FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Využití ontologické analýzy pro zajištění sémantické interoperability marketingových dat
Student:	Jana Martínková
Vedoucí:	doc. Ing. Robert Pergl, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Informační systémy a management
Katedra:	Katedra softwarového inženýrství
Platnost zadání:	do konce letního semestru 2021/2022

Pokyny pro vypracování

Téma přispívá k projektu Datového inkubátoru dat pro marketingové analýzy. Cílem práce je ontologická analýza klíčových domén a jejich propojení s datovými sadami tak, aby byla umožněna jejich sémantická interoperabilita.

1. Seznamte se s projektem Datového inkubátoru, problematikou sémantické interoperability, Unified Foundational Ontology, jazykem OntoUML a nástrojem OpenPonk.
2. Ve spolupráci s vedoucím vyberte několik klíčových domén a souvisejících datových sad.
3. Vytvořte ontologické konceptuální modely těchto domén.
4. Propojte ontologické konceptuální modely s datovými sadami a vytvořte pravidla pro jejich mapování.
5. Zdokumentujte své řešení a přínos pro zajištění sémantické interoperability.

Elektronicky schválil/a Ing. David Buchtela, Ph.D. dne 11. února 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Využití ontologické analýzy pro zajištění sémantické interoperability marketingových dat

Jana Martínková

Katedra softwarového inženýrství
Vedoucí práce: doc. Ing. Robert Pergl, Ph.D.

12. května 2021

Poděkování

Ráda bych poděkovala doc. Ing. Robert Perglovi, Ph.D. za cenné rady, věcné připomínky a vstřícnost při konzultacích a vypracování bakalářské práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené.

V Praze dne 12. května 2021 . . . . . .

Odkaz na tuto práci

Martínková, Jana. *Využití ontologické analýzy pro zajištění sémantické interoperability marketingových dat.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2021.

Abstrakt

Tato bakalářská práce se zabývá využitím ontologické analýzy k zajištění významové propojitelnosti dat.

V řešení byl využit ontologický jazyk konceptuálního modelování OntoUML a veškeré modely byly tvořeny v platformě OpenPonk. Výsledkem práce jsou konceptuální ontologické modely vybraných domén, které jsou propojené s určitými datovými sadami. Vytvoření konceptualizace domény předchází ontologická analýza, která spočívá v přesném vymezení pojmů domény. Tato činnost vyžaduje nejčastěji součinnost s doménovými experty. Provázání ontologického modelu s atributy datových sad bylo stanoveno na obecnější a přesnější úrovni, pro zajištění co nejlepší souvislosti.

Celkem bylo zpracováno 8 datových sad, ze kterých vzniklo 7 ontologických modelů. Výsledky práce jsou určené pro projekt vyvíjející datovou platformu pro efektivní zpracování dat. V závěru práce je zdůrazněna dosažená sémantická interoperabilita mezi heterogenními datovými sadami.

Klíčová slova konceptuální model, ontologická analýza, projekt Nest BDA, sémantická interoperabilita, FAIR data, OntoUML, UFO

Abstract

This bachelor thesis deals with the use of ontological analysis to ensure semantic interoperability of data.

The ontological language of conceptual modeling OntoUML was used in the solution and all models were created in the OpenPonk platform. The result of the work are conceptual ontological models of selected domains, which are connected with certain data sets. The creation of the conceptualization of the domain is preceded by ontological analysis, which consists in the precise definition of the terms domain. This activity most often requires the cooperation with domain experts. The linking of the ontological model with the attributes of the datasets was determined at a more general and precise level, to ensure the best possible coherence.

A total of 8 data sets were processed, from which 7 ontological models were created. The results of the work are intended for a project developing a data platform for efficient data processing. At the end of the work, the achieved semantic interoperability between heterogeneous data sets is emphasized.

Keywords konceptual model, ontological analysis, project Nest BDA, semantic interoperability, FAIR data, OntoUML, UFO

Obsah

Seznam obrázků

Úvod

Vyhledávání informací a relevantních dat na internetu se stalo téměř samozřejmostí pro každého z nás. Internet obsahuje velké množství užitečných informací, které lze použít k dalšímu výzkumu. Jejich použití je ale leckdy velmi obtížné. Využitím moderních technologií je možné data strojově zpracovávat, ale příprava relevantních dat do strojově zpracovatelné podoby je obvykle časově velmi náročná. Je třeba data získat, ověřit jejich původ, splnit licenční podmínky užití, převést je do jednotného formátu, a navíc zajistit jejich interoperabilitu. V případě, že by i tato příprava dat mohla probíhat strojově, pak by se veškeré vyhledávání i užití značně urychlilo. Tato práce je zaměřena na zajištění sémantické interoperability dat.

Výsledek této práce je určen pro projekt Nest BDA, který má za cíl vytvořit datovou platformu umožňující lepší přístupnost k datům, jejich koncentraci na jednom místě a umožnění interoperability bez nutné lidské činnosti. V projektu Nest BDA se jedná primárně o data z odvětví marketingu, médií a komunikace. Toto téma jsem si zvolila, protože mě velmi zajímá ontologická analýza a chtěla jsem přispět k urychlení zpracování a využití dat.

V této práci se zabývám ontologickou analýzou klíčových domén a sémantickým propojením relevantních datových sad, které budou použity v Nest BDA. Na vybraných doménách a s nimi souvisejících datových sadách bude popsána ontologická analýza domény, následné propojení se souvisejícími sadami a zvýrazněna dosažená sémantická interoperabilita dat.

Kapitola 1

Cíl práce

Hlavním cílem bakalářské práce je ontologická analýza definovaných klíčových domén a následné propojení s dostupnými souvisejícími datovými sadami, za účelem zajištění jejich sémantické interoperability v rámci projektu Nest Big Data Arena, který realizuje společnost Remmark, a.s.

Cílem rešeršní části práce je vytyčení cílů projektu Nest Big Data Arena, popis FAIR principů o efektivním publikování dat, definice sémantické interoperability a ontologického konceptuálního modelování. Popis ontologie UFO, konceptuálního jazyka OntoUML a nástroje OpenPonk.

Cílem praktické části práce je vytvoření ontologických konceptuálních modelů vybraných domén a jejich následné propojení s datovými sadami za pomoci vydefinovaných pravidel pro mapování. Na propojených modelech zdůraznit přínos spočívající v jejich sémantické interoperabilitě a popsat postup zpracování datových sad v projektu Nest BDA.

Kapitola 2

Teoretická část

V teoretické části této práce je charakterizován cíl efektivního publikování dat projektu Nest Big Data Arena a jakým způsobem jej bude na úrovni konceptualizace dostupných dat dosaženo. Je popsána iniciativa FAIR a její principy, které jsou v postupu využívány. Dále je vysvětlena důležitost ontologie v informačních systémech, hlavně v odvětví konceptuálního modelování. Je přiblížena ontologie UFO a modelovací jazyk OntoUML, který je na ní založen. Nakonec je představen nástroj OpenPonk, ve kterém se tvoří konceptuální modely potřebné pro tento projekt.

2.1 Projekt Nest BDA

Realizátorem projektu Nest Big Data Arena, je česká společnost Remmark, a.s., která se specializuje na informační kampaně hlavně v oblasti veřejného sektoru. Projekt je podporován Evropskými strukturálními a investičními fondy prostřednictvím Operačního programu Praha - pól růstu ČR v rámci 3. výzvy Pražského voucheru na inovační projekty. [1]

Cílem tohoto projektu je vytvoření datové platformy, která bude propojovat informace, data a výzkumy z oblasti marketingu, medií a komunikace. Největším přínosem této platformy je koncentrace dat na jednom místě, zviditelnění propojitelnosti dostupných informací a zvýšení uživatelské dosažitelnosti dat. V minulosti byly roztroušené veřejně dostupné nespolupracující informace a data uchovávána v různých úložištích a pro uživatele tak nebyla dostatečně přístupná. [2]

U dostupných datových sad v platformě budou uvedena relevantní odvětví, pro která jsou informace v sadě podstatná a metodika popisující data a práci s nimi. Platforma bude poskytovat několik funkčních částí, z nichž první bude přehled všech datových sad. Další úsek bude umožňovat jejich vyhledávání a filtrování pomocí formátu, původu, klíčových slov a výrazů, které jsou obsaženy v popisu sady. Poslední částí bude možnost zadání konkrétních dotazů, které se nad obsahem datové platformy zodpoví. [3]

Cílovou skupinou platformy jsou uživatelé, kteří plánují marketingové aktivity. Může se jednat o veřejnou správu, média nebo vývojářské skupiny, které chtějí získat data jak pro potřeby dalšího vývoje, tak ke spolupráci s různými aplikacemi. Uživatelé budou na různé technické úrovni, primárně rozděleni na 4 typy [4]:

1. Běžný uživatel

   Základní uživatel, datový laik, který využívá platformu bez odborných znalostí informačních technologií. Tento typ uživatele ví, jaké informace chce vyhledat a s pomocí jakých klíčových slov. Jedná se například o novináře, zástupce médií, studenty, pedagogy nebo zaměstnance veřejné správy.

2. Běžný analytik

   Týká se analytiků v mediální agentuře, kteří mají dobré odborné znalosti v oblasti informačních technologií a umí pracovat s programy na zpracování mediálních a marketingových dat. Tento typ uživatelů data využívá v rámci své pracovní náplně. Platformu budou využívat k tvorbě podkladů pro mediální strategie, prezentací, analýzy či psaní analytických článků. Mezi tento typ uživatelů se řadí zkušení analytičtí novináři, odborní pedagogové, edukovaní studenti, nákupčí a plánovači v mediálních a reklamních agenturách nebo marketingových odděleních.

3. Pokročilý analytik

   Jedná se o zkušenější analytiky s odbornější znalostí informačních technologií, kteří dokáží pracovat i s primárními daty. Představitelé tohoto typu uživatele budou platformu využívat k tvorbě marketingových strategií, prezentací a analýz. Jedná se například o senior plánovače a tvůrce strategií, senior zaměstnance marketingových společností, datové novináře, junior zaměstnance výzkumných agentur nebo analytiky v médiích.

4. Profesionál

   Profesionál je typ uživatele kvalifikovaného v oboru informačních technologií, který spravuje či vyvíjí softwarové systémy. Na základě znalosti veřejných API platformy ji může využít pro vývoj svého vlastního nástroje, nebo rozšíření již existujícího, pro potřebu managementu či marketingového oddělení.

Při zajištění propojitelnosti a opětovné použitelnosti dat se v projektu vycházelo z principů iniciativy FAIR, která se zabývá efektivním publikováním dat.

2.2 FAIR

FAIR je iniciativa, která vznikla v návaznosti na konferenci Jointly designing a Data FAIRPORT, kterou zorganizoval v roce 2014 správce nizozemské datové struktury, institut Dutch Techcentre for Life Sciences (DTL), ve spolupráci s Netherlands eScience Center a the Lorentz Center. Výsledkem konference, během které byly projednávány pravidla pro udržitelné a efektivní publikování dat, byla dohoda o vytvoření principů zajišťujících nalezitelnost (findability), dostupnost (accessibility), provázanost (interoperability) a opětovnou použitelnost (reusability) dat. Z těchto vytyčených cílů tak vznikl i název iniciativy FAIR. [5]

Principy nesuplují způsob implementace ani neurčují technologie, které by se k dosažení cíle měli využívat. Jedná se o návody či postupy, jejichž aplikováním je možné dosáhnout znovupoužitelnosti a dosažitelnosti dat, jak pro člověka tak pro stroje, neboli datové agenty.

Základní principy, které se dále dělí na jednotlivé požadavky, je možné využít v potřebné míře a složení. Ani citlivá data tak nejsou překážkou pro úspěšnou FAIRifikaci dat (viz 2.2.2). [6]

Pojem „datový objekt“, který je ve FAIR principech využíván, je zobecněný pojem „digitální objekt“, jenž dle definice [7] sestává z dat, metadat a globálního identifikátoru. „FAIR datový objekt“ je potom datový objekt, který se řídí principy FAIR.

Spojení „strojově čitelný datový objekt“ je v tomto případě definován jako datový objekt, který poskytuje datovému agentovi dostatek informací, byť se s ním agent nikdy nesetkal, aby dokázal [6]:

1. Získat obecnou identifikaci objektu, jak strukturální, tak sémantickou.
2. Vyhodnotit, zda se jedná o data obsahující informace, které jsou v kontextu se zadaným vyhledáváním a jsou pro problematiku relevantní.
3. Určit, zda je licenčně možné použít tento datový objekt, případně jej používat s definovaným omezením.
4. Učinit vhodné kroky, které by provedl člověk.

Základními principy jsou vyhledatenost, přístupnost, interoperabilita a znovupoužitelnost. Tyto principy se dále dělí na dílčí požadavky. [8]

2.2.1 Vyhledatelnost

První FAIR princip říká, že datový objekt by měl být jednoduše vyhledatelný pro člověka i pro datového agenta, protože strojově čitelné objekty jsou nezbytné pro automatické objevení dat.

1. (Meta)data mají přiřazeny globálně unikátní trvalé identifikátory

   Jedná se o nejdůležitější požadavek všech FAIR principů, protože v ostatních bodech se využívá faktu existence takovéhoto identifikátoru.

   Globálně unikátní a trvalé identifikátory se přiřazují každému datovém objektu. Proto je možné propojit jednotlivé prvky objektu a zároveň zamezit nejednoznačnosti. Díky nim mohou stroje smysluplně interpretovat data. Navíc přispívají ke správné citaci při jejich opětovném použití. [8]

   Globálně unikátní identifikátor, vytvořený autorizovaným generátorem, se skládá ze dvou částí. První identifikuje lokální autorizovaný generátor či platformu, kde se dataset vyskytuje. Druhá část je lokálně unikátní řetězec, který identifikuje přímo datový objekt. [7]

   Dlouhodobé udržování těchto identifikátorů je finančně i časově náročné, takže mají tendenci se v průběhu času zneplatnit. Tento požadavek FAIR principu se snaží o jejich trvalost, alespoň v nějaké míře.

2. Datový objekt s bohatě popsanými metadaty

   Podle myšlenky FAIR principů by autor dat neměl předpokládat, že ví, kdo a za jakým účelem bude chtít data použít. Metadata by tak měla být popisná rozsáhlá včetně informací o kontextu, stavu dat a hlavně by měla obsahovat identifikátory dat, jež popisují. Bohatě popsaná metadata by

měla poskytnout dostatek informaci pro datové agenty, kteří tak budou moci převzít provedení rutinních a zdlouhavých třídění a prioritizování relevantních úkolů, což nyní musí vykonávat lidé.

3. Metadata jasně a explicitně zahrnují identifikátor dat, která popisují

   Protože jsou metadata a popisovaná data většinou oddělené soubory, pak pomyslné spojení mezi souborem metadat a datasetem by mělo být zprostředkováno uvedením globálního identifikátoru ve všech částech datového objektu.

4. (Meta)data jsou registrována nebo indexována v prohledávatelném zdroji

   Přítomnost identifikátorů a rozsáhlé popisy metadat nezaručí vyšší nalezitelnost datových zdrojů. Data by měla být publikována v dostupném zdroji, jinak mohou zůstat nevyužitá jen proto, že nikdo neví o jejich existenci. [8]

V projektu Nest BDA je využíván unikátní řetězec, tzv. Identifier. Ten je explicitně uveden jak v datech, tak jejich metadatech i popisných modelech. Proto je na vyšších úrovních datové platformy umožněna spolupráce jednotlivých prvků datového objektu bez nutnosti jejich přímé fyzické provázanosti.

2.2.2 Přístupnost

Druhým FAIR principem je znalost přístupu k datům. Jakmile jsou data nalezena, pak je pro konzumenta, ať člověka nebo datového agenta, důležitá znalost technické dostupnosti.

1. (Meta)data lze získat se znalostí identifikátoru za pomoci standardizovaného komunikačního protokolu

   Tento požadavek principu uvádí, že FAIR data by měla být dostupná i bez specializovaných nástrojů či komunikačních metod. Nutná je pouze znalost identifikátoru požadovaného datového objektu.

   a) Protokol je bezplatný a univerzálně implementovatelný
      Aby se maximalizovalo znovupoužití dat, měl by použitý protokol být bezplatný, open source a také implementovatelný na většinu zařízení. Tento požadavek principu ovlivňuje výběr úložiště, kde se data budou publikovat.

   b) Protokol podporuje autentizaci a autorizaci
      FAIR data by měla mít jasně specifikované podmínky získání oprávnění (viz 1a). Nejlépe definované podmínky přístupnosti k datům dokáže datový agent buď sám splnit, nebo na ně dovede uživatele upozornit. Často se využívá nutnosti vytvoření uživatelského účtu pro dané úložiště, které umožní ověření např. vlastníka datového

objektu a nastaví mu jeho specifická práva. Tento požadavek principu také ovlivňuje výběr úložiště, kde se budou data publikovat.

2. Metadata by měla být dostupná i v případě, že data již nejsou dostupná

   Data s postupem času degradují a udržování jak jejich kvality, tak i dostupnosti je finančně náročné. Při vyhledávání pak neplatné odkazy směřující na neexistující, či neaktuální data zdržují. Udržování a uchovávání metadat je jednodušší i finančně méně náročné, a proto by metadata měla přetrvávat i v případě, že data již dostupná nejsou. [8]

2.2.3 Interoperabilita

Aby byla data globálně využitelná bez nutnosti ručního zpracování je nutné zajistit jejich interoperabilitu jak technickou tak sémantickou. Interoperabilita spočívá v možnosti propojení systémů na různých rovinách komunikace tak, aby došlo k téměř automatickému spojení (viz 2.3).

