Soubor metadat v tomto kroku tedy budou tvořit prvky takto identifikované v předchozím kroku 3 při tvorbě sémantického modelu společně s informacemi, které jsou o datové sadě uvedeny „stranou“ - v popiscích, komentářích nebo i jen ústně předávané podobě v rámci vědeckého týmu. Podrobněji o různých druzích metadat pojednává kapitola 2.3.4 této práce věnovaná FAIR principu R1 výše.

Nad rámec toho se k těmto metadatům řadí i technické informace které k záznamům přidává cílový repozitář, tzv. data repository metadata. Dále, pokud při sestavování sémantického modelu byla identifikováno organizační zařazení primárních dat do datových sad či katalogů, je nutné doplnit o příslušná inverzní metadata i tyto entity. Pro základní orientaci je vhodné k nahlédnutí doporučit systémy organizace dat vycházející z ontologie Dublin Core [41] a Darwin Core [42], které jsou pro oblast vědeckého výzkumu relevantní. Je třeba též zdůraznit, že na způsob sémantického vyjádření metadat jsou z hlediska FAIR principů kladeny stejné nároky jako na data samotná.

Krok 7. – Publikovat ve FAIR datovém zdroji

Posledním krokem procesu FAIRifikace je pomocí syntézy výstupů předchozích kroků připravit datový soubor a publikovat jej (jinak též „nasadit“) ve FAIR datovém zdroji.

FAIR datovým zdrojem se rozumí konkrétní technické zařízení, systém či služba, které zajistí uchování a přístup k datům jak pro interní potřeby, tak příp. i pro internetové vyhledavače či širší vědeckou komunitu v souladu s FAIR principem F4 (zde předmětný „searchable resource“). Principy FAIR skupiny A navíc na takový zdroj kladou technický požadavek, aby se tak dělo za využití standardních a otevřených komunikačních protokolů podporujících standardizované metody pro autentizaci a řízení přístupu.

Ve fázi nasazení či publikace obvykle též dochází ke konečnému přiřazení identifikátorů (URI / IRI) pro data a metadata, pokud již nejsou definovány v předchozích krocích. Jelikož FAIR klade opodstatněné požadavky na globální unikátnost a spolehlivou dostupnost takových identifikátorů, je velmi vhodné při jejich sestavování využít tzv. směrovací službu, která je spolehlivá a v odvětví zavedená. Základním principem těchto služeb je úplatné či bezúplatné poskytnutí IRI adres či prefixu v prostoru směrovací služby, která poté spolehlivě odkáže požadavky na tyto adresy do cílového umístění. Typickým příkladem takové služby je ve vědeckých kruzích notoricky známý komerční systém DOI spravovaný nadací IDF (International DOI Foundation), který je hojně využíván zejména pro identifikaci vědeckých publikací. Mezi zavedené služby, které poskytují persistentní identifikátory dále patří PURL (spravovaný v současnosti neziskovou organizací Internet Archive) a W3ID (spravovaný konsorciem W3C).

Volba datového zdroje se obecně odvíjí od možností, jaké má vědecký tým k dispozici v rámci své komunity a instituce. V současnosti jsou dostupné zejména oborově zaměřené repozitáře, mezi které patří v oblasti přírodních věd např. GBIF, které však trpí zásadním nedostatkem v tom, že de facto vynucují velmi rigidní datovou strukturu. Existují i oborově neutrální služby, jako je např. rovněž hojně využívaný Mendeley Data, jejichž univerzálnost je však dána zejména obětováním určitého množství sémantické informace dat, protože z podstaty umožňují „pouze“ uchování surových datasetů doplněných o sémanticky čitelná metadata.

V době psaní této práce není k dispozici prakticky použitelné, skutečně univerzální řešení, které by umožnilo uchování, zpracování, kolaborativní editaci, sdílení a publikaci vědeckých dat ve FAIR podobě, což je možné přičíst skutečnosti, že FAIR je relativně nový fenomén. Tímto problémem se mj. v evropském kontextu zabývá organizace ELIXIR, v rámci které je

na Ústavu aplikované informatiky Přírodovědecké fakulty Jihočeské univerzity v Českých Budějovicích vyvíjeno interdisciplinární řešení zejména pro nálezová data, projekt UniCatDB [43], na kterém se podílí i autor této práce, která má s tímto projektem souvislost.

2.5.2 Diskuse k problémům a alternativní postupy

Výše nastíněný postup z praktického hlediska trpí dle autorova názoru určitými nedostatky, které je vhodné pro zdárnou implementaci korigovat či doplnit:

Postup je chápán jako jednorázová operace, tedy jako proces (vstup-výstup) a nikoliv jako produkt (postupné zlepšování). Vypůjčíme-li si terminologii řízení projektů, lze problém připodobnit k rozdílu mezi vodopádovým vs. agilním vývojem – při realizaci typu vodopád se snadno může stát, že dojdou zdroje (čas, peníze) nebo skončíme s výstupem který nereflektuje úpravu zadání, která nastane v průběhu realizace (budeme to muset zahodit a předělat znovu).

Autor považuje za vhodné chápat FAIRová data jako agilně se vyvíjející produkt, nikoliv pouze jako výstup procesu, který je z definice jednorázový. Je nutné umožnit postupnou adopci, zejména s ohledem na to, že použité ontologie se v průběhu času budou zcela jistě vyvíjet. Iterativně-agilní přístup navíc umožní rozumně rozložit vynakládané úsilí, resp. náklady, což v důsledku zvýší pravděpodobnost úspěšné implementace FAIR alespoň v minimální „životaschopné“ podobě (v terminologii agile MVP – Minimum viable product).

Dále se v existujících metodikách tiše se předpokládá, že původní data buď trvale změní formát nebo se zahodí. To je ovšem v rozporu se způsobem, jakým v kontextu této práce sledované vědecké skupiny s daty aktuálně pracují – převažují operace na tabulkovými (zejména CSV) nebo plaintext daty (např. FASTA), které obvykle realizují Python / R skripty, Excel makra, Jupyter notebooky a další obdobné nástroje. Bio-informatické zpracování dat, které se mj. realizuje např. pomocí služeb Metacentra nebo ELIXIR Container orchestration [44], má charakter pipeliningu – zřetězeného zpracování. Je nesporné, že k určité formě transformace či normalizace musí v prvotní fázi dojít, nutně pak pokud data nejsou digitalizována vůbec. Proces FAIRifikace však vždy ultimátně skončí převedením do RDF tripletů, a i když je tato sémanticky správná podoba žádaná, má potenciál silně nabourat veškeré stávající workflow vědeckých skupin (a díky čemuž budou vyvstávat tendence se celému procesu bránit).

Z toho důvodu autor považuje za důležité použít technické řešení které je implementováno jako sémantická vrstva nad daty samotnými, která umožní vhodným způsobem zachování zdrojové podoby dat, příp. dá k dispozici mechanismus pro zpětné zobrazení sémantických

dat do takové podoby, a to minimálně po přechodnou dobu, dokud se sémantické formě nepřizpůsobí ostatní workflow a používané nástroje.

Tyto závěry nejsou v konfliktu se smyslem FAIR principů [9], a proto není problém z nich při úvahách o způsobu realizace FAIRifikace počítat. V rámci praktické části této práce bylo vytvořeno proof-of-concept řešení, které v sobě mj. zahrnuje možný způsob implementace těchto závěrů.

3 Praktická část

3.1 Analýza a hodnocení stávající datové základny

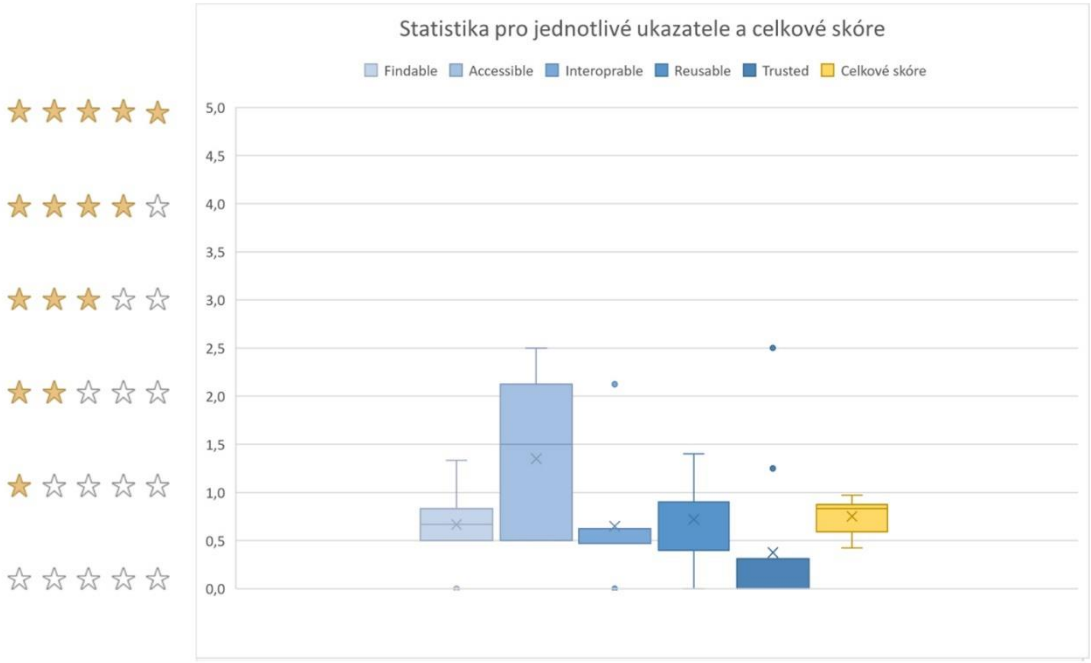
Pro získání podkladů pro analýzu stávající datové základny bylo provedeno dotazníkové šetření. Připravený dotazník vychází z metodiky CSIRO 5-star Data Rating self-assesment testu, který organizace CSIRO vyvinula [12]. Tato metodika je podrobně popsána v kapitole 2.2 této práce. Formulář byl rozšířen o možnost přidání textových komentářů k jednotlivým otázkám pro získání referencí na konkrétně používané nástroje, technologie a standardy. Metodicky bylo dotazníkové šetření doplněno o očištění od duplicit – odpovědi které se týkají jedné a té samé datové sady se nezapočítávají. Naopak v případě, kdy jeden výzkumný tým pracuje s více datovými sadami, které se svou povahou významně liší, byly týmy vyzvány, aby vyplnili dotazník pro každou z nich.

Celé znění formuláře je k dispozici v Příloze 1 a detaily výpočtu v Příloze 2 této práce. Vzhledem k tomu, že účelem dotazníku není hodnotit a vzájemně porovnávat práci jednotlivců, výzkumných týmů nebo kateder, jsou výsledky anonymizovány. Příslušnost výsledků ke katedrám a jednotlivcům kteří uvedli svá jména je však autorovi známa.

3.1.1 Dotazníkové šetření a přehled výsledků

Dotazníkové šetření bylo konzultováno s vedením instituce a jeho provedení laskavě zaštítil proděkan pro vědu a výzkum PřF JU, doc. RNDr. Milan Předota, Ph.D. Šetření probíhalo v období března až dubna roku 2020. Před ostrým vydání dotazníku proběhla pilotáž, na základě které byl původní překlad dotazníku dle metodiky CSIRO 5-star Data Rating doplněn o příklady více ilustrující význam jednotlivých možností.

Šetření bylo úspěšné, byly získány odpovědi od deseti (10) výzkumných skupin z pěti (5) kateder Přírodovědecké fakulty. Kompletní anonymizované výsledky jsou zachyceny v Příloze 3 a zpracované výsledky v Příloze 4 této práce. Na Obrázku 11 níže je vykreslen výtah z této přílohy – statistika pro jednotlivé ukazatele a celkové skóre.



Obrázek 11: Výtah z Přílohy 4 - výsledky dotazníkového šetření

3.1.2 Vyhodnocení dotazníkového šetření a obecné závěry

Výsledky a z nich vyplývající zjištění obecně potvrzují hypotézu vyslovenou v úvodu této práce a rovněž se významně neodchylují od obdobných výzkumů prováděných na jiných institucích [1].

Celkové skóre u žádné ze zkoumaných datových sad nedostačuje na udělení ani jedné hvězdičky proto **je bohužel nutné stav všech datových sad hodnotit v kontextu FAIR jako zcela nevyhovující (!).** V dílčích ukazatelích však lze pozorovat určitou snahu o zlepšení a některé jednotlivé datové sady v nich dosahují až průměrně dobrých výsledků. Ze statistického hlediska však jde de facto o odlehlé hodnoty (viz Obrázek 11). Obecně lze říct, že výsledky jsou tím lepší, čím větší je množství zpracovávaných strukturovaných dat, které vytváří tlak na využití chytřejších způsobů zpracování – nejlepších výsledků dosahují skupiny, které pracují s GIS daty.

Z hlediska ukazatele dohledatelnosti (*Findable*) jsou data prakticky neodkazovatelná – pokud již jsou uložena v nějakém sdíleném prostředku je jejich identifikace pouze zvyková. Pouze v jednom případě se používají lokální relační databáze. Citujme některé slovní komentáře:

„Měření jsou organizována podle jména pracovníka provádějícího analýzy a dále pak podle použité metody a data analýzy.“

„Posíláme si název adresáře s daty.“

„Obvykle mám spíše doplňková data, která uchovávám v oddělených excelových souborech.“

Jistou výjimkou je ukazatel přístupnosti (*Accessible*), jehož výsledky zlepšuje snaha data sdílet – ve většině případů (5) se tak však děje pomocí sdíleného disku či FTP (např. se využívají služby Metacentra CESNET), pouze jedna skupina poskytuje data na webu jako soubor ke stažení. Ve všech případech však zcela schází jakýkoliv strojově čitelný popis sémantiky – jde o izolované soubory, většinou tabulková data.

Z hlediska ukazatele interoperability (*Interoperable*) je podoba dat nevyhovující pro obecné zpracování, pouze v jediném případě jsou používány otevřené strukturované formáty (GeoJSON, JSON, GML). Ostatní skupiny používají buď proprietární formáty nástrojů, se kterými pracují, nebo Excel. Co v těchto formátech zaznamenat nelze (což jsou typicky metadata) je obvykle uloženo v prostých TXT souborech vedle dat a popsáno slovně volnou formou. Citujme některé komentáře:

„Formáty chromatografických software Chromeleon a Xcalibur.“

„Používáme buď aplikace běžně používané odborníky v naší komunitě => standardní výstupní soubory, nebo vlastní programu s popisky sloupců v datových souborech.“

„Data jsou v excelu.“

Většina respondentů navíc použití proprietárních formátů nechápe jako zásadní problém – dosud jim vždy postačovala kompatibilita v rámci jejich velmi úzké oborové skupiny. Pochopení pro interoperabilitu přímo úměrně roste s tím, jak často pro svoji práci využívají data z externích zdrojů – tedy až v momentě kdy sami potřebují integrovat nebo využít data někoho jiného.

Z hlediska ukazatele znovupoužitelnosti (*Reusable*) trpí datové sady zásadním problémem v tom, že know-how vedoucí k jejich vytvoření zůstává v hlavách individuálních pracovníků, kteří je pořídili a logicky se tedy ztratí v případě, že tito pracovníci z nějakého důvodu z výzkumné skupiny odejdou. Některé vědecké skupiny jsou si tohoto problému vědomi a snaží se je řešit svépomocí prostřednictvím vlastního systému pojmenování či zavádí úzus pro formát „metadat“ doprovodného popisu – jedna vědecká skupina např. používá systém, kdy ke každému CSV souboru je ve stejné složce přiložen TXT soubor se stejným názvem, kde první řádek je copy-paste vložený příkaz z příkazové řádky, který byl použit k vygenerování tohoto CSV souboru. Tyto zvyková pravidla však nebývají nikde popsána, a

to ani volnou slovní formou (jako guidelines, interní wiki apod.). Absence uchování dat v jakékoliv databázi (až na jednu výjimku) implikuje nepřítomnost jakéhokoliv indexování, používané kódy či zkratky jsou až na výjimky rovněž zvykové. Výběr ze slovních komentářů:

„Vlastní systém pojmenovávání adresářů. Z názvu adresáře jsou patrné typicky 2-4 hodnoty parametrů, jejichž vliv studujeme.“

„V rámci GIS je automaticky nastaven OGC/ISO, resp. využívá jako minimum Dublin core. Každopádně, ale reálně používám minimálně. V "rámci rychlosti zpracování" pro analýzu dat, kterou zpracovávám sám, mám občas základní popis dat ve formě strukturovaného textu (markdown text, často v podobě jupyter notebooku, který je vázán na projekt – často pro GIS analýzu využívám python, proto jupyter notebook).“

„Vlastní kódy, ale pochopitelné většině blízkých vědců. Tato data ale stejně prakticky nepředáváme nikomu mimo skupinu, ostatní se spokojují s finálními grafy v článcích.“

Z hlediska ukazatele důvěryhodnosti (*Trusted*) pouze jeden výzkumný tým svou datovou sadu aktualizuje pravidelně, jeden nepravidelně. Ostatní (8) vnímají svá data jako jednorázový výstup, který dále nerozvíjí. Žádná výzkumná skupina nesleduje informace o tom, jak jsou data využívána (citovanost publikací, které na těchto datech stavěli se nepočítá, protože se netýká dat samotných). V jednotlivých případech však byla vyjádřena snaha tuto problematiku výhledově řešit.

3.1.3 Problémy, které brzdí zlepšení

Ohlasy na dotazníkové byly všeobecně pozitivní – drtivá většina respondentů na sebe zanechala kontakt, což umožnilo na dotazník navázat osobními nestrukturovanými rozhovory. Respondenti obecně byli velice otevření, zajímali se o možnosti, jak své stávající procesy v daném aspektu zlepšit. Výběr ze závěrečných komentářů:

„Pokud je cílem i podpořit lepší nakládání s daty, ocenil bych zveřejnění prázdného dotazníku s uvedenými možnostmi formátů, úložišť atd., aby člověk tušil, jaké jsou nabízené možnosti.“

„Mnoho dotazů, na které jsem v tomto dotazníku narazila, vystihuje mé současné přemýšlení, neb právě uvažuji nad tím, jakou formou svá narůstající data do budoucna uchovávat, abych je mohla mezinárodně prezentovat a byla připravena pro další mezinárodní spolupráce.“

Z odpovědí lze vysledovat, že snahy výzkumníků brzdí zejména absence systematické podpory. Těžiště jejich vlastní odbornosti se zcela přirozeně nachází jinde, a proto při řešení způsobu práce s daty volí stejně přirozeně minimální funkční řešení, které jim poskytne nezbytné vlastnosti s vynaložením minimálních nákladů (časových i finančních), pro ilustraci citace:

„U spousty otázek byla většina možností cílena na organizované formy nakládání s daty, které zásadně překračují naši praxi – a přesto jsme docela dobře schopni data dohledat, případně i od jiného člena skupiny. Jinak je většinou jeden člen skupiny plně zodpovědný za svá data.“

Dále je často zmiňována obava z úniku primárních dat. Po zmínce o „otevřených datech" velká část respondentů zpozorní, protože si tento termín nejdříve spojí s představou, že svá data budou nuceni zdarma poskytnout komukoliv a přijdou tak o své léta budované know-how.

Při rozhovorech byly dokonce popisovány situace, kdy předčasným uvolněním primárních dat tato stihnul vytěžit někdo jiný a závěry následně publikoval dříve, než by to stihnuli původní pořizovatelé těchto primárních dat, čímž je připravil o plody jejich třeba i mnohaleté práce. Takové situaci se zároveň těžko následně brání, neboť je ve své podstatě lege artis – původní zveřejněná data byla řádně ocitována. **Kromě absence technické a metodické podpory je dle názoru autora toto zcela zásadní faktor, který bude brzdit jakékoliv snahy o zlepšení podoby datových základen vědeckých skupiny**, přičemž vinu za tento stav lze dle názoru autora přičíst zejména způsobu hodnocení práce vědeckých pracovníků spočívající ve sledování počtu vydaných publikací. Je velmi důležité opět zdůraznit, že „mít FAIR data“ neznamená „vzdát se kontroly nad daty“ a že naopak principy FAIR přímo vyžadují po technickém řešení umožnění řízení přístupu k datům prostředky autentizace a autorizace. Tato spíše sociologická rovina problému zcela překračuje rámec této práce, nicméně jde o velmi zajímavé zjištění, doplňme citací:

„…Nicméně s aktuální zkušeností, co mám, je ochota lidí sdílet data, ale i jen metadata velmi omezená. V oblasti GIS dat hodně diskutované téma a mnoho existujících řešení.“

K oběma výše zmíněných problémů lze hledat řešení v samostatné oblasti tzv. data stewardship, kdy je v rámci organizace poskytnuta vědeckým pracovníkům manažerská a metodická podpora od pracovníků v roli tzv. datového stevarda (data steward). V této souvislosti je možné pro další studium problematiky odkázat na projekt Data Stewardship Wizard [45] a související práci jeho zakladatele, doc. Roberta Pergla z FIT ČVUT v Praze.

Posledním problémem, který je důležité zmínit je, že z praktického pohledu mají výzkumné týmy zavedeny určité postupy a nástroje pro zpracovávání dat, či obecně vybudované workflow, které by si logicky přáli při příp. FAIRifikaci dat zachovat. Při hledání technického řešení budou tedy logicky preferovat takové, které jim jejich stávající postupy nenaruší – tedy umožní uchovat data také v původní podobě nebo ve struktuře tomu blízké. Z výsledů dotazníku i následných četných konzultací vyplývá, že většina v současnosti používaných způsobů ukazuje na tabulková data a za hypotetický široce akceptovatelný formát by mohlo být považován CSV, se kterým si většina vědců dokáže již nyní poradit jak na vstupu, tak na výstupu. Formát CSV by tedy mohl být vhodným mezistupněm na cestě od uzavřených proprietárních formátů.

3.2 Návrh technického řešení (Proof of conecept)

Na základě závěrů z teoretické části této práce, a s ohledem na zjištění z předchozí kapitoly, je v reakci na identifikované problémy autorem navrženo technické řešení, jehož popisem se zabývá tato kapitola.

