FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Systém pro agregaci a zobrazení dat o insolvenčních řízeních
Student:	Bc. Pavel Tůma
Vedoucí:	Ing. Marek Sušický
Studijní program:	Informatika
Studijní obor:	Webové a softwarové inženýrství
Katedra:	Katedra softwarového inženýrství
Platnost zadání:	Do konce letního semestru 2021/22

Pokyny pro vypracování

Insolvenční řízení je zákonem řízený složitý proces, který je dopodrobna zdokumentovaný v celém svém průběhu ve veřejném systému ISIR prostřednictvím PDF dokumentů.
- Seznamte se s problematikou insolvenčních řízení, popište celý proces a možnosti získávání dat.
- Navrhněte, implementuje a otestujte systém pro zpracování a přehledné zobrazení agregovaných dat o insolvencích, který bude zahrnovat:
  - modul pro získání dat a extrakci dat z oficiálních PDF formulářů MSp
  - serverová a klientská část včetně GUI
- Extrahujte data z PDF formulářů od 1.1.2019 a nad takto získanými daty nabídněte tvorbu základních statistik. Bude tak možné porovnat doby řízení u různých správců, porovnávat velikosti insolvenci apod. V práci není vyžadováno použití pokročilých metod pro zpravování obrazu a důraz je kladen na průběh oddlužení.
- Dbejte maximálně na srozumitelnost a snadné ovládání.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Michal Valenta, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 19. listopadu 2020

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Systém pro agregaci a zobrazení dat o insolvenčních řízeních

Bc. Pavel Tůma

Katedra softwarového inženýrství
Vedoucí práce: Ing. Marek Sušický

5. května 2021

Poděkování

Děkuji vedoucímu své diplomové práce Ing. Markovi Sušickému za návrh zajímavého tématu, vstřícnost při konzultacích a cenné rady při tvorbě 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 principu 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 předpisu. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisu, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programu, 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 zpusobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelum). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným zpusobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 5. května 2021 . . . . . .

Odkaz na tuto práci

Tůma, Pavel. *Systém pro agregaci a zobrazení dat o insolvenčních řízeních.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2021.

Abstrakt

Tato práce se zabývá analýzou, návrhem a implementací systému pro agregaci statistických dat o insolvenčních řízeních v České republice. V úvodní části je popsán proces insolvenčního řízení a možnosti, jak o tomto procesu automatizovaně získávat data. Návrh systému je členěn na modul pro sběr a agregaci dat a webové uživatelské rozhraní pro prezentaci výsledných statistik. Návrh modulu pro získávání dat se zaměřuje na způsob extrakce údajů z veřejně dostupných PDF formulářů, které jsou generovány z oficiálních šablon Ministerstva spravedlnosti. V práci je dále popsána realizace a způsoby testování výsledného řešení.

Výsledné statistiky může využít veřejnost pro analýzu různých aspektů insolvenčních řízení se zaměřením na oddlužení. Zveřejněný modul pro extrakci dat může být využit pro další analýzy dat insolvenčního procesu. V příloze práce lze nalézt zdrojový kód všech částí systému a jeho uživatelskou dokumentaci.

Klíčová slova insolvenční řízení, insolvenční rejstřík, oddlužení, extrakce dat, statistiky, návrh, implementace

Abstract

This thesis deals with the analysis, design and implementation of a system for the aggregation of statistical data on insolvency proceedings in the Czech Republic. The introductory part describes the process of insolvency proceedings and the possibilities of automated data acquisition regarding this process. The system design is divided into a module for data acquisition and aggregation and a web user interface for the presentation of the resulting statistics. The design of the data acquisition module focuses on data extraction from publicly available PDF documents, which are generated from official templates provided by the Ministry of Justice. The thesis also describes the implementation and testing of the resulting solution.

The resulting statistics can be used by the public to analyze various aspects of insolvency proceedings with a focus on debt relief. The published data extraction module can be used for further data analysis of the insolvency process. The attachment contains a source code of all parts of the system and a documentation reference.

Keywords insolvency proceedings, insolvency register, debt relief, data extraction, statistics, design, implementation

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Může se stát, že se osoba ocitne v platební neschopnosti a není schopna plnit své závazky vůči věřitelům. K této situaci může dospět například změnou životní situace, jako je zvýšení životních výdajů nebo rovněž snížení nebo dokonce ztráta příjmů. To je v dnešní době obzvláště aktuální, neboť aktuálně probíhající virová pandemie má negativní vliv na mnoho odvětví podnikání, jako je cestovní ruch, provoz restauračních zařízení, kulturní a rekreační zařízení a maloobchodní prodej.

Insolvenční řízení je typ soudního procesu, jehož cílem je volba způsobu řešení této situace tak, aby věřitelé dlužníka byli uspokojeni co nejvyšší možnou mírou. Tento proces se řídí insolvenčním zákonem (č. 182/2006 Sb.), který je v účinnosti od roku 2008. Jednou z možných forem insolvenčního řízení je proces oddlužení. Vyznačuje se tím, že pokud dlužník řádně plní povinnosti dané zákonnými podmínkami tohoto procesu a po celou dobu trvání oddlužení (zpravidla 5 let) usiluje o maximální uspokojení věřitelů, na konci procesu jsou mu zbývající dosud neuhrazené závazky zcela odpuštěny, a dlužník je tak zbaven dluhů.

Cílem práce je navrhnout a implementovat systém, který bude sbírat data o insolvencích, zpracovávat je a poskytovat nad získanými daty statistiky o insolvenčních řízeních v České republice, se zaměřením na řízení probíhající formou oddlužení. Výsledné statistiky může využít veřejnost za účelem větší informovanosti o vývoji i aktuálním stavu situace v oblasti insolvencí.