1. (Meta)data využívají formální přístupný sdílený a univerzálně použitelný jazyk pro reprezentaci znalostí

   Datový objekt by měl být strojově čitelný, bez toho, aby stroj musel znát speciální algoritmy, překladače nebo mapování. Každý stroj by tak měl mít alespoň znalost formátů pro výměnu dat s jiným systémem.

2. (Meta)data využívají slovníky, které jsou v souladu s FAIR principy

   Slovník používaný k popisu datového objektu je potřeba dokumentovat a také označit globálním unikátním trvalým identifikátorem, protože se jedná o součást datového objektu. Aby byl význam dat stále srozumitelný musí být slovník přístupný všem, kteří mají přístup k datovému objektu.

3. (Meta)data obsahují odkazy na jiná (meta)data

   Mezi (meta)datovými zdroji je cílem vytvořit co nejvíce vazeb a odkazů k rozšíření kontextových znalostí o datech. Pro větší konkrétnost je možné v odkazu uvést, zda se jedná o doplňující informace potřebné k přečtení datového objektu, o navazující datový objekt nebo o potřebný datový objekt k plnohodnotnému přečtení. Všechny datové objekty je nutné citovat pomocí globálních trvalých unikátních identifikátorů. [8]

V projektu Nest BDA je dosaženo interoperability na sémantické úrovni v rámci datové platformy díky využití strukturálních modelů. Ty zajišťují významové sjednocení pojmů a uspořádání domén, jenž jsou popisovány datovými sadami. Nejedná se o slovník ve formě seznamu vysvětlující pojmy z datového objektu, ale o přesnější a jasnější popisy v podobě modelů (viz 2.4). S jejich pomocí je splněn i požadavek FAIR principů o odkazech mezi datovými objekty.

2.2.4 Opětovné použití

Poslední FAIR princip spočívá v opětovném použití dat, čehož se dosáhne správným popisem metadat a datasetů tak, aby mohly být kombinovány v různém složení.

1. (Meta)data jsou popsána rozsáhlým množstvím relevantních atributů

   Připojením relevantních popisných štítku budou datové objekty snadněji objevitelné, čímž se zvýší i jejich znovupoužitelnost. Tento popis by měl jak člověku, tak stroji pomoci rozhodnout, zda jsou právě tyto data v prohledávaném kontextu užitečné. Může se jednat o výrobce, značku stroje či senzoru, který data vytvořil, účel vytváření dat, určité omezení, období generování či sběru dat atd. Autor by neměl, jak je zmíněno i v jednom z předchozích FAIR principů, předpovídat identitu a potřebu konzumenta dat. Z tohoto důvodu by měly být metadata co nejrozsáhlejší a poskytovat i informace, které se mohou zdát irelevantní.

   a) (Meta)data jsou vydávána s jasnou licencí pro přístup a využívání dat

      Podmínky použití dat by měly být srozumitelné lidem i datovým agentům. Jejich nejednoznačnost může vést k omezení opětovného použití dat organizacemi, které se snaží licenční omezení dodržovat.

   b) (Meta)data jsou spojena s podrobným původem

      Aby bylo možné data při jejich použití správně citovat je nutné jasně zveřejnit, ideálně ve strojově čitelné podobě, odkud data pochází a koho, příp. jak, citovat. Měl by být uveden i způsob postupu vytvoření dat.

   c) (Meta)data splňují standardy doménové komunity

      Pokud v komunitě dané domény existují určité standardy či postupy pro uložení, popsání a sdílení dat, pak je třeba je dodržovat. Pro konzumenty je snazší používat datové soubory s jimi známými typy, formáty a organizací dat. Měla by se také užívat společná slovní zásoba a doménově profesní výrazy. Pokud se autor rozhodne pro odchýlení od standardu, pak by taková informace měla být zmíněna v metadatech. [8]

V projektu Nest BDA se využívá vyhledávání datových sad podle jejich klíčových slov a výrazů využitých v popisu dat, včetně přesného popisu původu dat. [3] V sémantických modelech se pro přesnost užívají i významově specifické výrazy dané domény, které byly konzultovány s oborovými specialisty.

2.2.5 FAIRifikační proces

FAIR data principy se mohou řídit datové objekty a podporující infrastruktury (např. vyhledávače). Požadavky na zjistitelnost a dostupnost lze aplikovat hlavně na metadata, naopak opětovná použitelnost a interoperabilita musí být zajištěna na úrovni dat.

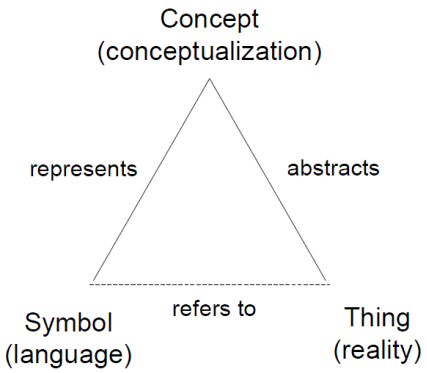
K dosáhnutí FAIR datových objektů lze postupovat tzv. FAIRifikačním procesem, jehož kroky jsou následující [9]:

1. Načtení dat – Získání dat, která se budou FAIRifikovat.
2. Analýza načtených dat – V tomto kroku probíhá identifikace a analýza dat, přičemž záleží na formátu jejich uložení. Zjišťuje se struktura dat, jaké koncepty a relace jsou v datech znázorněny.
3. Sémantický model – Pro datovou sadu se definuje její sémantický model, který jednoznačně popisuje význam a vztah entit.
4. Propojení dat – Díky vytvoření sémantického modelu datové sady se entity stanou propojitelné. Každý objekt bude mít svůj jednoznačný identifikátor a skrze něj se lze dotazovat.
5. Přiřazení licence – Datovému objektu se přiřadí licence omezující jeho použití. Tato informace je nutná i v případě, že se jedná o data s plně otevřeným přístupem.
6. Definování metadat pro datovou sadu – Pro splnění většiny FAIR principů jsou klíčovou součástí důkladně popsaná metadata.
7. Nasazení datového zdroje – Publikování či zpřístupnění FAIR datového objektu včetně licence a metadat, aby mohla být data vyhledatelná i v případě, že vyžadují ověření a autorizaci.

FAIRifikační proces popisuje náplň konání metadata týmu, pod vedení doc. Ing. Roberta Pergla, Ph.D., který v rámci projektu Nest BDA zajistil metodologický popis dat z dostupných datových sad, což představuje první a druhý bod procesu. Kroky 3 a 4 spočívaly ve vytvoření sémantických modelů, které sloužily k pochopení významu a struktury informací a hlavně k propojení oddělených datasetů na základě smyslu pojmů v nich obsažených.

2.3 Sémantická interoperabilita

Interoperabilita informačních systémů spočívá v jejich propojení na několika různých úrovních. Jedná se například o spojení na komunikační rovině, které umožňuje vzájemný přístup k informacím. Mezi další úrovně patří syntaktická interoperabilita, díky níž je možné jednotně zpracovávat sdílená data, protože jsou uložena ve standardních syntaktických strukturách. [10]



Obrázek 2.1: Ullmanův trojúhelník. [11]

Interoperabilita na sémantické úrovni dle [12] je schopnost kombinovat datové prvky různých modelů, schémat či databází na významové hladině. Umožňuje vyhledávat informace v heterogenních datových úložištích bez nutnosti přípravy kompatibilnosti dat na sémantické rovině. Využitím standardů a přesných definic pojmů lze předejít situacím, kdy jeden termín popisuje více objektů, nebo naopak objekt je v různých schématech pojmenován odlišně.

Jedinou možností, jak validně sémantickou interoperabilitu zajistit, je využití schémat, které popisují objekty obsažené v doméně datové sady. Principy FAIR tento popis označují jako „slovník“ (viz 2.3). Ty ale, ve významu seznamu použitých pojmů, nejsou plně dostačující. K vytvoření schémat dat je nutné nejdříve pochopit vztahy mezi objekty v reálném světě. [13] Tyto sémantické popisy se nazývají konceptuální modely.

2.4 Konceptuální modelování

Konceptuální modelování je činnost, jejíž cílem je získat formální popis relevantních fyzických a sociálních aspektů části reality. Jedná se o deskripci domény a aktivit, které se v ní odehrávají. S výsledným popisem by měli souhlasit všichni aktéři, kteří se v dané doméně vyskytují a rozumí ji. Popis zároveň slouží jako komunikační a edukační nástroj pro přiblížení domény a vysvětlení aktivit v ní. [14]

Výsledný model popisuje abstrakci relevantních prvků v doméně konceptualizace. [15] Ullmanův trojúhelník (viz obrázek 2.1) vyjadřuje vztah mezi konceptualizací, jazykem konceptualizace a abstrakcí domény, která je konceptualizována. Přerušovaná čára mezi částí reality a symboly modelovacího jazyka značí, že je vždy tento vztah zprostředkován konceptualizací. [16]

K vytvoření a uchování konceptuálního modelu je nezbytný jazyk, který

bude konceptualizaci reprezentovat stručně, úplně a jednoznačně. Jazyk může být doménově specifický, jako například jazyk DEMO, nebo doménově nezávislý, to je například jazyk OntoUML [15], který byl využit v projektu Nest BDA.

Velmi často se stává, že ve vícero modelech je jeden reálný objekt nazván rozdílnými termíny, nebo naopak jedno označení je použité pro dva odlišné koncepty. Tato nejednoznačnost je původcem významové nekompatibilnosti a zdroje tak mezi sebou nejsou schopni sémantické interoperability. [17] Pro určitou soudržnost definovaných pojmů by základem konceptuálního modelování dle [18] měla být ontologie (viz 2.6).

2.5 Ontologie

Ontologie je filosofická disciplína, jejíž vznik pramení z Aristotelovy První filosofie. [19] V souboru spisů, také známém jako Metafyzika, se Aristoteles zabýval otázkou bytí a jsoucna. [20]

Termínem Metafyzika jej pojmenoval Andronikos z Rhodu. Název původně vznikl z důvodu umístění souboru spisů v knižním katalogu za spisy fyzikální. Později dostal název podstatnější význam. Bytí je dle Aristotela skryté za jevy, které se projevují a je tedy pro vnímání člověka nepřístupné. Metafyzika v tomto významu označuje zkoumání oblasti, která se skrývá „za fyzikou“, tedy za projevenými úkazy. [21]

Aristotelova První filosofie zkoumá jsoucno, zabývá se otázkami a vyslovuje teorie o tom, co ve světě je skutečné, co je podstatou reality a jaká je vlastní povaha věcí. Popisuje skutečnost, že každá věc, vlastnost, nebo činnost nějak existuje a je tedy jsoucnem. Jsoucno, které není viditelné ani představitelné, je základní substancí bytí. [20]

Ontologie, jakožto filosofické odvětví, řeší mimo jiné otázku původu bytí a zkoumá, zda je základem světa hmota, idea, či božský princip. Aristoteles pojednává o primárních látkách, mezi které řadí vodu, vzduch, zemi a oheň. Na příkladu masa ukazuje, že předměty se neskládají pouze z těchto počátečních hmot, ale obsahují ještě něco jiného. Rozložením masa, dle Aristotela, vznikne oheň a země. Pokud je ale oheň a země spojeno, ne nutně vznikne maso. Tím Aristoteles poukazuje na další složku, která nemá původ v základních hmotách. Popisuje ji jako příčinu vzniku objektu a nazývá to jeho podstatou. Druhové přiřazení pak spočívá v porovnání těchto podstat. [19]

Ontologie, na rozdíl od jiných filosofických odvětví, které zkoumají objekty spadající do studované domény, se zabývá mezikategoriálními vztahy a postavením objektů v rámci domény. [20]

Tato filosofická disciplína využívá takových pojmů jako je část, kompozice, systém, relace, stav, událost, změna, možnost, proces, prostor a čas, které se využívají i v informačních technologiích, zejména v oblasti konceptuálního modelování (viz 2.4). [17]

Dnešní význam ontologie spočívá v úplné definici myslitelných a vnímatelných objektů včetně relací mezi nimi. Proto by se jakýkoli reprezentační systém měl řídit základy ontologie. To je možné pozorovat v různých odvětví jako je znalostní, softwarové nebo i podnikové inženýrství. [15]

2.5.1 Ontologie v informačních technologiích

V souvislosti s informačními technologiemi byl pojem ontologie poprvé použit Mealym v roce 1967 ve studii o základech zpracování dat. [22] Postup analýzy dat pomyslně rozděluje na tři různé části. První je existence konceptů, neboli pojmů, v reálném světe. Další částí je lidské vnímání existence těchto konceptů a poslední je způsob syntaktického zachycení jejich postavení. V otázce vnímání existence věcí bez ohledu na jejich reprezentaci odkazuje právě na filosofickou ontologii.

Pojem ontologie má dnes v informačních technologiích odlišný význam od toho filosofického. Jedná o formální strukturu, neboli konceptuální model, který obsahuje pojmy a vztahy mezi nimi. Jedno z možných rozdělení ontologie konceptů, které zmiňuje i Guizzardi [17], je do 4 různých vrstev (viz obrázek 2.2):

1. Top-level ontologie – Popisují obecné koncepty, které nejsou závislé na doméně nebo zkoumané otázce. Jedná se například o čas, objekt, událost, akci, nebo prostor.

2. Doménové ontologie – Charakterizují pojmy spojené s obecnou doménou. Specifikují významné pojmy specializace v rámci top-level ontologie. Může jít například o doménu léčiv, nebo automobilového průmyslu.

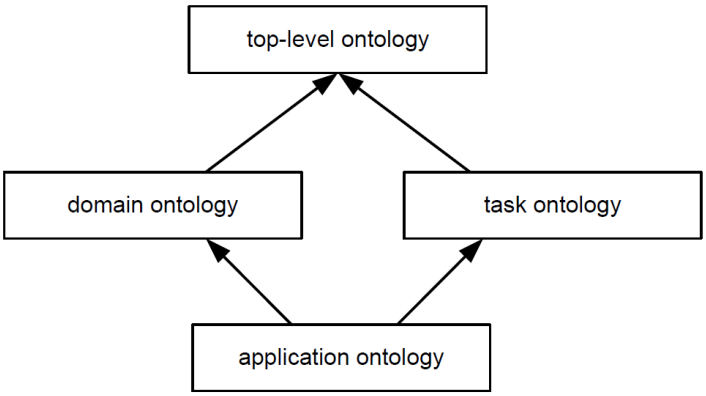
3. Ontologie problematiky – Vystihují výrazy spojené s obecnou problematikou, například diagnostikou či prodejem, a uvádí tyto koncepty v souvislosti s top-level ontologií.

4. Aplikační ontologie – Definuje pojmy, které závisí jak na určité doméně, tak na problematice. Představované koncepty nejčastěji odpovídají rolím (viz 2.8.4.1), které hrají entity domény ve vztahu s prováděním určité činnosti.

Dále jsou přiblíženy výskyty ontologie v různých odvětvích informačních technologií. které spočívají v motivaci vybudovat co nejpřesnější strukturu domén, jejichž se zaměřená problematika dotýká.