3.2.1 Funkční a nefunkční požadavky

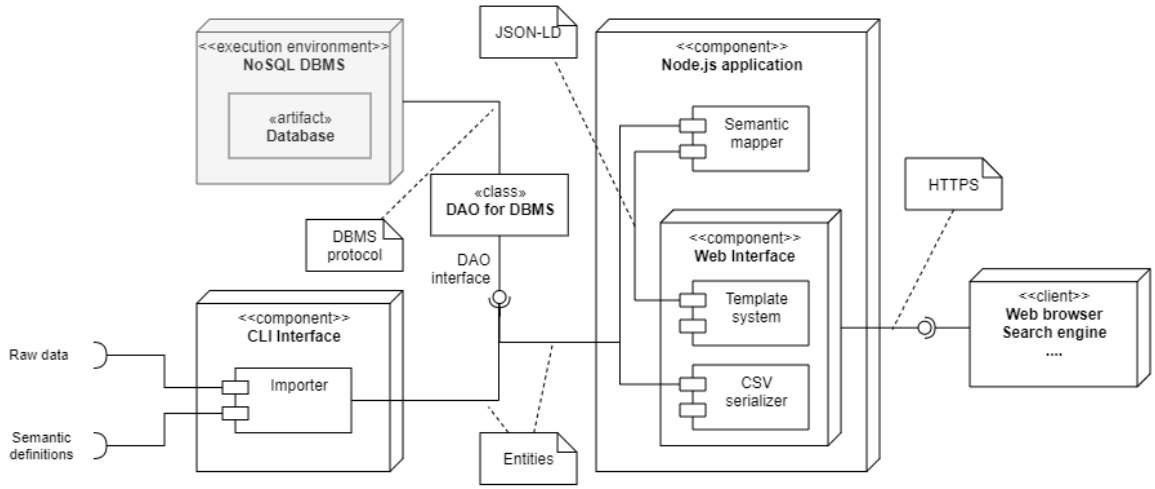
- Umožní import datových entity – surová data v CSV + jejich sémantický popis v definovaném formátu
- Při importu entitám přiřadí UUID nebo volitelně použije existující UUID obsažené v importovaných datech
- Uchová entity v nerelační databázi dokumentového typu (nebo bude takové úložiště alespoň emulovat)
- Bude mapovat uchované entity na sémantickou RDF formu (data + metadata)
- Poskytne HTTP(S) rozhraní (endpoint) pro publikaci sémantické formy, a to:
    - V lidsky čitelné podobě (webový portál)
    - Ve strojově čitelné podobě (JSON-LD)
- Poskytne plnou funkčnost při nulové znalosti povahy vstupních dat
- V lidsky čitelné podobě vykreslí k entitám základní informace (pouze na základě sémantiky dat) s tím, že bude možné podporu dalších jednoduše doprogramovat
- Umožnit reexport surových dat (zobrazení entit zpět do CSV)
- Bude podporovat kontejnerizované nasazení a využití externích směrovacích služeb pro persistentní URL

- Obecně bude podporovat postupnou FAIRifikaci a zachová se správně i v situaci, kdy některým prvkům dat schází sémantický popis, ať již částečně nebo zcela.
- Bude splňovat všechny v současnosti měřitelné požadavky FAIR (úspěšně projde validátorem)

3.2.2 Technologie a návrh hlavních komponent

Návrh řešení vychází ze stanovených požadavků a svou povahou odpovídá tzv. mikroslužbě. Volba technologií a zvolená architektura (či spíše do jisté míry absence architektonického formalismu) reflektuje experimentální podstatu aplikace – účelem je relativně rychle vyvinout agilní řešení, které umožní aplikovat, demonstrovat a ověřit závěry vzešlé z předchozích částí této práce a ukázat tak cestu, kterou se lze ubírat při realizaci FAIR datových zdrojů. Z toho důvodu v návrh absentují prvky, které by byly typické pro plné produkční nasazení jako je např. robustní logování, kešování či propracovaná CI/CD pipeline. Na druhé straně se však toto řešení nedělá příliš velké kompromisy po stránce dodržení postupů v odvětví obvyklé dobré praxe (best-practices) a jeho flexibilní podoba nebrání dalšímu rozvoji a pozdějšímu povýšení na produkční úroveň plně škálovatelné mikroslužby a integraci do většího celku – jednotlivé prvky jsou volně vázané, avšak s formálně a striktně popsaným rozhraním (v celé aplikaci je důsledně využit TypeScript).

Řešení je založeno na technologii Node.js, která je určena pro realizaci svižných, asynchronních a škálovatelných síťových aplikací [46][47]. Řešení bylo v návrhu rozděleno do tří základních komponent, viz Obrázek 12.



Obrázek 12: Schéma hlavních komponent

Hlavní komponentou je samotná Node.js aplikace postavená nad frameworkem Express, což je minimalistický webový framework poskytující základní nástroje pro tvorbu aplikací s HTTP(S) rozhraním. Ten je doplněn dalšími podpůrnými knihovnami, zejména pak template systémem EJS pro vykreslování lidsky čitelného výstupu na základě sémantických dat a middlewarem Helmet pro realizaci základního zabezpečení.

Samostatnou komponentou je CLI interface (rozhraní příkazového řádku) postavený nad frameworkem Vorpal, jehož funkcí je přijímání a import dat. Poskytované CLI rozhraní funguje jak v interaktivním, tak neinteraktivním režimu, proto je možné jej obsloužit buď manuálně nebo programaticky napojením na jiné systémy, příp. zapojit do existujících pipeline pro zpracování dat.

Obě zmíněné komponenty komunikují s NoSQL (nerelační) databází dokumentového typu prostřednictvím pomocné třídy DAO (data access object), které implementuje příslušný interface. Volba konkrétního databázového systému (DBMS) je v tomto případě prakticky libovolná, za podmínky že zvolený databázový systém umožní uložení JSON objektů (např. v podobě JSON, YAML nebo BSON dokumentů), jejich zpětné reprodukování na základě jejich ID a umožní indexaci nad vybranými poli těchto dokumentů. Takto laxní podmínky byly zvoleny záměrně tak, aby řešení umožnilo integraci s širokou škálou dostupných dokumentových či hybridních řešení, jako je např. MongoDB, ElasticSearch, RavenDB, PostgreSQL a další. Předpokládá se, že v případě praktického využití již pravděpodobně bude určité storage řešení nasazeno a při příp. volbě nového je ponechán prostor pro zohlednění specifických požadavků integrátora. Nemusí jít přitom nutně o lokální databázový systém – díky jednoduchým požadavků je možné místo něj konzumovat datové API. Pro samotnou integraci pak postačí implementovat jediný jednoduchý interface DAO a tuto implemntaci předat ostatním komponentám k užívání. V dodaném řešení je pro názornost přiložena minimalistická implementace DAO využívající jako úložiště prostý JSON soubor.

3.2.3 Formát pro popis sémantiky a logika mapování

Klíčovým prvkem celého navrženého řešení je způsob vyjádření sémantické informace uchovávaných dat. Jak je dobře patrné ze schématu na Obrázku 12, vstupují surová data a sémantický popis do aplikace odděleně a odděleně jsou i uchovány v databázovém systému.

Za účelem vyjádření sémantického popisu musí být tedy sestrojen vlastní výrazový aparát, jehož účelem je popsat transformaci surových dat do sémantické formy RDF formátu JSON-LD.

Požadavky

Vzhledem k velmi proměnlivé povaze popisovaných dat je nutné podporovat vyjádření i netriviálních konstrukcí. Na základě analýzy dostupných zdrojů byly definovány požadavky, které musí takto univerzální výrazový aparát splňovat a definovány konstrukce, které musí být schopny zaznamenat alespoň:

1. IRI prefixy použitých ontologií a odkazů na externí subjekty
2. jeden nebo více IRI RDF tříd subjektu který entita představuje a určit jeden jako typ základní
3. pro každý prvek CSV dat:
   3.1. definovat predikáty, které jeho hodnota, příp. každý element z kolekce, je-li hodnota kolekcí, představuje vzhledem k subjektu, který představuje entita, a to pomocí jednoho nebo více IRI
   3.2. umožnit definovat projekci jeho hodnoty, příp. každého elementu z kolekce, je-li hodnota kolekcí, na jinou hodnotu, kterou může být:
      3.2.1. jeden nebo více RDF literálů (číslo nebo řetězec), nebo
      3.2.2. jeden nebo více IRI, nebo
      3.2.3. jeden nebo více expandovaných IRI objektů, nebo
      3.2.4. libovolná kombinace předchozích možností
   3.3. umožnit nahrazení jeho hodnoty, příp. každého elementu z kolekce, je-li hodnota kolekcí, jedním nebo více prázdnými uzly RDF, přičemž pro každý takový uzel definovat:
      3.3.1. IRI řetězec, které představuje predikát, jenž tento uzel pojí se subjektem, který představuje entita
      3.3.2. jeden nebo více IRI, které představují predikát, jenž tento uzel pojí s nahrazovanou hodnotou
      3.3.3. jeden nebo více IRI RDF tříd subjektu který nulový uzel představuje
      3.3.4. projekci hodnoty stejným způsobem jako v bodě 3.2
      3.3.5. jednu nebo více statických RDF vět s predikátem ve formě IRI, jejichž objekt může být:
         3.3.5.1. jeden nebo více RDF literálů (číslo nebo řetězec), nebo
         3.3.5.2. jeden nebo více IRI, nebo
         3.3.5.3. jeden nebo více expandovaných IRI objektů, nebo
         3.3.5.4. libovolná kombinace předchozích možností

4. pro subjekt, který entita představuje, jednu nebo více statických RDF vět s predikátem ve formě IRI, jejichž objekt může být:
   4.1. jeden nebo více RDF literálů (číslo nebo řetězec), nebo
   4.2. jeden nebo více IRI, nebo
   4.3. jeden nebo více expandovaných IRI objektů, nebo
   4.4. libovolná kombinace předchozích možností
5. u statických vět, které popisuje předchozí bod 4, určit, zda mají sémantickou roli dat nebo metadat
6. všechny přítomné IRI umožnit zapsat v prefixové či absolutní formě

Implementace

Výrazový aparát, který splňuje výše uvedené požadavky je v podobě TypeScript definice[6] zachycen na Obrázku 13.

```typescript
export type ExpandedIri = { [key: string]: string } & {
    "@id": string
}

export type RawValue = string | number | ExpandedIri

export type NamespaceAliases = {
    [key: string]: string
}

export type StaticIdentity = {
    identity: string,
    value: RawValue | RawValue[]
}

export type PrimitiveMapping = {
    valueIdentity: string | string[],
    valueProjection?: { [key: string]: RawValue | RawValue[] }
}

export type NestedMapping = PrimitiveMapping & {
    relationIdentity: string,
    relationType: string | string[],
    staticIdentities?: StaticIdentity[]
}

export type AttributeMapping = string | PrimitiveMapping | NestedMapping

export type SemanticMappingDefinition = {
    namespaceAliases: NamespaceAliases,
    baseType: string,
    additionalTypes: string[],
    mappings: {
        [key: string]: AttributeMapping | AttributeMapping[]
    },
    staticIdentities?: StaticIdentity[],
    additionalMetadata?: StaticIdentity[]
}
```

Obrázek 13: TypeScript definice soustavy typů výrazového aparátu pro popis sémantiky

Základním prvkem je typ `SemanticMappingDefinition`, který agreguje jednotlivé elementy definice. Podmínku bodu 1 realizuje jeho atribut `namespaceAliases` jehož hodnotou

[6] Vzhledem k netriviálnímu chování popisované struktury by bylo kontraproduktivní pokoušet se ji vyjádřit schématický. Definice používá některé mírně pokročilé konstrukce jazyka TypeScript pro jejichž lepší pochopení lze čtenáři doporučit prostudování sekce *Union and Intersection Types* dokumentace jazyka TypeScript [48].

je objekt stejnojmenného typu `NamespaceAliases`. Pro vyjádření typů RDF tříd z podmínky bodu 2 jsou určeny atributy `baseType` a `additionalTypes`.

Mapování jednotlivých prvků CSV dat (podmínky bodu 3) uchovává atribut `mappings`, jehož obsahem je objekt, kde klíč odpovídá názvu sloupce CSV dat a hodnota uchovává jednu nebo více definic příslušného mapování – prvků odpovídající typu `AttributeMapping`. Prvek mapování může být řetězec predikátu (podmínka bodu 3.1) nebo objekt odpovídající typu `PrimitiveMapping` či `NestedMapping`.

Typu `PrimitiveMapping` umožňuje volitelně definovat projekce hodnot (podmínky bodu 3.2 a podřízených), přičemž definici predikátů přesouvá do atributu `valueIdentity` (zachování podmínky bodu 3.1). Typ `NestedMapping` rozšiřuje `PrimitiveMapping` (čímž je implicitně splněna podmínky bodů 3.3.2 a 3.3.4), umožňuje vytvářet podřízené prázdné RDF uzly (podmínka bodu 3.2) a definovat v atributu `relationEntity` vazební predikát k nadřazenému uzlu subjektu (podmínka bodu 3.3.1). RDF třídy nově vytvářeného uzlu uchovává atribut `relationType` (podmínka bodu 3.3.3).

Jak u jednotlivých předpisů typu `NestedMapping`, tak u základního top-level typu `SemanticMappingDefinition` je volitelně možné v atributech `staticIdentities` a `additionalMetadata` definovat pole s objekty odpovídající typu `StaticIdentity`, které umožňují zapsat statické RDF věty v roli dat, resp. metadat (podmínky bodů 3.3.5 a podřízených, 4 a 5) – v těch atribut `identity` uchovává IRI predikátu a atribut `value` jím připojený RDF objekt (RDF literál či další subjekt).

Podmínka bodu 6 je splněna implicitně neboť obě formy IRI jsou pro zjednodušení uchovány jako řetězce ve formátu odpovídajícímu konvencím *<prefix>:<zbytek IRI>* s prefixem odděleným od zbytku IRI znakem dvojtečky, tedy stejným způsobem jako používá JSON-LD.

Možné nedostatky a jejich řešení

Výše definovaná a popsaná struktura umožňuje vytvořit poměrně komplexní RDF strukturu, aniž by bylo nutné zasahovat do podoby zdrojových dat jako takových (!) a popsat *téměř* jakoukoliv sémantiku. Ukázka sémantického popisu reálných dat sestaveného na základě výše popsané implementace viz část 3.3 této práce.

Jedinou slabinou, kterou se v průběhu testování podařilo identifikovat je absence možnosti agregační transformace, tedy slučování CSV sloupců napříč daty do nulových uzlů. Jelikož se ale v takovém případě jedná de facto o nové a samostatné datové entity, nebyla by

implementace takové transformace nejvhodnější. Situace, kdy by byla agregační transformace nutná by ukazovala na nevhodnou podobu primárních dat, ve kterých se míchá více analytických typů entit s tím, že korekce tohoto stavu již nutně znamená netriviální zásah do zdrojových dat, což by porušilo stanovenou podmínku.

3.2.4 Princip funkce – jednotlivé komponenty v detailu

Import

Pro možnost importu aplikace na vstupu přijímá přes CLI realizované pomocí komponenty importéru dva soubory: data ve formátu CSV a sémantický popis ve formátu YAML nebo JSON.

CSV data jsou očekávána ve variantě Excel-kompatibilní (oddělené středníkem, kódování UTF-8 BOM, první řádek představuje popis sloupců). Na nativní typy se překládají pouze číselné hodnoty, pokusy o převod data a času jsou explicitně zakázány a tyto hodnoty se záměrně ponechávají jako originální řetězce. V případě, kdy je v CSV souboru přítomno více sloupců s totožným názvem, sdruží se hodnoty těchto sloupů do pole. Výsledkem CSV parsování je pole objektů, kde klíč tvoří název sloupce a hodnota může být buď řetězec, číslo nebo pole řetězců nebo čísel.

Sémantický popis přijímaný při importu současně s CSV daty může být ve formátu JSON nebo YAML. Tyto formáty jsou obsahově a významově ekvivalentní, rozdíl je pouze v syntaxi. Podpora YAML byla zařazena z toho důvodu, že tento formát je standardem v definicích datových pipelines (a také je vhodnější při příp. ruční zápis), JSON je tradiční a běžně používaný formát pro strojový výstup. Formát sémantického popisu byl kvůli specifickým potřebám definován vlastní a je formálně popsán typovou definicí TypeScript `SemanticMappingDefinition` popsaným v předchozí části 3.2.3 této práce. V momentě importu dochází pouze k parsování JSON nebo YAML souboru a rutinní kontrole struktury oproti výše zmíněné TypeScript definici. Tato definice je samostatně přenositelná do vývoje jiných externích aplikací, kde ji lze využít jako podklad pro generování sémantické definice ve formátu přijímaném importérem.

Po parsování CSV dat a JSON/YAML sémantiky provede importér přiřazení lokálního identifikátoru typu UUIDv4 a vzniknou entitu zapíše do databáze. Pokud již data již UUIDv4 obsahují, je volitelně možné (pomocí přepínače `-g`) určit název sloupce který se má jako lokální identifikátor použít. Importér v tom případě provádí zápis stylem *upsert (update+insert)* – tedy aktualizuje existující záznamy se stejným identifikátorem (*update*) a

dosud neexistující vloží (insert). Při této operaci uloží importér společně se záznamem časovou značku vytvoření či aktualizace záznamu.

Použití importéru je popsáno v nápovědě dostupné přes CLI standardním příkazem `help`, jehož výpis je pro ilustraci přiložen na Obrázku 14.

```
app-cli$ help

  Commands:

    help [command...]                  Provides help for a given command.
    exit                               Exits application.
    import [options] [csv] [semantics]  Imports data from CSV according to given JSON or
                                          YAML semantic description

app-cli$ help import

  Usage: import [options] [csv] [semantics]

  Imports data from CSV according to given JSON or YAML semantic description

  Options:

    --help       output usage information
    -g <column>  Use GUID from column of data
```

Obrázek 14: Popis použití CLI (výstup příkazů help)

Datový model

Datový model struktur uchovávaných aplikací v databázi je triviální – tvoří jej kolekce entit jediné třídy jejíž podoba je popsána definicí TypeScript interface na Obrázku 15. Entitu představuje objekt, který sestává pouze ze čtyř atributů:

- ***id*** uchovávající UUIDv4 entity,
- ***attributes*** které představují deserializované CSV,
- ***semantics*** kde je uložen sémantický popis,
- ***meta*** kde jsou uloženy provozně technické informace – nejde o metadata záznamu, ta se odvozují od sémantického popisu! (aktuálně pouze datum založení a poslední úpravy, v budoucnu je zde připravené místo pro uchování oprávnění a dalších provozních či telemetrických údajů)

```
import {SemanticMappingDefinition} from "@entities/SemanticMappingTypes";

export type Attribute = string | number | string[] | number[];

export type EntityAttributes = {[key:string]: Attribute}

export type EntityMeta = {
    created: string,
    updated: string,
}

export interface IEntity {
    // structural
    id: string | null,

    // raw data layer
    attributes: EntityAttributes,

    // semantic layer
    semantics: SemanticMappingDefinition

    // metadata layer
    meta: EntityMeta
}
```

Obrázek 15: TypeScript definice struktury uchovávané aplikací v databázi

Volba této podoby datového modelu vychází z požadavků a má tyto pozitivní praktické důsledky:

- Je zachována původní podoba vstupních dat, což umožňuje zpětný reexport.
- Každý záznam je zcela nezávislý, protože si s sebou nese všechny informace včetně definice sémantiky.
- Ač to na první pohled není patrné, neomezuje datový model libovolné propojování a provazování záznamů s pomocí patřičných sémantických predikátů (lze tvořit i složité stromy, a orientované grafy).

Nevýhodou tohoto způsobu jsou zvýšené nároky na výpočetní kapacitu (pro získání sémantické RDF podoby je nutné vždy provést mapování atributů a sémantiky) a úložnou kapacitu (možné duplicitní uložení sémantiky). Obě je však v možné později řešit jednak na úrovni zvoleného DBMS (automatická deduplikace, indexace RDF podoby), příp. vřazením cache v rámci komponenty Node.js aplikace (v úvahu připadá např. Redis nebo Elasticsearch).

Web interface

Hlavní komponenta Node.js aplikace poskytuje několik endpointů webového rozhraní (v kontextu použitého frameworku *Express* nazývaných jako routy). Routy jsou prefixovány

pomocí konfigurace zavedeným modelem *.env* (čti „dotenv“) souborů dosazením proměnné *IRI_PREFIX_OF_SELF*, které jsou oddělené pro produkční a vývojové prostředí (což v důsledku umožňuje mj. v sémantický datech detekovat odkazy aplikace sama na sebe a registrovat persistentní URL prefix externích služeb).

Přehled dostupných rout a jejich účel je pro základní přehled vypsán v Tabulce 2. Je zde uveden také typ těla HTTP odpovědi (*Content-type*). Routy poskytují jak lidsky čitelnou podobu HTML, tak strojově čitelnou sémantickou podobu JSON-LD a tabulkovou CSV. Je však nutné zdůraznit, že i ve všech případech, kdy je obsahem odpovědi HTML jsou sémantická data v podobě JSON-LD vložena (embedována) pomocí tzv. data bloku dle specifikace [49]. Volitelné parametry uvedené v tabulce jsou technicky URL GET parametry (označované též jako query parametry). Všechny routy reagují na HTTP dotazy typu GET a OPTIONS – slouží tedy pouze pro čtení.

Routa	**Content-type**	**Volitelné parametry**
Obsah / popis		
Hlavní routy		
IRI_PREFIX_OF_SELF/entity	text/html	id[] type[]
Výpis seznamu všech entit v databázi. Volitelnými parametry lze omezit na jeden nebo více konkrétních sémantický typy nebo jednotlivých		
IRI_PREFIX_OF_SELF/entity/csv	text/csv	id[] type[]
Reexport všech entit v databázi do původní CSV podoby. Volitelnými parametry lze omezit na jeden nebo více konkrétních sémantický typy nebo jednotlivých		
IRI_PREFIX_OF_SELF/entity/<id>	text/html	
Mapovaná metadata + data jedné konkrétní entity.		
IRI_PREFIX_OF_SELF/entity/<id>/data	application/ld+json	
Mapovaná data jedné konkrétní entity.		
IRI_PREFIX_OF_SELF/entity/<id>/csv	text/csv	
Reexport jedné konkrétní entity do původní CSV podoby.		
Pomocné routy		
IRI_PREFIX_OF_SELF/	text/html	
Homepage		
IRI_PREFIX_OF_SELF/type	text/html	
Výpis seznamu základních sémantických typů entit v databázi.		
IRI_PREFIX_OF_SELF/persistence-policy	text/html	
Poskytuje dokument "Persistence policy" aplikace.		