Cílem analytické části práce je seznámení se s průběhem insolvenčního řízení, konkrétně definice úpadku a popis způsobů jeho řešení. Dále bude provedena analýza existujících řešení poskytujících statistiky o insolvencích a následně budou analyzovány možné zdroje, které bude možné využít pro sběr dat nutných pro tvorbu statistik. V závěru analytické části budou stanoveny cíle na výsledné řešení formou funkčních a nefunkčních požadavků. Cílem praktické části práce je návrh a implementace systému pro sběr a zpracování dat o insolvenčních řízeních a prezentace statistik formou webového uživatelského rozhraní. Poslední část práce se zabývá testováním implementovaného řešení.

Kapitola 1

Analýza

1.1 Insolvenční řízení

Insolvenční řízení je zvláštní druh soudního řízení, jehož předmětem je dlužníkův úpadek a způsoby jeho řešení. Základním cílem insolvenčního řízení je uspořádaní majetkových vztahů mezi dlužníkem a jeho věřiteli [1]. V České republice se tento proces v současné právní úpravě řídí zejména zákonem č. 182/2006 Sb., o úpadku a způsobech jeho řešení (insolvenční zákon, [2]). Jednou ze zásad insolvenčního řízení je podle insolvenčního zákona to, aby žádný z účastníků řízení nebyl nespravedlivě poškozen nebo nedovoleně zvýhodněn a aby se dosáhlo rychlého a co největšího uspokojení věřitelů.

1.1.1 Úpadek dlužníka

Úpadek nebo hrozící úpadek dlužníka je základním předpokladem pro podání insolvenčního návrhu, tj. návrhu na zahájení insolvenčního řízení. Insolvenční zákon [2] definuje nutné podmínky úpadku dlužníka takto: dlužník má více věřitelů, má peněžité závazky alespoň 30 dnů po lhůtě splatnosti a tyto závazky není schopen plnit. V případě právnických osob nebo podnikajících fyzických osob je dlužník v úpadku také pokud je předlužen – má více věřitelů a celková hodnota jeho závazků převyšuje hodnotu jeho majetku [2]. O hrozící úpadek jde tehdy, pokud lze důvodně předpokládat, že dlužník nebude schopen včas uhradit větší část svých peněžitých závazků [2].

O tom, zda je dlužník skutečně v úpadku, rozhoduje insolvenční soud v počáteční fázi každého insolvenčního řízení [2]. Pokud insolvenční soud dojde k závěru, že dlužník je v úpadku nebo mu úpadek hrozí, vydává rozhodnutí o úpadku [1].

1.1.2 Účastníci insolvenčního řízení

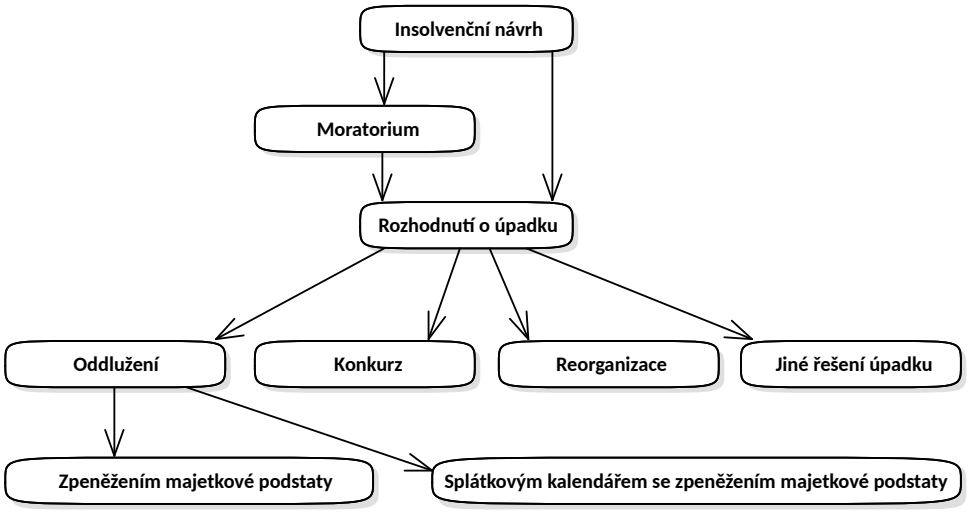
Typy subjektů, které figurují v insolvenčním řízení, jsou uvedeny v § 9 insolvenčního zákona [2]. Klíčovým subjektem řízení je dlužník, vůči kterému je řízení vedeno a věřitelé, kteří vůči němu uplatňují svá práva. V řízení dále figuruje insolvenční správce a insolvenční soud. V určitých případech vstupuje do řízení i státní zastupitelství nebo likvidátor dlužníka.

Dlužník je fyzická nebo právnická osoba, která má nesplněné závazky vůči svým věřitelům. Do insolvenčního řízení dlužník vstupuje, pokud jeho situace naplní definici úpadku dle ustanovení § 3 insolvenčního zákona. Zákon také uvádí určité typy subjektů, které nemohou vstoupit do insolvenčního řízení v pozici dlužníka – jde například o stát, Českou národní banku nebo veřejnou vysokou školu. [2]

Věřitel je osoba, které náleží plnění nějaké pohledávky od osoby dlužníka [3]. Věřitelé do řízení vstupují podáním tzv. přihlášky pohledávky. Zájmy většiny věřitelů v řízení prosazují věřitelské orgány – schůze věřitelů a věřitelský výbor. Věřitelský výbor musí být ustanoven, pokud se do řízení přihlásí více než 50 věřitelů. Pokud se věřitelů přihlásí méně, funkci věřitelského výboru nahrazuje zástupce věřitelů, kterého volí schůze věřitelů [2]. „Věřitelský výbor chrání společný zájem věřitelů a v součinnosti s insolvenčním správcem přispívá k naplnění účelu insolvenčního řízení.“ [2] Postavení věřitelů v insolvenčním řízení se liší dle toho, zda jsou zajištění nebo nezajištění. Zajištěný věřitel se vyznačuje tím, že jeho pohledávka je zajištěna majetkem dlužníka [1].