2.5.1.1 Ontologie v doménovém inženýrství

Údržba softwarového systému je velmi nákladná, což je značnou motivací pro znovupoužití softwaru i na jiných úrovních, než je programovací kód. Pojmem doménové inženýrství je označován proces, který se skládá primárně



Obrázek 2.2: Klasifikace typů ontologií: Šipky znázorňují závislosti mezi vrstvami ontologie. [23]

z doménové analýzy a designu. Výsledkem doménové analýzy je model domény, který definuje objekty, jež považují odborníci domény za důležité a přínosné, relace mezi nimi a události, ve kterých se vyskytují. Přínos takového modelu slouží mimo jiné k uchování znalostí o doméně pro příští využití, vysvětlení nejasností a k predikci skutečností, které nejsou přímo viditelné. [17]

2.5.1.2 Ontologie ve znalostním inženýrství

Znalostní systémy se do své znalostní báze snaží začlenit kroky, které napodobují rozhodovací činnost experta. Nejnákladnější částí ve vývoji expertního systému je proces získávání vědomostí a proto byl později navrhnut jiný způsob pojetí znalostních bází. Dle něj by neměla nahrazovat myšlení lidí, ale měla by obsahovat objektivní realitu. Dobře strukturovaná a reálná doména znalostí bude vhodnější pro opětovnou použitelnost ve znalostních systémech různých problematik. [17]

2.5.1.3 Ontologie a sémanticky web

Obsah dnešních webových zdrojů, který je strojově čitelný, zcela závisí na lidské schopnosti rozumět sdělení. [17] Další úroveň webu, označená jako sémantický web, by umožnila i strojům chápat informace na webových stránkách. Záměr této nadstavby spočívá v rozšíření dnešní webové struktury uzlů o logickou vrstvu, která bude využívat definovaná pravidla, vyvozovat závěry, vybírat postupy a odpovídat na zadané otázky. [24] Docílit toho lze pomocí anotací webových zdrojů metadaty reprezentující pojmy, data i pravidla, jak nad daty uvažovat. [17] Přítomnost této logické nadstavby v podobě strukturálních modelů přinese i další zlepšení. Bude možné vyhledávat v rámci

omezení webových zdrojů, které se odkazují na daný koncept v modelu, místo zadaných a leckdy významově nepřesných klíčových slov. [18]

2.6 Unified Foundational Ontology

V roce 2002 Guizzardi a kol. zahájili výzkumnou analýzu konceptuálních modelovacích jazyků s cílem vyvinout ontologický základ pro tyto jazyky. Výzkum byl motivován myšlenkou, že explicitní definice základů a dodržování určitého ontologického závazku je pro konceptuální modelování zásadní. Jakýkoli pokus o rozvoj základů pro konceptuální modelování by měl brát v potaz lidské vědomosti i jazykové schopnosti. Původní záměr výzkumného týmu bylo propojit základy ontologií General Formalized Ontology (GFO) z Německa a Descriptive Ontology for Linguistic and Cogintive Engineering (DOLCE) z Itálie, což bylo možné z toho důvodu, že obě mají ontologický základ v kategorii čtyř tříd. Proto dostala výsledná ontologie název Unified Foundational Ontology. [18]

Kategorické rozdělení do čtyř tříd je idea pocházející z Aristotelova spisu Kategorie, kde jsou jsoucna klasifikována na čtyři třídy. Prvním rozdělením je rozlišováno to, co vypovídá o nějakém podmětu a naopak to co o žádném nevypovídá. Výraz, který o něčem vypovídá, či něco označuje se nazývá obecninou (univerzálem) a jeho opak se nazývá jednotlivinou (individuem). Jako příklad lze uvést pojem „člověk“, který jasně vystihuje konkrétní objekt a jedná se tedy o univerzál. Druhé rozdělení odlišuje to, co nutně potřebuje podmět k existenci a to, co podmět k existenci nepotřebuje. Nutná jsoucna nazývá Aristoteles esenciální, naopak ty, které jsou nahodilá nazývá akcidentální. [25]

GFO a DOLCE ale neobsahovaly explicitní definici entit a vztahů, které jsou k validnímu popsání domény potřeba, a tak bylo nutné ontologii doplnit o teorie z odvětví filosofické formální ontologie, kognitivní vědy, lingvistické logiky a filosofické logiky. [15,18] UFO slouží jako báze pro doménově nezávislé ontologické jazyky, na kterých lze stavět ontologie pro specifické domény. [26]

UFO ontologie je rozdělena do tří částí, podle aspektů reality [18]:

1. UFO-A – Jedná se o ontologii endurantů, která se zabývá strukturálními aspekty konceptuálního modelování. Je založena na idee kategorického rozdělení do čtyř tříd.

2. UFO-B – Ontologie perdurantů, která se zabývá událostmi, jejich časovému uspořádání a procesy.

3. UFO-C – Ontologie postavena na UFO-A a UFO-B, která se zabývá sociálními aspekty.

Ontologie UFO je využívána v odvětví analýzy, re-designu, k propojení různých modelovacích jazyků a také jako základ pro ontologie specifických domén. [15,18]

2.7 Unified modeling language

Unified modeling language (UML) je standardní modelovací jazyk, který se využívá v oblasti business analýzy, softwarové architektury a vývoje, ke specifikaci procesů a struktury systémů. Jazyk byl vyvinut Grady Boochem, Ivar Jacobsonem a Jim Rumbaughem a v roce 1997 byl přijat Object Management Group jako standard.

UML obsahuje několik druhů notací pro specifické diagramy. V základu se UML specifikace dělí na strukturální a behaviorální diagramy. [27] Mezi strukturální diagramy je zařazen UML Class diagram, který tvoří notační základ pro jazyk OntoUML. [26] Z toho důvodu je níže blíže popsaná notace UML Class diagramu.

2.7.1 UML Class diagram

UML Class diagram se řadí mezi strukturální diagramy jazyka UML, ty popisují uspořádání části systému na různých úrovních. Class diagram pak zachycuje strukturu tříd popisovaného systému. Znázorňuje jednotlivé třídy, rozhraní, jejích vlastnosti, omezení a vazby mezi nimi.

Základním prvkem UML Class diagramu je třída. Třída definuje vzor pro objekty stejného typu, určuje jejich vlastnosti a omezení. [27] Objekt je potom instancí určité třídy. Má explicitně definované hodnoty vlastností a je angažován v relacích určených třídou. Notace třídy má tři části: název, definici atributů a výčet metod a funkcí, kterých nabývá. [26]

Existuje několik druhů relací mezi třídami, například asociace, generalizace a závislost (dependency). [27]

Asociace je relace mezi dvěma třídami, která slouží k zachycení stavu, kdy jedna třída je sémanticky propojena s jinou. U asociací lze určit [26]:

1. Násobnost, neboli multiplicitu – Údaj, který se přiděluje oběma stranám relace. Násobnost vymezuje rozsah intervalu, kterého může nabývat počet instancí cílové třídy ve vztahu k jedné zdrojové třídě.

2. Název relace – Relace se pojmenuje z pohledu jednoho z aktérů.

3. Název rolí – Angažované třídy mohou mít určené jméno v souvislosti s danou relací.

Generalizace, nazývaná také jako dědičnost, je klasifikační vztah mezi obecnou a jí podřadnou, specifickou třídou. [27] Tato relace se využívá v situaci, kdy jedna obecná nadtřída má více podružných podtříd. Vazba směrem od podtřídy k nadtřídě se nazývá generalizace, obráceně pak specializace. [26] Instance podtřídy je zároveň instancí nadtřídy, protože dědí všechny její vlastnosti a nabývá i jejích funkcí a metod. [27]

Pro využití notace v OntoUML je důležitá možnost uskupení podtříd do více množin generalizace, které spojuje stejný význam specifikace. V rámci

množiny lze stanovit, zda je disjunktní anebo vyčerpávající. Pojem *isCovering* značí vyčerpávající množinu, což znamená, že instance nadtřídy musí být vždy zároveň i instancí jedné z podtříd. Pojem *isDisjoint* označuje disjunktní podtřídy, tedy instance nadtřídy může být instancí nejvýše jedné podtřídy. Pokud má nadtřída více generalizačních množin, pak její instance může být instancí z více generalizačních množin zároveň. [26]

Dependency, nazývaná také jako závislost, je vztah mezi dvěma třídami, které se v rámci této relace označují jako dodavatel a klient. Závislost vyjadřuje situaci, kdy klient je závislý na dodavateli, který mu něco poskytuje a není tak pro klienta možné bez dodavatele existovat. [27]

2.8 OntoUML

OntoUML je ontologický jazyk konceptuálního modelování, který je založený na ontologii UFO a konstruován pomocí notací UML Class diagramu. [15, 28] Základy pro OntoUML a UFO stanovil Giancarlo Guizzardi s cílem vytvořit jednotný nástroj pro konstrukci ontologicky správných konceptuálních modelů. [29] V minulosti byl aplikován v různých odvětvích jako je doména oleje a pohonných hmot, logistika, doména zpráv informačního managementu a také v Data Modeling Guide od Ministerstva obrany Spojených států amerických. [26]

Díky skutečnosti, že OntoUML vychází z ontologie UFO, je OntoUML více expresivnější a přesnější, než samotné UML. Výhodnou této notace je sémantická preciznost, díky které je uživatel veden k co nejsprávnější konceptualizaci pojmů v doméně. [29]

Základním prvkem modelu jsou instance, které jsou definovány jako cokoli myslitelného či pozorovatelného, jedná se tedy prvky reálného světa. [26] Z ontologického UFO základu jsou zde použity pojmy z Kategorie od Aristotela (viz 2.6) a to individuum, což lze připodobnit reálnému prvku a univerzál. Univerzály jsou obecné klasifikace individuí. Jsou reprezentovány jednotlivými typy konstruktů notace, což jsou definující vzory, kterými jejich instance, individua, nabývají. [29] Jelikož OntoUML využívá notací UML Class diagramu, tak jsou tyto typy konstruktů reprezentovány třídami a jednotlivé druhy se rozlišují pomocí jejich stereotypů. [26]

2.8.1 Princip identity

Ke správné konstrukci hierarchie pojmů vyskytujících se v doméně vede uživatele tzv. princip identity a individualizace. Princip identity a individualizace vypovídá o totožnosti se sebou samým a rozdílech, které odlišují dva výskyty těch samých objektů. [30] To přirozeně plyne z Aristotelovi idei o podstatě prvku reality (viz 2.5). [29] Guizzardi popisuje princip identity na příkladu psa, domácího mazlíčka, kterého majitelé pojmenovali. Majitel ho rozpozná

pomocí jména, rasy, výšky, váhy, barvy, ale nic z toho plně neurčuje jeho princip identity. Může totiž existovat jiný pes s totožným jménem, stejné rasy, výšky, váhy i barvy a v případě, že by tyto vlastnosti a informace plně definovali jeho princip identity, by to znamenalo, že není možné tyto dva psy odlišit. [28, 29] Tyto vlastnosti se nazývají podmínky identity, jsou nutnou součástí principu identity, ale ne postačující.

Princip identity musí být použitelný pro všechny myslitelné prvky, proto nelze považovat ani u lidí jejich rodná čísla za onu individuální informaci, protože nejsou aplikovatelná například přes všechny státy. Je nutné, aby byl udržitelný i napříč časem. Kdyby se tak nedělo, pak by skutečnost že ze štěněte vyrostl pes znamenala, že se jedná o jiné zvíře. Lidské vnímaní přirozeně dokáže rozlišit a odhalit tento princip identity u reálných prvků. Aristoteles tuto informaci, či zdroj nazval podstatou, neboli substancí (viz 2.5).

OntoUML typy univerzálů, které poskytují svým instancím identitu, se nazývají *Sortal*. [28] Protože každé individuum vyskytující se v modelu musí nabývat unikátní identity je nutné, aby bylo přímou nebo nepřímou instancí stereotypu poskytujícího identitu. Proto každé individuum musí být v konceptuálním modelu instancí univerzálu typu Sortal. Typy, které princip identity neposkytují se nazývají *Non-sortal*. [29]

2.8.2 Rigidita

Rigidita je vlastnost určitých typů univerzálů jazyka OntoUML plynoucí z modální logiky. Tento pojem definuje zda je individuum instancí takového typu za všech možných situací a v jakékoli chvíli své existence. [28, 29] Modální logika označuje tyto situace, nazývané i světy, jako realitu v daném čase a prostoru. Například člověk, je člověkem v každé fázi svého života a žádná mimořádná událost během jeho existence nemůže tuto skutečnost změnit. Ontologie UFO definuje několik možností rigidity [29]:

1. Rigidní – Individua, která jsou instancí rigidního typu, jsou jeho instancí za všech okolností po dobu své existence.

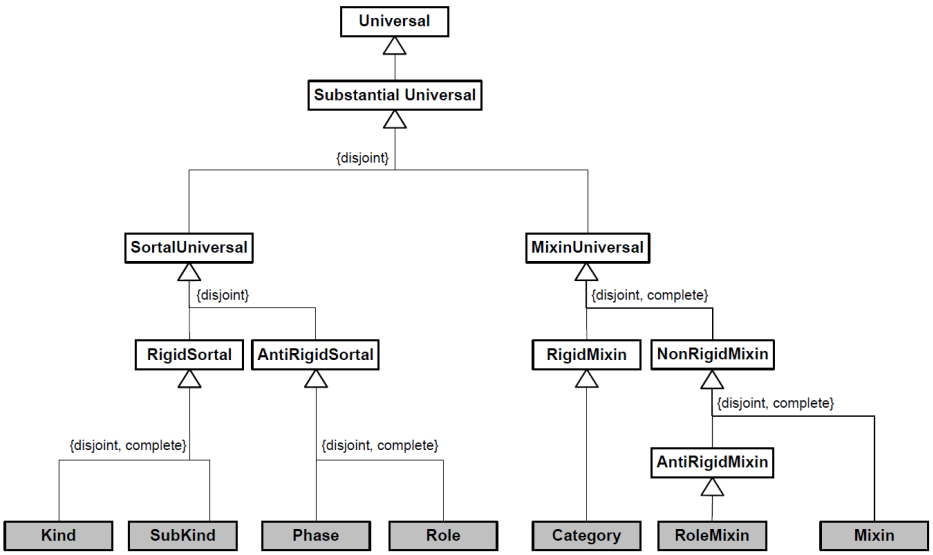
2. Non-rigidní – Je logická negace rigidity. Univerzál je non-rigidní, pokud existuje alespoň jedno individuum, které může přestat býti jeho instancí.

3. Anti-rigidní – Typ, pro který platí, že všechna individua, která jsou jeho instancí, mohou během své existence přestat býti jeho realizací a neovlivní to jejich trvání. Jedná se o druh non-rigidity.

4. Semi-rigidní – Typ univerzálu, který je non-rigidní, ale není anti-rigidní.

2.8.3 Generalizace

V UFO a tedy i OntoUML zajišťuje relace generalizace dědění principu identity i vlastností mezi typy. Na rozdíl od UML, kde jsou relace generalizace



Obrázek 2.3: Rozdělelní Substančních univerzálů: Univerzály typu Sortal se dále dělí na rigidní a anti-rigidní. Univerzály typu Non-sortal, které jsou na schéma uvedené jako „MixinUniversal“ se dále dělí dle rigidity. Na schéma je znázorněno, že anti-rigidia je typem non-rigidity. [31]

vždy rigidní, mohou být v OntoUML jak rigidní tak anti-rigidní. Individua jsou po celou dobu své existence instancemi nadtypu, který definuje jejich princip identity, ale mohou libovolně tvořit instance typů, které se řídí stejným principem identity. V případě, že má nadtyp více množin generalizace, pak je možné, a v případě že je množina generalizace vyčerpávající tak i nutné, aby individuum bylo instancí podtypů z různých množin generalizace. [26]

2.8.4 Typy univerzálů

Univerzály se dělí na *Substanční*, též označované jako Objekty, Substances či Substantials a *Aspekty*, také nazývané jako Accidents, Tropes nebo Moments. Substanční univerzály, objekty, jsou nezávislé konstrukty, které jsou existenčně nezávislé na ostatních objektech. Dále se dělí na *Sortaly* a *Non-sortaly* (viz obrázek 2.3). Aspekty jsou konstrukty, které jsou existenčně závislé na nositeli, se kterými jsou spojeny vazbou nazvanou *inherence*. [28] Také poskytují svým instancím princip identity, ale na rozdíl od substančních univerzálů jsou existenčně závislé na jiném individuu, jejich nositeli, který je zároveň částí jeho identity. Typy univerzálů, neboli stereotypy, se v OntoUML zapisují do špičatých závorek. [26] Toto označení má původ v jazyce XML.

2.8.4.1 Sortal

Pojmem *Sortal* jsou označovány typy individuí, které mají z našeho vnímání svoji identitu. Do této skupiny stereotypů se řadí Kind, Subkind, Phase, Role, Relator, Quantity a Collective. Ne všechny stereotypy spadající pod pojem Sortal poskytují identitu, ale všechny ji vlastní. Tyto stereotypy ji tak pomocí relace generalizace dědí od typů poskytující identitu.

«Kind»

Stereoptyp *Kind* je rigidní Sortal poskytující identitu. Kvůli tomu se tak nikdy nemůže stát podtypem jiného Sortalu, který poskytuje identitu, protože by pak nastala situace, kdy by jeho instance získala identitu dle dvou principů, což není možné. Protože je Kind rigidní, tak všechny individua, které jsou jeho instancí musí být jeho instancí po celou dobu své existence. Z toho samého důvodu nemůže být podtypem jiného anti-rigidního stereotypu.