Tabulka 2 - Přehled poskytovaných endpointů a jejich rout

Hlavní, entitní routy sloužící k získávání jednotlivých záznamů byly doplněny o routy poskytující výpisy (indexy). Ty jsou jednak nutné pro to, aby byly jednotlivé záznamy objevitelé pro webové crawlery, jednak díky volitelným parametrům umožňují realizovat získávání souborů entit, jejichž podoba může být různými způsoby vyjádřena pomocí predikátů sémantických informací (např. „dataset má součásti“, „akademická práce má

seznam spoluautorů“, „k nálezu přísluší více měření“ apod.). Pokročilé filtrování na endpointu v tomto případě není vůbec potřeba, neboť sémantická RDF data jsou sama filtrem – kolekcí IRI prvků, které danému predikátu odpovídají. V kombinaci s možností získat výstup také ve formátu CSV je možné realizovat v lidsky čitelném rozhraní takříkajíc „na jeden klik“ funkce typu „stáhni mi celý dataset“, a to pouze a jen (!) na základě sémantické informace bez jakékoliv předchozí úpravy systému. Samotná CSV serializace je triviální operací, inverzní k CSV deserializaci popsané výše v úvodu části 3.2.4 této práce.

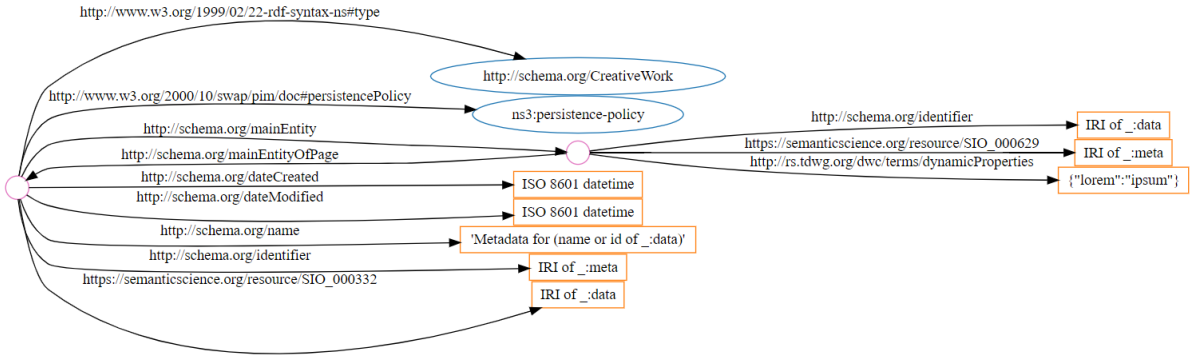
Kromě hlavních rout je pro úplnost zmínit ještě routy pomocné, které plní převážně funkci zlepšení uživatelského zážitku. Jde především o homepage a seznam základních sémantických typů entit, které jsou v řešení uloženy (tvoří se na základě reflexe sémantických dat).

Zajímavostí je speciální routa pro politiku udržitelnosti (Persistence policy), což je prostý dokument či článek, který popisuje jak se tvůrci a provozovatelé řešení samotného staví či jaký závazek mají toto řešení a data v něm dlouhodobě udržet. Tato informace nevztahuje k žádné entitě ale aplikaci samotné. Dostupnost takového dokumentu není přísně vzato z technický důvodů nutná, vyžaduje jej však metodika FAIR.

Mapper

Pomyslným jádrem celého řešení je komponenta sémantického mapperu. Tvoří ji kolekce tzv. arrow funkcí, které slouží k jednosměrnému převodu entity datového modelu uloženého v databázovém systému na její sémantické vyjádření v objektu formátu JSON-LD. Modul exportuje dvě funkce `mapDataToJsonLd` a `mapMetadataToJsonLd`, které vrací objekty reprezentující data, resp. metadata entity předané na jejich vstupu. Připomeňme, že entita obsahuje surová data deseriaizovaná z řádku CSV + sémantický popis zavedený v předchozí podkapitole.

Interní logika komponenty vychází z modelu sémantického popisu a východisek teoretické části této práce. Nad rámec prostého mapování na základě sémantického popisu modul sestavuje a doplňuje během mapování IRI a minimální technické predikáty nutné pro korektní provázání objektu dat a metadat z ontologií *Schema.org* (http://schema.org/, prefix `sorg`), *Semanticscience Integrated Ontology* (https://semanticscience.org/resource, prefix `sio`) a *Darwin Core* (http://rs.tdwg.org/dwc/terms, prefix `dwc`). Tyto predikáty tvoří v podstatě minimální kostru RDF grafu, který mapováním vznikne – její zobecněná reprezentace zakreslená schematicky a zapsaná pomocí RDF N-Triples je zachycena na Obrázku 16 (subjekty dat a metadat zde reprezentují prázdné uzly `_:data` resp. `_:meta`).




```
_:meta <http://schema.org/identifier> "IRI of _:meta"
_:meta <https://semanticscience.org/resource/SIO_000332> "IRI of _:data"
_:meta <http://schema.org/mainEntity> _:data
_:meta <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/CreativeWork>
_:meta <http://schema.org/name> "'Metadata for (name or id of _:data)'"
_:meta <http://schema.org/dateCreated> "ISO 8601 datetime"
_:meta <http://schema.org/dateModified> "ISO 8601 datetime"
_:meta <http://www.w3.org/2000/10/swap/pim/doc#persistencePolicy>
↪<http://purl.org/dktr/fair/persistence-policy>
_:data <http://schema.org/identifier> "IRI of _:data"
_:data <http://schema.org/mainEntityOfPage> _:meta
_:data <https://semanticscience.org/resource/SIO_000629> "IRI of _:meta"
_:data <http://rs.tdwg.org/dwc/terms/dynamicProperties> "{\"lorem\":\"ipsum\"}"
```

Obrázek 16: Zobecněná reprezentace minimální kostry RDF grafu dat a metadat

Subjekty reprezentující data a metadata jsou oboustranně provázány predikáty `sorg:mainEntity` / `sorg:mainEntityOfPage` a ekvivalentními `sio:SIO_000629` („is subject of“) / `sio:SIO_000332` („is about“). IRI obou subjektů je dodatečně specifikováno predikátem `sorg:identifier`.

Kromě toho je subjektu, který reprezentuje metdatdata, přiřazena třída `sorg:CreativeWork` umožňující specifikovat provozní metadata (`sorg:dateCreated` a `sorg:dateModified`) a také automaticky generovaný název ve tvaru *‘Metadata for [název nebo id subjektu dat]‘* (`sorg:name`), přičemž zástupka v názvu se dynamicky dovozuje ze sémantických prvků samotných dat, které odpovídají některému predikátu typu „má jméno“ z nejběžněji používaných ontologií Schema.org, Darwin Core a Dublin Core.

Subjektu, který reprezentuje metadata, je dále přiřazen speciální predikát pro provázání na Presistence policy dokument aplikace – datového zdroje (IRI je automaticky generováno na základě konfiguračního parametru *IRI_PREFIX_OF_SELF*).

Mapper se dokáže též vypořádat se situací, kdy se v datech entity objeví atributy, které nemají sémantický popis. Výchozí postup RDF by diktoval takové atributy zcela ignorovat – to však vzhledem k požadavku na podporu postupné FAIRifikace nelze akceptovat. Z toho

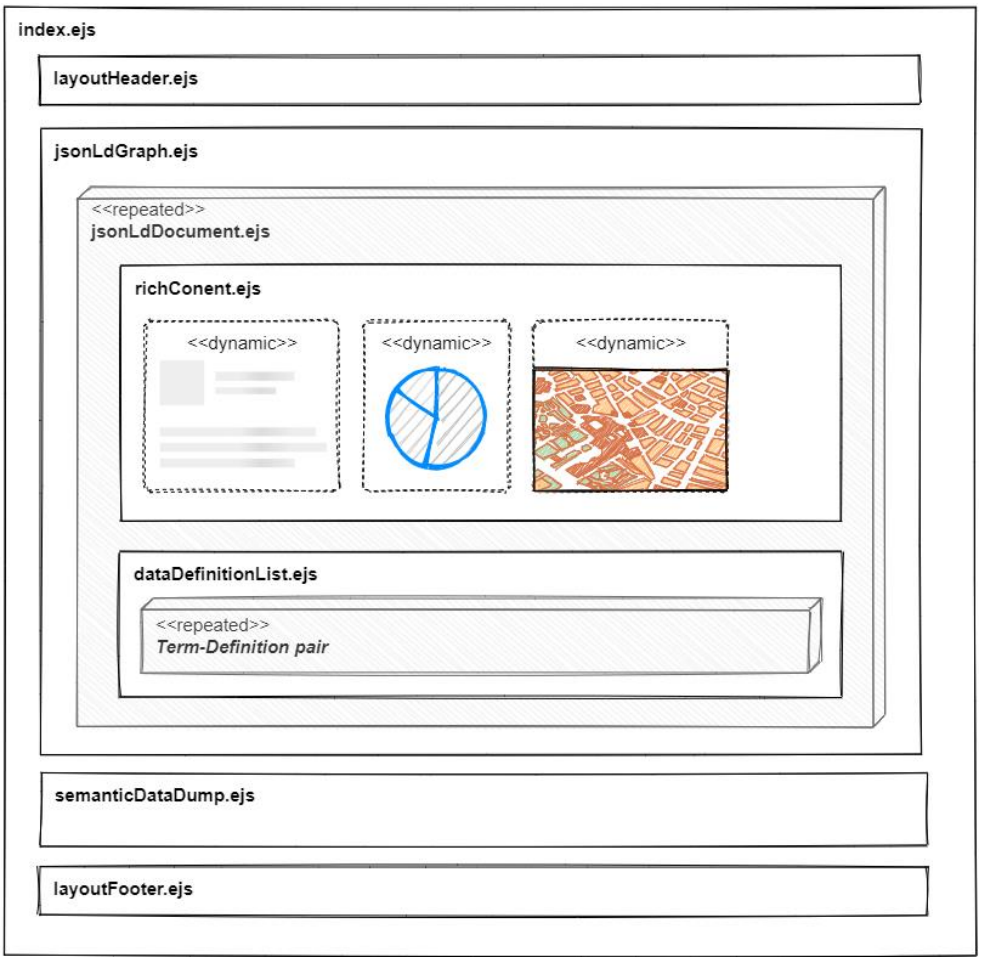
důvodu jsou atributy dat, kterým schází sémantický popis, sdruženy do JSON řetězce a přiřazeny k subjektu, který reprezentuje data, pomocí predikátu `dwc:dynamicProperties`.

Z čistě technického pohledu vychází implementace interní logiky komponenty z programovacího modelu *map-reduce-filter* [50] a kombinuje imperativní a deklarativní styl. Pro detailní studium algoritmu lze čtenáře odkázat přímo na zdrojový kód komponenty v souboru `/src/shared/mapper.ts` který je součástí netextové přílohy této práce.

Systém dynamických šablon

Návrh struktury šablon (template system) je v rámci řešení plně dynamický. Pro vykreslení všech obrazovek existuje jediná šablona (`index.ejs`), do které se předávají pouze a jen (!) sémantická data ve formátu JSON-LD. Je třeba zdůraznit, že tímto způsobem jsou kromě samotných entit předávány i ostatní prvky grafického rozhraní, a to i včetně statického obsahu jakou jsou bloky textu, tlačítka, záhlaví apod. – všechny tyto prvky se totiž dají sémanticky popsat stejným způsobem jako data a metadata entit samotných (např. pomocí tříd ontologie *Schema.org* `sorg:WebPage` nebo `sorg: WPHeader`). Z hlediska šablon tak není rozdílu mezi vlastními a cizími strukturami a vykreslovací logika se řídí výhradně sémantickými predikáty, které jsou jim v JSON-LD objektu předány. Celé front-end řešení má tedy nulovou předchozí znalost dat, která mu na vstupu přijdou v důsledku čehož je dokonale univerzální a je schopno korektně prezentovat jakákoliv sémantická data, alespoň v základní podobě.

Tohoto silně dynamického chování je docíleno s pomocí hierarchie podmíněně vykreslovaných dílčích šablon (tzv. partials). Hlavní komponenty této hierarchie zachycuje mockup diagram na Obrázku 17.



Obrázek 17: Mockup a hierarchie systému šablon – základní prvky

Základní kostru hlavní šablony (`index.ejs`) tvoří:

- **záhlaví a zápatí** (`layoutHeader.ejs` a `layoutFooter.ejs`) - obsahují základní stavební prvky HTML a statické prvky designu, vkládají skripty a stylopisy a také zpřístupňují sémantická data pro strojové čtení pomocí tzv. databloku v tagu <head>,
- **oblast dat** (`jsonLdGraph.ejs`) - v cyklu předává jednotlivé prvky JSON-LD grafu k dynamickému vykreslení (každý pomocí `jsonLdDocument.ejs`)
- **výpis sémantických dat** (`semanticDataDump.ejs`) - slouží pouze pro demonstrační účely a přívětivě ukazuje surová data „pod povrchem“

Šablony oblasti dat (`jsonLdDocument.ejs`) již dynamicky rozhodují o své podobě na základě sémantiky předaného JSON-LD dokumentu, zejména pak detekcí predikátu určující ontologickou třídu. V diagramu na Obrázku 17 je zobrazena situace, kdy tento dokument netvoří známý prvek uživatelského rozhraní, ale (třeba i šabloně neznámý) subjekt dat či metadat entity. V takovém případě se dále připojí dílčí šablony definičního seznamu a bohatého obsahu.

Šablona definičního seznamu (`dataDefinitionList.ejs`) se vykresluje v plném rozsahu v každé situaci. Tvoří ji cyklus, který vykreslí definici párů predikát-subjekt. U známých predikátů se šablona pokusí doplnit jejich lidsky čitelné pojmenování, případně nabídne náhledové okénko definice, kterou predikát představuje (vzdáleným dotazem na IRI predikátu, pokud tento má lidsky čitelné vyjádření). U objektů, jejichž podstatu tvoří odkaz, se dynamicky vytváří interní i externí prokliky, přičemž šablona je schopna rozpoznat kvalifikované IRI na jiné, i externí subjekty.

Šablona bohatého obsahu (`richContent.ejs`) je nejzajímavějším prvkem celého návrhu. Její obsah zcela závisí na tom, jaké sémantické predikáty příslušný JSON-LD dokument obsahuje. Bohatý obsah tvoří kolekce sémantických karet, inspirovaný svou podobou např. *Google rich results* [51]. Jde o modulární systém, kdy každá karta hledá v sémantickém obsahu predikáty, kterým rozumí a je schopna jejich obsah přívětivě vykreslit.

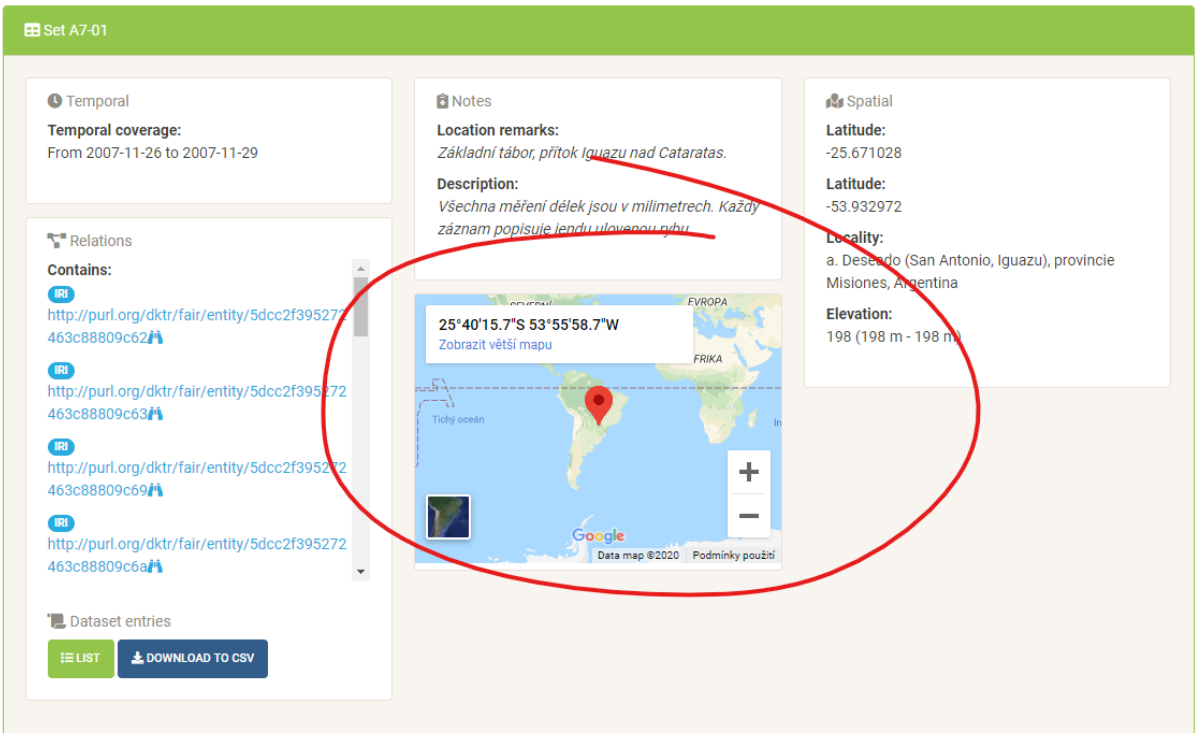
Příkladem takové karty je *mapa*, která zareaguje, pokud se v sémantických datech objeví predikáty z ontologií *Schema.org* nebo *Darwin Core* popisující zeměpisné souřadnice nebo slovní popis lokality, protože obojí umí vykreslit s pomocí embedded Google mapy. Nezkrácený (!) zdrojový kód této karty je na Obrázku 18 a její vykreslení na Obrázku 19.

Díky dynamické povaze šablon je možné postupně přidávat další karty, či těm stávajícím rozšiřovat sémantický repertoár společně s tím, jaké specifické ontologie jsou momentálně populární, čímž je možné pružně reagovat jak na změny v demografii uživatelské základny, tak zajistit dopřednou kompatibilitu s ontologiemi, které buď teprve vzniknou, nebo si svou popularitu získají v budoucnu. Je vhodné zde připomenout, že absence podpory karet bohatých výsledků nemá žádný negativní vliv na schopnost řešení prezentovat i zcela neznámou sémantiku.

Řešení v okamžiku uzávěrky této práce obsahuje sémantické karty schopné porozumět vybraným údajům: časovým, lokačním, mapovým, licenčním, taxonomickým, o měřeních, kvantitativním, vyjadřujícím vazby a vztahy, údajům o osobách a představujících poznámky.

```
<%
const lat = j['http://schema.org/latitude'] || j["http://rs.tdwg.org/dwc/terms/decimalLatitude"];
const lon = j['http://schema.org/longitude'] || j["http://rs.tdwg.org/dwc/terms/decimalLongitude"];
const verbatimLocality = j['http://rs.tdwg.org/dwc/terms/verbatimLocality'];
%>
<% if((lat && lon) || verbatimLocality) { %>
    <div class="card">
        <% const mapQ = (lat && lon) ? lat+','+lon : verbatimLocality;  %>
        <iframe
                width='100%'
                height='250px'
                id='mapcanvas'
                src='https://maps.google.com/maps?q=<%=encodeURIComponent(mapQ)%>&Road-
map&z=1&ie=UTF8&iwloc=&output=embed'
                frameborder='0'
                scrolling='no'
                marginheight='0'
                marginwidth='0'
        >
            <div style='overflow:hidden;'>
                <div id='gmap_canvas' style='height:100%;width:700px;'></div>
            </div>
        </iframe>
    </div>
<% } %>
```

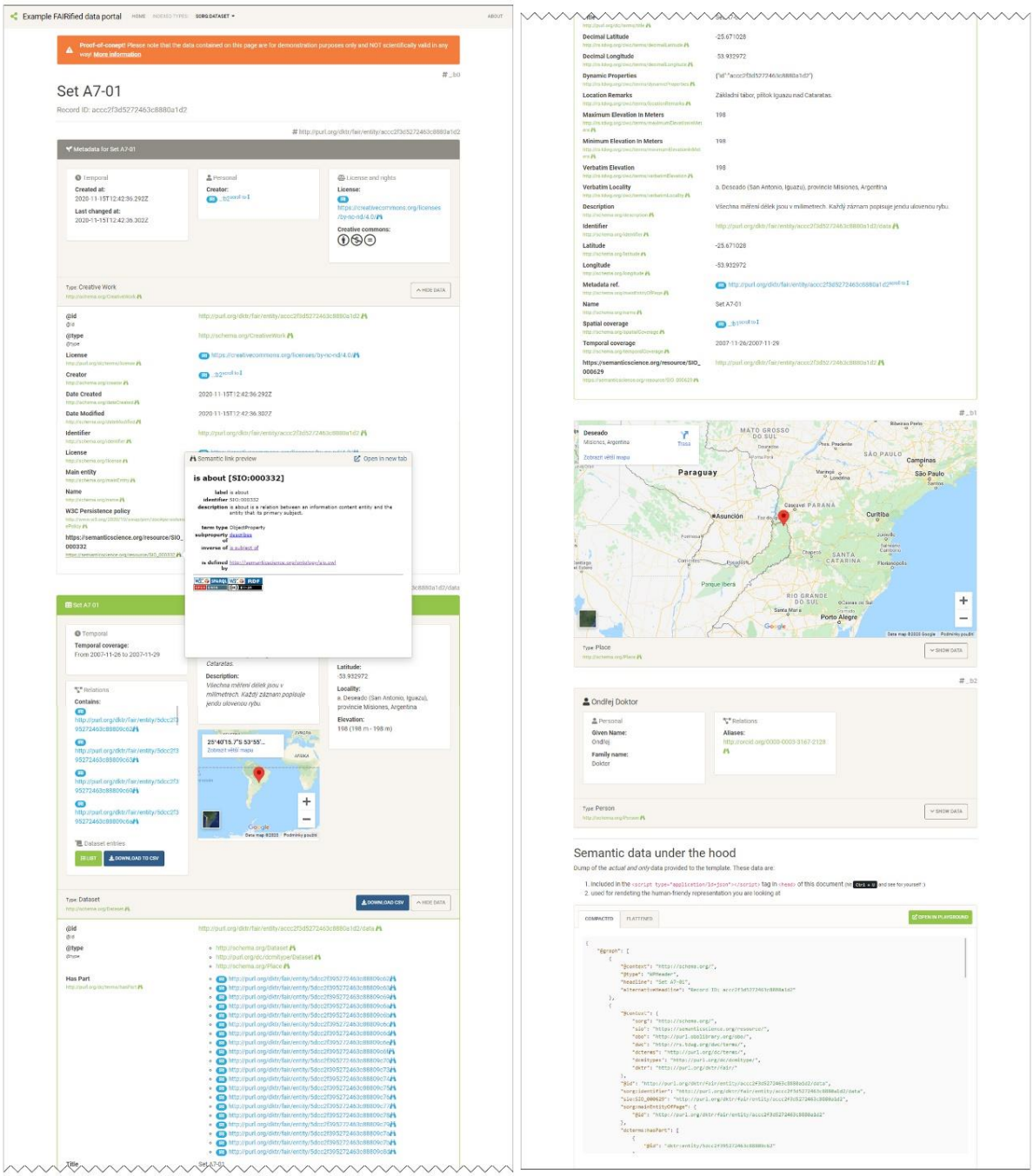
Obrázek 18: Ukázka definice sémantické karty – mapa



Obrázek 19: Ukázka vykreslení sémantické karty – mapa

Uživatelské rozhraní

Grafika uživatelského rozhraní je založena na front-endovém frameworku Bootstrap s upraveným barevným tématem, který je mj. plně responsivní. Realizovaná podoba plně vychází z výše popsaného systému dynamických šablon.