Insolvenční správce je fyzická nebo právnická osoba, kterou lze považovat za administrátora celého insolvenčního řízení [1]. Činnosti mu ukládá insolvenční zákon a zákon o insolvenčních správcích. Mezi jeho úlohy patří například sepsání majetku dlužníka, přezkoumání přihlášených pohledávek nebo uspokojování věřitelů z prostředků dlužníka. Insolvenční správce musí mít pro výkon své činnosti povolení od Ministerstva spravedlnosti, které získá po složení zkoušky insolvenčního správce a splnění dalších zákonem definovaných požadavků [1]. Insolvenčního správce pro konkrétní insolvenční řízení ustanovuje insolvenční soud výběrem ze seznamu insolvenčních správců [2].

Insolvenční soud vydává v insolvenčním řízení rozhodnutí, jejichž vydání ukládá nebo předpokládá zákon [2]. Mezi jeho úlohy patří také výkon dohlédací činnosti nad postupem ostatních subjektů v řízení a rozhodování o záležitostech týkajících se průběhu insolvenčního řízení [2]. V prvním stupni se jedná vždy o krajský soud, v jehož obvodu má dlužník bydliště nebo sídlo. V insolvenčních řízeních rozhoduje v prvním stupni jediný soudce (tzv. samosoudce) [2].



Obrázek 1.1: Diagram průběhu insolvenčního řízení (upraveno dle [6])

1.1.3 Průběh insolvenčního řízení

Insolvenční řízení začíná doručením insolvenčního návrhu na krajský soud, ve kterém má dlužník bydliště nebo sídlo. Návrh může podat jak dlužník, tak věřitel [4] a požadavky na návrh specifikuje § 103 a § 104 insolvenčního zákona [2]. Soud do 3 dnů od obdržení návrhu oznámí zahájení insolvenčního řízení zveřejněním vyhlášky v insolvenčním rejstříku. Soud také do řízení ustanoví insolvenčního správce. Od okamžiku zveřejnění vyhlášky nesmí být proti dlužníkovi prováděny exekuce a věřitelé mohou své pohledávky uplatnit pouze podáním přihlášky pohledávky do insolvenčního řízení [5].

V řízení může být v této fázi vyhlášeno moratorium. Jedná se o období, ve kterém nelze vydat rozhodnutí o úpadku. Jeho účelem je poskytnout dlužníkovi čas na vypořádání se s věřiteli ještě před proběhnutím insolvenčního řízení. Návrh na jeho vyhlášení podává dlužník a délka moratoria je maximálně 3 měsíce [6].

Soud následně zkoumá, zda-li je dlužník v úpadku nebo mu úpadek hrozí. Pokud tak shledá, vydává rozhodnutí o úpadku, čímž zahajuje druhou fázi insolvenčního řízení, ve které bude úpadek řešen [6]. O tom, který způsob řešení úpadku bude zvolen rozhoduje insolvenční soud dle okolností příslušného insolvenčního řízení. Možné průběhy insolvenčního řízení jsou znázorněny na diagramu 1.1.

1.1.4 Způsoby řešení úpadku

Insolvenční zákon [2] uvádí 4 možné způsoby řešení úpadku: konkurz, reorganizace, oddlužení a zvláštní způsoby pro určité subjekty nebo druhy případů stanovené zákonem.

1.1.4.1 Konkurz

Podstatou konkurzu je poměrné uspokojení zjištěných pohledávek věřitelů z výnosu zpeněžení majetku dlužníka. Neuspokojené pohledávky nebo jejich části po skončení řízení nezanikají [2]. Okamžikem vyhlášení konkurzu připadá právo nakládat s majetkem dlužníka na insolvenčního správce. Majetek se zpeněžuje např. prodejem movitých věcí a nemovitostí nebo veřejnou dražbou [6].

1.1.4.2 Reorganizace

Reorganizace je určena pro řešení úpadku podnikatelů s ročním obratem alespoň 50 000 000 Kč. Vyznačuje se tím, že během ní zůstává podnikatelská činnost dlužníka nepřerušena, je-li tato činnost v mezích reorganizačního plánu. Toto řešení je často pro věřitele výhodné, protože pokud je reorganizace úspěšná, je větší pravděpodobnost, že se jejich pohledávky podaří uspokojit ve větší míře, než při řešení konkurzem, kdy by bylo podnikání dlužníka zastaveno [6].

1.1.4.3 Oddlužení

Oddlužení je způsob řešení úpadku fyzických osob, jejichž dluhy nepocházejí z podnikání, nebo právnických osob, které nejsou považovány za podnikatele. Oddlužení se vyznačuje tím, že po jeho splnění dlužníkovy závazky vůči věřitelům zanikají a to i v případě, že se je nepodařilo uspokojit v plné míře. Nesmí však jít o plnění nižší než 30% z celkového dluhu dlužníka [6].