«Subkind»

Stereotyp *Subkind* je taktéž rigidní Sortal, který nejčastěji definuje speciální případy stereotypu Kind, nebo jiného Subkindu. Neposkytuje přímo identitu, ale dědí ji pomocí relace generalizace od nadřazeného stereotypu s principem identity. [29] Protože se jedná o rigidní stereotyp, není možné aby jiný anti-rigidní typ byl jeho nadtypem. [28] Díky tomu, že stereotypy Subkind i Kind jsou rigidní, pak i relace generalizace mezi nimi je rigidní, což znamená, že není možné aby individuum, které je instancí Subkindu přestalo býti instancí Kindu, jelikož mu Kind určuje princip identity. [26]

«Phase»

Stereotyp *Phase* je anti-rigidní Sortal představující stádia Sortalu, které jsou definována jejich vnitřním stavem či vlastností. [26,29] Neposkytuje identitu, ale dědí ji od nadřazeného Sortalu. Phase vždy musí být součástí množiny generalizace, není tedy možné, aby existovalo pouze jedno stádium Sortalu reprezentované pouze pomocí jedné Phase. Množině generalizace lze stanovit, zda je disjunktní anebo vyčerpávající pomocí pojmů *disjoin* a *complete*. [28]

«Role»

Stereotyp *Role* je anti-rigidní Sortal zachycující specializace Sortalu, kterých nabyly ve vztahu s jinou entitou. Neposkytuje identitu, ale tak jako Phase či Subkind ji dědí od nadřazeného Sortalu. Jedná se tedy o role v kontextu relace, která zároveň figuruje jako „pečetidlo“ skutečnosti, že se jedná o roli a proto musí mít multiplicitu minimálně jedna. Všechny anti-rigidní Sortaly, tedy i Role a Phase, musí dědit identitu od právě jednoho Sortalu, který ji poskytuje. [29]

2.8.4.2 Non-sortal

Pojmem *Non-sortal* se označuje skupina stereotypů, jejichž instance mohou nabývat identit dle různých principů. Reprezentují vlastnosti, které jsou společné pro více univerzálů s rozdílnými principy identity a tak není možné tuto vlastnost definovat v nadtypu těchto univerzálů. [29] Z tohoto důvodu je zakázáno, aby jakýkoli Sortal byl jejich nadtypem, protože by to znamenalo, že by musely následovat jeho princip identity. [28]

OntoUML rozlišuje několik druhů těchto Non-sortalů v závislosti na jejich rigiditě. [29]

«Mixin»

Stereotyp *Mixin* je semi-rigidní Non-Sortal. Pojem semi-rigidní, kterým je označován pouze tento sterotyp, označuje situaci, kdy je univerzál rigidní pro nějaké individua a naopak anti-rigidní pro jiná. [28]

Jedná se o stereotyp, který představuje společné vlastnosti rigidních i anti-rigidních univerzálů, které se řídí rozdílným principem identity. [26] Jelikož se jedná o semi-rigidní stereotyp, tak není možné aby byl podtypem anti-rigidního nebo i rigidního stereotypu. Z toho plyne, že jeho nadtypem může být vždy pouze jiný stereotyp Mixin. [28]

«Category»

Stereotyp *Category* je rigidní Non-sortal, který se využívá k definici společných vlastností rigidních univerzálů, nejčastěji stereotypu Kind, které se řídí rozdílnými principy identity. [29] Protože se jedná o rigidní stereotyp, tak není možné, aby byl podtypem jakéhokoli anti-rigidního stereotypu. [28]

«RoleMixin»

Stereotyp *RoleMixin* je anti-rigidní Non-sortál, který je podobný anti-rigidnímu Sortalu Role s tím rozdílem, že se aplikuje pro instance s rozdílnými principy identity. [26]

«PhaseMixin»

Stereotp *PhaseMixin* je anti-rigidní Non-sortal, který je ekvivalentní anti-rigidnímu Sortalu Phase, ale stejně tak jako RoleMixin, popisuje společné vlastnosti stereotypů Phase, které se řídí rozdílným principem identity. [28]

2.8.4.3 Aspekty

Aspekty jsou prvky závislé na jiném univerzálu, který má v relaci inherence název nostiel. Relace inherence, která je irreflexivní, asymetrická a není tranzitivní, je vztah mezi nositelem a jeho aspektem. V relaci inherence je možné jako nositele označit aspekt a přiřadit mu jeho jiný aspekt. V případě zániku nositele jeho aspekt také automaticky přestává existovat. V jazyce OntoUML se rozlišují dva stereotypy Aspektů a to Quality a Mode. [29]

«Mode»

Stereotyp *Mode* je rigidní Aspekt, který reprezentuje vnitřní nestrukturované vlastnosti nositele, často se jedná o jejich stavy. [28] Se svým nositelem je propojen vazbou s názvem *Charakterization*, která musí mít multiplicitu 1 : 1, či 1 : N. Minimálně 1 je z toho důvodu, že nositel se podílí na identitě aspektu, zároveň nositel může nabývat více módů stejného typu. [26]

«Quality»

Stereotyp *Quality* je rigidní Aspekt představující, podobně jako stereotyp Mode, vnitřní vlastnosti nositele, které jsou ale strukturované. Tak jako stereotyp Mode je Quality se svým nositelem spojen vazbou *Charakterization* s multiplicitou 1 : 1. Není tak možné, aby nositel disponoval více než jednou hodnotou, kterou Quality reprezentuje. [26]

UFO rozlišuje tři druhy stereotypu Quality [28]:

- Perceivable quality – Reprezentuje hodnoty měřitelné přístrojem, například výška nebo rychlost.
- Non-perceivable quality – Představuje vlastnosti, které nejsou přístroji měřitelné, například měna.
- Nominal quality – Znázorňuje kvality, které nelze měřit a často se využívají jako označení pro individua jež jsou nositeli tohoto aspektu. Jedná se například o jméno, označení ISBN atd.

2.8.4.4 Asociace

V OntoUML jsou relace rozděleny na formální a materiální. Formálními relacemi jsou ty, které propojují instance na základě jejich vnitřních vlastností. Jedná se buď o relaci porovnání na základě jejich měřitelných kvalit, nebo o vztah celek-část (viz 2.8.4.5). Pro materiální relace existuje tzv. pečetidlo, které aktéry spojuje. [26] Toto pečetidlo představující vlastnosti relace [28] může být fyzický objekt, jako například manželská smlouva, nebo abstraktní jako telefonní hovor. V OntoUML se pro ono pečetidlo používá konstrukt Relator. [26]

«Relator»

Stereotyp *Relator* je rigidní Sortal, který poskytuje identitu. Umožňuje spojení dvou a více individuí materiální relací. Je nutné aby byl aktérem alespoň v jedné vazbě *Mediation* s celkovou násobností minimálně 2. [28]

2.8.4.5 Vztah celek–část

Teorií relace celek–část se zabývá filosofický obor, který se nazývá Merologie. Tato teorie postavila formální základy pro reprezentaci vztahů mezi částmi

i jejich vlastnostmi v rámci této relace. Ze své podstaty jsou tyto relace irreflexivní a anti-symetrické. Obecně ale nelze určit, zda jsou tyto vztahy mezi částí a celkem tranzitivní. Guizzardi pro příklad uvádí dvě rozdílné situace.

1. Ruka je součástí paže, ta je ale zároveň součástí člověka. Znamená to tedy, že ruka je i součástí člověka? Ano, relaci lze tedy v tomto případě označit za tranzitivní.

2. Město Praha je hlavním městem České republiky a je tedy její součástí. Česká republika je součástí Evropské unie. To by, pokud by relace byla tranzitivní, mělo znamenat, že Praha je také součástí Evropské unie. Toto tvrzení ale není pravdivé. Praha i Česká republika mají v Evropské unii rozdílnou roli a tak relaci celek–část nelze obecně považovat za tranzitivní.

Ontologie UFO definuje sekundární charakteristiky relací celek–část a to sdílitelnost, povinnost a neoddělitelnost. [29]

2.8.4.6 Sdílitelnost

V jazyce UML se možnost sdílení aktérů ve vztahu celek–část rozlišuje pomocí pojmů agregace a kompozice. V notaci UML jsou tyto skutečnosti realizovány pomocí tzv. diamantu. V případě kompozice, což je situace kdy objekt představující část nemůže existovat bez celku, je diamant plný. Naopak k realizaci agregace, kdy část může existovat nezávisle na celku, je diamant prázdný. Toto rozdělení se ukázalo jako nedostatečné a OntoUML tak převzalo detailnější dělení z ontologie UFO, ale zachovalo notaci pomocí diamantu umístěného u univerzálu, který značí celek.

V jazyce OntoUML prázdný diamant reprezentuje sdílitelnou část. To znamená, že instance části může být najednou součástí více celků bez ohledu na jejich typ. Naopak plný diamant představuje situaci, kdy část není sdílitelná a její instance tak nemůže být součástí více celků stejného typu v jednu chvíli, ale je dovoleno aby byla současně prvkem více celků různého typu. [26]

2.8.4.7 Povinnost

Další sekundární charakteristikou relace celek–část je povinnost jak z pohledu celku tak z pohledu části. Povinnost části z hlediska celku se dělí na volitelnou, povinnou a esenciální.

- **Volitelná část** znamená, že celek může existovat bez ohledu na přítomnost části. Část tak nemusí být v celku přítomna vůbec, nebo je možné že během své existence celek změní instanci této části. Tato charakteristika se v jazyce OntoUML zobrazuje pomocí multiplicity relace minimálně 0 na straně části.

- **Povinná část** vyjadřuje, že celek nemůže existovat bez přítomnosti části. V případě, že se jedná o tzv. generickou závislost, tak nezáleží na určité instanci části, ale je důležitá její přítomnost. Proto je možné, že instance části se během existence celku mění. Není ale možné aby, byť jen omezený čas, existoval celek bez přítomnosti části. Tato skutečnost se v modelu reprezentuje také pomocí multiplicity relace, která musí být větší jak 1 na straně části.

- **Esenciální část**, také nazývána jako existenční závislost, je druhým typem povinnosti. Jedná se o situaci, kdy celek je závislý přímo na instanci dané části a tuto instanci není možné změnit, protože se podílí na jeho identitě. Tato charakteristika se označuje pomocí pojmu *essential* a zároveň je nutné aby multiplicita na straně části byla větší jak 1. Z toho je viditelné, že existenční závislost, neboli esencialita, implikuje povinnost části.

Povinnost celku z pohledu části se rozlišuje na volitelnou, povinnou a neoddělitelnou.

- **Volitelný celek** znamená, že část může existovat i sama o sobě a během své existence je schopna měnit celky, kterých je součástí. V modelu se realizuje pomocí multiplicity relace minimálně 0 na straně celku.

- **Povinný celek** vyjadřuje skutečnost, kdy část není schopna existovat sama o sobě a je genericky závislá na libovolné instanci celku. To znamená, že během své existence se nemůže vyskytovat bez celku, ale jeho instance se může jakkoliv měnit. Podobně jako v povinnosti z pohledu celku se generická závislost z pohledu části také zobrazuje pomocí multiplicity relace větší jak 1, v tomto případě ale na straně celku.

- **Neoddělitelnost**, také nazýváno jako existenční závislost, celku značí situaci, kdy část vyžaduje existenci dané instance celku a tuto instanci není možné změnit, protože se podílí na její identitě. V OntoUML se realizuje pomocí označení relace pojmem *inseparable*, zároveň s multiplicitou větší jak 1 na straně celku.

Pojmy *essential* a *inseparable*, označující existenční závislost buď celku nebo části, jsou použitelné v případě, že se jedná o relaci, které se účastní pouze rigidní typy. V případe anti-rigidních typů mohou totiž individua přestat býti instancemi univerzálu a to by bylo v rozporu s jejich existenční závislostí. Z tohoto důvodu je v ontologii UFO definován pojem *immutable*, neboli neměnnost. Tento pojem označuje relaci celek–část, ve které figuruje anti-rigidní univerzál, který v každé situaci za nichž je instancí anti-rigidního typu zároveň plní svou definovanou roli v relaci celek–část. [26]

Ze studií kognitivních věd vyšlo najevo, že lidé vnímají více typů celků a zároveň i vazeb celek–část, které byly v OntoUML popsány.

Funkční celek

Jedná se o nejběžnější reprezentaci vztahu celek–část. Členové funkčního celku mohou být instance různých stereotypů a v rámci celku mohou mít různé úlohy. Funkční celek je možné reprezentovat libovolným stereotypem univerzálu (proto ani není zapsán ve špičatých závorkách), ale nejčastější se jedná o rigidní Sortal Kind. Vztah mezi dvěma funkčními celky se nazývá *ComponentOf*. Obecně se nejedná o tranzitivní relaci, záleží na kontextu použití. [29]

«Collective»

Stereotyp *Collective* je rigidní Sortal, který označuje celek, jenž je složený z členů, individuí domény libovolného stereotypu, které nejsou stejného typu jako celek, ale hrají v něm stejnou roli. Mezi členy existuje maximální sjednocující relace, která je zároveň původem principu identity, který tento Sortál poskytuje. [28] Stejně jako rigidní Sortaly Kind, či Relator poskytuje svým instancím princip identity. [29]

Relace mezi Collectivem a jeho členy, která není tranzitivní, se nazývá *MemeberOf*. Tento vztah vyžaduje minimálně dva členy přítomné v Colletivu. [28]

K zachycení vztahu mezi dvěma stereotypy Collective se užívá vazba *SubCollectionOf*. Jedná se o situaci, kdy členové jednoho kolektivu jsou zároveň členové druhého kolektivu. Tato relace, na rozdíl od relace MemberOf, je tranzitivní, což znamená, že prvky podkolektivu jsou současně prvky nadkolektivu. [26] Využívá se například k propagaci právních povinností a závazků.

«Quantity»

Stereotyp *Quantity* je rigidní Sortal představující celek, jehož části jsou stejného typu jako on sám. Tento rigidní Sortal reprezentuje maximálně spojité hmoty či materiály, které jsou nekonečně dělitelné. Nejčastěji představuje kapalné, sypké hmoty a materiály. [26, 28] Tak jako Collective i Quantity poskytuje instancím princip identity.

Vztah mezi dvěma stereotypy Quantity, který se nazývá *SubQuantityOf*, značí skutečnost, kdy jedna instance tohoto typu může být obsažena v jiném maximálně spojitém objektu (například ve víně je obsažen alkohol), ale jedná se o jeho esenciální část. [29]

Vztah mezi nádobu či formu, ve které se instance tohoto stereotypu nachází a instancí samotnou se nazývá *Containment*. [28]

2.9 OpenPonk

OpenPonk je bezplatná, open source platforma implementovaná v jazyce Pharo, která se využívá v oblasti konceptuálního modelování, tvoření diagramů, si-

mulace, generovaní zdrojových kódů atd. OpenPonk je vyvíjen v Centru pro konceptuální modelování a implementace. [32]

Platforma, dříve nazývána DynaCASE, poskytuje základ pro výzkumné skupiny, které mají za cíl realizovat vlastní specifické a nestandardní modelovací notace, transformace modelů nebo konkrétní algoritmy a nemají dostatečné kapacity na vývoj samostatného nástroje. [33]

OpenPonk obsahuje mimo jiné i několik již implementovaných rozšíření a modulů, které přinášejí možnost využití standardních notací a algoritmů. [33] Jedno z rozšíření je notace OntoUML pro konceptuální modelování, které je v řešení této práce použito. [32] V případě rozšíření OntoUML se zde vyskytuje framework pro verifikaci OntoUML modelů, který kontroluje, zda všechny entity a vztahy v modelu jsou vytvořeny dle definovaných pravidel jazyka OntoUML. [34]

Souběžně s touto prací je vyvíjena detekce návrhových antivzorů, která bude poskytovat další úroveň kontroly OntoUML modelů. Na realizaci detekce návrhových antivzorů pracuje bc. Marek Bělohoubek v rámci své diplomové práce s názvem Rozšíření možností modelování v OntoUML na platformě OpenPonk, která bude obhajována v červnu roku 2021.

Pro účely tvoření konceptuálních modelů v jazyce OntoUML byl nástroj OpenPonk uplatněn i v projektu Nest BDA.

Kapitola 3

Praktická část

V projektu Nest BDA v rámci metadata týmu, pod vedením doc. Ing. Roberta Pergla, Ph.D., jsem pracovala na ontologické analýze klíčových domén. Cílem celého metadata týmu bylo zajistit dostatečně přesné popsání dostupných datových sad včetně zajištění jejich homogenity, jak syntaktické tak sémantické. Syntaktickou interoperabilitu zajišťuje popis datových sad, na který plynule navazuje část konceptuálního modelování zajištující sémantickou interoperabilitu. Moje práce spočívala v zajištění sémantické interoperability, tedy v ontologické analýze klíčových domén a vytváření konceptuálních modelů spojujících doménu a příslušné heterogenní datové sady tak, aby bylo možné propojit data z různých odvětví reality.

Postup mé práce k výslednému modelu propojující datové sady s rozdílnými významy a termíny lze rozdělit do 3 fází:

1. Vytvoření ontologického modelu domény.

2. Provázání datových sad s ontologickými modely.

3. Vytvoření upřesňujících pravidel určující mapování atributů dat.

Modely byly zpracovány v nástroji OpenPonk (viz 2.9), který umožňuje modelování diagramů různých notací. Platforma OpenPonk dovoluje různé úpravy a rozšíření například v podobě popisů v rámci modelu či obsahu celých struktur. Toho bylo využito zejména v následné práci s modely na technické úrovni, která již nebyla součástí mé činnosti. Pro potřeby mé části práce v rámci metadata týmu byl využit ontologický jazyk konceptuálního modelování OntoUML (viz 2.8), který díky sémantické preciznosti dokáže uživatele vést k co nejsprávnější konceptualizaci pojmů v doméně.

Celkem jsem zpracovala 8 datových sad, ze kterých vzniklo 6 ontologických modelů s více jak 320 ontologickými entitami. Některé datové sady bylo možné propojit již v ontologické části. Jedná se o datové sady, které se týkají téměř shodné oblasti reality, nebo naopak velmi odlišné a na propojení ontologických

modelů jsem chtěla zdůraznit možnou sémantickou interoperabilitu na datech s rozdílným zaměřením.

Vytvořila jsem celkem 7 ontologických modelů, které jsou dostupné v příloze B, z čehož 6 má původ v datových sadách:

1. Model `stavebni-sporeni` popisující obsah datové sady Vývoj stavebního spoření [35]
2. Model `vyrobek-sklizen` spojující dvě datové sady a to Průměrné spotřebitelské ceny vybraných výrobků [36] a Sklizeň zemědělských plodin podle krajů [37]
3. Model `radio` reprezentující datovou sadu Radio, která měřila poslechovost rádia [38]
4. Model `duchody` popisující datové sady Přehled o počtu důchodů [39] a Počet vyplacených důchodů [40]
5. Model `penze` reprezentující datovou sadu Penzijní připojištění a penzijní doplňkové spoření [41]
6. Model `nakupy` představující datovou sadu Internetové nákupy jednotlivců [42]

Poslední model `uzemi` nemá základ v žádné datové sadě, ale slouží jako reference pro zjednodušení dodržování pojmenování a stereotypů k zajištění sémantické interoperability (viz 3.3).