Obrázek 20 Náhled uživatelského rozhraní (full-size sceenshot)

3.2.5 Kontejnerizace a nasazení

Celé řešení je navržené jako zcela standardní *Node.JS* + *TypeScript* aplikace kontejnerizovaná pomocí technologií *Docker* a *Docker Compose*. Díky kontejnerizaci je celé prostředí zapouzdřené, což umožňuje vývoj a nasazení na jakékoliv podporované platformě (Linux, macOS, and Windows, podporovaný cloud), aniž by bylo nutné jakkoliv zasahovat do prostředí hostitele nebo by toto bylo nasazením jakkoliv ovlivněno – veškeré závislosti se instalují přímo do kontejneru při jeho *buildu* (sestavení). Výsledkem procesu buildu je *image* (obraz), který je buď přímo nebo pomocí *container registry service* možné okamžitě nasadit na cílový server – fyzický, virtuální nebo v cloudové službě.

Připraveny jsou oddělené *dockerfile* konfigurace pro vývoj a produkci. Obě jsou založeny na image `node:14-alpine`, přičemž HTTP endpoint je vystaven na portu TCP 8081 (směrování na standardní porty činí proxy server / load balancer cílového prostředí). Ve vývojové konfiguraci je k dispozici vzdálený debugging na portech TCP 9229 (komponenta hlavní aplikace) a TCP 9230 (komponenta CLI) s automatickou rekompilací pomocí nástroje *nodemon*. Zaveden je též linter *eslint*. Vývoj, build a řízení závislostí řídí pro celé řešení standardní nástroj *npm* (řešení je zároveň npm balíčkem). Pro usnadnění práce vývojáře je připraven *Makefile* se zkratkami pro základní příkazy související s vývojem a nasazením, podrobný popis použití je k dispozici ve standardním souboru `README.md`.

Důležitým technickým požadavkem FAIR je výhradní užití globálně unikátních persistentních identifikátorů, které je blíže popsáno v části 2.5.1, Krok 7. – Publikovat ve FAIR datovém zdroji, této práce. Pro účely nasazení řešení autor zaregistroval ve službě PURL prefix http://purl.org/dktr/fair, který je také uveden v produkční konfiguraci v proměnné *IRI_PREFIX_OF_SELF*.

Výše uvedený prefix je v době uzávěrky této práce nasměrován na produkční demo prostředí, které bylo autorem zřízeno v cloudové službě *Microsoft Azure*. Sestavený produkční kontejner aplikace je uložen v privátním kontejnerovém registru služby *Docker Hub* odkud služby Azure přebírají a automaticky nasazují poslední build verzi. Kompozice alokovaných Azure zdrojů je triviální – pro účely nasazení byl vytvořena Resource Group a App Service Plan s platformou Linux umístěný v regionu West Europe v tieru F1: Free, v rámci něj pak vlastní single-container App Service nasměrovaná na výše zmíněný kontejnerový registr. Automatická podpora HTTPS je samozřejmostí. Zajímavostí je, že díky této kombinaci zdrojů a dalších výše zmíněných technologií a služeb jsou náklady na provoz nulové.

3.3 Ukázková aplikace zvoleného řešení na datové sadě

Vybaveni teoretický aparátem, metodikou FAIRifikace a technickým řešením z předchozích kapitol je již možné tyto aplikovat na ukázkové datové sadě. Postup je proveden dle doporučení popsaného v části 2.5 této práce. Pro účely této demonstrace byla připravena datová sada, která vychází ze vzorků reálně zpracovávaných dat, která autor během své činnosti s laskavým svolením získal od vědeckých skupin působících na sledované instituci.

Upozornění: *Vzhledem k citlivému problému oprávněných obav z úniku primárních dat byla na podkladových datech provedena úprava obdobná anonymizaci, která je z vědeckého pohledu znehodnotila – byl proveden scrambling, náhodné permutace, nahrazení vybraných hodnot za náhodné, odstranění všech osobních údajů a nahrazení identifikátorů za smyšlené a náhodně generované hodnoty. Takto upravená data tedy nemají sama o sobě žádnou vědeckou hodnotu a informace v nich obsažené jsou z odborného pohledu dané vědecké oblasti záměrně nesprávné! I přesto je však možné je považovat za reprezentativní vzorek skutečných dat, pouze svého druhu „vědecky anonymizovaný", neboť vědecká nesprávnost dat nemá vliv na samotnou strukturu a obecnou sémantiku těchto dat.*

3.3.1 Krok 1 – Získat data

Situace: Na začátku procesu máme k dispozici CSV soubor s tabulkou dvaceti záznamů a TXT soubor, který byl uložen vedle něj ve stejné složce v souborovém systému osobního počítače vědce – ichtyologa, který ústně data popsal jako „tabulku úlovků ryb z jedné expedice".

Na Obrázku 21 je vyobrazen náhled CSV souboru a opis TXT souboru, kompletní opis je k dispozici v Příloze 5 této práce a originální podobě také ve složce `/fixtures/demo` v netextové příloze této práce.

id	document Name	genus	species	specimen_old_id	determination_note	river_name	fishingmethod	standard_length	head_depth	fishingmethod
5dcc2f395272463c88809c62	Pimelodus maculatus (9	Pimelodus	maculatus	http://example.com/o		Deseado	gillnets	37	24	hook&line
5dcc2f395272463c88809c63	Geophagus brasiliensis	Geophagus	brasiliensis	http://example.com/o		Deseado	gillnets	66	17	hook&line
5dcc2f395272463c88809c69	Crenicichla iguassuensis	Crenicichla	iguassuensis	http://example.com/o		Deseado	gillnets	79	28	hook&line
5dcc2f395272463c88809c6a	Crenicichla iguassuensis	Crenicichla	iguassuensis	http://example.com/o		Deseado	gillnets	69	26	hook&line
5dcc2f395272463c88809c6b	Crenicichla lepidota (10	Crenicichla	lepidota	http://example.com/o		Deseado	gillnets	90	18	hook&line
5dcc2f395272463c88809c6c	Crenicichla tapii x tuca	Crenicichla	tapii	http://example.com/o	Crenicichla tapii x tuca ? th	Deseado	gillnets	78	19	hook&line
5dcc2f395272463c88809c6d	Crenicichla tapii (104)	Crenicichla	tapii	http://example.com/o		Deseado	electrofishing	100	24	
5dcc2f395272463c88809c6e	Crenicichla tapii (105)	Crenicichla	tapii	http://example.com/o		Deseado	electrofishing	78	30	
5dcc2f395272463c88809c6f	Crenicichla tesay (106)	Crenicichla	tesay	http://example.com/o		Deseado	electrofishing	53	17	
5dcc2f395272463c88809c70	Crenicichla tesay (107)	Crenicichla	tesay	http://example.com/o		Deseado	castnet	64	15	hook&line
5dcc2f395272463c88809c73	Gymnogeophagus meri	Gymnogeopha	meridionalis	http://example.com/o		Deseado	castnet	77	25	hook&line
5dcc2f395272463c88809c74	Gymnogeophagus meri	Gymnogeopha	meridionalis	http://example.com/o		Deseado	castnet	79	16	hook&line
5dcc2f395272463c88809c75	Crenicichla tesay (112)	Crenicichla	tesay	http://example.com/o		Deseado	gillnets	89	17	hook&line
5dcc2f395272463c88809c76	Crenicichla tesay (113)	Crenicichla	tesay	http://example.com/o		Deseado	gillnets	48	20	hook&line
5dcc2f395272463c88809c77	Crenicichla tesay (114)	Crenicichla	tesay	http://example.com/o		Deseado	gillnets	69	29	hook&line
5dcc2f395272463c88809c78	Crenicichla iguassuensis	Crenicichla	iguassuensis	http://example.com/o		Deseado	gillnets	40	28	hook&line
5dcc2f395272463c88809c79	Crenicichla iguassuensis	Crenicichla	iguassuensis	http://example.com/o		Deseado	gillnets	50	27	hook&line
5dcc2f395272463c88809c7a	Crenicichla lepidota (11	Crenicichla	lepidota	http://example.com/o		Deseado	gillnets	77	22	hook&line
5dcc2f395272463c88809c7b	Crenicichla lepidota (11	Crenicichla	lepidota	http://example.com/o		Deseado	gillnets	86	25	hook&line
5dcc2f395272463c88809c8d	Gymnogeophagus meri	Gymnogeopha	meridionalis	http://example.com/o	or Geophagus	Deseado	gillnets	58	19	hook&line

```
Set A7-01
Datum: 27.11.2007 - 29.11.2007

Lokalita: a. Deseado (San Antonio, Iguazu), provincie Misiones, Argentina
GPS: -25.671028,-53.932972
Popis: Základní tábor, přítok Iguazu nad Cataratas.
Nadmoř. výška: 198 m

Všechna měření délek jsou v milimetrech. Každý záznam popisuje jendu ulovenou rybu.
```

Obrázek 21: Náhled výchozích souborů CSV a TXT

Pro další kroky je z technických důvodů nutné vydestilovat z TXT souboru jednotlivé atributy a převést je do tabulkové podoby – dalšího CSV souboru. Vydestilovanou tabulku v transponované podobě zachycuje Obrázek 22. Poslední řádek TXT souboru velmi volnou formou popisuje sémantiku CSV záznamů – tuto informaci vezme v úvahu v dalších krocích.

jmeno	Set A7-01
obdobi	2007-11-26/2007-11-29
lokalita	a. Deseado (San Antonio, Iguazu), provincie Misiones, Argentina
gps_lat	-25.671028
gps_lon	-53.932972
popis_lokality	Základní tábor, přítok Iguazu nad Cataratas.
nadmorska_vyska	198
poznamky	Všechna měření délek jsou v milimetrech. Každý záznam popisuje jendu ulovenou rybu.

Obrázek 22: Tabulka atributů vydestilovaných z TXT souboru (transponováno)

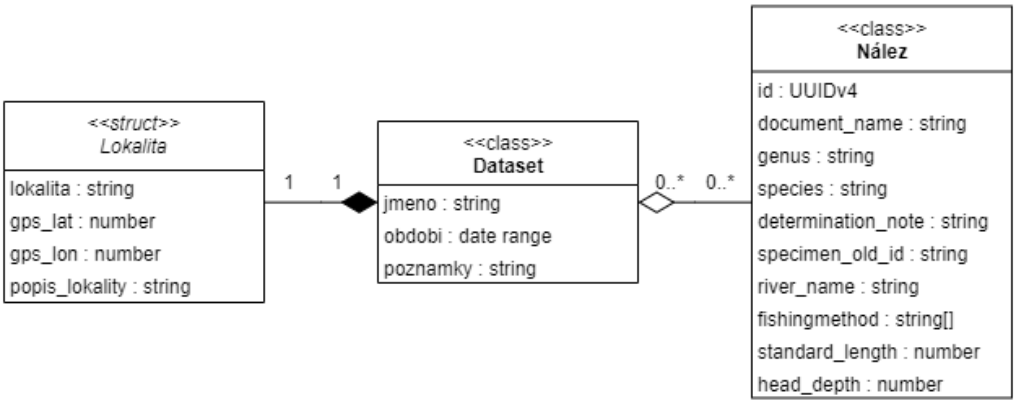
3.3.2 Krok 2. – Analyzovat získaná data

V předaných datech lze identifikovat dvě hlavní analytické entity – třídy, které pracovně můžeme nazvat jako *Nález* a *Dataset*, mezi nimiž je relace. Tato relace má podstatu sdílené agregace s násobností 0 .. * na obou stranách. Volná vazba je zde namístě, protože obě entity nejsou závislé (Dataset může být prázdný, Nález může být součástí více Datasetů nebo existovat samostatně). V datech vydestilovaných z TXT souboru je rovněž možné

identifikovat skupinu atributů, která by dohromady mohli tvořit entitu představující *Lokalitu*. Jelikož vstupní nepočítají s tím, že by existovala samostatná entita lokality, lze tuto pomocnou entitu považovat za strukturu která je pevně svázaná s entitou Datasetu – jejich relace má tedy podstatu agregace s násobností 1 na obou stranách.

Datové typy atributů je možné relativně snadno dovodit ze sloupců vstupních dat – ve většině případů se jedná o primitivní datové typy řetězce a čísla. Povšimněme si však duplicitních sloupců `fishingmethod`, které je možné společně interpretovat jako pole řetězců. V CSV soboru s nálezy je přítomný identifikátor odpovídající formátu UUIDv4, který je později možné znovupoužít pro entitu Nález, pro entitu Dataset identifikátor schází a v pozdějších krocích jej bude nutné doplnit nebo generovat.

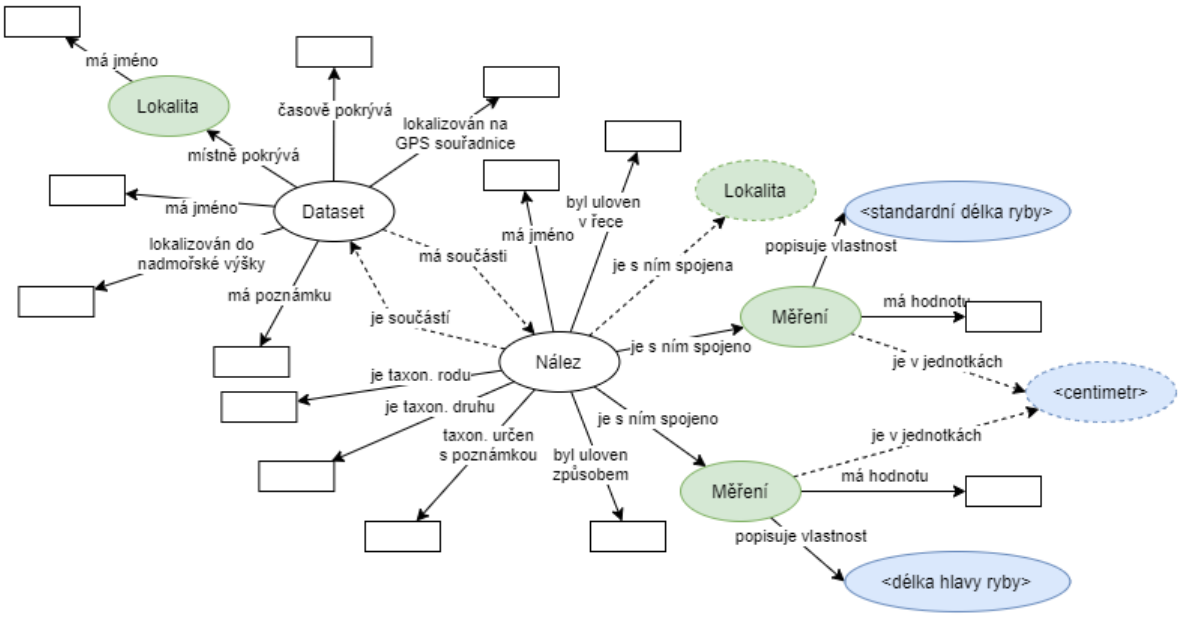
Výsledek strukturální a datové analýzy v podobě UML class diagramu je zachycen na Obrázku 23.



Obrázek 23: Class diagram analytických entit

3.3.3 Krok 3. – Definovat sémantický model

Jako vhodný začátek pro realizaci kroku byl sestaven zjednodušený sémantický model vycházející z interpretace vstupních dat prostředky přirozeného jazyka, který je schématický znázorněn na Obrázku 24.



Obrázek 24: Zjednodušený sémantický model dat

Schéma je sestrojeno na základě analytických entit z předchozího kroku, které zde představují subjekty a objekty (znázorněny jako elipsy bez výplně), primitivní atributy utvořili prosté literály (znázorněny jako obdélníky). Z hlediska sémantického modelování však vyvstala potřeba rozšíření modelu o další, umělé subjekty Měření a Lokalita – ty tvoří v sémantickém modelu prázdné uzly (označeny zeleně).

Za pozornost stojí zejména subjekty Měření – k nim se kromě literálu jejich numerické hodnoty dále vztahují predikáty určující, jakou vlastnost popisují a v jakých jsou jednotkách. Lze předpokládat, že objekty těchto vztahů bude možné zapsat pomocí IRI na externí subjekty standardních ontologií pro měrné jednoty a oblast ichtyologie (označeny modře).

Jednotky těchto měření, a také některé další elementy modelu, však nejsou explicitně uvedeny v datech samotných – jsou statické a je nutné je dovozovat z externího popisu dat. Tyto prvky jsou ve schématu označeny přerušovanou čarou. To se týká mj. samotné vazby mezi Datasetem a Nálezem, který je nutné po přiřazení chybějících identifikátorů korektně specifikovat.

Postupem popsaným v teoretické části práce je pro formální sémantický popis modelu možné dohledat následující zavedené ontologie doplněné specifickou ontologií dané vědecké oblasti:

- Schema.org
- Dublin Core
- DCMI Metadata Terms
- Darwin Core Terms
- Darwin Core IRI-value Term Analogs
- Open Biological and Biomedical Ontology (OBO)
- Semanticscience Integrated Ontology (SIO)
- Fish Ontology (FISHO)

Konečný sémantický model je možné v této fázi rovnou zapisovat pomocí formátu sémantického popisu navrženého řešení[7].

Schématický popis entity Dataset

Vytvoříme nový soubor, ve kterém nejprve definujeme prefixy použitých ontologií:

```yaml
# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases:
  sorg: http://schema.org/
  obo: http://purl.obolibrary.org/obo/
  dwc: http://rs.tdwg.org/dwc/terms/
  dcterms: http://purl.org/dc/terms/
  dktr: http://purl.org/dktr/fair/
```

Dále určíme třídy ontologie a popíšeme jednotlivé atributy vhodnými predikáty. U atributu `lokalita` dochází ke vzniku prázdného uzlu třídy `sorg:Place`, který je s entitou Datasetu spojen predikátem `sorg:spatialCoverage`, přičemž příslušná hodnota dat z tabulky bude připojena ke vzniklému prázdnému uzlu predikátem `sorg:name`:

[7] Pozn: V následujícím textu bude tedy pro názornost sestavován tento popis souběžně s tím, jak bude sémantický model krok za krokem postupně budován. Pro přehlednost jsou ve výpisech YAML kódu irelevantní uzly sbaleny a nahrazeny znaky `(...)`

```yaml
# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases: (...)
baseType: sorg:Dataset
additionalTypes:
  - sorg:Place
mappings:
  jmeno:
    - sorg:name
    - dcterms:title
  obdobi: sorg:temporalCoverage
  lokalita:
    - dwc:verbatimLocality
    - relationIdentity: sorg:spatialCoverage
      relationType: sorg:Place
      valueIdentity: sorg:name
  gps_lat:
    - sorg:latitude
    - dwc:decimalLatitude
  gps_lon:
    - sorg:longitude
    - dwc:decimalLongitude
  popis_lokality: dwc:locationRemarks
  nadmorska_vyska:
    - dwc:minimumElevationInMeters
    - dwc:maximumElevationInMeters
    - dwc:verbatimElevation
  poznamky: sorg:description
```

Schématický popis entity Nález

Postup je obdobný jako u předchozí entity – nejprve definujeme prefixy ontologií, určíme třídu a popíšeme primitivní atributy:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases:
  sorg: http://schema.org/
  obo: http://purl.obolibrary.org/obo/
  sio: https://semanticscience.org/resource/
  dwc: http://rs.tdwg.org/dwc/terms/
  dwciri: http://rs.tdwg.org/dwc/iri/
  dcterms: http://purl.org/dc/terms/
  fisho: http://bioportal.bioontology.org/ontologies/FISHO#
  dktr: http://purl.org/dktr/fair/
baseType: dwc:Occurrence
additionalTypes:
  - sorg:Thing
mappings:
  documentName:
    - sorg:name
    - dwc:recordNumber
    - dcterms:title
  genus: dwc:genus
  species: dwc:specificEpithet
  specimen_old_id: sorg:sameAs
  determination_note: dwc:identificationRemarks
```

Atribut `river_name`, který z pohledu sémantiky představuje místopisný popis, je mapován vícenásobně – jednak v oborově specifické Fish Ontology (FISHO) predikátem `fisho:FISHO_0000055` který představuje vodní plochu nebo vodní tok, jednak samostatným

prázdným uzlem třídy `dwc:Location` a odpovídajícími predikáty z obecné ontologie Darwin Core. K Nálezu je tento prázdný uzel připojen v souladu s doporučením ontologie predikátem `dcterms:relation`, který vyjadřuje obecný blíže neurčený vztah, neboť ontologie nedává k dispozici konkrétnější predikát, který by bylo možné vyjádřit vztah „byl uloven v řece“. Do tohoto prázdného uzlu jsou navíc přidány statické predikáty a hodnoty vydestilované z popisu dat v původním TXT souboru:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings:
  (...)
  river_name:
    - fisho:FISHO_0000055
    - relationIdentity: dcterms:relation
      relationType: dcterms:Location
      valueIdentity: dwc:waterBody
      staticIdentities:
        - identity: dwc:decimalLatitude
          value: -25.671028
        - identity: dwc:decimalLongitude
          value: -53.932972
        - identity: dwc:verbatimLocality
          value: a. Deseado (San Antonio, Iguazu), provincie Misiones, Argentina
        - identity: dwc:higherGeography
          value: South America | Argentina | Misiones | San Antonio
        - identity: dwc:municipality
          value: San Antonio
        - identity: dwc:stateProvince
          value: Misiones
        - identity: dwc:country
          value: Argentina
        - identity: dwc:countryCode
          value: AR
        - identity: dwc:continent
          value: South America
        - identity: dwc:locationRemarks
          value: Základní tábor, přítok Iguazu nad Cataratas.
        - identity: dwc:minimumElevationInMeters
          value: 198
        - identity: dwc:maximumElevationInMeters
          value: 198
        - identity: dwc:verbatimElevation
          value: 198 m
```