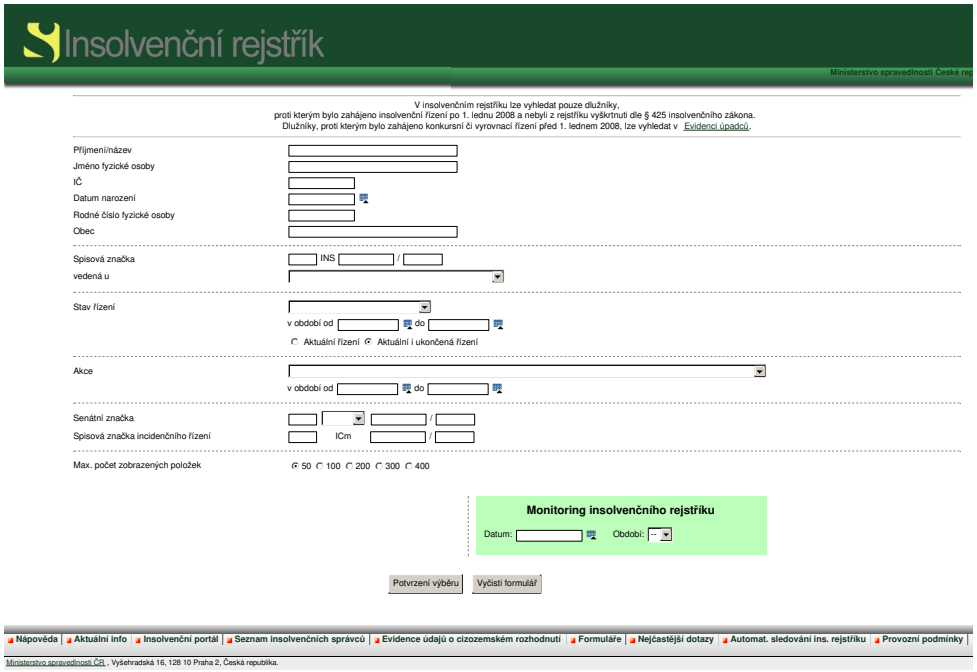
Oddlužení se provádí buď zpeněžením majetkové podstaty dlužníka podobným způsobem, jako je tomu u konkurzu, nebo plněním splátkového kalendáře se zpeněžením majetkové podstaty. V případě splátkového kalendáře je dlužník povinen po dobu 5 let[1] odvádět ze svého příjmu část prostředků, které jsou přerozdělovány mezi věřitele dle jejich postavení [6].

Dlužník musí během oddlužení poskytovat maximální součinnost a vynakládat veškeré úsilí k plnému uspokojení pohledávek svých věřitelů. Pokud je například dlužník nezaměstnaný, musí o získání příjmu usilovat. Pokud je zjištěn nepoctivý záměr dlužníka nebo neplní-li dlužník své povinnosti (např. zatajuje některé příjmy), může insolvenční soud oddlužení zrušit. V případě zrušení oddlužení je úpadek řešen konkurzem, nebo je řízení úplně zastaveno [8].

1.1.5 Insolvenční rejstřík

Insolvenční rejstřík je informačním systémem veřejné správy, ve kterém jsou evidovány spisy všech probíhajících i ukončených insolvenčních řízení [2]. Tento systém provozuje Ministerstvo spravedlnosti a jeho veřejná část je

[1] V době tvorby práce je v projednávání novela zákona, která v určitých případech umožňuje dobu trvání oddlužení 3 roky. V případě schválení by vešla v platnost od 1. 7. 2021. [7]



Obrázek 1.2: Aktuální podoba webového rozhraní insolvenčního rejstříku [9]

k dispozici na adrese isir.justice.cz [9]. Dle ustanovení § 419 insolvenčního zákona [2] je insolvenční rejstřík „veřejně přístupný, s výjimkou údajů, o kterých tak stanoví tento zákon. Každý má právo do něj nahlížet a pořizovat si z něj kopie a výpisy.“ Z rejstříku jsou dlužníci vyškrtnuti 5 let po skončení insolvenčního řízení a všechny související údaje jsou dle § 425 znepřístupněny.

Spisy insolvenčních řízení jsou v rejstříku členěny do oddílů (Řízení do úpadku, Řízení po úpadku, Incidenční spory, Ostatní a Přihlášky). V každém oddílu jsou chronologicky vypsány události, které v řízení nastaly a datum jejich zveřejnění. U většiny těchto záznamů je k dispozici ke stažení dokument ve formátu PDF [9].

1.2 Existující řešení

V této podkapitole jsou popsány existující služby nebo subjekty, které poskytují statistiky o insolvenčních řízeních nebo jinak zpracovávají data související s insolvencemi.

1.2.1 Mapa insolvencí

Mapa insolvencí je dostupná na portálu insolcentrum.cz, který od roku 2019 provozuje společnost InsolCentrum ve spolupráci s Hospodářskou komorou České republiky [10]. Na mapě jsou barevně odlišeny územní celky České re-

publiky podle toho, kolik dlužníků je v dané oblasti v insolvenčním řízení. V závislosti na přiblížení mapy lze tato data zkoumat na úrovni krajů až po jednotlivé okresy a města.

Kromě mapy jsou na webu InsolCentra prezentovány i další statistiky o insolvencích s rozdělením dle způsobu řešení úpadku. Pro oddlužení jsou k dispozici statistiky jako graf častých věkových kategorií dlužníka, průměr procentuální návratnosti dluhu věřitelům a procentuální srovnání nákladů insolvenčního řízení s částkou skutečně vyplacenou věřitelům. Obdobné typy statistických výstupů jsou k dispozici i pro způsob řešení úpadku konkurzem a reorganizací [11]. Nevýhodou je, že Mapa insolvencí není průběžně aktualizována a poslední verze vychází z dat k datu 1. 1. 2019 [11].

1.2.2 Insolvency report

Agentura Surveilligence vydává od roku 2018 na začátku každého měsíce veřejně dostupnou přehledovou zprávu zvanou Insolvency Report [12]. V ní jsou prezentována data o insolvencích v České a Slovenské republice za poslední kalendářní měsíc. Jako zdroj dat pro Českou republiku je použit insolvenční rejstřík a administrativní registr ekonomických subjektů.

V úvodu těchto statistických zpráv jsou shrnující informace o počtu nových insolvenčních návrhů, konkurzů a reorganizací. Je uveden počet vyhlášených moratorií, kolik reorganizací bylo přeměněno na konkurz nebo jaké je typické stáří firem v konkurzu. V úvodu je graf počtů insolvenčních řízení po jednotlivých měsících za poslední 4 roky, přičemž do grafu jsou pro srovnání vyneseny i počty obdobného procesu na Slovensku. V části s detailnějšími statistikami pro Českou republiku je možné nalézt graf měsíčního počtu konkurzů, reorganizací a četnost moratorií. Data s počty insolvenčních návrhů a konkurzů jsou srovnána s údaji o subjektech dostupných v registru firem, jako je počet zaměstnanců, doba existence subjektu nebo velikost základního kapitálu. Je k dispozici seznam insolvenčních správců seřazený dle počtu konkurzních řízení, do kterých byli přiřazeni. V závěru je vyobrazen přehled počtu přihlášek bankovních a zahraničních věřitelů za poslední měsíc [13].