V příloze B jsou uloženy jak náhledy modelů ve formátu PNG tak soubory ve formátu OPP, které je možné otevřít v platformě OpenPonk od verze v2.1.0.

3.1 Ontologický model

Ontologický model pojmů a jejich vztahů se zabývá pouze reálnými aspekty bez ohledu na jakékoli technické informace datových sad. Doména, či zkoumaná oblast, vychází z obsahu dat a tím je vymezena část reality, která bude modelována. Vytvoření ontologického modelu se dělí do dvou etap a to prvotní analýza domény a následné ontologické modelování konceptů.

3.1.1 Analýza datové sady

Analýza datové sady spočívá v určení oblasti reality a jejích významných prvků. V této části je důležité zjistit přesné významy a definice použitých pojmů v realitě. Tato činnost vyžaduje vyhledávání v zákonech, významových slovnících či součinnost s doménovými experty.

Analýza plynule navazuje a zároveň vychází z popisu datové sady. Popis datových sad nebylo primární náplní mé práce v rámci metadata týmu, tuto část zaštiťovala Ing. Jana Freeman, Ph.D. Protože práce na konceptuálních modelech vzniká z těchto popisů, tak je důležité je alespoň přiblížit.

Popis datové sady primárně slouží pro dvě potřeby. První je, jak už bylo zmíněno, zachycení důležitých pojmů pro účely ontologického a datového modelování. Druhá spočívá v zaznamenání struktury formátu a důležitých technických údajů pro potřeby dalšího technického zpracování. Popis datových sad se tak vypořádá i se syntaktickou heterogenitou. Dokument, který je human-readable, se pro navazující technologie generuje i do machine-actionable formátu.

Popis datové sady obsahuje části jako jsou základní informace, popis obsahu, metodické vysvětlivky, struktura dat a náhledy. V základních informacích je vždy uveden unikátní kód, tzv. *Identifier*, který označuje tuto datovou sadu a všechny její potřebné části. Nejvýznamnější části pro následné ontologické zpracování je popis obsahu a struktura dat.

Z popisu obsahu jsem čerpala obecné informace o čem sada pojednává a jakých domén se může dotýkat. Důležitější z pohledu ontologického modelování je část popisující strukturu dat (viz obrázek 3.1). Z ní jsem získávala údaje o jednotlivých atributech dat a jejich významu. Díky této části je možné zaměřit přesnější oblast reality, která bude modelována. Struktura dat udává přesné názvy atributů použité v datech, jejich datový typ, stručný popis významu těchto hodnot a jejich formát. V případě, že datový objekt obsahuje již svá metadata, je vhodné čerpat hodnoty i z nich, abych dostala co nejpřesnější oblast reality, kterou datová sada popisuje. I tak je ale nutné vytvořit popis nový, který bude vyhovovat potřebám projektu.

Názvy atributů, jejich datové typy a formáty se využívají v části datového modelování a mapování atributů. V analýze domény jsou nejdůležitější částí popisy. V nich se velmi často objevují odkazy na číselníky (viz obrázek 3.1),

10	Název	Datový typ	Popis	Formát
11	Idhod	Unikátní identifikátor údaje Veřejné databáze ČSÚ	Využije se v případě dotazu ke konkrétnímu údaji, tj. údaji týkajícímu se sklizni a hektarovému výnosu plodiny v daném území	numerický
12	hodnota	Zjištěná hodnota	V numerickém formátu: kolik bylo sklizeno plodiny v tunách, v případě, že se jedná o důvěrný údaj, je sloupec prázdný	numerický
13	stapro_kod	Kód statistické proměnné	5906 tj. Sklizeň zemědělských plodin 5908 tj. Hektarový výnos sklizně zemědělských plodin	numerický
14	mj_cis	Kód číselníku měřících jednotek	V této DS použit číselník 78 (vybrané měřící jednotky)	numerický
15	mj_kod	Kód položky z číselníku měřících jednotek (78)	V této DS použit pouze kód 20103, tj. jednotka tuny	numerický
16	druhzplod_cis	Kód číselníku pro druh zemědělské plodiny	V této DS použity číselníky: 208 tj. Druh zemědělské plodiny 209 tj. Druh zemědělské plodiny - agregace	numerický
		Kód položky číselníku pro druh		

Obrázek 3.1: Příklad struktury dat: Jedná se o strukturu dat datové sady týkající se sklizně zemědělských plodin. [37]

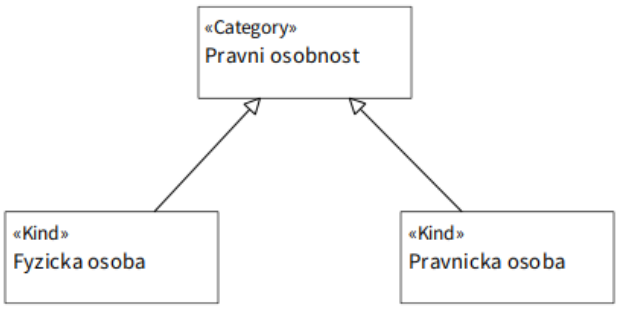
seznamy hodnoty, kterých může atribut nabývat. I tyto informace jsou důležité pro plnohodnotnou analýzu dat.

U jednotlivých pojmů je důležité znát jejich přesnou definici, aby nedocházelo k nesprávnému či víceznačnému označení. Pro některé prvky reality je nutné zjišťovat jejich právní dělení a označení či vymezení zákonem. S tímto jsem se potýkala hlavně u datových sad z veřejných databází popisující odvětví financí, sociálního zabezpečení, mzdy a náklady práce nebo sčítání lidu, domů a bytů. Například v datové sadě týkající se vývoje stavebního spoření (viz model `stavebni-sporeni` v příloze B) jsem čerpala ze sbírky zákonů. [43] U datových sad z odvětví marketingu a komunikace jsem naopak využívala znalosti doménových expertů ze společnosti Remmark, a.s.

3.1.2 Konceptuální model datové sady

Konceptuální modelování datové sady, potažmo domény které se týká, začíná určením základních pojmů, které poskytují z našeho vnímaní princip identity. Nejčastěji se jedná o osobu, územní určení nebo různé objekty jako třeba zboží. Dále se určují navazující anti-rigidní Sortaly, jako je Role nebo Phase, poté stereotypy patřící do Non-sortalů a nakonec Aspekty. Díky využití jazyka OntoUML a jeho ontologické a sémantické preciznosti jsou pravidla notace návodná pro co nejsprávnější popis reality.

Na příkladu zjednodušené datové sady týkající se stavebního spoření (viz model `stavebni-sporeni` v příloze B) návodně popíšu postup tvorby konceptuálního modelu. Z analýzy této datové sady vyvstaly informace popisující tuto doménu, které jsem čerpala primárně z jejího popisu a sbírky zákona [43]:



Obrázek 3.2: Model reprezentující právní osobnost (viz model `stavebni-sporeni` v příloze B).

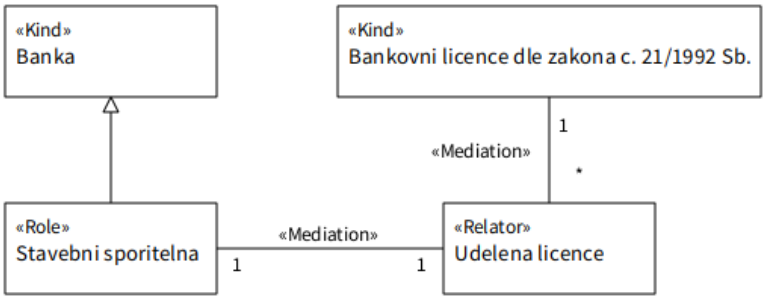
Stavební spoření je druh spoření nejčastěji za účelem investice do nemovitosti. Může jej uzavřít jak fyzická, tak právnická osoba se stavební spořitelnou. Výše cílové částky spoření i úroková sazba jsou určeny ve smlouvě stavebního spoření.

Z analýzy vyplývá, že rigidními základy modelu budou fyzická osoba, právnická osoba a stavební spořitelna. Tyto pojmy musí být pro správnou konceptualizaci uvedeny do kontextu domény. Fyzická osoba je dle občanského zákona [44] pojem odlišující člověka od jiných právních subjektů. Nabývá právní osobnosti svým narozením a zaniká smrtí. Právnická osoba je pojem definován jako organizace, které zákon uzná právní osobnost. Z těchto poznatků vyplývá, že právní osobnost je nadřazený termín pro fyzickou i právnickou osobu a tak je možné říci, že stavební spoření si může sjednat subjekt disponující právní osobností (viz obrázek 3.2). Právní osobnost je reprezentována stereotypem `Category`, protože představuje vlastnost dvou rigidních typů, které se řídí různým typem identity.

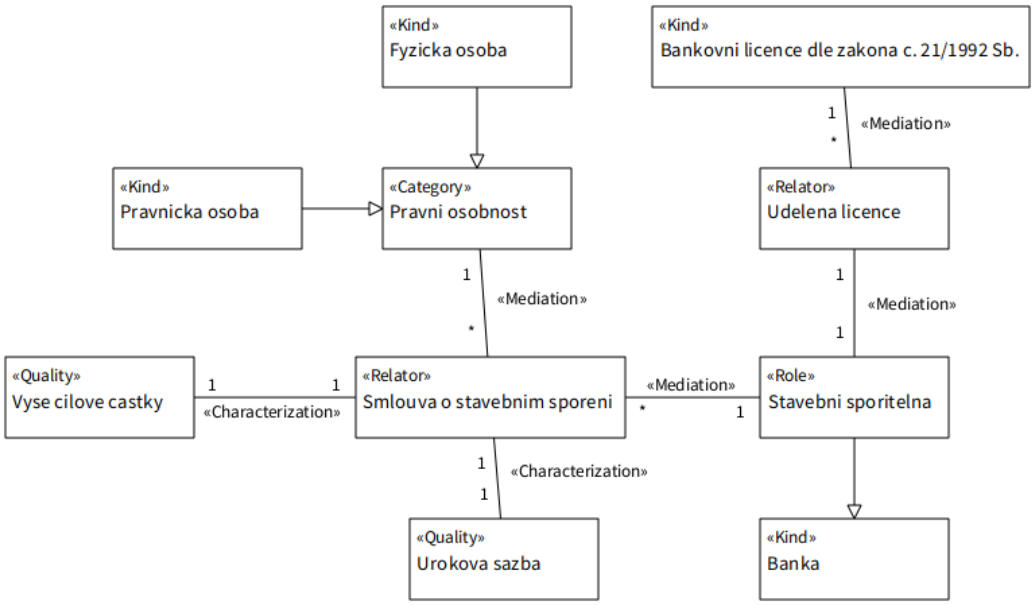
Právní osobnosti nabývá i stát, pokud figuruje jako subjekt v právních vztazích. Stát jako účastník právních vztahů dle zákona č. 219/2000 [45] je právnická osoba. Nastává tak situace, kdy jeden objekt může v různých kontextech být konceptualizován odlišnými způsoby. Stát v kontextu právního vztahu je obsažen v modelu, který se týká penzijního připojištění (viz model `penze` v příloze B).

Stavební spořitelna je banka, která disponuje povolením ve smyslu zákona č. 21/1992. [46] Proto se jedná o stereotyp `Role` získávající identitu od banky, která je reprezentována stereotypem `Kind` (viz obrázek 3.3). Samozřejmě by bylo možné dále specifikovat koncepty zákoníků, do kterých tato licence patří, ale je nutné určit jisté omezení oblasti konceptualizace. Mohla by tak také vzniknout snaha o vytvoření konceptuálního modelu celého světa.

Vztah mezi stavební spořitelnou a právnickou či fyzickou osobu je materiální relace, která je zprostředkována smlouvou o stavebním spoření. Platí, že jedna právnická či fyzická osoba může mít neomezené množství smluv



Obrázek 3.3: Model reprezentující stavební spořitelnu (viz model `stavebni-sporeni` v příloze B).



Obrázek 3.4: Model reprezentující uzavření smlouvy o stavebním spoření (viz model `stavebni-sporeni` v příloze B).

o stavebním spoření. Aspekty smlouvy jsou pak jak cílová výše, tak úroková sazba (viz obrázek 3.4).

K uchování sémantické interoperability mezi modely datových sad je nezbytné dodržovat jednotné pojmenování i druhy stereotypů pro konstrukty reprezentující stejné reálie. Díky tomu se pak i z technického hlediska jedná o schodný prvek a všechny jejich vazby či navázané informace se sjednotí. Z tohoto důvodu jsem v rámci metadata týmu vytvořila modely reprezentující často využívané konstrukty reality, jako třeba model politicky–správního dělení území (viz model `uzemi` v příloze B) nebo právních definicí subjektů a

vztahů mezi nimi, který je využit výše.

3.1.2.1 Rozšířené popisy vazeb

Pro nároky části analytické platformy, která se zabývá experimentálním dotazováním nad obsahem složeným z datových sad, byla aplikována další vrstva zjednodušených popisů vazeb v ontologickém modelu. Tato zjednodušená označení definoval vývojový tým projektu Nest BDA. Jedná se o pojmenování jednotlivých vazeb mezi entitami modelu. Tyto názvy mají sémantický význam vztahu bez nutnosti znalosti stereotypů ontologických entit, které tento vztah definují přesněji. Existují tři pojmy, kterými se vazby popisují:

1. Vazba IS

   Pojmem *IS* se označují vazby, které spojují podřazené entity. Týká se vyjádření hierarchie mezi entitami, které jsou nejčastěji spojeny pomocí ontologické vazby generalizace a následují tak stejný princip identity. Jedná se například o podřadnost hierarchických pojmů v doméně zvířat. Koala západní je druh koaly, druh koala zase spadá do řádu koalovité. Mezi těmito pojmy je podřadná vazba, kterou je nutné označit notací IS.

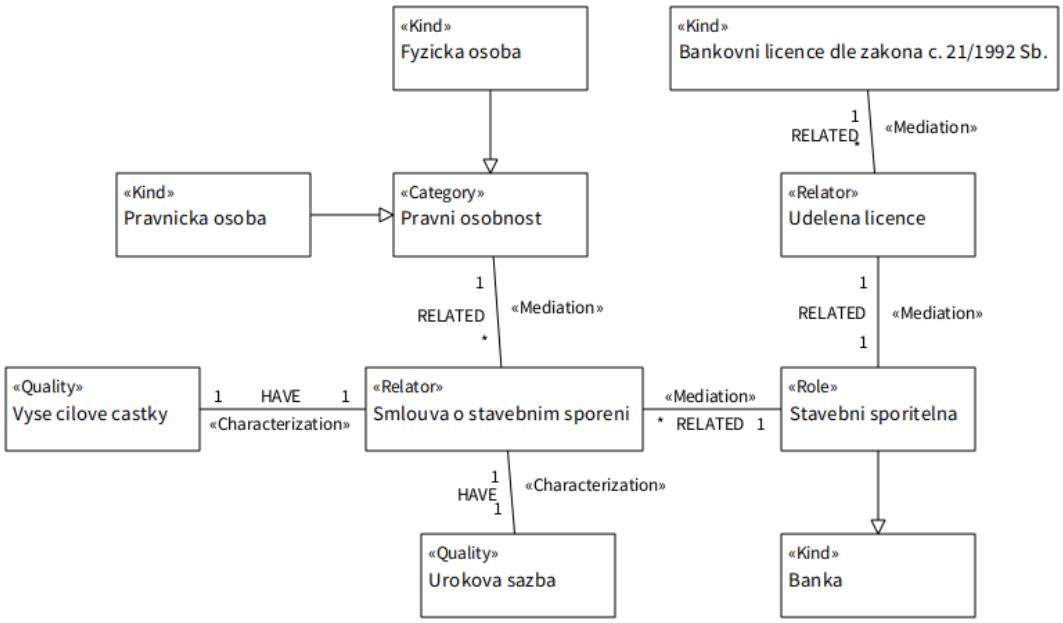
2. Vazba HAVE

   Označení *HAVE* se využívá u vazeb, které reprezentují relaci mezi individuem a jeho vnitřní vlastností, stavem či fyzickou dispozicí. Typicky se jedná o vazbu charakterization mezi nositelem a jeho aspektem, ať už stereotypem Quality, nebo Mode. Může se týkat také vztahu mezi entitou a jejím vnitřním stavem, který je reprezentován stereotypem Phase. Například skutečnost, že koala má huňatou srst může být označena jako HAVE. Název HAVE je ale poněkud zavádějící, naznačuje totiž že instance entity „něco má“, což často nutí k nadbytečnému a nesprávnému užití tohoto označení.

3. Vazba RELATED

   Pojmenování *RELATED* je označení pro relace spojující individuum a vlastnosti dané situacemi, ve které se instance nachází. Nejčastěji se jedná o vazby představující materiální relace, nebo vztahy mezi instancí a rolí kterých může nabývat. Každý stereotyp Role totiž musí mít „pečetidlo“, které stvrzuje skutečnost, že individuum této role opravdu dosáhlo, což lze považovat za onu situaci.