Atribut `fishingmethod` popisující způsob, jakým byla ryba ulovena je zcela oborově specifický, proto je popsán odpovídajícími predikáty ontologie FISHO. Pro sémanticky korektní reprezentaci hodnot tohoto atributu, které jsou přítomné ve zdrojových datech, je třeba tyto mapovat na jejich sémantické ekvivalenty – IRI ontologií popisovaných způsobů lovu. Toto mapování je realizováno direktivou `valueProjection`, přičemž z hlediska sémantické definice je lhostejno, zda je atribut původní dat jeden řetězec nebo pole řetězců – komponenta mapperu si později poradí s oběma situacemi:

```
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings:
  (...)
  fishingmethod:
    valueIdentity: fisho:FISHO_0000133
    valueProjection:
      gillnets:
        "@id": fisho:FISHO_0000025
      electrofishing:
        "@id": fisho:FISHO_0000368
      castnet:
        "@id": fisho:FISHO_0000259
```

Dále je možné popsat jednotlivá měření, které se k Nálezu vztahují. Ta jsou obdobně jako atribut `river_name` realizována jednak oborově specifickým predikátem ontologie FISHO, jednak jako prázdné uzly připojené opět obecným predikátem `dcterms:relation`. Měrná jednotka je vyjádřena jak v prosté formě (ontologie Darwin Core), tak strojově čitelné formě (ontologie SIO a Units of measurement ontology (UO) která je podřízena ontologii OBO). Tímto způsobem je popsán atribut `standard_length`, který není pouhým obecným měřením, ale vyjadřuje ichtyologický terminus technicus:

```
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings:
  (...)
  standard_length:
    - fisho:FISHO_0000064
    - relationIdentity: dcterms:relation
      relationType: dwc:MeasurementOrFact
      valueIdentity: dwc:measurementValue
      staticIdentities:
        - identity: dwc:measurementType
          value: standard length
        - identity: dwc:measurementUnit
          value: cm
        - identity: sio:SIO_000221
          value:
            "@id": obo:UO_0000016
```

U atributu `head_depth` je situace analogická s tím rozdílem, že toto měření je již obecné a není přímo součástí ustálené ichtyologické terminologie. Díky této obecnosti je na rozdíl od měření `standard_length` možné měření délky hlavy popsat obecnými anatomickými referencemi. Zcela přiléhavý je pak dle popisu zejména predikát `obo:VT_0000038` z Vertebrate trait ontology (VT), cit.: *„The distance from point to point along the longest axis of the portion of the body containing the brain and organs of sight, hearing, taste, and smell.“*. Kromě něj jsou připojeny i ekvivalentní obecné predikáty oborově specifické ontologie FISHO:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings:
  (...)
  head_depth:
    relationIdentity: dcterms:relation
    relationType:
      - dwc:MeasurementOrFact
      - fisho:FISHO_0000027
    valueIdentity: dwc:measurementValue
    staticIdentities:
      - identity: dwc:measurementType
        value: head depth
      - identity: dwc:measurementUnit
        value: cm
      - identity: sio:SIO_000221
        value:
          "@id": obo:UO_0000016
      - identity: sio:SIO_000563
        value:
          "@id": fisho:FISHO_0000457
      - identity: sio:SIO_000563
        value:
          "@id": obo:VT_0000038
```

Pro použitou třídu `dwc:Occurence` je doporučeno vhodným způsobem uvést počet nalezených jedinců. Tuto informaci nemáme přímo v datech samotných, ale můžeme ji pouze dovozovat z připojeného popisu v souboru TXT – z něj vyplývá, že každý datový záznam představuje nález právě jednoho jedince. Tuto informaci připojíme ve formě statických hodnot pomocí odpovídajících predikátů:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings:
  (...)
staticIdentities:
  - identity: dwc:organismQuantity
    value: 1
  - identity: dwc:organismQuantityType
    value: individuals
```

3.3.4 Krok 4. – Transformovat data do odkazovatelné podoby

Díky využití navrženého řešení je problém transformace do odkazovatelné podoby automaticky vyřešen – dojde k němu importem do aplikace, viz kapitola 3.2.4 této práce.

V sémantickém popisu je však nutné explicitně specifikovat oboustranné vazby mezi Datasetem a jednotlivými Nálezy. V modelovém řešení je záměrně ukázána situace, kdy jeden druh záznamů obsahuje znovupoužitelné GUID a druhý nikoliv. V takovém případě je nutné buď chybějící GUID doplnit do dat ručně nebo provést import dat tak jak jsou – při importu jsou chybějící ID generovány. Jelikož je logika importu navržena s prací v režimu *upsert,* je možné po doplnění či změně sémantického popisu provést opakovaný import a data budou aktualizována.

Provedeme tedy import Datasetu (předpokládejme např. že, mu byl přiděleno ID `accc2f3d5272463c8880a1d2`). Sémantický popis poté aktualizujeme přidáním vazeb – statických predikátů ve formě předepsané použitými ontologiemi:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases:
  (...)
  dktr: http://purl.org/dktr/fair/
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities: (...)
additionalMetadata:
  - identity: dwciri:inDataset
    value:
      "@id": dktr:entity/accc2f3d5272463c8880a1d2
  - identity: dcterms:isPartOf
    value:
      "@id": dktr:entity/accc2f3d5272463c8880a1d2

# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases:
  (...)
  dktr: http://purl.org/dktr/fair/
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities:
  - identity: dcterms:hasPart
    value:
      - "@id": dktr:entity/5dcc2f395272463c88809c62
      - "@id": dktr:entity/5dcc2f395272463c88809c63
      - "@id": dktr:entity/5dcc2f395272463c88809c69
      - "@id": dktr:entity/5dcc2f395272463c88809c6a
      - "@id": dktr:entity/5dcc2f395272463c88809c6b
      - "@id": dktr:entity/5dcc2f395272463c88809c6c
      - "@id": dktr:entity/5dcc2f395272463c88809c6d
      - "@id": dktr:entity/5dcc2f395272463c88809c6e
      - "@id": dktr:entity/5dcc2f395272463c88809c6f
      - "@id": dktr:entity/5dcc2f395272463c88809c70
      - "@id": dktr:entity/5dcc2f395272463c88809c73
      - "@id": dktr:entity/5dcc2f395272463c88809c74
      - "@id": dktr:entity/5dcc2f395272463c88809c75
      - "@id": dktr:entity/5dcc2f395272463c88809c76
      - "@id": dktr:entity/5dcc2f395272463c88809c77
      - "@id": dktr:entity/5dcc2f395272463c88809c78
      - "@id": dktr:entity/5dcc2f395272463c88809c79
      - "@id": dktr:entity/5dcc2f395272463c88809c7a
      - "@id": dktr:entity/5dcc2f395272463c88809c7b
      - "@id": dktr:entity/5dcc2f395272463c88809c8d
```

3.3.5 Krok 5. – Přiřadit licenci

Jak je podrobně popsáno v kapitole 2.5.1, části Krok 5, této práce, samotná volba licence je především manažerským rozhodnutím. Z technického pohledu je vhodné volit takovou, která má strojově čitelné vyjádření. Autor v tomto případě zvolil licenci Creative Commons BY-NC-ND verze 4. Informaci o licenci technicky vyjádříme vhodným predikátem v sekci pro metadata:

```
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities: (...)
additionalMetadata:
    (...)
  - identity: dcterms:license
    value:
      "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: sorg:license
    value:
     "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/

# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities: (...)
additionalMetadata:
    (...)
  - identity: dcterms:license
    value:
      "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: sorg:license
    value:
     "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
```

3.3.6 Krok 6. – Definovat metadata

Technická metadata (identifikátory, čas vytvoření, čas poslední úpravy záznamu, provázání dat a metadat, odkaz na persistence policy) automaticky přidá aplikace sama. Metadata přiřazená v předchozích krocích tak doplníme pouze o identifikaci tvůrce (autora) dat:

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities: (...)
additionalMetadata:
    (...)
  - identity: sorg:creator
    value:
      "@type": sorg:Person
      sorg:sameAs: http://orcid.org/0000-0003-3167-2128
      sorg:givenName: Ondřej
      sorg:familyName: Doktor
      sorg:name: Ondřej Doktor

# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases: (...)
baseType: (...)
additionalTypes: (...)
mappings: (...)
staticIdentities: (...)
additionalMetadata:
    (...)
  - identity: sorg:creator
    value:
      "@type": sorg:Person
      sorg:sameAs: http://orcid.org/0000-0003-3167-2128
      sorg:givenName: Ondřej
      sorg:familyName: Doktor
      sorg:name: Ondřej Doktor
```

3.3.7 Krok 7. – Publikovat ve FAIR datovém zdroji

Kompletní opis sémantického popisu, který byl postupně sestaven v předchozích krocích je k dispozici v Příloze 6 této práce a v originální podobě také ve složce `/fixtures/demo` v netextové příloze této práce.

Všechny potřebné údaje jsou připraveny a samotná realizace nasazení je díky použitému řešení triviální – provedeme import do aplikace následujícími příkazy v prostředí kontejneru běžící aplikace:

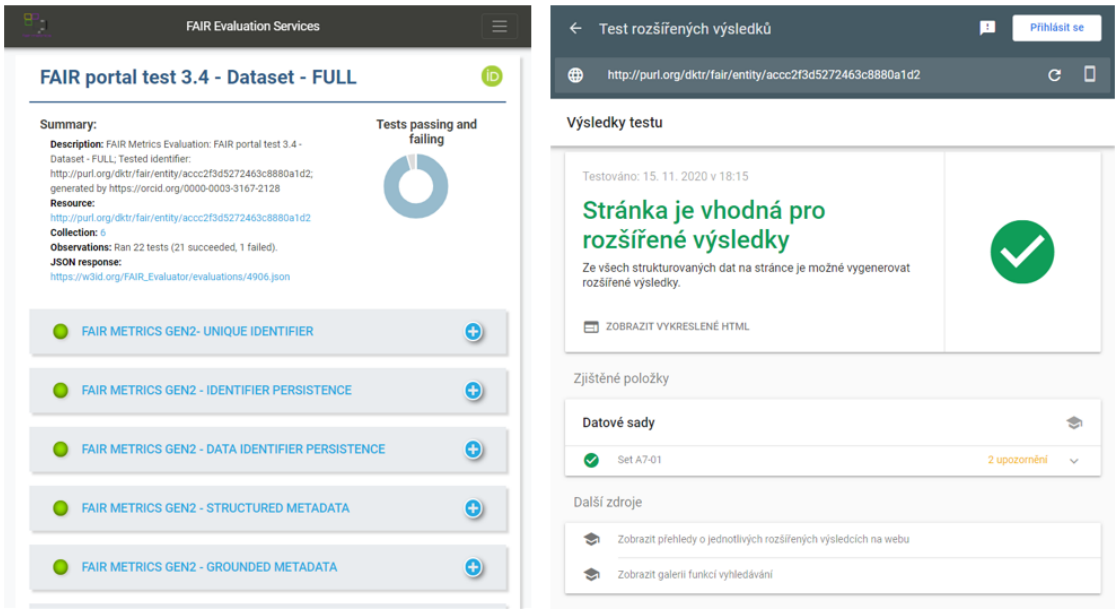
```
npm run cli

import -g id /code/fixtures/demo/records.csv /code/fixtures/demo/records-mapping.yaml
import -g id /code/fixtures/demo/datasets.csv /code/fixtures/demo/datasets-mapping.yaml
```

3.4 Ověření – Vyhodnocení ve FAIR Maturity Evaluation Service

Díky tomu, že do aplikace byla v předchozím kroku přidána ukázková data, je možné celé technické řešení otestovat a ověřit tak, že zvolené postupy a realizace praktické části práce produkuje výstup, který je skutečně možné označit přívlastkem FAIR. K tomu účelu byly použity tyto nástroje:

- **FAIR Evaluation Service** – nástroj vyvinutý skupinou FAIRMetrics [11], který je stručně popsán v kapitole 2.2 této práce
  (k dispozici na adrese: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd)
- **Google Rich Resulsts Test** – test rozšířených výsledků známého webového vyhledavače, který mj. v současné době umí v základní podobě zpracovávat sémanticky popsané datasety [18].
  (k dispozici na adrese: https://search.google.com/test/rich-results)



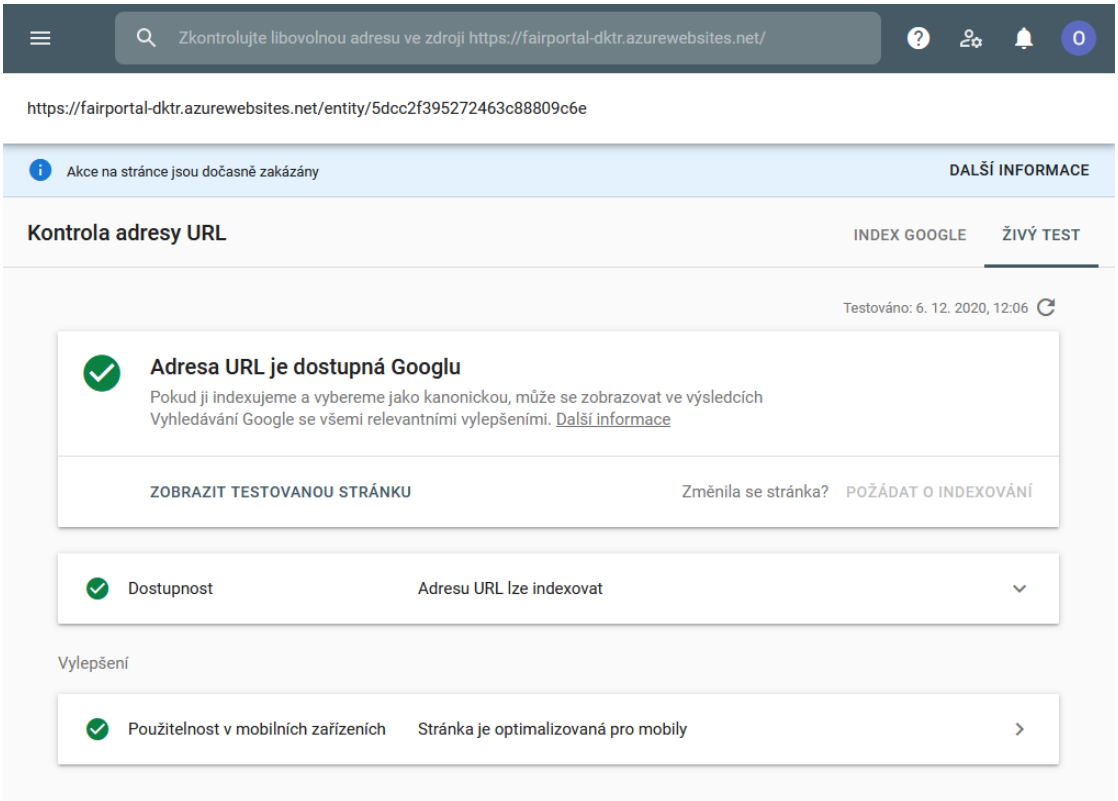
Obrázek 25: Náhled výsledků testů

Všechny provedené testy byly úspěšné. Kompletní výsledky testů jsou veřejně dostupné online, uložené v jednotlivých službách:

- Zkušební nález:
  - https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/evaluations/4907
- Zkušební dataset:
  - https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/evaluations/4906
  - https://search.google.com/test/rich-results?id=elW5s6NAGLaGa7zP-uaP-Q

Ve službě *FAIR Evaluation Service* byla provedena kompletní sada všech testů, které jsou aktuální k dispozici a testují všechny aspekty FAIR kromě principů R1.2 a R1.3 (soulad s komunitními standardy dané vědecké oblasti), které vzhledem k jejich povaze nelze testovat univerzálním nástrojem. **Z celkového počtu 22 provedených testů prošel v době uzávěrky této práce neúspěšně pouze jeden** – *Searchable in major search engine* (princip F4), který dle svého popisu zkoumá pouze to, zda je adresa daného datového zdroje indexována v jediném konkrétním vyhledavači, a to v Microsoft Bing.

Autor bohužel ze své pozice nemůže ovlivnit rychlost indexování webových vyhledavačů, proto byl pro vyloučení technického problému dodatečně proveden test nástrojem *Google Search Console*, který ověřil že nasazená **aplikace je pro indexování plně způsobilá, dostupná, a navíc k tomu ještě optimalizovaná i pro zobrazení na mobilních zařízeních** (stránky které optimalizované nejsou mohou být v indexování penalizovány) – úspěšný výsledek testu je doložen screenshotem na Obrázku 26.



Obrázek 26: Screenshot testu dostupnosti pro webové vyhledavače

Ve službě Google Rich Resulsts Test byl test datasetu rovněž úspěšný – sémantické vyjádření datasetu se daří načíst a korektně interpretovat. Nástroj však vydal doporučení, že informace o licenci a autorovi by měli být přímo součástí entity datasetu. Tyto informace jsou aktuálně přítomné v propojené samostatné entitě metadat, dle požadavků FAIR, které jsou

v této věci s metodikou společnosti Google [18] v mírném, rozporu. Technicky však není problém v budoucnu vyhovět i těmto požadavkům přidáním těchto informací redundantně do entity dat i metadat, navržené řešení toto plně podporuje – rozhodnutí zcela závisí na autorech, kam a jaké informace na základě sémantického popisu umístí a příp. pro jaké médium budou chtít svá data optimalizovat.

Celkově je možné považovat výsledek testů navrženého řešení za velmi úspěšný a je možné jej rozhodně označit jako zcela FAIR-compliant.

4 Závěr

Tato práce se zabývala problematikou otevřených dat (OpenData) v akademické a vědecko-výzkumné sféře se zaměřením na FAIR Data Principles a způsob strojově čitelného vyjádření sémantiky dat, mimo jiné také ve vztahu k současné datové základně výzkumných skupin na Přírodovědecké fakultě Jihočeské univerzity v Českých Budějovicích.

Diskutovaná problematika se jeví jako zajímavé a perspektivní téma, které dle rešerše relevantních zdrojů bude dále nabývat na významu, a to jednak jako nutná podmínka pro udržitelnost efektivní vědecké práce v kontextu neustále narůstajícího objemu dat, jednak vzhledem k tendencím podmínit budoucí financování vědecko-výzkumných projektů existencí kvalitního data management plánu.

Z rešerše provedené v teoretické části této práce dále vyplývá, že pro řešení výzev spojených s implementací otevřených dat existují kvalitní metodiky podporované mnoha významnými institucemi, které přináší obecné odpovědi na hlavní otázky – jak mají data vypadat, jak lze tohoto stavu obecně dostáhnout a jak stav zhodnotit. Zcela zásadní je v tomto kontextu pojem FAIR a související metodiky FAIR Data Principles, které jsou uznávány širokou vědeckou komunitou a podporovány mimo jiné i Evropskou komisí.

Vzhledem k technologicky neutrální povaze FAIR principů však v současnosti zůstávají stále nezodpovězené určité otázky, které jsou klíčové pro jejich praktickou aplikaci a nápravu nevyhovujícího stavu. Z toho důvodu byl v teoretické části této práce definován doporučený postup pro proces tzv. FAIRifikace dat, který vychází kormě samotných FAIR principů z doporučených postupů skupiny GO FAIR, přičemž jsou zde diskutovány praktické problémy a jejich možná řešení, které autor práce identifikoval při rozhovorech s vědeckými pracovníky sledované instituce.

Práce se rovněž v teoretické části podrobněji věnuje technickým aspektům FAIRifikace, zejména způsobu řešení sémantického strojově čitelného popisu dat, kdy byl jako vhodný prostředek identifikován Resource Description Framework (RDF) a technologie JSON-LD.

Za účelem zjištění institucionálního stavu datových sad provedl autor v praktické části práce s použitím uvedených metodik a pod záštitou vedení fakulty dotazníkové šetření v rámci vědeckých skupin působících na Přírodovědecké fakultě Jihočeské univerzity v Českých Budějovicích, na které následně navázal osobními rozhovory a individuálními konzultacemi nad reálně zpracovávanými datovými sadami.

Výsledky tohoto šetření jsou bohužel poměrně znepokojivé a zpracovávaná data rozhodně nelze označit jako FAIR. Všechny datové sady trpí zásadními nedostatky ve všech sledovaných parametrech. Z odpovědí lze vysledovat, že snahy výzkumníků brzdí zejména absence systematické podpory – technologické i metodické (tzv. data stewardship). Bylo také zjištěno, že úspěšné implementaci FAIR principů brání kromě technických důvodů také v různé míře opodstatněné obavy z úniku primárních dat a ztráty cenného know-how v silně konkurenčním vědeckém prostředí, a tedy že problém implementace FAIR má i socioekonomickou rovinu. Šetření se přesto setkalo s pozitivním ohlasem a prakticky všichni respondenti projevili ochotu, zájem o téma a chtějí být informováni o praktických možnostech pro zlepšení v této oblasti.

Jako odpověď na identifikovaný problém absence vhodného nástroje pro uchování a zpřístupnění dat v sémantické formě, vyvinul autor v praktické části práce proof-of-koncept řešení, které umožňuje převést téměř jakákoliv data do FAIR podoby, zabezpečuje jejich dohledatelnost, globální unikátnost záznamů, umožňuje jejich doplnění o bohatá metadata, zajišťuje jejich zpřístupnění jak v lidsky, tak ve strojově čitelné podobě pomocí vhodných protokolů se zachováním komplexní sémantiky dat, přičemž toto vše činí způsobem, který umožňuje zachovat v co nejvyšší možné míře původní podobu dat a tedy i stávající workflow vědeckých skupiny, čímž podporuje nasazení procesu FAIRifikace postupným a nenásilným způsobem.