1.2.3 Komerční poskytovatelé dat

Společnost CRIF – Czech Credit Bureau poskytuje službu CRIBIS, která slouží k sledování kredibility podnikatelů a fyzických osob v České republice. Taková služba má význam převážně pro firmy, kterým umožní prověřit své obchodní partnery. Většina výstupů této aplikace je podmíněna nákupem prémiového členství, ale i v bezplatné verzi aplikace dostupné na adrese informaceofirmach.cz je možné vyhledat subjekt a aplikace zobrazí kromě jiných informací i to, zda je daný subjekt účasten v insolvenčním řízení [14]. Kromě této aplikace vydává CRIF tiskové zprávy týkající se aktuální situace ohledně probíhajících insolvenčních řízeních v České republice. Mezi statistické výstupy

v těchto zprávách patří například vývoj počtu nových insolvenčních návrhů po měsících za posledních několik let a jejich kategorizace dle typu dlužníka – obchodní společnosti, podnikající a nepodnikající fyzické osoby [15].

Nástroje pro kontrolu kredibility subjektu poskytuje i společnost Bisnode prostřednictvím placených produktů jako jsou Bisnode Artemis, Bisnode Kerberos, Bisnode MagnusWeb, Bisnode Risk Portfolio, D&B BIR – Kreditní zprávy a další [16]. Zde je zaměření převážně na to, zda insolvence u konkrétního subjektu existuje, veřejné statistiky o insolvencích poskytovány nejsou.

Služba Creditcheck [17] je dalším komerčním nástrojem pro kontrolu kredibility. Na svých stránkách uvádí [18], jak data z insolvenčního rejstříku využívá. Každou minutu provádí kontrolu rejstříku a v případě nového insolvenčního řízení zobrazí u subjektu červený semafor v podnikovém informačním systému zákazníka nebo zašle upozornění emailem.

Existuje mnoho dalších komerčních služeb, které nabízejí funkcionalitu sledování zadaných subjektů v insolvenčním rejstříku, patří mezi ně např. Monitoring Insolvence 2008 [19], Monitor Justice [20], Sledování insolvence [21] nebo GRiT – Monitoring insolvence [22]. Jde však převážně spíše o redistribuci existujících záznamů v insolvenčním rejstříku než o tvorbu statistik.

1.2.4 Statistiky zadlužení

Dále existuje řada organizací zveřejňující statistiky související obecně se zadlužením. Ačkoliv nejde přímo o statistiky insolvenčních řízení, bývají tato data se statistikami o insolvencích často srovnávána. Jedná se například o statistiky míry zadlužení domácností a míry úvěrů v selhání, které poskytuje ČNB [23], statistiky míry příjmové chudoby domácností, které poskytuje ČSÚ [24], nebo údaje podílu populace s dluhy po splatnosti, které poskytuje statistický úřad Evropské unie [25].

1.3 Zdroje dat o insolvenčním řízení

V této sekci budou popsány dostupné možnosti strojového čtení dat z insolvenčního rejstříku. Těchto poznatků bude využito v návrhu praktické části pro získání dat pro agregaci statistik.

1.3.1 Webová služba insolvenčního rejstříku

Insolvenční rejstřík má kromě webového uživatelského rozhraní i webovou službu pro strojový přístup k datům. Tato služba je veřejně přístupná a provozuje ji Ministerstvo spravedlnosti ve spolupráci s CCA Group a.s. Rozhraní webové služby využívá protokol SOAP. Prostřednictvím této služby je možné získat všechny údaje o věcech insolvenčního řízení, které jsou dostupné v uživatelském rozhraní insolvenčního rejstříku [26].

Databáze insolvenčního rejstříku ukládá informace formou událostí, ke kterým v průběhu řízení dochází. Tyto akce jsou unikátně identifikovány sekvenčně generovaným číslem vytvořením v době zápisu události do systému. Webová služba umožňuje pouze jeden typ dotazu a to filtrovací dotaz na vybraný rozsah událostí. Není k dispozici žádný jiný filtr např. dle názvu dlužníka, jako je tomu v uživatelském rozhraní rejstříku. Služba je takto navržena z důvodu zajištění vysoké dostupnosti – předpokládá vytvoření kopií databáze na straně uživatelů. Po vytvoření vlastní kopie databáze uživatelé služby odesílají do služby požadavek obsahující identifikátor poslední stažené události a aktualizují si svoji kopii jen o nově přidané události do rejstříku.

Každá událost obsahuje informace jako spisovou značku řízení, ke kterému náleží, typ události z číselníku události, dodatečná strukturovaná data v závislosti na typu události a URL adresu dokumentu přiloženého k události, je-li dokument zveřejněn.

Webová služba poskytuje ve strojově čitelném formátu jen základní informace o stavu, průběhu a účastnících řízení. Detailnější údaje jako např. peněžní výše závazků dlužníka jednotlivým věřitelům nebo informace o aktuálním uspokojení věřitelů služba neposkytuje, ale je možné je vyčíst z přiložených dokumentů, na které služba odkazuje.

1.3.2 Údaje z přiložených dokumentů

Dokumenty přikládané k jednotlivým událostem jsou zveřejňovány ve formátu PDF. Jejich struktura se liší v závislosti na typu události, ke které dokument náleží. V případech, kdy písemnost byla na insolvenční soud doručena poštou, může jít i o naskenované dokumenty, někdy i s ručně psaným obsahem. Dokumenty často obsahují prostý text vyjádření některého ze subjektů figurujícího v řízení – jde např. o dokumenty typu Sdělení, Zpráva, Usnesení, Oznámení, Návrh, Přípis, Protokol, Výzva, Žádost, Odvolání, Dotaz, Stížnost, Nařízení, Vyrozumění, Souhlas, Referát, Pokyn nebo Vyhláška [9].