Tyto názvy jsem definovala u jednotlivých vazeb ontologických modelů. Pojmenování lze libovolně kombinovat a skládat. V případě, že jedna vazba vyhovuje více názvům, pak se jednotlivá označení oddělují středníkem. Na modelu stavebního spoření (viz obrázek 3.5) jsou viditelné názvy vazeb. U relací generalizace v nástroji OpenPonk nejsou názvy vazeb zobrazené.



Obrázek 3.5: Model reprezentující uzavření smlouvy o stavebním spoření doplněný o vazby IS, HAVE, RELATED (viz model `stavebni-sporeni` v příloze B).

3.2 Datové modelování

Během datového modelování jsou do ontologického modelu domény postupně přidávány spojitosti s datovými sadami. V této části jsou důležité přesné názvy atributů dat, které jsou uvedené v popisu datových sad. Provázání s daty se dělí do dvou fází a to vytvoření tzv. *Data entit* a formování mapovacích pravidel.

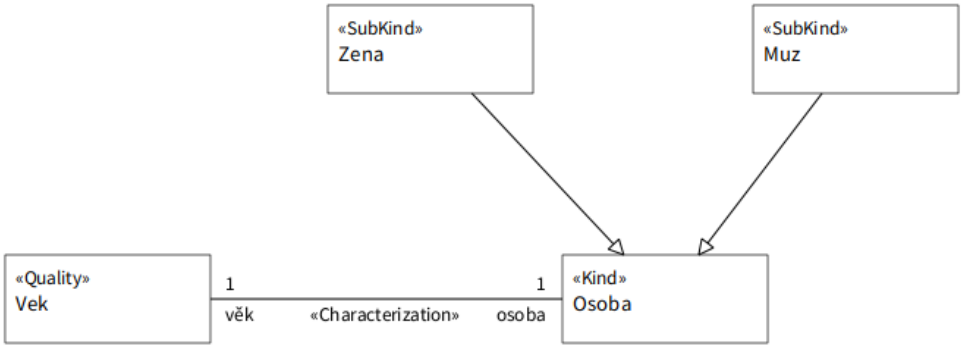
3.2.1 Data entity

V této části postupu přípravy datové sady jsem rozšiřovala vytvořený ontologický model domény o entity reprezentující uspořádání atributů datové sady, či sad. Tímto je získán obecný náhled na význam informací obsažených v datové sadě vzhledem ke konceptuálnímu modelu, tedy k jejich smyslu v realitě popisované domény.

Nejdříve bylo nutné rozdělit strukturu informací datové sady dle sémantické spojitosti s ontologickým modelem tak, aby spojitost byla co nejpřesnější. Je samozřejmé, že pro každý atribut datové sady lze nalézt nejlepší, nebo alespoň nějaké, umístění, protože ontologický model popisuje právě doménu, která je v datové sadě obsažena. Pro příklad uvádím datovou sadu, jejichž data obsahují informace o pohlaví osoby a jejím věku.

pohlavi_kod	vek_kod
F	39
M	53
M	64
F	24

Tabulka 3.1: Data obsahující informace o pohlaví osoby a jejím věku



Obrázek 3.6: Model reprezentující pohlaví osoby a její věk.

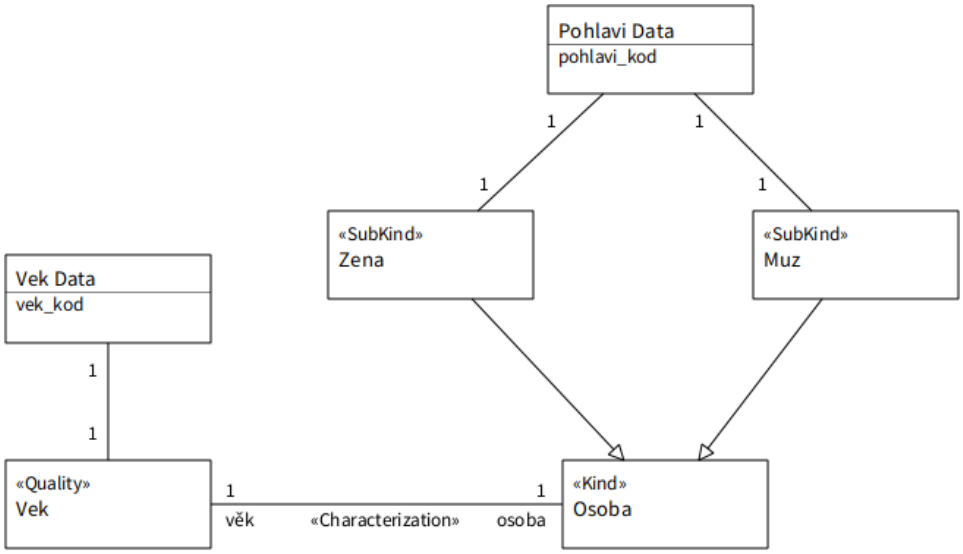
V datové sadě jsou kódy atributů definovány jako `pohlavi_kod` odkazující na pohlaví osoby, kde hodnota `M` značí, že se jedná o muže a hodnota `F` označuje ženu. A atribut `vek_kod` obsahuje věk osoby (viz tabulka 3.1).

K této datové sadě je vytvořen ontologický model popisující její doménu (viz obrázek 3.6).

Atribut `pohlavi_kod`, označující pohlaví osoby, se váže jak s entitou Osoba, tak s entitami `Zena` a `Muz`. V tuto chvíli je nutné rozhodnout, která ontologická entita modelu nejpřesněji tento atribut vystihuje. V případě, že by atribut byl spojen s příliš obecným pojmem, pak by nastala jistá ztráta přesnosti, kterou dobře rozpracovaný ontologický model nabízí. Z tohoto důvodu je entita Osoba příliš generální označení a jedná se o příliš obecnou spojitost. Proto `pohlavi_kod` souvisí s entitami `Zena` a `Muz`.

Atribut `vek_kod`, představující věk osoby, se váže k entitě `Vek`. Tato spojitost s ontologickým modelem je zjevnější.

K tomuto prvotnímu propojení ontologického modelu s atributy datové sady jsem využívala tzv. *Data entity*. Nejedná se o konstrukt či stereotyp zavedený jazykem OntoUML, ale byl vytvořen dodatečně metadata týmem pro účely použití v projektu Nest BDA. Technicky se jedná o konstrukt v OpenPonku typu třída, jehož název vždy obsahuje označení *Data*. Pro každou entitu ontologického modelu, se kterou se významově pojí nějaký atribut datové sady se vytvoří Data entita. Zároveň musí platit, že všechny atributy se v těchto Data entitách vyskytují právě jednou. Z toho plyne že nelze pro



Obrázek 3.7: Model reprezentující pohlaví osoby a její věk doplněný o Data entity.

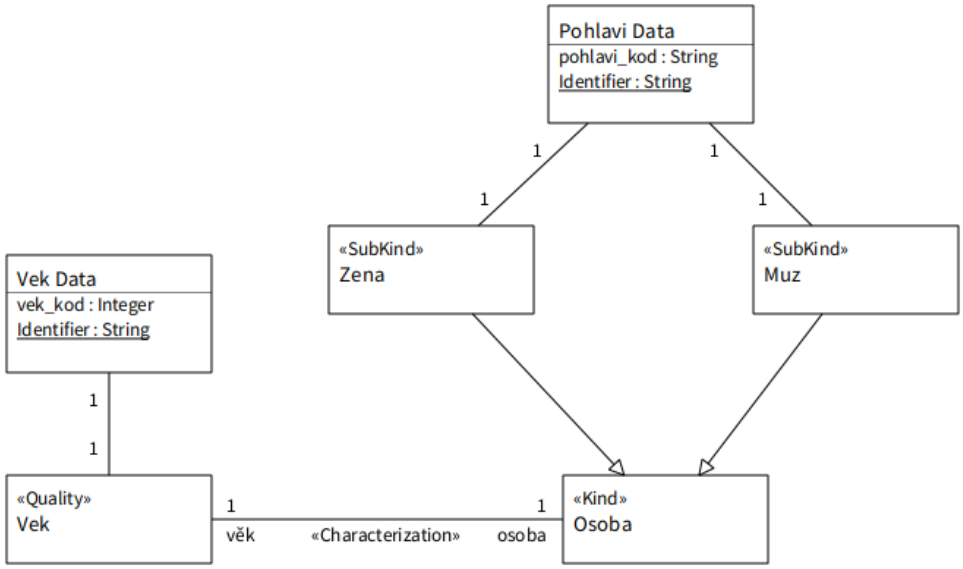
entity `Zena` a `Muz` vytvořit dvě různé Data entity, protože by obě obsahovaly atribut `pohlavi_kod`, což je v rozporu s možným počtem výskytů atributů. To znamená, že se vytvoří pouze jedna Data entita, která bude spojena s oběma entitami Žena i Muž a pojmenuje se generickým označením obou pojmů. Logické odlišení významu se následně určí pomocí zpřesňujících pravidel pro mapování atributů (viz 3.2.2).

Pro atribut `vek_kod` je realizace přímočařejší. Pro entitu `Vek` se vytvoří Data entita Věk, která bude obsahovat atribut `vek_kod` a bude spojena pouze s entitou `Vek`. Přidané Data entity do ontologického modelu jsou viditelné na modelu (viz obrázek 3.7). Tímto způsobem se zvýrazní ontologické entity, které vypovídají o informacích, které datová sada poskytuje.

Pro další použití na technické úrovni je nutné atributům definovat jejich datové typy, v případě zmíněných atributů se jedná o String a Integer, což je dohledatelné v popisu datové sady.

Aby byla zajištěna provázanost jak s daty datové sady tak jejím popisem, je třeba zaručit, aby bylo zřetelné odkud (z jaké sady) atributy pochází. Proto se do Data entit vkládá Identifier používaný v projektu Nest BDA. (viz obrázek 3.8).

Díky Data entitám v ontologických modelech je na první pohled patrné, k jakým prvkům reality jsou dostupná data a naopak k jakým částem by do budoucna bylo dobré zajistit lepší či alespoň nějaké informace. Například v datové sadě, která se týká stavebního spoření nejsou dostupné informace o stavebních spořitelnách či výši úrokových sazeb ujednaných ve smlouvě (viz



Obrázek 3.8: Model reprezentující pohlaví osoby a její věk doplněn o Identifier.

model `stavebni-sporeni` v příloze B).

3.2.2 Mapování atributů

Pro přesnější provázání datové sady a ontologického modelu, které je obecně znázorněno pomocí Data entit, jsem využívala mapovacích pravidel, pomocí kterých jsem pro jednotlivé atributy dat definovala preciznější vztah k ontologickým entitám. Syntaxe těchto předpisů vznikla v rámci týmové práce skupiny metadata, pod vedením doc. Ing. Roberta Pergle, Ph. D., projektu Nest BDA. Díky pravidlům bylo docíleno plné provázanosti mezi datovou sadou a ontologickým modelem, který ji popisuje. Pravidla se konstruují pomocí jednoduchých logických celků, které jsou blíže popsány.

3.2.2.1 Formulace pravidel

Formát pravidla je složen ze dvou částí. První vyjadřuje atributy, či množinu hodnot tohoto atributu, kterých se dané pravidlo týká. Druhá část vyjadřuje o jakých entitách atributy vypovídají.

```
"atribut" -> "entita";
```

Základní pravidlo popisuje holou souvislost mezi atributem a danou entitou, bez specifikovaných podmínek. Tuto formu pravidla jsem využívala v případě, kdy Data entita je spojena pouze s jednou ontologickou entitou

a její hodnoty mají jednoznačný význam, jako je atribut `vek_kod` a entita `Vek` (viz obrázek 3.8). V případě atributu `vek_kod` a entity Věk vypadá pravidlo následovně.

```
"vek_kod"  -> "Vek";
```

3.2.2.2 Pravidla s podmínkou

Pokud se jedná o situaci, kdy atribut nelze spojit přímo s jednou entitou, či má více významů na základě nabývajících hodnot, pak je nutné využít v pravidlech podmínky založené na hodnotě atributu. Jedná se o případy atributů jako je `pohlavi_kod`, který může nabývat hodnot `M` a `F`. Podle své hodnoty rozlišuje, zda se jedná o ženu nebo muže. Tento atribut je v Data entitě `Pohlavi`, která je spojena jak s ontologickou entitou `Zena` i `Muz` (viz obrázek 3.8). Proto je potřeba zanést do pravidel toto rozlišení na základě kterého bude zřejmé, jaké entity se určitý údaj týká. Zda se jedná o informace, v tomto případě věku (viz tabulka 3.1), o muži či ženě.

Toto omezení se uvádí do první části k atributu ve formě hodnoty, které se má atribut rovnat. V případě textové hodnoty, jako je to mu i v atributu `pohlavi_kod`, je přirovnání následující.

```
"pohlavi_kod" = "F" -> "Zena";
"pohlavi_kod" = "M" -> "Muz";
```

Další možností je číselná hodnota, tak jako tomu bylo v datové sadě týkající se sklizně zemědělských plodin podle krajů (viz model `vyrobek--sklizen` v příloze B). Atribut statistické proměnné pojmenované `stapro_kod` nabýval hodnoty 5906 v případě, že se jednalo o hodnotu, která odpovídala sklizni plodiny. Naopak, pokud se jednalo o hektarový výnos plodiny, tak atribut měl hodnotu 5908. Ontologický model tak rozděloval pojem sklizeň na výnos a sklizené množství plodiny. Pravidla pro atribut `stapro_kod` tak přiřazovala v podmínce atributu číselnou hodnotu.

```
"stapro_kod" = 5906 -> "Sklizeno";
"stapro_kod" = 5908 -> "Vynos";
```

V datové sadě týkající se přehledu počtu důchodů (viz model `duchody` v příloze B) bylo potřeba rozlišit spojení na základě regulárního výrazu. Atribut referenčního území s názvem `referencni_oblast_kod` nabývá hodnot, které začínají řetězcem „OK“ v případě, že se jedná o údaj v rámci okresu, začíná řetězcem „VC“ když se jedná o kraj anebo „SP“ pokud údaj vypovídá o obci. Pravidla, která jsem vytvořila tak měla podobu:

```
"referencni_oblast_kod" =  /OK+/ -> "Okres";
 "referencni_oblast_kod" = /VC+/ -> "Kraj";
 "referencni_oblast_kod" = /SP+/ -> "Obec";
```

V datové sadě, která popisovala přehledy počtu druhů důchodů (viz model `duchody` v příloze B), uváděl atribut `druh_duchodu_kod` označení kombinace druhů důchodů ke kterým se údaj vztahoval. Atribut `druh_duchodu` pak popisoval slovně o jakou kombinaci druhu důchodů se jedná, to znamenalo že i oba tyto atributy musely být v konjunkci. V rámci metadata týmu pojektu Nest BDA se tak rozšířily mapovací pravidla o logické spojky AND a OR. Mapování, která jsem vytvářela pak měla tuto podobu:

```
"druh_duchodu_kod" = "PK_IPVM" AND "druh_duchodu" =
"Invalidni duchod prvniho stupne vyplaceny v soubehu s vdoveckym
        duchodem" -> "1.stupen" AND "Vdovecky";
```

3.2.2.3 Podmíněná pravidla

V datové sade která se týkala průměrné měsíční hrubé mzdy a mediánu mezd bylo potřeba vytvořit pravidlo, které by zohledňovalo atribut, který udával význam hodnoty jiného atributu. Proto byla metadata týmem vytvořena syntaxe tzv. podmíněného pravidla, které říká že pokud má atribut A nějakou hodnotu, pak platí následující pravidlo pro atribut(y) B.

```
("atribut_A" = "hodnota atributu A") -> (pravidlo atributu_B);
```

V případě datové sady týkající se průměrných mezd atribut `spkvantil_kod` označoval, zda celý údaj v atributu hodnota vypovídá o mediánu či průměru mezd. Pravidlo tak vyjadřuje, že pokud `spkvantil_kod` obsahuje hodnot 5958, pak atribut `hodnota` vypovídá o mediánu mezd.

```
("spkvantil_kod" = 5958) -> ("hodnota" -> "Median");
```

3.3 Zajištění sémantické interoperability

Mezi heterogeními datovými sadami byla díky ontologickým modelům umožněna sémantická interoperabilita. Přes společné prvky datových sad je možné provázat různé odvětví reality. Nejčastěji vyskytujícím se prvkem v datových sadách a tedy i v ontologických modelech je osoba. Na obrázku (viz obrázek C.1) je viditelná entita osoby v modelu týkajícím se penzijního připojištění.

Na obrázku (viz obrázek C.2) lze spatřit entitu osoba v modelu souvisejícího s poslechovostí rozhlasu. Přes osobu pak lze data provázat.

Další častou společnou oblastí je území. Pro koncepci území jsem vytvořila samostatný model k zachování správného pojmenování částí a stereotypů entit (viz model `uzemi` v příloze B). Model území slouží pouze jako reference a přímo se k němu nevztahuje žádná datová sada. Prvky územního modelu jsou použity například v modelech, které se týkají sklizně zemědělských plodin (viz obrázek C.3) a poslechovosti rádia (viz obrázek C.4).

Modely popisující ceny výrobků (viz obrázek C.5) a nákupy na internetu (viz obrázek C.6) obsahují shodně entity prodej, či prodejna. V tomto případě se jedná o podobné odvětví reality a nalezení propojujících prvků je snazší.