Toto technické řešení bylo při uplatnění zjištění z teoretické části práce aplikováno na zvolené reprezentativní datové sadě derivované z reálných dat poskytnutých vědeckými skupinami, bylo nasazeno v cloudovém prostředí Azure (dostupné pod persistentním identifikátorem http://purl.org/dktr/fair) a bylo úspěšně plně validováno relevantními technickými nástroji jako plně FAIR-compliant.

Vzhledem k tematické souvztažnosti této práce s aktuálně probíhajícím projektem univerzální mezioborové nálezové databáze UniCatDB, na kterém autor spolupracuje, je po konzultaci se členy projektu plánována budoucí integrace vyvinutého technického řešení jako modulu v rámci tohoto projektu.

Na základě výše uvedených skutečností je možné konstatovat, že cílů definovaných v zadání této práce bylo dosaženo v plném rozsahu.

Odkazy a literatura

[1] ROCHE, Dominique G., Loeske E. B. KRUUK, Robert LANFEAR a Sandra A. BINNING. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biology [online]. 2015, 13(11) [cit. 2020-02-24]. DOI: 10.1371/journal.pbio.1002295. ISSN 1545-7885. Dostupné z: https://dx.plos.org/10.1371/journal.pbio.1002295

[2] GANTZ, John a David REINSEL. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. 2012.

[3] EUROPEAN COMMISSION. Turning FAIR into reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data [online]. Luxembourg: Publications Office of the European Union, 2018 [cit. 2020-01-08]. ISBN 978-92-79-96546-3. Dostupné z: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

[4] ELIXIR Position Paper on FAIR Data Management in life sciences. ELIXIR: A distributed infrastructure for life-science information [online]. 2017 [cit. 202002-24]. Dostupné z: https://www.elixir-czech.cz/news/elixir-position-paper-on-fairdata-management-in-life-sciences-sep-2017

[5] SOYLU, Ahmet, Felix MÖDRITSCHER a Patrick DE CAUSMAECKER. Ubiquitous web navigation through harvesting embedded semantic data: A mobile scenario. Integrated Computer-Aided Engineering [online]. 2012, 19(1), 93-109 [cit. 2020-02-24]. DOI: 10.3233/ICA-2012-0393. ISSN 18758835. Dostupné z: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/ICA2012-0393

[6] BERNERS-LEE, Tim. Linked Data. In: Design Issues: Architectural and philosophical points [online]. W3C, 2016, 2006-07-27 [cit. 2020-02-01]. Dostupné z: https://www.w3.org/DesignIssues/LinkedData.html

[7] 5-star Open Data: Information around Tim Berners-Lee's 5-star Open Data Plan [online]. [cit. 2020-02-01]. Dostupné z: https://5stardata.info/en/

[8] OECD Principles and Guidelines for Access to Research Data from Public Funding [online]. Francie: OECD Publishing, 2007 [cit. 2020-02-01]. DOI:

10.1787/9789264034020-en-fr. ISBN 9789264034020. Dostupné z: https://www.oecd.org/sti/inno/38500813.pdf

[9] WILKINSON, Mark D., Michel DUMONTIER, IJsbrand Jan AALBERSBERG, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data [online]. 2016, 3(1) [cit. 2020-02-01]. DOI: 10.1038/sdata.2016.18. ISSN 2052-4463. Dostupné z: http://www.nature.com/articles/sdata201618

[10] WILKINSON, Mark D., Susanna-Assunta SANSONE, Erik SCHULTES, Peter DOORN, Luiz Olavo BONINO DA SILVA SANTOS a Michel DUMONTIER. A design framework and exemplar metrics for FAIRness. Scientific Data [online]. 2018, 5(1) [cit. 2020-02-01]. DOI: 10.1038/sdata.2018.118. ISSN 2052-4463. Dostupné z: http://www.nature.com/articles/sdata2018118

[11] WILKINSON, Mark D., Michel DUMONTIER, Susanna-Assunta SANSONE, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific Data [online]. 2019, 6(1) [cit. 2020-02-01]. DOI: 10.1038/s41597019-0184-5. ISSN 2052-4463. Dostupné z: http://www.nature.com/articles/s41597019-0184-5

[12] CSIRO 5-star Data Rating Tool [online]. Australia: CSIRO, 2017 [cit. 2020-02-24]. Dostupné z: https://oznome.csiro.au/5star/

[13] HASNAIN, Ali a Dietrich REBHOLZ-SCHUHMANN. Assessing FAIR Data Principles Against the 5-Star Open Data Principles. GANGEMI, Aldo, Anna Lisa GENTILE, Andrea Giovanni NUZZOLESE, Sebastian RUDOLPH, Maria MALESHKOVA, Heiko PAULHEIM, Jeff Z PAN a Mehwish ALAM, ed. The Semantic Web: ESWC 2018 Satellite Events [online]. Cham: Springer International Publishing, 2018, 2018-08-02, , 469-477 [cit. 2020-02-01]. Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-98192-5_60. ISBN 978-3-319-98191-8. Dostupné z: http://link.springer.com/10.1007/978-3-319-98192-5_60

[14] Rating systems and maturity models for data publication and sharing. In: Oznome Confluence [online]. Australia: CSIRO, 2018 [cit. 2020-03-03]. Dostupné z: https://confluence.csiro.au/display/OZNOME/Rating+systems+and+maturity+models+for+data+p

[15] Oznome Data ratings. In: Oznome Confluence [online]. Australia, CSIRO, 2018 [cit 2020-03-03]. Dostupné z: https://confluence.csiro.au/display/OZNOME/Data+ratings

[16] Find FAIR Data tools. In: Dutch Techcentre for Life Sciences [online]. Utrecht [cit. 2020-03-09]. Dostupné z: https://www.dtls.nl/fair-data/find-fair-data-tools/

[17] NOY, Natasha. Making it easier to discover datasets. In: Search: News about Google Search [online]. USA: Google Blog, 2020, Sep 5, 2018 [cit. 2020-03-09]. Dostupné z: https://www.blog.google/products/search/making-it-easier-discover-datasets/

[18] Dataset. In: Google Developers: Google Search Reference [online]. USA: Google, 2020, 2020-02-10 [cit. 2020-03-09]. Dostupné z: https://developers.google.com/search/docs/data-types/dataset

[19] FAIR Data Point design specification [online]. Dutch Techcentre for Life Sciences, 2019 [cit. 2020-03-09]. Dostupné z: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec

[20] ŠTENCEK, Jiří. Užití sémantických technologií ve značkovacích jazycích. Praha, 2009. Dostupné také z: https://theses.cz/id/t77n2t/. Bakalářská práce. Vysoká škola ekonomická v Praze. Vedoucí práce Marek Nekvasil.

[21] Data on the Web Best Practices [online]. W3C, 2017 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/dwbp/

[22] RDF 1.1 Concepts and Abstract Syntax [online]. W3C, 2017 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/rdf11-concepts/

[23] Skolemisation. In: RDF Working Group Wiki [online]. W3C, 2011 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/2011/rdf-wg/wiki/Skolemisation

[24] OWL 2 Web Ontology Language Document Overview (Second Edition) [online]. W3C, 2012 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/owl2-overview/

[25] EMBL-EBI Ontology Lookup Service [online]. Cambridgeshire, UK: European Molecular Biology Laboratory, 2020 [cit. 2020-12-07]. Dostupné z: https://www.ebi.ac.uk/ols/index

[26] Ontology nad Terminology. FAIRsharing.org [online]. FAIRsharing, 2020 [cit. 2020-12-07]. Dostupné z:

https://fairsharing.org/standards/?q=&selected_facets=expanded_onto_disciplines_exact:Ontology%20and%20Terminology

[27] FAIRsharing: bsg-s000593: Schema.org. FAIRsharing.org [online]. FAIRsharing, 2020 [cit. 2020-12-07]. Dostupné z: https://fairsharing.org/FAIRsharing.hzdzq8

[28] RDF 1.1 N-Triples: A line-based syntax for an RDF graph [online]. W3C, 2014 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/n-triples/

[29] RDF 1.1 N-Quads: A line-based syntax for an RDF graph [online]. W3C, 2014 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/n-quads/

[30] RDF 1.1 Turtle: Terse RDF Triple Language [online]. W3C, 2014 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/turtle/

[31] Notation3 (N3): A readable RDF syntax [online]. W3C Team Submission, 2011 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TeamSubmission/n3/

[32] RDF 1.1 XML Syntax [online]. W3C, 2014 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/rdf-syntax-grammar/

[33] Understand how structured data works. Google Search Central [online]. [cit. 2020-12-07]. Dostupné z: https://developers.google.com/search/docs/guides/intro-structured-data

[34] JSON-LD 1.1: A JSON-based Serialization for Linked Data [online]. W3C, 2020 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/json-ld/

[35] Compaction Algorithms. In: JSON-LD 1.1 Processing Algorithms and API [online]. 2018 [cit. 2020-12-07]. Dostupné z: https://json-ld.org/spec/FCGS/json-ld-api/20180607/#compaction-algorithms

[36] FAIRification Process [online]. GO FAIR [cit. 2020-12-07]. Dostupné z: https://www.go-fair.org/fair-principles/fairification-process/

[37] FOWLER, Dan, Jo BARRATT a Paul WALSH. Frictionless Data: Making Research Data Quality Visible. International Journal of Digital Curation [online]. 2017, 12(2), 274-285 [cit. 2020-12-07]. ISSN 1746-8256. Dostupné z: doi:10.2218/ijdc.v12i2.577

[38] Frictionless Data [online]. [cit. 2020-12-07]. Dostupné z: https://frictionlessdata.io/

[39] Creative Commons [online]. Creative Commons [cit. 2020-12-07]. Dostupné z: https://creativecommons.org/

[40] Creative Commons: Nejčastěji používané licence. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2020 [cit. 2020-12-07]. Dostupné z: https://cs.wikipedia.org/wiki/Creative_Commons#Nej%C4%8Dast%C4%9Bji_pou%C5%BE%C3%ADvan%C3%A9_licence

[41] DCMI Metadata Terms [online]. Dublin Core™ Metadata Initiative, 2020 [cit. 2020-12-07]. Dostupné z: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[42] *Darwin Core* [online]. Darwin Core Maintenance Interest Group, 2020 [cit. 2020-12-07]. Dostupné z: https://dwc.tdwg.org/

[43] UniCatDB: Universal Catalog Database for biological findings [online]. Budweis, Czechia: UAI FSci USB, 2020 [cit. 2020-12-07]. Dostupné z: https://www.unicatdb.org

[44] Container Orchestration. ELIXIR [online]. Cambridgeshire, UK: ELIXIR [cit. 2020-12-07]. Dostupné z: https://elixir-europe.org/about-us/commissioned-services/container-orchestration

[45] Data Stewardship Wizard [online]. [cit. 2020-12-07]. Dostupné z: https://ds-wizard.org/

[46] TILKOV, Stefan a Steve VINOSKI. Node.js: Using JavaScript to Build High-Performance Network Programs. IEEE Internet Computing [online]. 2010, 14(6), 80-83 [cit. 2020-12-07]. ISSN 1089-7801. Dostupné z: doi:10.1109/MIC.2010.145

[47] CHANIOTIS, Ioannis K., Kyriakos-Ioannis D. KYRIAKOU a Nikolaos D. TSELIKAS. Is Node.js a viable option for building modern web applications? A performance evaluation study. Computing [online]. 2015, 97(10), 1023-1044 [cit. 2020-12-07]. ISSN 0010-485X. Dostupné z: doi:10.1007/s00607-014-0394-9

[48] Advanced Types. In: The TypeScript Handbook [online]. Microsoft, 2020 [cit. 2020-12-07]. Dostupné z: https://www.typescriptlang.org/docs/handbook/advanced-types.html

[49] Embedding JSON-LD in HTML Documents. JSON-LD 1.1: A JSON-based Serialization for Linked Data [online]. W3C, 2020 [cit. 2020-12-07]. Dostupné z: https://www.w3.org/TR/json-ld11/#embedding-json-ld-in-html-documents

[50] MAHESH. Inner workings of Map, Reduce & Filter in JavaScript. In: JavaScript In Plain English [online]. Medium, 2019-07-11 [cit. 2020-12-07]. Dostupné z: https://medium.com/javascript-in-plain-english/inner-workings-of-map-reduce-filter-f06ba87f2509

[51] Explore the search gallery. Google Search Central [online]. Google, 2020 [cit. 2020-12-07]. Dostupné z: https://developers.google.com/search/docs/guides/search-gallery

Příloha 1 – Elektronický dotazníkový arch

Faculty
of Science

Stav datových sad výzkumných skupin na PřF JU ve vztahu k FAIR Data Principles

Vážení,
s laskavým svolením a podporou pana doc. Předoty, proděkana pro vědu a výzkum PřF JU, si Vás dovolujeme oslovit v rámci šetření o stavu, povaze a kvalitě datových sad, které vytváří a používají výzkumné vědecké týmy na PřF JU ve vztahu k tzv. FAIR Data Principles, tedy dlouhodobé udržitelnosti a znovu využitelnosti dat.

Rádi bychom Vás ubezpečili, že smyslem tohoto šetření NENÍ hodnotit Vaši práci. Naší snahou je zmapovat aktuálně používané způsoby práce s daty v různých vědeckých skupinách, přičemž očekáváme že paleta Vašich přístupů k zpracování dat bude velmi pestrá.

Vyplnění vám zabere: cca 30 min.
Povinných otázek: 17

Zpracovává: Bc. Ondřej Doktor, DiS., PhDr. Miloš Prokýšek, Ph.D., ÚAI PřF JU
Všechny výsledky budou anonymizovány.

* Povinné

Identifikace

1

Název projektu, výzkumné skupiny nebo jiná identifikace *

2

Charakterizujte jednou větou vámi spravovanou nebo vytvářenou datovou sadu. Pokud máte více datových sad, vyberte tu hlavní, která tvoří jádro vašeho výzkumu. *

Např. "obrazové snímky z dronu zkoumající výskyt kůrovce", "nálezová data mapující výskyt různých druhů sladkovodních ryb"

3

Organizační zařazení v rámci instituce *

Pokud váš team nebo projekt přesahuje do více organizačních složek, vyberte tu, ke které přísluší hlavní řešitel nebo vedoucí.

- ◯ PřF JU - Katedra biologie ekosystémů
- ◯ PřF JU - Katedra botaniky
- ◯ PřF JU - Katedra experimentální biologie rostlin
- ◯ PřF JU - Katedra medicínské biologie
- ◯ PřF JU - Katedra molekulární biologie a genetiky
- ◯ PřF JU - Katedra parazitologie
- ◯ PřF JU - Katedra zoologie
- ◯ PřF JU - Ústav aplikované informatiky
- ◯ PřF JU - Ústav fyziky
- ◯ PřF JU - Ústav chemie
- ◯ PřF JU - Ústav matematiky
- ◯ PřF JU - Centrum polární ekologie
- ◯ PřF JU - Laboratoř archeobotaniky a paleoekologie
- ◯ [ ] Jiné

Sdílení

4

Jsou data dostupná kromě vlastníka či tvůrce i jiným uživatelům? *

U této otázky je lhostejné, zda jsou data dostupná veřejně, či pouze úzké skupině jednotlivců. Zajímá nás pouze způsob, jak se k nim lze dostat. Sdílení nutně neznamená, že jsou data poskytována veřejnosti a/nebo zdarma!

- [ ] Ne (máme je jen u sebe)
- [ ] Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje)
- [ ] Jako soubor ke stažení (máme je na webu jako excel, word, zip apod.)
- [ ] V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.)
- [ ] Máme webové API rozhraní, ale nemáme jej formálně popsané (popis v prostém textu a nebo se vás někdo musí zeptat, když se chce na vaše API napojit)
- [ ] Máme webové API rozhraní a máme jej formálně popsané (popis ve strojově zpracovatelné podobě - OpenApi/Swagger, WSDL apod.)
- [ ] Máme API rozhraní, máme jej formálně popsané a jeho podoba odpovídá standardu (např. JSON:API, SensorThings API, OGC)

5

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud využíváte nějakou službu nebo podporujete nějaký standard, napište jaký. (nepovinné)

Citovatelnost

6

Jsou data označena formálním identifikátorem? *

Příklad: Chcete kolegovi poslat odkaz na nějaká vaše data. Půjde to? A pokud ano, co mu pošlete?

- ◯ Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)
- ◯ Název sdíleného adresáře/souboru s daty ("je to ve složce \\SDILENY-DISK\vyzkum\iterace1")
- ◯ Místní identifikátor ("je to v naší databázi 'mereni' v tabulce 'hodnoty' pod klíčem 'A425B'")
- ◯ Webová adresa ("teď je to na http://example.com/mereni?stranka=12 (http://example.com/mereni?stranka=12), ale za rok už to tam být nemusí")
- ◯ Persistentní webový identifikátor - PURL, DOI, Handle System, apod. ("vždy to najdeš přes https://doi.org/10.1109/5.771073 (https://doi.org/10.1109/5.771073)")

7

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaký máte ve vlastních identifikátorech systém? Pokud používáte persistentní identifikátory, jaké? (nepovinné)

Anotace metadaty

METADATA jsou doprovodné informace k datům samotným, která dokumentují zejména postup, který vedl k jejich pořízení a další související okolnosti.

8

Jsou data označena metadaty? *

- ○ Ne, metadata se systematicky neuchovávají (nosíme je v hlavě nebo píšeme bokem podle potřeby, každý si to dělá po svém)
- ○ Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)
- ○ Základní metadata máme strukturovaná dle standardu (např. Dublin Core), Pokud máte metadata v Excelu, pak volte jednu z předchozích možností
- ○ Také specializovaná metadata máme strukturovaná dle standardu (e.g. Darwin Core, ISO 19115/19139, vědecký profil schema.org (http://schema.org))
- ○ Máme bohatá metadata využívající více standardních slovníků RDF (např. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)

9

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Máte nějaké zvyklosti pro uchování metadat? Pokud používáte nějaký standard, napište jaký. (nepovinné)

Dohledatenost

Data jsou indexovaná - máme je zavedena v nějakém systému (katalogu), který nám umožní vyhledávání v nich dle zadaných kritérií stylem "dotaz -> výsledek".

10

Jsou data indexována ve vyhledávacím systému? *

- ◯ Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)
- ◯ Nahráváme je do místního nebo interního systému (např. máme SQL server nebo sdílenou databázi v Accessu)
- ◯ Nahráváme je do komunitního nebo regionálního systému (např. GBIF, Mendeley Data)
- ◯ Data jsou dobře indexovaná sama o sobě a dáváme je k dispozici přes webové API rozhraní (obecné datové vyhledavače si je mohou zaindexovat sami, POZOR: Obecným datovým vyhledavačem není Google, Seznam, Bing a spol.!)

11

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. V čem data indexujete? Pokud využíváte nějaký systém nebo službu, jakou? (nepovinné)

Zpracovatelnost

12

Jsou data ve běžném nebo komunitou podporovaném standardním formátu? *

- ◯ Nestandardní formát (může s nimi pracovat jen konkrétní software nebo zařízení, např. Word, PDF, RTF, ofocené stránky textu jako obrázek, proprietární formát nějakého stroje, atd.)
- ◯ Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)
- ◯ Více standardních formátů

13

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaké formáty používáte? (nepovinné)

Použitelnost

14

Jsou data strukturována s pomocí schématu nebo datového modelu podporovaného komunitou? *

- ◯ Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)
- ◯ Vlastní schéma nebo datový model (formalizováno pomocí DDL, XSD, DDI, RDFS, JSON-Schema, data-package apod.)
- ◯ Schéma nebo datový model používaný a definovaný komunitou. Obvykle veřejně online dostupný.

15

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký strukturovaný formát, jaký? Jaká schémata příp. používáte? (nepovinné)

Srozumitelnost

16

Jsou všechna pole dat podpořena jednoznačnými definicemi jejich obsahu? *

- ◯ Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")
- ◯ Vlastní kódy/zkratky doplněné podrobnými slovními vysvětlivkami (uvádíme zkratku "Amer." a udržujeme vlastní seznam zemí)
- ◯ Používáme kódy/zkratky standardizované v rámci komunity, ale nemáme je propojené odkazy (tedy např. zkratku US podle normy ISO-3166-1)
- ◯ Některá pole jsou propojena odkazem na standardní, externě spravované definice. Metadata obsahují strojově čitelný odkaz na příslušnou definici.
- ◯ Všechna pole jsou propojena odkazem na standardní, externě spravované definice

17

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaká standard názvosloví či ontologii, jaké? (nepovinné)

Provázanost

18

Jsou (meta)data provázána s pomocí veřejných identifikátorů (URL, URI, PURL apod.)? *

- ◯ Ne, nejsou provázána
- ◯ Příchozí odkazy - někdo jiný může odkazovat na naše (meta)data, a to na i na konkrétní záznam např. pomocí URL odkazu (URI, PURL apod.)
- ◯ Příchozí + odchozí odkazy - na naše (meta)data lze odkazovat a ve vlastních (meta)datech se odkazujeme na (meta)data z dalších zdrojů pomocí URL (URI, PURL apod.)

19

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký druh křížového provázání odkazy, jaký a jak? Používáte k formalizaci provázání nějaký standard? (nepovinné)

Licencování

20

Jsou právní podmínky pro znovupoužití dat dostupné a jasně vyjádřené? *

- ◯ Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)
- ◯ Používáme vlastní licenci. (Např. na webových stránkách uvádíme, že publikace je možná pouze se svolením autora, nebo máme přiložen vlastní text, např. ve Wordu.)
- ◯ Standardní licence s odkazem (např. Creative Commons, GNU GPL, MIT, BSD, Apache, Public Domain)

21

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pod jakou licencí svá data poskytujete a proč? (nepovinné)

Údržba a původ

Následující otázky se zaměřují zejména na administrativní aspekty a správu vašich dat.

22

Udržitelnost - Máte závazek udržet data (nebo alespoň metadata) dostupná v dlouhodobém horizontu? *

Za dlouhodobý horizont lze považovat např. období po skončení trvání projektu.