Určité typy dokumentů mají podobu formulářů generovaných ze šablon, které poskytuje insolvenčním správcům Ministerstvo spravedlnosti na stránkách insolvenčního rejstříku [27]. Jde o dokumenty, které se v řízení vyskytují často a vždy obsahují stejné typy informací, a využití poskytnutých šablon tak insolvenčním správcům šetří čas. Mezi typy dokumentů, pro které jsou tyto formuláře k dispozici patří např. Přihláška pohledávky, Seznam přihlášených pohledávek, Soupis majetkové podstaty, Konečná zpráva, Zpráva pro oddlužení a o přezkumu, Zpráva o plnění oddlužení nebo Zpráva o splnění oddlužení [27]. Příklad podoby těchto formulářů je na obrázku 1.3. Obrázek 1.3a znázorňuje úvodní stranu Přihlášky pohledávky s údaji o dlužníkovi a věřiteli. Obrázek 1.3b znázorňuje tabulku měsíčního výkazu plnění ve formuláři Zpráva o plnění oddlužení. V dokumentech generovaných z těchto šablon je grafické rozložení textových prvků a názvy polí stejné. Bude tak možné využít

nástroje pro převod PDF na text a následnou extrakci jednotlivých polí dle jejich popisných textů.

Ne vždy jsou dokumenty vytvořeny pomocí poskytnutých šablon. Insolvenční správce může např. pro určitý typ dokumentu používat vlastní šablonu nebo vůbec šablony nevyužít a dokument vytvořit pomocí běžného tabulkového nebo textového editoru. Ve většině případů jsou oficiální předlohy použity, což bude dále v práci přesně otestováno.

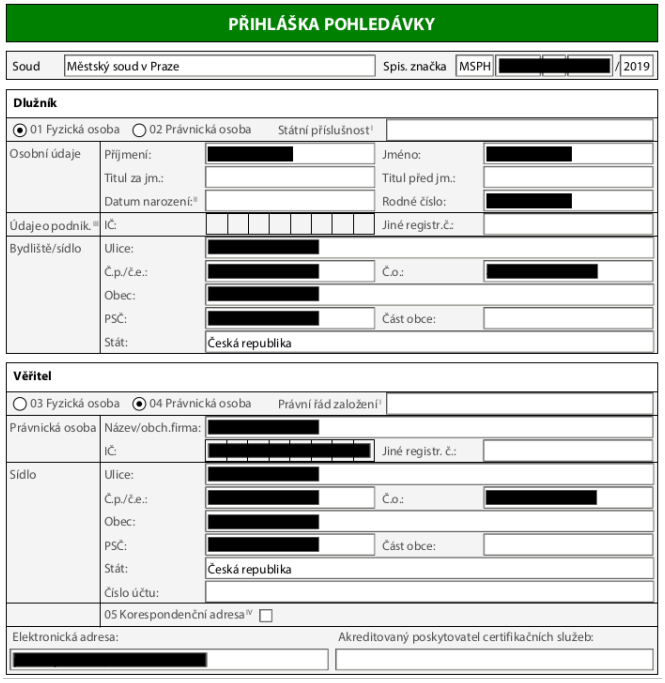
1.3.3 Volba typů dokumentů pro extrakci dat

Pro účely extrakce dat jsem ze seznamu dostupných formulářů [27] zvolil níže uvedených 5 typů tak, aby byly zahrnuty formuláře, které poskytují detailní informace o průběhu oddlužení a zároveň jsou v rejstříku zveřejňovány dostatečně často – tj. v běžném průběhu každého řízení řešícího úpadek formou oddlužení je zveřejněna alespoň jedna událost s přiloženým dokumentem tohoto typu.

Přihláška pohledávky je dokument, prostřednictvím kterého věřitel přihlašuje svoji pohledávku do řízení. Na úvodní straně se nacházejí údaje o dlužníkovi a o věřiteli, následují detaily jednotlivých pohledávek, kterých může být v jedné přihlášce více. U každé pohledávky je uveden typ pohledávky, výše jistiny, důvod vzniku, celková výše a vlastnosti pohledávky (např. zda-li je podřízená, peněžitá, podmíněná, splatná nebo v cizí měně). U pohledávky mohou být také uvedeny informace o vykonatelnosti, druh příslušenství (např. úroky, soudní poplatky a jiné náklady) a jeho výše nebo další okolnosti.

Přehledový list sestavuje insolvenční správce po uplynutí lhůty pro podání přihlášek a poté co provede přezkumné jednání. V úvodní části dokumentu se nachází shrnutí s celkovou výší přihlášených pohledávek včetně rozdělení na zajištěné a nezajištěné. Hlavní částí dokumentu je přehledová tabulka, ve které je seznam přihlášek a u každé je evidováno pořadové číslo a číslo věřitele, celková výše pohledávky, kolik zbývá k uspokojení, jaká část je vykonatelná a nevykonatelná, případně jaká částka z pohledávky byla popřena, odmítnuta, podmíněna nebo duplicitní.