Naopak v modelu (viz obrázek C.7) popisující druhy důchodů a jejich čerpání se zobrazují entity jako důchodce, či důchod. Tyto entity se také vyskytují v modelu (viz obrázek C.8), který se týká poslechovosti rádia. Díky této spojitosti tak lze data z rozdílných oblastí reality propojit.

Entita představující období sběru informací, kterou v modelech uvádím pod názvem `Referencni obdobi`, se vyskytuje téměř ve všech modelech. Například v modelu, který se týká stavebního spoření (viz obrázek C.9). nebo penzijního připojištění (viz obrázek C.10).

Díky dodržování jednotného pojmenování a určení stereotypů entit představující totožné reálné objekty je možné datové sady propojit přes významově stejné prvky reality.

Závěr

Cílem mé práce bylo vypracovat ontologickou analýzu klíčových domén a sémantické propojení mezi modely domén a vybranými datovými sadami. Datové sady byly definovány pro účely projektu Nest BDA, kde jsem působila v rámci metadata týmu pod vedením doc. Ing. Roberta Pergla, Ph.D. Cílem bylo vyhotovit ontologickou analýzu vybraných domén a vytvořit jejich konceptuální modely. Následně modely propojit se souvisejícími datovými sadami a zvýraznit dosaženou sémantickou interoperabilitu.

Výsledkem mé práce jsou konceptuální modely relevantních domén propojené s datovými sadami. Vytvoření konceptuálního modelu předcházela analýza domény konceptualizace, která spočívala v určení významných a přesně definovaných pojmů domény. Konceptuální modely byly vytvořeny v jazyce OntoUML v platformě OpenPonk. Pro záměry experimentálního dotazování nad obsahem platformy projektu Nest BDA byly navíc přidány zjednodušené popisy vazeb.

K významovému provázání konceptů reality a atributů datových sad byly využity Data entity, konstrukty vytvořené pro účely projektu Nest BDA. Pro přesnější definici propojení atributů datových sad a konceptů doménových modelů byly použity mapovací pravidla, jejichž syntaxe byla stanovena metadata týmem.

Díky dodržování jednotného pojmenování a využití ontologického modelovacího jazyka, bylo možné modely propojit přes významově shodné koncepty reality a zajistit tak sémantickou interoperabilitu datových sad.

V budoucnu bude určitě nutné stále rozšiřovat množství zpracovaných datových sad pro projekt Nest BDA. V postupu vytváření konceptuálních modelů v platformě OpenPonk bude v budoucnosti umožněna detekce návrhových antivzorů, čímž by se zvýšila přesnost a kvalita vytvořených konceptuálních modelů.

Literatura

[1] *Remmark* [online]. REMMARK, a.s. [cit. 2021-04-08]. Dostupné z: http://www.remmark.cz/

[2] Remmark a.s. [online]. *Datová platforma pro oblast marketingu, médií a komunikace*. 2020. [cit. 2021-04-01]. [neveřejně dostupný dokument].

[3] Remmark a.s. [online]. *NABÍDKA ZNALOSTÍ OD POSKYTOVATELE ZNALOSTÍ/SLUŽBY. Datová platforma pro marketingovou komunikaci*. 2020. [cit. 2021-04-01]. [neveřejně dostupný dokument].

[4] Remmark a.s. [online]. *Datová platforma pro marketing - typový uživatel a úkoly*. 2020. [cit. 2021-04-01]. [neveřejně dostupný dokument].

[5] Data FAIRport. Data FAIRport conference: JOINTLY DESIGNING A DATA FAIRPORT. In: *Data FAIRport* [online]. [cit. 2021-04-01]. Dostupné z: https://www.datafairport.org/component/content/article/8_news/9_item1/index.html

[6] Wilkinson, M.; Dumontier, M.; Aalbersberg, I.; aj. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [online]. 2016, 3(1). [cit. 2021-04-01]. ISSN 2052-4463. Dostupné z: doi:10.1038/sdata.2016.18

[7] Kahn, R.; Wilensky, R. A framework for distributed digital object services. *International Journal on Digital Libraries* [online]. 2006, 6(2), 115-123 [cit. 2021-04-01]. ISSN 1432-5012. Dostupné z: doi:10.1007/s00799-005-0128-x

[8] GO FAIR. FAIR Principles. In: *GO FAIR* [online], 2016. [cit. 2021-04-01]. Dostupné z: https://www.go-fair.org/fair-principles/

[9] GO FAIR. FAIRification Process. In: *GO FAIR* [online], 2016. [cit. 2021-04-01]. Dostupné z: https://www.go-fair.org/fair-principles/fairification-process/

[10] Guizzardi, G. Ontology, Ontologies and the "I" of FAIR. *Data Intelligence* [online]. 2020, 2(1-2), 181-191. [cit. 2021-04-07]. ISSN 2641-435X. Dostupné z: doi:10.1162/dint_a_00040

[11] Guizzardi, G. Ullmann's Triangle. In: Ontological Foundations for Structural Conceptual Models [online]. In: *CTIT PhD Thesis Series*. Enschede: Telematica Instituut / CTIT, 2005. [cit. 2021-04-02]. ISSN 1381-3617. ISBN 90-75176-81-3. Dostupné z: http://doc.utwente.nl/50826/

[12] Ceblová, L. Sémantická interoperabilita. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha : Národní knihovna ČR, 2003. [cit. 2021-04-07]. Dostupné z: https://aleph.nkp.cz/F/?func=directdoc_number=000000555local_base=KTD

[13] Guizzardi, G. Ontological Patterns, Anti-Patterns and Pattern Languages for Next-Generation Conceptual Modeling. In: *International Conference on Conceptual Modeling* [online]. Cham: Springer International Publishing, 2014. s. 13-27 [cit. 2021-04-07]. ISBN 978-3-319-12205-2. Dostupné z: doi:10.1007/978-3-319-12206-9_2

[14] Mylopoulos, J. Conceptual Modelling and Telos. In: Loucopoulos, P.; Zicari, R. *Conceptual Modeling, Databases, and Case: An Integrated View of Information Systems Development* [online]. United States, New York: John Wiley Sons, Inc., 1992. [cit. 2021-04-01]. ISBN 978-0-471-55462-2. Dostupné z: http://www.cs.toronto.edu/ jm/2507S/Readings/CM+Telos.pdf

[15] Pergl, R. *Conceptualisation: Chapters from Harmonising Enterprise and Software Engineering* [online]. Praha, 2018. [cit. 2021-04-01]. Dostupné z: doi:10.13140/RG.2.2.27388.08325. Habilitační práce. ČVUT v Praze, Fakulta informačních technologií, Katedra softwarového inženýrství.

[16] Guizzardi, G. On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models. In: *Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference DBIS'2006* [online]. Amsterdam: IOS Press, 2007. [cit. 2021-04-01]. 18-39. ISBN 9781586037154. Dostupné z: https://www.researchgate.net/publication/221278057_On_Ontology_ontologies_Conceptualizations_Modeling_Languages_and_MetaModels

[17] Guizzardi, G. *Ontological Foundations for Structural Conceptual Models* [online]. Enschede, Nizozemsko: Centre for Telematics and Information Technology, University of Twente, 2005. [cit. 2021-04-01]. ISBN 90-75176-81-3. Dostupné z: https://www.researchgate.net/publication/215697579_Ontological_Foundations_for_Structural_Conceptual_Models

[18] Guizzardi, G.; Wagner, G.; Almedia, J.P.A.; Guizzardi, R. Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story. *Applied Ontology* [online]. 2015, 10(3-4), 259-271. [cit. 2021-04-01]. ISSN 18758533. Dostupné z: doi:10.3233/AO-150157

[19] Šmajs, J.; Krob, J. *Úvod do ontologie: Bytí, prostor, čas, pohyb, evoluce, struktura, systém, řád, informace, vesmír, kultura, člověk : (Skriptum filoz. fak. MU)* [online]. 2. opr. a rozš. vyd. Brno: Masarykova univerzita, 1994. [cit. 2021-04-01]. ISBN 80-210-0879-2. Dostupné z: https://www.phil.muni.cz/fil/eo/skripta/uvod_do_ontologie.pdf

[20] Aristoteles. *Metafyzika*. 2. vyd. Přeložil Antonín Kříž. Praha: Petr Rezek, 2003. ISBN 80-86027-19-8.

[21] Heřt, J. Metafyzika.In: *Sisyfos* [online]. Občanské sdružení Český klub skeptiků Sisyfos, 2007. [cit. 2021-04-01]. Dostupné z: https://www.sisyfos.cz/clanek/983-metafyzika

[22] Mealy, G. Another Look at Data. In: *Managing Requirements Knowledge, International Workshop on* [online]. Anaheim: 1967. [cit. 2021-04-01]. Dostupné z: doi: 10.1109/AFIPS.1967.112

[23] Guizzardi, G. A classification of different types of ontology. In: Ontological Foundations for Structural Conceptual Models [online]. In: *CTIT PhD Thesis Series*. Enschede: Telematica Instituut / CTIT, 2005. [cit. 2021-04-02]. ISSN 1381-3617. ISBN 90-75176-81-3. Dostupné z: http://doc.utwente.nl/50826/

[24] Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *SCIENTIFIC AMERICAN* [online]. 2001, 5. [cit. 2021-04-01]. Dostupné z: https://www.scientificamerican.com/magazine/sa/2001/05-01/

[25] Vancura, M. Systém kategorií u Aristotela. *E-LOGOS* [online]. 2009, 16(1), 1-22. [cit. 2021-04-01]. ISSN 1211-0442. Dostupné z: https://e-logos.vse.cz/artkey/elg-200901-0022_System-kategorii-u-Aristotela.php

[26] Rybola, Z. *Towards OntoUML for Software Engineering: Transformation of OntoUML into Relational Databases* [online]. Praha, 2017. [cit. 2021-04-08]. Dostupné z: http://147.32.232.248/sites/default/files/PhDThesis-Rybola.pdf.

Disertační práce. České vysoké učení technické v Praze, Fakulta informačních technologií. doc. Ing. Karel Richta, CSc.

[27] *UML diagrams* [online]. uml-diagrams.org. [cit. 2021-04-02]. Dostupné z: https://www.uml-diagrams.org/

[28] *OntoUML specification* [online]. Centrum pro konceptuální modelování a implementace, Fakulta informačních technologií, České vysoké učení technické v Praze. [cit. 2021-04-02]. Dostupné z: https://ontouml.readthedocs.io/en/latest/index.html

[29] Guizzardi, G. Ontological Foundations for Structural Conceptual Models [online]. In: *CTIT PhD Thesis Series*. Enschede: Telematica Instituut / CTIT, 2005. [cit. 2021-04-02]. ISSN 1381-3617. ISBN 90-75176-81-3. Dostupné z: http://doc.utwente.nl/50826/

[30] Identity, Principle of. In: *New Catholic Encyclopedia* [online]. Encyclopedia.com. [cit. 2021-04-08]. Dostupné z: https://www.encyclopedia.com/religion/encyclopedias-almanacs-transcripts-and-maps/identity-principle

[31] Guizzardi, G.Ontological Distinctions in a Typology of Substantial Universals. In: Ontological Foundations for Structural Conceptual Models [online]. In: *CTIT PhD Thesis Series*. Enschede: Telematica Instituut / CTIT, 2005. [cit. 2021-04-02]. ISSN 1381-3617. ISBN 90-75176-81-3. Dostupné z: http://doc.utwente.nl/50826/

[32] *OpenPonk modeling platform* [online]. Centre for Conceptual Modelling and Implementation, 2020. [cit. 2021-04-01]. Dostupné z: https://openponk.org/

[33] Uhnák, P.; Pergl, R. The OpenPonk modeling platform. *Proceedings of the 11th edition of the International Workshop on Smalltalk Technologies* [online]. New York, NY, USA: ACM, 2016. s. 1-11. [cit. 2021-04-01]. ISBN 9781450345248. Dostupné z: doi:10.1145/2991041.2991055

[34] Bělohoubek, M. *OntoUML Models Verification for the OpenPonk platform* [online]. Praha, 2019. [cit. 2021-04-01]. Dostupné z: https://dspace.cvut.cz/handle/10467/83191. Bakalářská práce. ČVUT v Praze, Fakulta informačních technologií, Katedra softwarového inženýrství. doc. Ing. Robert Pergl, Ph.D.

[35] Ministerstvo financí České republiky. *Vývoj stavebního spoření* [online]. Poslední změna 02.03.2021. [cit. 2021-04-18]. Dostupné z: https://www.mfcr.cz/cs/soukromy-sektor/stavebni-sporeni/vyvoj-stavebniho-sporeni

[36] Český statistický úřad. *Průměrné spotřebitelské ceny vybraných výrobků - potravinářské výrobky* [online]. Poslední změna 14.12.2018. [cit. 2021-04-18]. Dostupné z: https://www.czso.cz/csu/czso/prumerne-spotrebitelske-ceny-vybranych-vyrobku-potravinarske-vyrobky

[37] Český statistický úřad. *Sklizeň zemědělských plodin podle krajů* [online]. 18.02.2021. [cit. 2021-04-18]. Dostupné z: https://www.czso.cz/csu/czso/sklizen-zemedelskych-plodin-podle-kraju

[38] STEM/MARK; MEDIAN. *RADIO PROJECT*. 3. a 4. čtvrtletí 2010. [cit. 2021-04-18]. [neveřejně dostupný dokument].

[39] Česká správa sociálního zabezpečení. Otevřená data. *Celkový počet důchodců, průměrná výše důchodu a průměrný věk důchodců podle roku, druhu důchodu a pohlaví za ČR, kraje a okresy* [online]. [cit. 2021-04-18]. Dostupné z: https://data.cssz.cz/-/duchodci-v-cr-krajich-okresech

[40] Česká správa sociálního zabezpečení. Otevřená data. *Počet vyplacených důchodů v České republice podle roku, druhu důchodu, měsíční výše důchodu a pohlaví* [online]. [cit. 2021-04-18]. Dostupné z: https://data.cssz.cz/-/vyplacene-duchody-dle-vyse

[41] Ministerstvo financí České republiky. *Vývoj penzijního připojištění a doplňkového penzijního spoření* [online]. Poslední změna 02.03.2021. [cit. 2021-04-18]. Dostupné z: https://www.mfcr.cz/cs/soukromy-sektor/soukrome-penzijni-systemy/iii-pilir-doplnkove-penzijni-sporeni-a-p/vyvoj-penzijniho-pripojisteni

[42] Eurostat. *Internet purchases by individuals (until 2019)* [online]. 27.01.2021. [cit. 2021-04-18]. Dostupné z: https://ec.europa.eu/eurostat/web/products-datasets/-/isoc_ec_ibuy

[43] Zákon č. 96/1993 Sb., o stavebním spoření a státní podpoře stavebního spoření a o doplnění zákona České národní rady č. 586/1992 Sb., o daních z příjmů, ve znění zákona České národní rady č. 35/1993 Sb.

[44] Zákon č. 89/2012 Sb., občanský zákoník.

[45] Zákon č. 219/2000 Sb., o majetku České republiky a jejím vystupování v právních vztazích.

[46] Zákon č. 21/1992 Sb., o bankách.