- ◯ Ne, jde o jednorázový výstup
- ◯ Pouze interně (můžeme je poskytnout na vyžádání později)
- ◯ Podle našich možností (např. na vlastní webové stránce projektu)
- ◯ Umístěním do obecného veřejného nebo institucionálního repozitáře (např. CKAN, GitHub)
- ◯ Umístěním do specializovaného repozitáře (např. GBIF)

23

Aktualizace - Jsou data součástí programu pravidelného sběru / zpracování dat s jasně definovanou údržbou a plánem aktualizace? *

- ◯ Ne, jde o jednorázově pořízenou datovou sadu
- ◯ Ano, příležitostné / nepravidelné aktualizace
- ◯ Ano, plánované pravidelné aktualizace

24

Kontrola kvality - Jsou data doprovázena nebo navázána na hodnocení kvality, je popsán jejich původ či metodika, která byla použita k jejich získání? *

- ◯ Ne, žádné informace o kvalitě nebo původu
- ◯ Ano, slovní popis původu a/nebo metodiky
- ◯ Ano, formální dokumentace původu (např. PROV-O)

25

Důvěryhodnost - Jsou data doprovázena nebo navázána na informace o tom, jak jsou data využívána, kým a jak často? *

- ◯ Ne, informace o použití se neshromažďují
- ◯ Ano, jsou dostupné statistiky použití
- ◯ Ano, navíc nás doporučuje/podporuje renomovaná organizace nebo skupina

26

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. (nepovinné)

Vaše další datové sady

27

Máte ještě jiné datové sady? *

- ◯ Ano, ale všechny mají STEJNÉ vlastnosti (pokud bych je popisoval, odpovědi na tento dotazník by byly stejné)
- ◯ Ano, ale některé mají JINÉ vlastnosti něž sada, kterou jsem v dotazníku právě popsal
- ◯ Ne, žádné další datové sady nemáme

Vaše další datové sady

28

Jaké máte datové sady, pro které by odpovědi na tento dotazník byly stejné? *

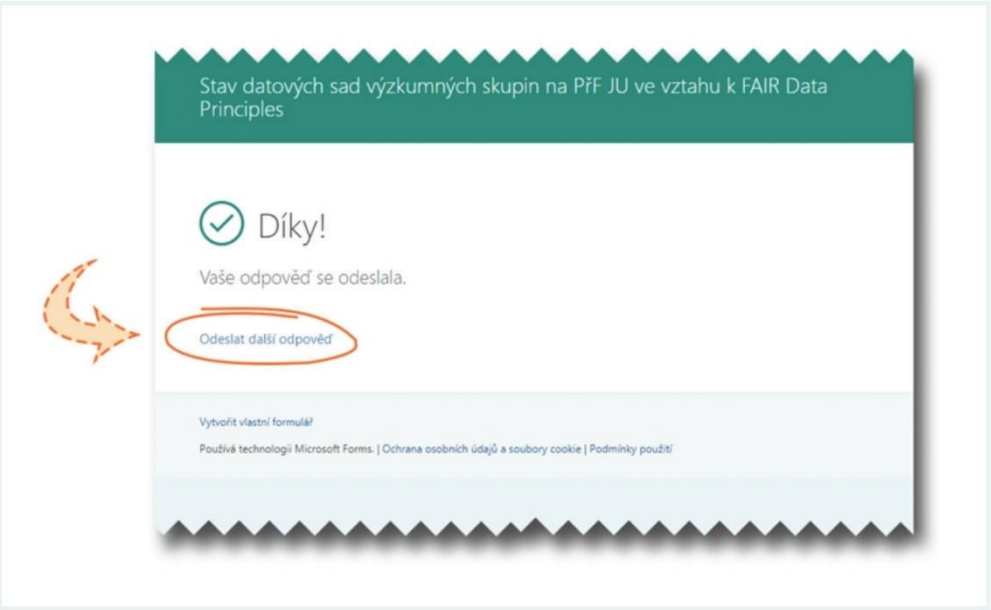
Stručně je prosím popište. Stačí každou jednou větou či názvem.

Máte další zajímavá data?

Najděte si, prosím, čas na vyplnění tohoto dotazníku i pro další datové sady

29

Další datovou sadu můžete vyplnit kliknutím na "Odeslat další odpověď" po dokončení dotazníku: *



☐ OK, chápu.

Závěr

Velmi děkujeme za váš čas, který jste věnovali vyplňování tohoto dotazníku. Pokud byste nám chtěli i nad rámec něj ještě něco sdělit, zde je pole pro poznámky. Pokud by vás zajímal výsledek, nechte nám na sebe kontakt.

30

Poznámky:

31

E-mail nebo kontakt pro zaslání výsledků:

Příloha 2 – Výpočet výsledků metodikou OZNOME

Pozn: Přítomné duplicity nejsou chybou - některé otázky dotazníku byly rozšířeny pro jemnější rozlišení, některé otázky se dle metodiky započítávaají co více ukazatelů. V případě více odpovědí u jedné otázky bereme pro výpočet v úvahu pouze tu s nejvyšší hodnotou.

Otázka	Odpoveď dotazníku	Odpovídá odpovědi v metodice OZNOME	Hodnota v jednotivých ukazatelích				
			F	A	I	R	T
Jsou data dostupná kromě vlastníka či tvůrce i jiným uživatelům?							
	Ne (máme je jen u sebe)	*No*		0			
	Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje)	*By individual arrangement*		1			
	Jako soubor ke stažení (máme je na webu jako excel, word, zip apod.)	*File download*		2			
	V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.)	*Institutional or community repository*		3			
	Máme webové API rozhraní, ale nemáme jej formálně popsané (popis v prostém textu a nebo se vás někdo musí zeptat, když se chce na vaše API napojit)	*Bespoke web service (informal API)*		3,5			
	Máme webové API rozhraní a máme jej formálně popsané (popis ve strojově zpracovatelné podobě - OpenApi/Swagger, WSDL apod.)	*Bespoke web service (OpenAPI/Swagger)*		4			
	Máme API rozhraní, máme jej formálně popsané a jeho podoba odpovídá standardu (např. JSON:API, SensorThings API, OGC)	*Standard web service API (e.g. OGC)*		5			
Jsou data označena formálním identifikátorem?							
	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)	*Not citeable*	0				
	Název sdíleného adresáře/souboru s daty ("je to ve složce \\SDILENY-DISK\vyzkum\iterace1")	*Local identifier*	2				
	Místní identifikátor ("je to v naší databázi 'mereni' v tabulce 'hodnoty' pod klíčem 'A425B'")	*Local identifier*	2				
	Webová adresa ("teď je to na http://example.com/mereni?stranka=12 (http://example.com/mereni?stranka=12), ale za rok už to tam být nemusí")	*Web address (URL - not guaranteed stable)*	3,5				
	Persistentní webový identifikátor - PURL, DOI, Handle System, apod. ("vždy to najdeš přes https://doi.org/10.1109/5.771073 (https://doi.org/10.1109/5.771073)")	*Persistent web identifier (URI)*	5				
Jsou data označena metadaty?							
	Ne, metadata se systematicky neuchovávají (nosíme je v hlavě nebo píšeme bokem podle potřeby, každý si to dělá po svém)	*No metadata*	0			0	
	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	*Abstract and keywords*	2			2	
	Základní metadata máme strukturovaná dle standardu (např. Dublin Core), Pokud máte metadata v Excelu, pak volte jednu z předchozích možností	*Basic metadata (e.g. Dublin Core)*	3			3	
	Také specializovaná metadata máme strukturovaná dle standardu (e.g. Darwin Core, ISO 19115/19139, vědecký profil schema.org (http://schema.org))	*Specialized metadata (e.g. Darwin Core, ISO 19115/19139, schema.org scientific data profile)*	4			4	
	Máme bohatá metadata využívající více standardních slovníků RDF (např. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)	*Rich metadata using multiple standard RDF vocabularies (e.g. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)*	5			5	
Jsou data indexována ve vyhledávacím systému?							
	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	*no*	0			0	
	Nahráváme je do místního nebo interního systému (např. máme SQL server nebo sdílenou databázi v Accessu)	*local or internal system only*	2			2	
	Nahráváme je do komunitního nebo regionálního systému (např. GBIF, Mendeley Data)	*community wide or jurisdictional system*	3,5			3,5	
	Data jsou dobře indexovaná sama o sobě a dáváme je k dispozici přes webové API rozhraní (obecné datové vyhledavače si je mohou zaindexovat sami, POZOR: Obecným datovým vyhledavačem není Google, Seznam, Bing a spol.!)	*highly ranked in general purpose index (Google, Bing etc)*	5			5	
Jsou data ve běžném nebo komunitou podporovaném standardním formátu?							
	Nestandardní formát (může s nimi pracovat jen konkrétní software nebo zařízení, např. Word, PDF, RTF, ofocené stránky textu jako obrázek, proprietární formát nějakého stroje, atd.)	*bespoke format (text, binary)*			0		
	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)	*one standard format, denoted by a MIME-type*			2,5		
	Více standardních formátů	*multiple standard formats*			5		
Jsou data strukturována s pomocí schématu nebo datového modelu podporovaného komunitou?							
	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	*no formal schema*			0	0	
	Vlastní schéma nebo datový model (formalizováno pomocí DDL, XSD, DDI, RDFS, JSON-Schema, data-package apod.)	*explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar*			2,5	2,5	

	Schéma nebo datový model používaný a definovaný komunitou. Obvykle veřejně online dostupný.	*community-shared schema or data model , available from a standard location*			5	5	
Jsou všechna pole dat podpořena jednoznačnými definicemi jejich obsahu?							
	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")	*local field codes or labels*			0		
	Vlastní kódy/zkratky doplněné podrobnými slovními vysvětlivkami (uvádíme zkratku "Amer." a udržujeme vlastní seznam zemí)	*labels with full text explanations*			2		
	Používáme kódy/zkratky standardizované v rámci komunity, ale nemáme je propojené odkazy (tedy např. zkratku US podle normy ISO-3166-1)	*community standard labels (e.g. CF Conventions, UCUM units)*			3		
	Některá pole jsou propojena odkazem na standardní, externě spravované definice. Metadata obsahují strojově čitelný odkaz na příslušnou definici.	*some fields linked to externally managed definitions*			4		
	Všechna pole jsou propojena odkazem na standardní, externě spravované definice	*all fields linked to standard, externally managed definitions*			5		
Jsou (meta)data provázána s pomocí veřejných identifikátorů (URL, URI, PURL apod.)? *							
	Ne, nejsou provázána	*no links*			0		
	Příchozí odkazy - někdo jiný může odkazovat na naše (meta)data, a to na i na konkrétní záznam např. pomocí URL odkazu (URI, PURL apod.)	*in-bound links from a catalogue or landing-page*			2,5		
	Příchozí + odchozí odkazy - na naše (meta)data lze odkazovat a ve vlastních (meta)datech se odkazujeme na (meta)data z dalších zdrojů pomocí URL (URI, PURL apod.)	*out-bound links to related data and definitions*			5		
Jsou právní podmínky pro znovupoužití dat dostupné a jasně vyjádřené?							
	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	*no license*				0	
	Používáme vlastní licenci. (Např. na webových stránkách uvádíme, že publikace je možná pouze se svolením autora, nebo máme přiložen vlastní text, např. ve Wordu.)	*license described in text*				2,5	
	Standardní licence s odkazem (např. Creative Commons, GNU GPL, MIT, BSD, Apache, Public Domain)	*link to a standard license (e.g. Creative Commons)*				5	
Udržitelnost - Máte závazek udržet data (nebo alespoň metadata) dostupná v dlouhodobém horizontu?							
	Ne, jde o jednorázový výstup	*once-off dump, no ongoing commitment*	0				
	Pouze interně (můžeme je poskytnout na vyžádání později)	*once-off dump, no ongoing commitment*	0				
	Podle našich možností (např. na vlastní webové stránce projektu)	*best effort, project website*	2				
	Umístěním do obecného veřejného nebo institucionálního repozitáře (např. CKAN, GitHub)	*public or institutional repository (e.g. CKAN, GitHub)*	3,5				
	Umístěním do specializovaného repozitáře (např. GBIF)	*certified repository*	5				
Aktualizace - Jsou data součástí programu pravidelného sběru / zpracování dat s jasně definovanou údržbou a plánem aktualizace?							
	Ne, jde o jednorázově pořízenou datovou sadu	*one-time dataset*					0
	Ano, příležitostné / nepravidelné aktualizace	*part of series - occasional/irregular update*					2,5
	Ano, plánované pravidelné aktualizace	*part of series - regular scheduled updates*					5
Kontrola kvality - Jsou data doprovázena nebo navázána na hodnocení kvality, je popsán jejich původ či metodika, která byla použita k jejich získání?							
	Ne, žádné informace o kvalitě nebo původu	*no quality or lineage information*				0	
	Ano, slovní popis původu a/nebo metodiky	*text lineage statement*				2,5	
	Ano, formální dokumentace původu (např. PROV-O)	*formal provenance trace (e.g. PROV-O)*				5	
Důvěryhodnost - Jsou data doprovázena nebo navázána na informace o tom, jak jsou data využívána, kým a jak často?							
	Ne, informace o použití se neshromažďují	*no information about usage*					0
	Ano, jsou dostupné statistiky použití	*usage statistics available*					2,5
	Ano, navíc nás doporučuje/podporuje renomovaná organizace nebo skupina	*Clearly endorsed by reputable organization or framework*					5

Výpočet výsledků

	F	A	I	R	T
1: Sečtěte hodnoty pro jednotlivé ukazatele:					
2: Spočtěte jejich aritmetický průměr - skóre jednoltivých ukazatelů:					
3: Spočtěte aritmetický průměr výsledků jednotlivých kategorií - celkové skóre:					

Příloha 3 – Anonymizovaná surová data

ID	1	4	6	8	9
Počáteční čas	5.3.20 11:47	6.3.20 10:40	11.3.20 15:31	10.4.20 10:40	10.4.20 13:39
Čas dokončení	5.3.20 12:19	6.3.20 11:19	11.3.20 17:58	10.4.20 10:54	10.4.20 13:54
E-mail	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Jméno					
Název projektu, výzkumné skupiny nebo jiná identifikace	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Charakterizujte jednou větou vámi spravovanou nebo vytvářenou datovou sadu. Pokud máte více datových sad, vyberte tu hlavní, která tvoří jádro vašeho výzkumu.	spektrální a termální snímky z dronu zkoumající napadení stromů kůrovcem	obrazové snímky z dronu zkoumající výskyt kůrovce, různé datové sady v rámci diplomových prací a dílčích výzkumných projektů (námi spravované jsou převážně prostorová data pro využití v GIS)	časová řada konfigurací molekulárního systému (=trajektorie), časový vývoj termodynamických veličin	Chromatogramy (záznamy chromatografických analýz).	Soubor fytocenologických snímků z různých člověkem narušených míst.
Organizační zařazení v rámci instituce	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Jsou data dostupná kromě vlastníka či tvůrce i jiným uživatelům?	V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.)	V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.)	V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.)	V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.);	Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje);

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud využíváte nějakou službu nebo podporujete nějaký standard, napište jaký. (nepovinné)		Pouze jako repozitář dat (projekt ANONYMIZOVÁNO, cesnet repozitář) - nejčastější přístup přes FTP protokol. Ostatní data většinou jen sdílené disky, případně sdílení ve skupině pomocí cloudovych reseni - owncloud (cesnet) nebo dropbox.	sdílený adresář na výpočetním klastru, vzájemná čtecí přístupová práva v linuxu.	Vzdálená plocha ve WIN10.	Na stránkách ANONYMIZOVÁNO máme databázi detailně popsanou, jsou zde definovaná pravidla pro její využití a na koho je třeba se obrátit, pokud je zájem o její využití.
Jsou data označena formálním identifikátorem?	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)	Místní identifikátor ("je to v naší databázi 'mereni' v tabulce 'hodnoty' pod klíčem 'A425B"')	Místní identifikátor ("je to v naší databázi 'mereni' v tabulce 'hodnoty' pod klíčem 'A425B"')	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaký máte ve vlastních identifikátorech systém? Pokud používáte persistentní identifikátory, jaké? (nepovinné)		Zatim pro identifikaci principialne jen adresářová struktura na libovolném úložišti. V rámci některých dílčích projektů, článků apod. mám dat v lokálních databázích (SQLite, PostgreSQL/PostGIS) a pak metadata v rámci aplikace (uložena opět v databázi, ale přístup využívám jen přes aplikaci - GIS). Jedná se především o metadata obsahu, ne o identifikátor sady v rámci nějakého systému.	Posíláme si název adresáře s daty	Měření jsou organizována podle jména pracovníka provádějícího analýzy a dále pak podle použité metody a data analýzy.	
Jsou data označena metadaty?	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Máte nějaké zvyklosti pro uchování metadat? Pokud používáte nějaký standard, napište jaký. (nepovinné)		Jak jsem psal v předchozím komentáři - občas metadata máme, ale je příslušné konkrétnímu projektu či paperu. V rámci GIS je automaticky nastaven OGC/ISO, resp. využívá jako minimum Dublin core. Každopádně, ale realně používám minimálně. V "rámci rychlosti zpracování" pro analýzu dat, kterou zpracovávám sám, mám občas základní popis dat ve formě strukturovaného textu (markdown text, často v podobě jupyter notebooku, který je vázán na projekt - často pro GIS analýzu využívám python, proto jupyter notebook).	Vlastní systém pojmenovávání adresářů. Z názvu adresáře jsou patrné typicky 2-4 hodnoty parametrů, jejichž vliv studujeme.	Slovní formou písemně v deníku přístroje v jistém minimálním požadovaném rozsahu.	
Jsou data indexována ve vyhledávacím systému?	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám, filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání pře Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. V čem data indexujete? Pokud využíváte nějaký systém nebo službu, jakou? (nepovinné)		V případě uložení do databáze (viz předešlé komentáře) pak samozřejmě základní indexace je.	hledáme adresáře/soubory dle kódů z názvů		
Jsou data ve běžném nebo komunitou podporovaném standardním formátu?	Nestandardní formát (např. Word, PDF, RTF, ofocené stránky textu jako obrázek, proprietární formát nějakého stroje, atd.)	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)	Nestandardní formát (může s nimi pracovat jen konkrétní software nebo zařízení, např. Word, PDF, RTF, ofocené stránky textu jako obrázek, proprietární formát nějakého stroje, atd.)	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaké formáty používáte? (nepovinné)	jedná se o specifické formáty dat, které se vztahuji k softwarům ke zpracování těchto dat.	GIS data - nejčastěji pracujeme s řešením firmy ESRI (formát geodatabase, proprietární uložení, ale konkurenční GIS nástroje umí číst, kromě toho používá standardní formáty geojson, TIFF, GPX, geopackages), v případě open source nástrojů využíváme standardizované formáty ((Geo)JSON, GML, geopackage apod.) Data projektu - proprietární raw soubory - snímky.	Používáme buď aplikace běžně používané odborníky v naší komunitě => standardní výstupní soubory, nebo vlastní programu s popisky sloupců v datových souborech.	Formáty chromatografických software Chromeleon a Xcalibur.	Data jsou v excelu.

Jsou data strukturována s pomocí schématu nebo datového modelu podporovaného komunitou?	Ne, nemáme definováno formální schéma nebo datový model	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký strukturovaný formát, jaký? Jaká schémata příp. používáte? (nepovinné)		Obě možnosti - první a druhá volba výše. V rámci zpracování GIS často jen csv soubory, rastrové formáty v obecných formátech (TIF apod.)			
Jsou všechna pole dat podpořena jednoznačnými definicemi jejich obsahu?	Vlastní kódy/zkratky doplněné podrobnými slovními vysvětlivkami	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")	Vlastní kódy/zkratky doplněné podrobnými slovními vysvětlivkami (uvádíme zkratku "Amer." a udržujeme vlastní seznam zemí)	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaká standard názvosloví či ontologii, jaké? (nepovinné)	Máme svoje vlastní zkratky		Vlastní kódy, ale pochopitelné většině blízkých vědců. Tato data ale stejně prakticky nepředáváme nikomu mimo skupinu, ostatní se spokojují s finálními grafy v článcích.		
Jsou (meta)data provázána s pomocí veřejných identifikátorů (URL, URI, PURL apod.)?	Ne, nejsou provázána	Ne, nejsou provázána	Ne, nejsou provázána	Ne, nejsou provázána	Ne, nejsou provázána
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký druh křížového provázání odkazy, jaký a jak? Používáte k formalizaci provázání nějaký standard? (nepovinné)					
Jsou právní podmínky pro znovupoužití dat dostupné a jasně vyjádřené?	Ne, není definována licence	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Používáme vlastní licenci. (Např. na webových stránkách uvádíme, že publikace je možná pouze se svolením autora, nebo máme přiložen vlastní text, např. ve Wordu.)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pod jakou licencí svá data poskytujete a proč? (nepovinné)	nn		Licenci nepoužíváme, ale brzy chceme zveřejnit výsledek - program a budeme řešit formu licence.		Na stránce ANONYMIZOVÁNO jsou ke stažení pravidla pro využití.
Udržitelnost - Máte závazek udržet data (nebo alespoň metadata) dostupná v dlouhodobém horizontu?	Ne, jde o jednorázový výstup	Podle našich možností (např. na vlastní webové stránce projektu)	Podle našich možností (např. na vlastní webové stránce projektu)	Pouze interně (můžeme je poskytnout na vyžádání později)	Pouze interně (můžeme je poskytnout na vyžádání později)

Aktualizace - Jsou data součástí programu pravidelného sběru / zpracování dat s jasně definovanou údržbou a plánem aktualizace?	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu	Ano, příležitostné / nepravidelné aktualizace
Kontrola kvality - Jsou data doprovázena nebo navázána na hodnocení kvality, je popsán jejich původ či metodika, která byla použita k jejich získání?	Ne, žádné informace o kvalitě nebo původu	Ano, slovní popis původu a/nebo metodiky	Ne, žádné informace o kvalitě nebo původu	Ano, slovní popis původu a/nebo metodiky	Ano, slovní popis původu a/nebo metodiky
Důvěryhodnost - Jsou data doprovázena nebo navázána na informace o tom, jak jsou data využívána, kým a jak často?	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. (nepovinné)					
Máte ještě jiné datové sady?			Ano, ale všechny mají STEJNÉ vlastnosti (pokud bych je popisoval, odpovědi na tento dotazník by byly stejné)	Ne, žádné další datové sady nemáme	Ne, žádné další datové sady nemáme