Zpráva pro oddlužení je dokument, který soudu předkládá insolvenční správce v souladu s § 398a insolvenčního zákona [2]. V dokumentu insolvenční správce zhodnotí předpokládané plnění věřitelům a pokud je doporučena forma oddlužení plněním splátkového kalendáře se zpeněžením majetkové podstaty, připojí také návrh distribučního schématu [2]. V první části dokumentu jsou údaje o hospodářské situaci dlužníka – je vypsán seznam příjmů dlužníka a jejich typy, případně zda dlužník přijímá důchod, finanční dary nebo rentu. Je uvedena forma bydlení



PŘIHLÁŠKA POHLEDÁVKY

Soud: Městský soud v Praze | Spis. značka: MSPH [redacted] / 2019

Dlužník

(●) 01 Fyzická osoba ( ) 02 Právnická osoba — Státní příslušnost[I]

Osobní údaje	Příjmení: [redacted]	Jméno: [redacted]	
	Titul za jm.:	Titul před jm.:	
	Datum narození:[II]	Rodné číslo: [redacted]	
Údaje o podnik.[III]	IČ:	Jiné registr.č.:	
Bydliště/sídlo	Ulice: [redacted]		
	Č.p./č.e.: [redacted]	Č.o.: [redacted]	
	Obec: [redacted]		
	PSČ: [redacted]	Část obce:	
	Stát: Česká republika		

Věřitel

( ) 03 Fyzická osoba (●) 04 Právnická osoba — Právní řád založení[I]

Právnická osoba	Název/obch.firma: [redacted]		
	IČ: [redacted]	Jiné registr. č.:	
Sídlo	Ulice: [redacted]		
	Č.p./č.e.: [redacted]	Č.o.: [redacted]	
	Obec: [redacted]		
	PSČ: [redacted]	Část obce:	
	Stát: Česká republika		
	Číslo účtu:		
	05 Korespondenční adresa[IV] ☐		

Elektronická adresa: [redacted] — Akreditovaný poskytovatel certifikačních služeb:

(a)

B. MĚSÍČNÍ VÝKAZ PLNĚNÍ SPLÁTKOVÉHO KALENDÁŘE

Rok	2020	2020	2020	2020	2020	2020
Měsíc	5	6	7	8	9	10
Příjem	23 983 Kč	22 776 Kč	21 816 Kč	24 178 Kč	22 496 Kč	25 911 Kč
Provedené srážky	11 382 Kč	10 578 Kč	9 938 Kč	10 936 Kč	9 816 Kč	12 092 Kč
ZM+NNB	6 908 Kč	6 908 Kč	6 908 Kč	6 908 Kč	6 908 Kč	6 908 Kč
Vyživované osoby	0	0	0	0	0	0
Nepostižitelné	12 601 Kč	12 198 Kč	12 454 Kč	13 242 Kč	12 680 Kč	13 819 Kč
Postižitelné	11 382 Kč	10 578 Kč	9 362 Kč	10 936 Kč	9 816 Kč	12 092 Kč
Vráceno dlužníkům	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Mimořádný příjem	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Darovací smlouva	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
K přerozdělení	**11 382 Kč**	**10 578 Kč**	**9 938 Kč**	**10 936 Kč**	**9 816 Kč**	**12 092 Kč**
- na odměnu IS	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč
- na výživné	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
- na jiné zapodstatové pohledávky	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
- ostatním věřitelům	10 293 Kč	9 489 Kč	8 849 Kč	9 847 Kč	8 727 Kč	11 003 Kč

\+ -

Věřitel	Dp.	%	Zjištěná pohledávka	Vyplaceno věřitelům					
[redacted]	☐	12,17 %	31 883 Kč	1 252,18 Kč	1 154,37 Kč	1 076,51 Kč	1 197,92 Kč	1 061,67 Kč	1 338,55 Kč
[redacted]	☐	1,83 %	4 800 Kč	188,52 Kč	173,79 Kč	162,07 Kč	180,35 Kč	159,83 Kč	201,52 Kč
[redacted]	☐	3,34 %	8 756,73 Kč	343,91 Kč	317,05 Kč	295,67 Kč	329,01 Kč	291,59 Kč	367,64 Kč
[redacted]	☐	44,09 %	115 542,13 Kč	4 537,81 Kč	4 183,36 Kč	3 901,2 Kč	4 341,19 Kč	3 847,42 Kč	4 850,83 Kč
[redacted]	☐	2,35 %	6 171,87 Kč	242,39 Kč	223,46 Kč	208,39 Kč	231,89 Kč	205,52 Kč	259,11 Kč
[redacted]	☐	2,03 %	5 318 Kč	208,86 Kč	192,55 Kč	179,56 Kč	199,81 Kč	177,08 Kč	223,27 Kč
[redacted]	☐	9,36 %	24 523,92 Kč	963,16 Kč	887,92 Kč	828,03 Kč	921,42 Kč	816,62 Kč	1 029,59 Kč
Celkem			Celkem přerozděleno věřitelům	168 088 Kč	177 577 Kč	186 426 Kč	196 273 Kč	205 000 Kč	216 003 Kč
			Míra uspokojení věřitelů	64,0 %	68,0 %	71,0 %	75,0 %	78,0 %	82,0 %
			Očekávaná míra uspokojení věřitelů	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
Měsíc oddlužení				25	26	27	28	29	30

(b)

Obrázek 1.3: Anonymizovaný příklad grafického rozložení dokumentů generovaných ze šablon poskytovaných Ministerstvem spravedlnosti [9]

a vyčísleny životní náklady dlužníka. Následuje přehled soupisu majetkové podstaty dlužníka, kde je vypsána výše dlužníkova majetku a je rozdělena dle typů majetku (nemovitý, movitý, finance, pohledávky, ostatní). Na základě těchto informací je vyčíslena tabulka s předpokládanou mírou uspokojení věřitelů a sestaven návrh insolvenčního správce, zda má být oddlužení povoleno a případně jakou formou.

Zpráva o plnění oddlužení se zveřejňuje v průběhu oddlužení a insolvenční správce v ní shrnuje výsledky své činnosti a aktuální stav uspokojení věřitelů, což mu ukládá § 412 odst. 2 insolvenčního zákona [2]. Podstatná část tohoto dokumentu je měsíční výkaz plnění splátkového kalendáře, ve kterém je detailně po jednotlivých měsících uveden příjem dlužníka za daný měsíc a kolik z tohoto příjmu bylo použito pro přerozdělení věřitelům. Druhá sekce tohoto výkazu obsahuje tabulku všech věřitelů a částku, která byla pro daný měsíc věřiteli vyplacena. Šablona pro tento typ dokumentu umožňuje tyto údaje zadat za období až 6 měsíců.