Příloha **A**

Seznam použitých zkratek

API Application Programming Interface

DOLCE Descriptive Ontology for Linguistic and Cognitive Engineering

GFO General Formalized Ontology

Nest BDA Nest Big Data Arena

UFO Unified Foundational Ontology

UML Unified Modeling Language

XML Extensible Markup Language

Příloha **B**

Obsah přiloženého CD

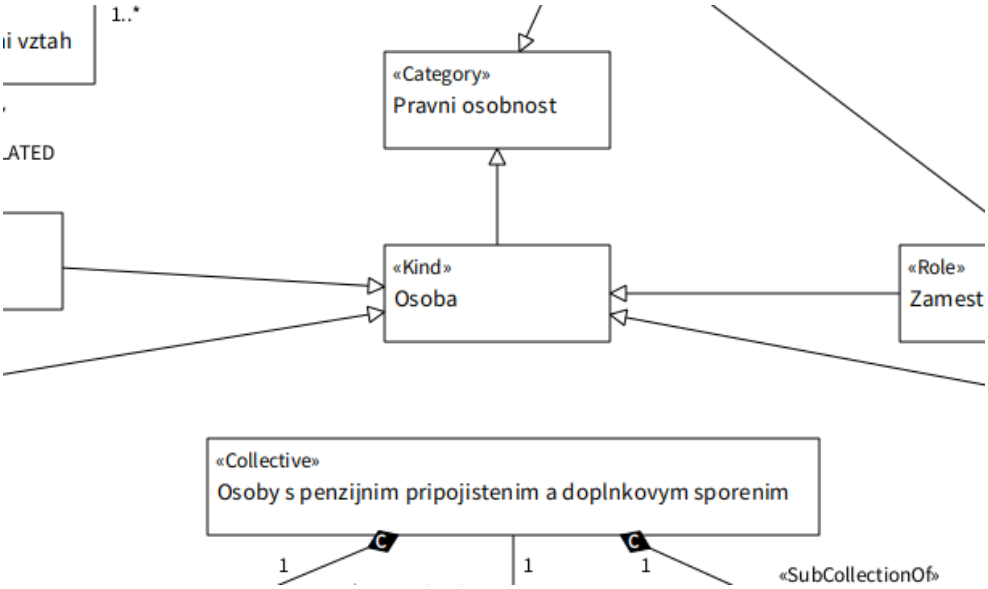
- `readme.txt` ..... stručný popis obsahu CD
- `models` ..... konceptuální modely
  - `stavebni-sporeni` ..... model stavebního spoření
  - `vyrobek-sklizen` ..... model cen výrobků a sklizně plodin
  - `radio` ..... model poslechovosti rozhlasu
  - `duchody` ..... model počtu důchodů a vyplacených důchodů
  - `penze` ..... model penzijního připojištění a spoření
  - `nakupy` ..... model internetových nákupů
  - `uzemi` ..... model území
- `BP_Martinkova_Jana_2021.pdf` ..... text práce ve formátu PDF
- `BP_Martinkova_Jana_2021.tex` .. zdrojová forma práce ve formátu LaTeX

Příloha C

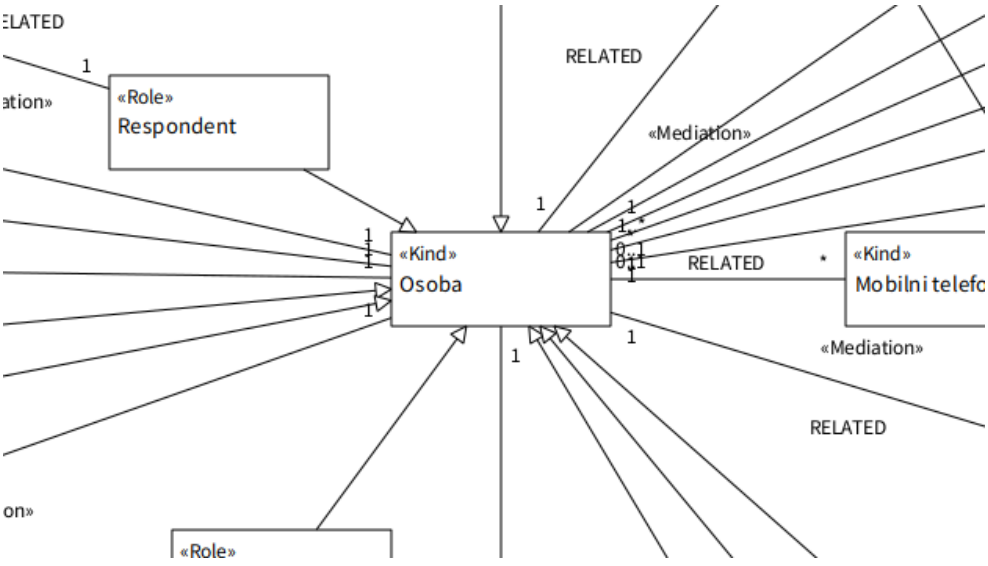
Zobrazení entit propojující modely

Příloha obsahuje výřezy modelů, které zobrazují společné entity představující reálné objekty. Přes tyto totožné prvky je možné rozdílné datové sady významově propojit.

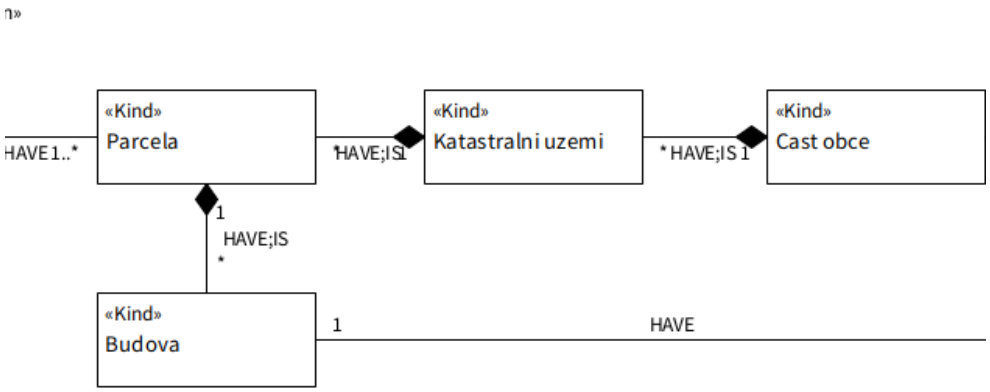
V příloze jsou postupně zobrazeny výřezy modelů obsahující entity `Osoba`, entity modelu `uzemi`, `Prodej`, `Duchod` a `Referencni obdobi`. Pro každou entitu jsou uvedeny dva vybrané ontologické modely datových sad, které danou entitu obsahují a tudíž jsou na základě tohoto významově stejného prvku propojitelné. Detailnější popis je uveden v 3.3.



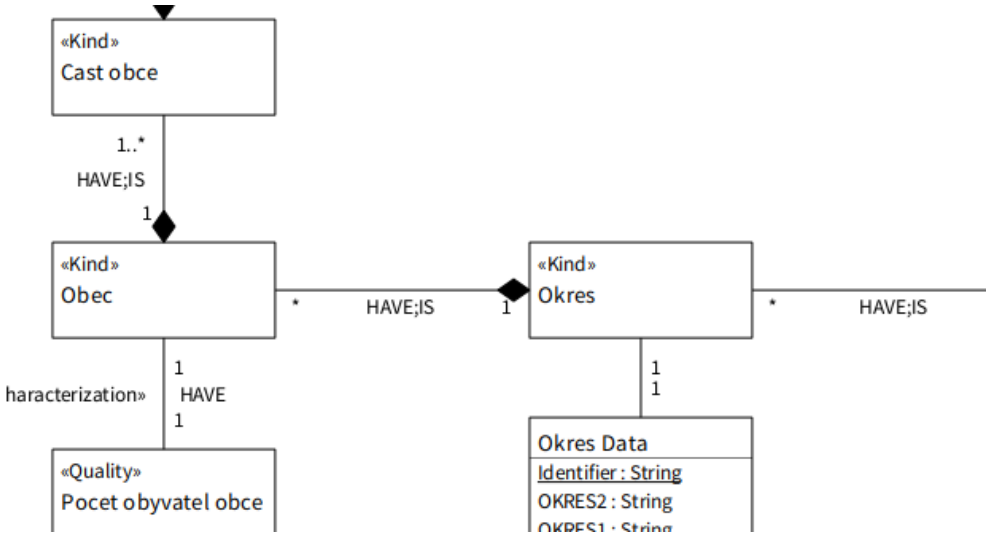
Obrázek C.1: Zobrazení entity `Osoba` v modelu `penze` (viz model `penze` v příloze B)



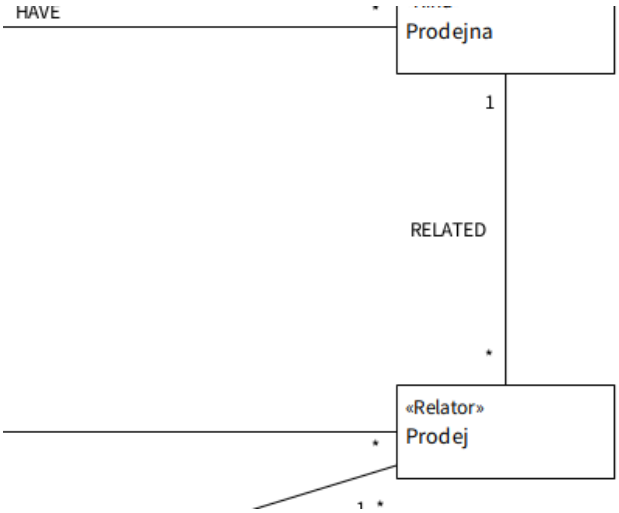
Obrázek C.2: Zobrazení entity `Osoba` v modelu `radio` (viz model `radio` v příloze B)



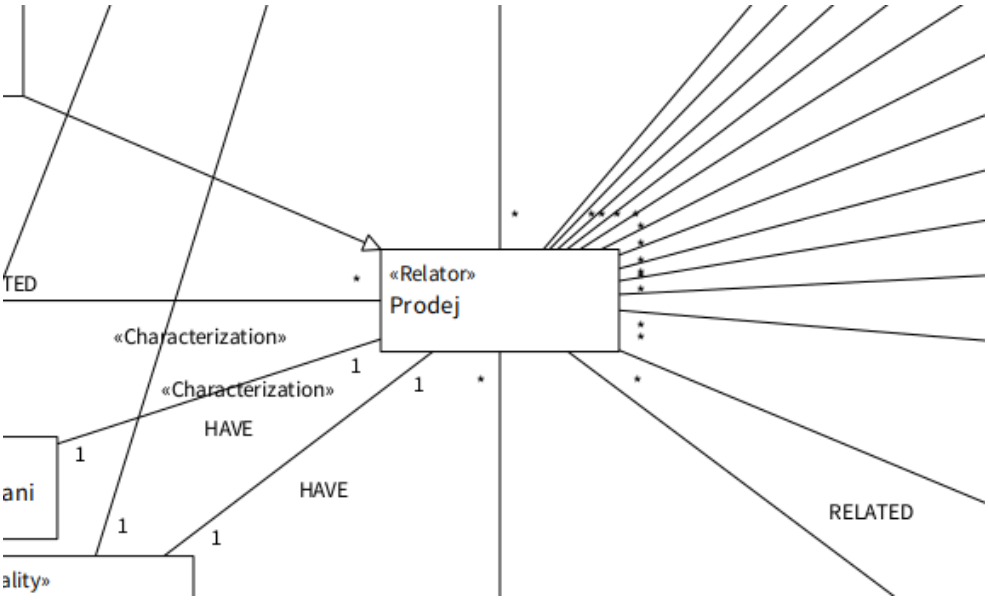
Obrázek C.3: Zobrazení entit území v modelu `vyrobek-sklizen` (viz model `vyrobek-sklizen` v příloze B)



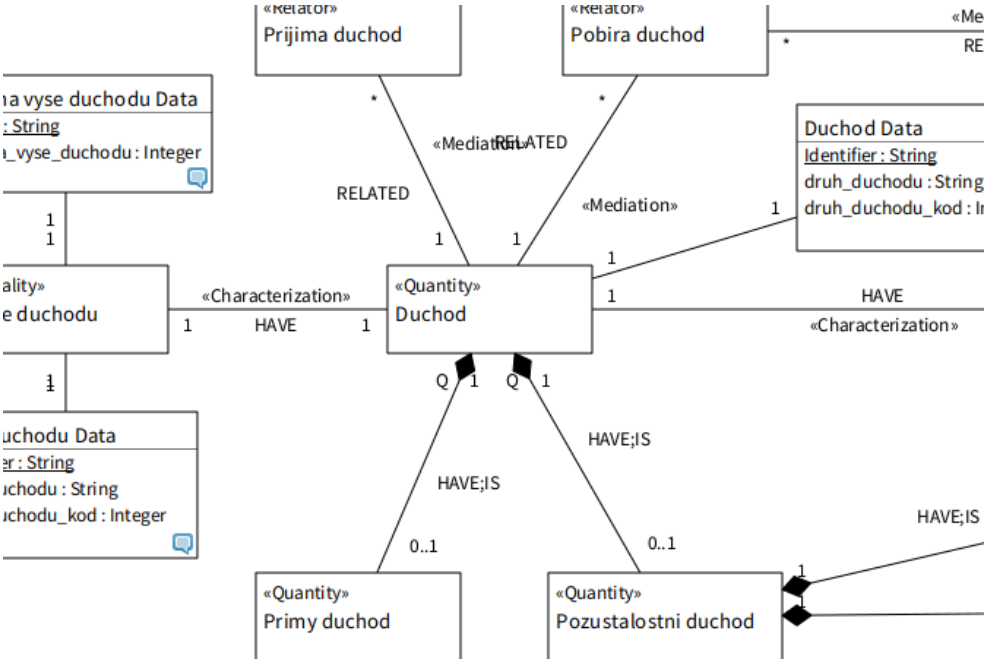
Obrázek C.4: Zobrazení entit území v modelu `radio` (viz model `radio` v příloze B)



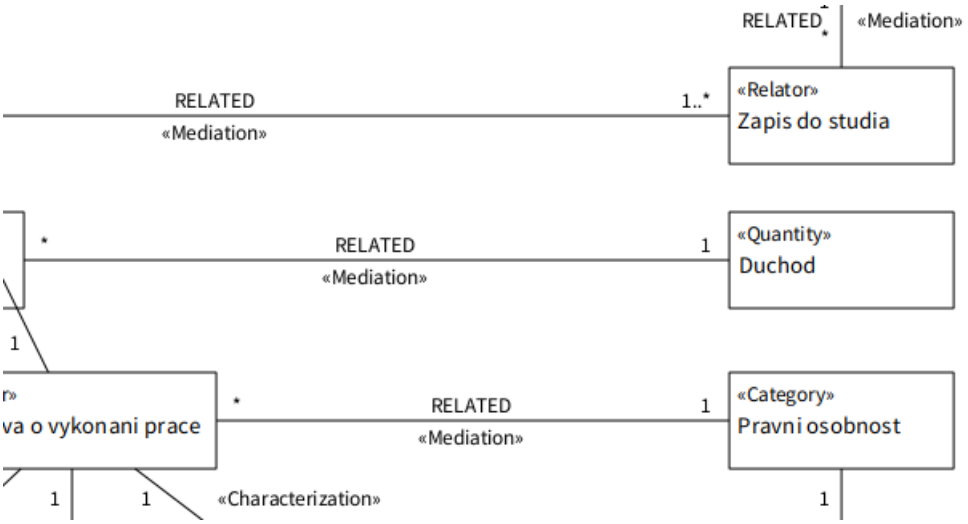
Obrázek C.5: Zobrazení entity `Prodej` v modelu `vyrobek-sklizen` (viz model `vyrobek-sklizen` v příloze B)



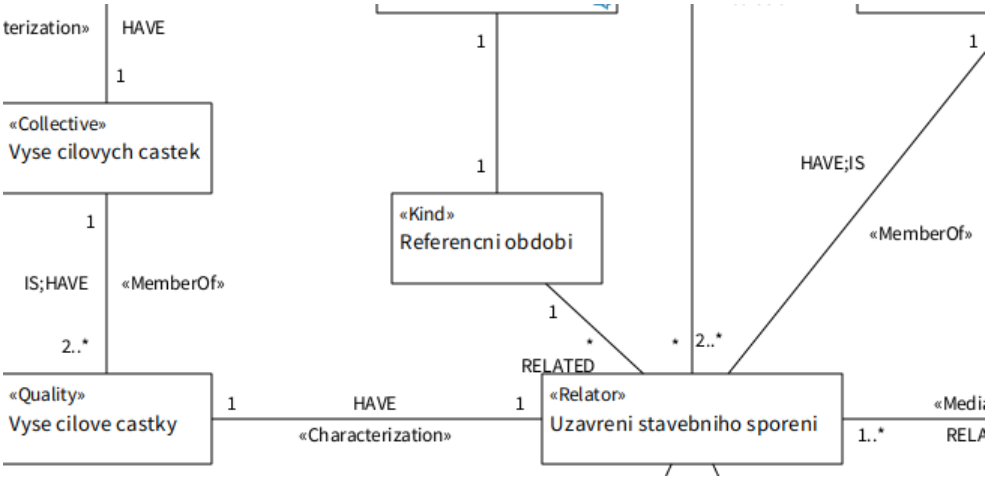
Obrázek C.6: Zobrazení entity `Prodej` v modelu `nakupy` (viz model `nakupy` v příloze B)



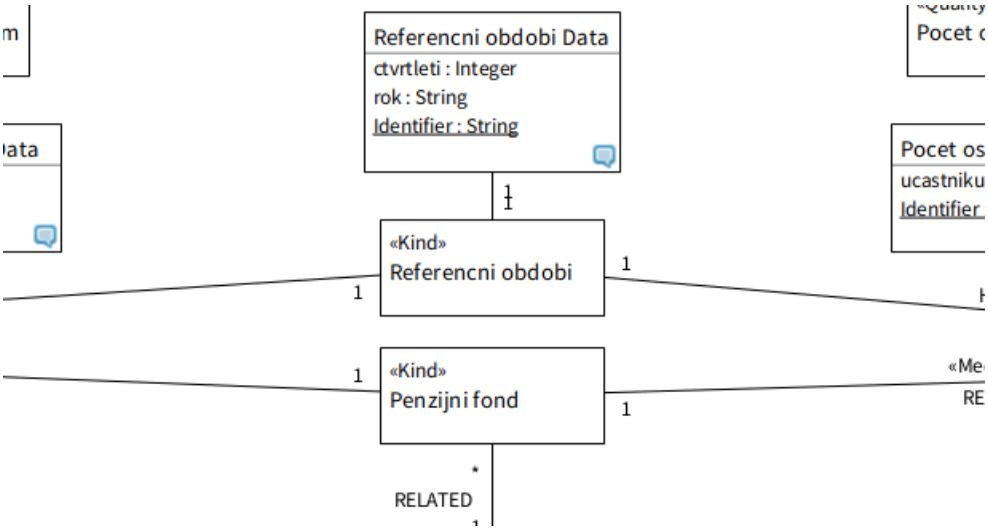
Obrázek C.7: Zobrazení entity `Duchod` v modelu `duchody` (viz model `duchody` v příloze B)



Obrázek C.8: Zobrazení entity `Duchod` v modelu `radio` (viz model `radio` v příloze B)



Obrázek C.9: Zobrazení entity `Referencni obdobi` v modelu `stavebni-soreni` (viz model `stavebni-soreni` v příloze B)



Obrázek C.10: Zobrazení entity `Referencni obdobi` v modelu `penze` (viz model `penze` v příloze B)