Jaké máte datové sady, pro které by odpovědi na tento dotazník byly stejné?		Již jsem v dotazníku odpovídal hromadně za více datových sad. Ucelená datová sada je v lastně jen projekt ANONYMIZOVÁNO - zde jde o pouhý repozitář, proprietárně získaných obrazových dat (raw, jpg, tif, asc). Další poznámky jsou obecně k datovým sadám, které využívám v GIS. Např. projekt (článek) mapování výskytu a preferencí dané druhu ptáků v ČB - potřebuje vlastní data za hnízda/ptáky, dále data o landuse území, hranice různých městských částí, typu zástavby apod. - jednorázový projekt, data uložena v geodatabázi, metadata jen základní (pořízení, autor, souřad. systémé, rozsah data apod). Takových datových sad je spousta, otázka je, nakolik jsou veřejně přístupná a tím teda indexovatelná... Centrální datový repositář nevyužíváme (zatím).	Pracujeme na několika projektech, liší se použitím volně dostupného SW pro geerování dat (simulací) a vlastních programů. Ale podstata dat je obdobná.		
Další datovou sadu můžete vyplnit kliknutím na "Odeslat další odpověď" po dokončení dotazníku:		OK, chápu.;			

Poznámky:	20 minut	Určitě zajímavé, otázka zda má vézt k nějakému centrálnímu repositáři nebo spíše jen metadatový server a rozhraní, což mi dává větší smysl. Nicméně s aktuální zkušeností, co mám, je ochota lidí sdílet data, ale i jen metadata velmi omezená. V oblasti GIS dat hodně diskutované téma a mnoho existujících řešení.	U spousty otázek byla většina možností cílena na organizované formy nakládání s daty, které zásadně překračují naši praxi - a přesto jsme docela dobře schopni data dohledat, případně i od jiného člena skupiny. Jinak je většinou jeden člen skupiny plně zodpovědný za svá data.	Jedná se pouze o několik stovek GB dat.	
E-mail nebo kontakt pro zaslání výsledků:	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*

ID	10	11	12	13
Počáteční čas	12.4.20 2:04	13.4.20 17:39	14.4.20 13:49	18.4.20 8:33
Čas dokončení	12.4.20 2:15	13.4.20 17:49	14.4.20 13:57	18.4.20 8:51
E-mail	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Jméno				
Název projektu, výzkumné skupiny nebo jiná identifikace	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Charakterizujte jednou větou vámi spravovanou nebo vytvářenou datovou sadu. Pokud máte více datových sad, vyberte tu hlavní, která tvoří jádro vašeho výzkumu.	velké textové soubory sekvenačních dat	spektroskopická data ve formě matice (čas, vlnová délka) charakterizující intenzitu signálu	RNAseq z drozofilích buněk/tkání	strukturní a termodynamické vlastnosti molekulárních systémů, trajektorie (= série konfigurací) systémů
Organizační zařazení v rámci instituce	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*
Jsou data dostupná kromě vlastníka či tvůrce i jiným uživatelům?	Ne (máme je jen u sebe);Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje);Jako soubor ke stažení (máme je na webu jako excel, word, zip apod.);	Ne (máme je jen u sebe);Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje);	Ne (máme je jen u sebe);Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje);	Na individuální žádost (můžeme je poslat emailem, dát na flashku když nás kontaktuje);V institucionálním nebo komunitním repozitáři (sdílený disk fakulty, skupina v univerzitním SharePointu, služby Metacentra nebo CESNETu apod.);

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud využíváte nějakou službu nebo podporujete nějaký standard, napište jaký. (nepovinné)				obdobné kódové značení názvů adresářů, standardní názvy souborů, otevřené adresáře v rámci vlastního linuxového klastru.
Jsou data označena formálním identifikátorem?	Název sdíleného adresáře/souboru s daty ("je to ve složce \\SDILENY-DISK\vyzkum\iterace1")	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)	Ne, nejsou odkazovatelná (odkaz mu poslat nemohu, někdo to má u sebe v počítači - mohu přiložit např. jako přílohu v e-mailu)	Název sdíleného adresáře/souboru s daty ("je to ve složce \\SDILENY-DISK\vyzkum\iterace1")

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaký máte ve vlastních identifikátorech systém? Pokud používáte persistentní identifikátory, jaké? (nepovinné)				
Jsou data označena metadaty?	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Slovní popis volnou formou nebo dle našich zvyklostí, ale nikoliv ve strukturovaném formátu (tedy metadata máme např. v textovém souboru nebo v Excelu vedle dat)	Ne, metadata se systematicky neuchovávají (nosíme je v hlavě nebo píšeme bokem podle potřeby, každý si to dělá po svém)

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Máte nějaké zvyklosti pro uchování metadat? Pokud používáte nějaký standard, napište jaký. (nepovinné)				
Jsou data indexována ve vyhledávacím systému?	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)	Ne (pokud něco hledáme musíme data ručně projít a najít si výsledek sám - filtrace v Excelu není indexování, vybírání dat jednoúčelovým skriptem není indexování, vyhledávání přes Ctrl+F v textovém souboru není indexování)

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. V čem data indexujete? Pokud využíváte nějaký systém nebo službu, jakou? (nepovinné)				
Jsou data ve běžném nebo komunitou podporovaném standardním formátu?	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)	Více standardních formátů	Jeden standardní formát (je možné načíst pomocí standardních aplikací pro zpracování dat např. CSV, JSON, XML, netCDF, JPG, TIFF, FASTA, DICOM, GPX, KML, atd.)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Jaké formáty používáte? (nepovinné)			FASTA, Geneious format, Excel	Používáme standardní formáty binárních kompresovaných souborů s trajektoriemi (jeden soubor až jednotky GB dat) i standardní formát souborů používaných volně dostupnými programovými balíky. U vlastních skriptů a jejich výstupů se jedná o srozumitelně čitelné textové soubory.

Jsou data strukturována s pomocí schématu nebo datového modelu podporovaného komunitou?	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)	Ne, nemáme definováno formální schéma nebo datový model (např. data jsou v CSV souboru, pouze v prvním řádku jsou názvy sloupců)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký strukturovaný formát, jaký? Jaká schémata příp. používáte? (nepovinné)				
Jsou všechna pole dat podpořena jednoznačnými definicemi jejich obsahu?	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")	Některá pole jsou propojena odkazem na standardní, externě spravované definice. Metadata obsahují strojově čitelný odkaz na příslušnou definici.	Ne, používáme vlastní kódy, zkratky nebo názvy (např. pro označení místa původu vzorku uvádíme "Spojené státy americké")

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaká standard názvosloví či ontologii, jaké? (nepovinné)				
Jsou (meta)data provázána s pomocí veřejných identifikátorů (URL, URI, PURL apod.)?	Ne, nejsou provázána	Ne, nejsou provázána	Ne, nejsou provázána	Ne, nejsou provázána
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pokud používáte nějaký druh křížového provázání odkazy, jaký a jak? Používáte k formalizaci provázání nějaký standard? (nepovinné)				
Jsou právní podmínky pro znovupoužití dat dostupné a jasně vyjádřené?	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)	Ne, není definována licence. (Když někomu data poskytnu, tak se domluvíme, jak je může použít.)
Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. Pod jakou licencí svá data poskytujete a proč? (nepovinné)				
Udržitelnost - Máte závazek udržet data (nebo alespoň metadata) dostupná v dlouhodobém horizontu?	Podle našich možností (např. na vlastní webové stránce projektu)	Pouze interně (můžeme je poskytnout na vyžádání později)	Pouze interně (můžeme je poskytnout na vyžádání později)	Pouze interně (můžeme je poskytnout na vyžádání později)

Aktualizace - Jsou data součástí programu pravidelného sběru / zpracování dat s jasně definovanou údržbou a plánem aktualizace?	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu	Ne, jde o jednorázově pořízenou datovou sadu
Kontrola kvality - Jsou data doprovázena nebo navázána na hodnocení kvality, je popsán jejich původ či metodika, která byla použita k jejich získání?	Ano, slovní popis původu a/nebo metodiky	Ano, slovní popis původu a/nebo metodiky	Ano, slovní popis původu a/nebo metodiky	Ne, žádné informace o kvalitě nebo původu
Důvěryhodnost - Jsou data doprovázena nebo navázána na informace o tom, jak jsou data využívána, kým a jak často?	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují	Ne, informace o použití se neshromažďují

Budeme rádi, pokud nám Vaše řešení alespoň velmi stručně popíšete. (nepovinné)				
Máte ještě jiné datové sady?	Ne, žádné další datové sady nemáme	Ano, ale všechny mají STEJNÉ vlastnosti (pokud bych je popisoval, odpovědi na tento dotazník by byly stejné)	Ne, žádné další datové sady nemáme	Ano, ale všechny mají STEJNÉ vlastnosti (pokud bych je popisoval, odpovědi na tento dotazník by byly stejné)

Jaké máte datové sady, pro které by odpovědi na tento dotazník byly stejné?		Data získaná z jiných měření (absorpční, fluorescenční, CD spektra, HPLC data, data z voltametrie, ...)		Obdobná data vzniklá spoluprací s jinými externími experimentátory, jiné studované systémy, výstupy vzniklé jiným SW (např. GROMACS vs LAMMPS)
Další datovou sadu můžete vyplnit kliknutím na "Odeslat další odpověď" po dokončení dotazníku:				

Poznámky:				Pokud je cílem i podpořit lepší nakládání s daty, ocenil bych zveřejnění prázdného dotazníku s uvedenými možnostmi formátů, úložišť, atd., aby člověk tušil, jaké jsou nabízené možnosti.
E-mail nebo kontakt pro zaslání výsledků:	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*	*ANONYMIZOVÁNO*

Příloha 4 – Výsledky a statistika šetření

Použitá metodika: CSIRO 5-Star Rating (Oznome Data ratings. In: Oznome Confluence [online]. Australia, CSIRO, 2018 [cit. 2020-03-03]. Dostupné z: https://confluence.csiro.au/display/OZNOME/Data+ratings)

Dataset ID	Findable	Accessible	Interoprable	Reusable	Trusted	Celkové skóre	Findable	Accessible	Interoprable	Reusable	Trusted	Celkové skóre
6	1,33	2,50	0,63	0,40	0,00	**0,97**	★☆☆☆☆	★★☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
4	0,67	2,50	0,63	0,90	0,00	**0,94**	★☆☆☆☆	★★☆☆☆	☆☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
9	0,67	0,50	0,63	1,40	1,25	**0,89**	★☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆	★☆☆☆☆	★☆☆☆☆
10	0,67	2,00	0,63	0,90	0,00	**0,84**	★☆☆☆☆	★★☆☆☆	☆☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
12	0,67	0,50	2,13	0,90	0,00	**0,84**	★☆☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
14	0,00	0,50	0,63	0,50	2,50	**0,83**	☆☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆
8	1,33	1,50	0,00	0,90	0,00	**0,75**	★☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
11	0,67	0,50	0,63	0,90	0,00	**0,54**	★☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
1	0,67	1,50	0,00	0,40	0,00	**0,51**	★☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	★☆☆☆☆
13	0,00	1,50	0,63	0,00	0,00	**0,43**	☆☆☆☆☆	★☆☆☆☆	★☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆

Příloha 5 – Demonstrační zdrojová data

Obsah CSV a TXT souboru

id	documentName	genus	species	specimen_old_id	determination_note	river_name	fishingmethod	fishingmethod	standard_length	head_depth
5dcc2f395272463c88809c62	Pimelodus maculatus (94)	Pimelodus	maculatus	http://example.com/old-fish-database/1		Deseado	gillnets	hook&line	37	24
5dcc2f395272463c88809c63	Geophagus brasiliensis (95)	Geophagus	brasiliensis	http://example.com/old-fish-database/2		Deseado	gillnets	hook&line	66	17
5dcc2f395272463c88809c69	Crenicichla iguassuensis (101)	Crenicichla	iguassuensis	http://example.com/old-fish-database/8		Deseado	gillnets	hook&line	79	28
5dcc2f395272463c88809c6a	Crenicichla iguassuensis	Crenicichla	iguassuensis	http://example.com/old-fish-database/9		Deseado	gillnets	hook&line	69	26
5dcc2f395272463c88809c6b	Crenicichla lepidota (102)	Crenicichla	lepidota	http://example.com/old-fish-database/10		Deseado	gillnets	hook&line	90	18
5dcc2f395272463c88809c6c	Crenicichla tapii x tuca ? (103)	Crenicichla	tapii	http://example.com/old-fish-database/11	Crenicichla tapii x tuca ? thick lips	Deseado	gillnets	hook&line	78	19
5dcc2f395272463c88809c6d	Crenicichla tapii (104)	Crenicichla	tapii	http://example.com/old-fish-database/12		Deseado	electrofishing		100	24
5dcc2f395272463c88809c6e	Crenicichla tapii (105)	Crenicichla	tapii	http://example.com/old-fish-database/13		Deseado	electrofishing		78	30
5dcc2f395272463c88809c6f	Crenicichla tesay (106)	Crenicichla	tesay	http://example.com/old-fish-database/14		Deseado	electrofishing		53	17
5dcc2f395272463c88809c70	Crenicichla tesay (107)	Crenicichla	tesay	http://example.com/old-fish-database/15		Deseado	castnet	hook&line	64	15
5dcc2f395272463c88809c73	Gymnogeophagus meridionalis (110)	Gymnogeophagus	meridionalis	http://example.com/old-fish-database/18		Deseado	castnet	hook&line	77	25
5dcc2f395272463c88809c74	Gymnogeophagus meridionalis (111)	Gymnogeophagus	meridionalis	http://example.com/old-fish-database/19		Deseado	castnet	hook&line	79	16
5dcc2f395272463c88809c75	Crenicichla tesay (112)	Crenicichla	tesay	http://example.com/old-fish-database/20		Deseado	gillnets	hook&line	89	17
5dcc2f395272463c88809c76	Crenicichla tesay (113)	Crenicichla	tesay	http://example.com/old-fish-database/21		Deseado	gillnets	hook&line	48	20
5dcc2f395272463c88809c77	Crenicichla tesay (114)	Crenicichla	tesay	http://example.com/old-fish-database/22		Deseado	gillnets	hook&line	69	29
5dcc2f395272463c88809c78	Crenicichla iguassuensis (115)	Crenicichla	iguassuensis	http://example.com/old-fish-database/23		Deseado	gillnets	hook&line	40	28
5dcc2f395272463c88809c79	Crenicichla iguassuensis (116)	Crenicichla	iguassuensis	http://example.com/old-fish-database/24		Deseado	gillnets	hook&line	50	27
5dcc2f395272463c88809c7a	Crenicichla lepidota (117)	Crenicichla	lepidota	http://example.com/old-fish-database/25		Deseado	gillnets	hook&line	77	22
5dcc2f395272463c88809c7b	Crenicichla lepidota (118)	Crenicichla	lepidota	http://example.com/old-fish-database/26		Deseado	gillnets	hook&line	86	25
5dcc2f395272463c88809c8d	Gymnogeophagus meridionalis ? (145)	Gymnogeophagus	meridionalis	http://example.com/old-fish-database/44	or Geophagus	Deseado	gillnets	hook&line	58	19

Set A7-01
Datum: 27.11.2007 - 29.11.2007

Lokalita: a. Deseado (San Antonio, Iguazu), provincie Misiones, Argentina
GPS: -25.671028, -53.932972
Popis: Základní tábor, přítok Iguazu nad Cataratas.
Nadmoř. výška: 198 m

Všechna měření délek jsou v milimetrech. Každý záznam popisuje jendu ulovenou rybu.

Příloha 6 – Výpis kompletního sémantického popisu z praktické ukázky

File - datasets-mapping.yaml

```yaml
# dataset-mapping.yaml (entita Dataset)
---
namespaceAliases:
  sorg: http://schema.org/
  obo: http://purl.obolibrary.org/obo/
  dwc: http://rs.tdwg.org/dwc/terms/
  dcterms: http://purl.org/dc/terms/
  dktr: http://purl.org/dktr/fair/
baseType: sorg:Dataset
additionalTypes:
  - sorg:Place
mappings:
  jmeno:
    - sorg:name
    - dcterms:title
  obdobi: sorg:temporalCoverage
  lokalita:
    - dwc:verbatimLocality
    - relationIdentity: sorg:spatialCoverage
      relationType: sorg:Place
      valueIdentity: sorg:name
  gps_lat:
    - sorg:latitude
    - dwc:decimalLatitude
  gps_lon:
    - sorg:longitude
    - dwc:decimalLongitude
  popis_lokality: dwc:locationRemarks
  nadmorska_vyska:
    - dwc:minimumElevationInMeters
    - dwc:maximumElevationInMeters
    - dwc:verbatimElevation
  poznamky: sorg:description
staticIdentities:
  - identity: dcterms:hasPart
    value:
      - "@id": dktr:entity/5dcc2f395272463c88809c62
      - "@id": dktr:entity/5dcc2f395272463c88809c63
      - "@id": dktr:entity/5dcc2f395272463c88809c69
      - "@id": dktr:entity/5dcc2f395272463c88809c6a
      - "@id": dktr:entity/5dcc2f395272463c88809c6b
      - "@id": dktr:entity/5dcc2f395272463c88809c6c
      - "@id": dktr:entity/5dcc2f395272463c88809c6d
      - "@id": dktr:entity/5dcc2f395272463c88809c6e
      - "@id": dktr:entity/5dcc2f395272463c88809c6f
      - "@id": dktr:entity/5dcc2f395272463c88809c70
      - "@id": dktr:entity/5dcc2f395272463c88809c73
      - "@id": dktr:entity/5dcc2f395272463c88809c74
      - "@id": dktr:entity/5dcc2f395272463c88809c75
      - "@id": dktr:entity/5dcc2f395272463c88809c76
      - "@id": dktr:entity/5dcc2f395272463c88809c77
      - "@id": dktr:entity/5dcc2f395272463c88809c78
      - "@id": dktr:entity/5dcc2f395272463c88809c79
      - "@id": dktr:entity/5dcc2f395272463c88809c7a
      - "@id": dktr:entity/5dcc2f395272463c88809c7b
      - "@id": dktr:entity/5dcc2f395272463c88809c8d
additionalMetadata:
```

File - datasets-mapping.yaml

```yaml
  - identity: dcterms:license
    value:
      "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: sorg:license
    value:
      "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: sorg:creator
    value:
      "@type": sorg:Person
      sorg:sameAs: http://orcid.org/0000-0003-3167-2128
      sorg:givenName: Ondřej
      sorg:familyName: Doktor
      sorg:name: Ondřej Doktor
```

File - records-mapping.yaml

```yaml
# record-mapping.yaml (entita Nález)
---
namespaceAliases:
  sorg: http://schema.org/
  obo: http://purl.obolibrary.org/obo/
  sio: https://semanticscience.org/resource/
  dwc: http://rs.tdwg.org/dwc/terms/
  dwciri: http://rs.tdwg.org/dwc/iri/
  dcterms: http://purl.org/dc/terms/
  fisho: http://bioportal.bioontology.org/ontologies/FISHO#
  dktr: http://purl.org/dktr/fair/
baseType: dwc:Occurrence
additionalTypes:
  - sorg:Thing
mappings:
  documentName:
    - sorg:name
    - dwc:recordNumber
    - dcterms:title
  genus: dwc:genus
  species: dwc:specificEpithet
  specimen_old_id: sorg:sameAs
  determination_note: dwc:identificationRemarks
  river_name:
    - fisho:FISHO_0000055
    - relationIdentity: dcterms:relation
      relationType: dcterms:Location
      valueIdentity: dwc:waterBody
      staticIdentities:
        - identity: dwc:decimalLatitude
          value: -25.671028
        - identity: dwc:decimalLongitude
          value: -53.932972
        - identity: dwc:verbatimLocality
          value: a. Deseado (San Antonio, Iguazu), provincie Misiones
, Argentina
        - identity: dwc:higherGeography
          value: South America | Argentina | Misiones | San Antonio
        - identity: dwc:municipality
          value: San Antonio
        - identity: dwc:stateProvince
          value: Misiones
        - identity: dwc:country
          value: Argentina
        - identity: dwc:countryCode
          value: AR
        - identity: dwc:continent
          value: South America
        - identity: dwc:locationRemarks
          value: Základní tábor, přítok Iguazu nad Cataratas.
        - identity: dwc:minimumElevationInMeters
          value: 198
        - identity: dwc:maximumElevationInMeters
          value: 198
        - identity: dwc:verbatimElevation
          value: 198 m
  fishingmethod:
```

File - records-mapping.yaml

```yaml
    valueIdentity: fisho:FISHO_0000133
    valueProjection:
      gillnets:
        "@id": fisho:FISHO_0000025
      electrofishing:
        "@id": fisho:FISHO_0000368
      castnet:
        "@id": fisho:FISHO_0000259
  standard_length:
    - fisho:FISHO_0000064
    - relationIdentity: dcterms:relation
      relationType: dwc:MeasurementOrFact
      valueIdentity: dwc:measurementValue
      staticIdentities:
        - identity: dwc:measurementType
          value: standard length
        - identity: dwc:measurementUnit
          value: cm
        - identity: sio:SIO_000221
          value:
            "@id": obo:UO_0000016
  head_depth:
    relationIdentity: dcterms:relation
    relationType:
      - dwc:MeasurementOrFact
      - fisho:FISHO_0000027
    valueIdentity: dwc:measurementValue
    staticIdentities:
      - identity: dwc:measurementType
        value: head depth
      - identity: dwc:measurementUnit
        value: cm
      - identity: sio:SIO_000221
        value:
          "@id": obo:UO_0000016
      - identity: sio:SIO_000563
        value:
          "@id": fisho:FISHO_0000457
      - identity: sio:SIO_000563
        value:
          "@id": obo:VT_0000038
staticIdentities:
  - identity: dwc:organismQuantity
    value: 1
  - identity: dwc:organismQuantityType
    value: individuals
additionalMetadata:
  - identity: dcterms:license
    value:
      "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: sorg:license
    value:
     "@id": https://creativecommons.org/licenses/by-nc-nd/4.0/
  - identity: dwciri:inDataset
    value:
      "@id": dktr:entity/accc2f3d5272463c8880a1d2
  - identity: dcterms:isPartOf
```

File - records-mapping.yaml

```
    value:
      "@id": dktr:entity/accc2f3d5272463c8880a1d2
  - identity: sorg:creator
    value:
      "@type": sorg:Person
      sorg:sameAs: http://orcid.org/0000-0003-3167-2128
      sorg:givenName: Ondřej
      sorg:familyName: Doktor
      sorg:name: Ondřej Doktor
```

Netextová příloha práce

Netextová příloha této práce (počítačový program) je dostupná pod tímto persistentním identifikátorem:

http://purl.org/dktr/master-thesis-repo

V době uzávěrky této práce tento identifikátor směruje na GitHib repozitář:

https://github.com/drml/jcu-fairportal/