Zpráva o splnění oddlužení je dokument sestavovaný insolvenčním správcem na konci procesu oddlužení. Dokument obsahuje shrnutí průběhu řízení a jeho výsledek, zejména konečné uspokojení věřitelů. Insolvenční správce v ní dále uvádí, zda dlužník řádně plnil všechny povinnosti podle insolvenčního zákona a zda doporučuje rozhodnout o splnění oddlužení. Zpráva obsahuje i sekci s detailem vyúčtování odměny a náhrady nákladů insolvenčního správce.

1.4 Způsoby strojového čtení dat z PDF

Specifikace formátu PDF byla vytvořena společností Adobe Systems a slouží pro přenositelnou reprezentaci elektronických dokumentů při zachování stejného grafického zobrazení nezávisle na prostředí, ve kterém byl dokument vytvořen, nebo ve kterém je zobrazován [28]. Formát PDF náleží do nejnižší úrovně 5-hvězdičkového schématu nasazení otevřených dat [29], neboť takto uložená data nejdou strojově zpracovávat bez vytvoření specializovaného scraperu.

Textové řetězce jsou v interní struktuře PDF uloženy pomocí textových objektů na přesně definovaných souřadnicích [28]. Text, který se v PDF zobrazí na jednom řádku, se často může skládat z mnoha samostatných textových objektů. Pro převod takového řádku na text je nutné zobrazit všechny textové objekty v PDF do virtuálního souřadnicového prostoru a následně dle jejich pozice aproximovat jejich příslušnost k řádkům a správně přidat mezery mezi slovy.

Mezi volně dostupné nástroje pro konverzi PDF dokumentu na text patří program pdftotext ze sady nástrojů Poppler utils [30] a program Ghost-

script [31]. Oba nástroje v sobě mají integrovaný prohlížeč PDF formátu a kromě textu umožňují jeho konverzi i do dalších formátů, jako je postscript nebo různé obrazové formáty. Z těchto dvou nástrojů poskytuje program pdftotext větší možnosti parametrizace, a proto jsem jej zvolil pro použití v praktické části práce.

1.5 Stanovení cíle práce

Na základě provedené analýzy stávajících řešení a požadavků vyplývajících ze zadání práce jsem stanovil následující souhrn požadavků na výsledné řešení. Odděleny jsou požadavky na nástroj pro získávání a zpracovávání dat z insolvenčního rejstříku a požadavky na aplikaci prezentující výsledné statistiky uživatelům.

1.5.1 Požadavky na nástroj pro extrakci dat

- **F1.1** Scraper bude podporovat extrakci dat z elektronických PDF formulářů vytvořených z oficiálních šablon. Podporované typy budou Přihláška pohledávky, Přehledový list, Zpráva pro oddlužení, Zpráva o plnění oddlužení, Zpráva o splnění oddlužení.
- **F1.2** Scraper bude podporovat formuláře ve verzích vyskytujících se od 1. 1. 2019, aby bylo možné přečíst většinu formulářů v období od tohoto data.
- **F1.3** Výstupem scraperu bude JSON dokument obsahující údaje nalezené v načteném PDF formuláři.
- **F1.4** Klient pro komunikaci s webovou službou insolvenčního rejstříku pro získání záznamů o insolvenčních řízeních a dokumentech v nich zveřejněných.
- **F1.5** Nástroj pro hromadné stahování PDF dokumentů publikovaných u událostí, které scraper umožňuje zpracovat.
- **F1.6** Možnost importu údajů z přečtených dokumentů do relační databáze.
- **F1.7** Konfigurace všech zmíněných funkcí prostřednictvím konfiguračního souboru.

Nefunkční požadavky

- **N1.1** Nástroje pro získávání dat budou realizovány jako konzolová aplikace spustitelná v prostředí s dostupným interpretem jazyka Python verze 3.7 a vyšší.

- **N1.2** Aplikaci půjde nainstalovat prostřednictvím nástroje pip, správce balíčků pro moduly programovacího jazyka Python.

1.5.2 Požadavky na aplikaci prezentující statistiky

Funkční požadavky

- **F2.1** Seznam nejčastějších věřitelů. Možnost seřazení seznamu dle těchto statistik: počet insolvencí, ve kterých figurují, celková výše přihlášených pohledávek, průměrná výše přihlášené pohledávky. Ze seznamu věřitelů budou vyřazeny nepodnikající fyzické osoby.

- **F2.2** Detail věřitele. V detailu věřitele budou zobrazeny statistiky o insolvenčních řízení, ve kterých věřitel figuruje – konkrétně typy těchto řízení, velikosti insolvencí a kraje ČR, ze kterých dlužníci věřitele nejčastěji pocházejí.

- **F2.3** Seznam insolvenčních správců. Možnost seřazení seznamu dle těchto statistik: počet insolvencí, velikosti insolvencí, celková odměna za všechna řízení a průměrná odměna za jedno řízení.

- **F2.4** Detail insolvenčního správce. V detailu správce budou zobrazeny statistiky o insolvenčních řízení, která spravuje – typy řízení, velikosti insolvencí (dle celkové přihlášené částky a dle počtu přihlášených pohledávek), výše popřených pohledávek. Detail správce umožní zobrazit také kraje ČR, ze kterých dlužníci nejčastěji pocházejí. Bude zobrazen výčet odměn správce za poslední skončená oddlužení.

- **F2.5** Zobrazení statistik o insolvencích pro jednotlivé kraje ČR – zobrazení počtu insolvencí, výše přihlášených pohledávek, průměrný věk dlužníka, úspěšnost oddlužení, počet zrušených oddlužení, průměrnou výši osvobození od dluhů v oddlužení a průměrné příjmy dlužníka v oddlužení.

- **F2.6** Zobrazení přehledových statistik formou vizualizací dat. Půjde o statistiky jak o insolvencích celkově, tak dle jednotlivých způsobů řešení úpadku (konkurz, reorganizace, oddlužení). Bude možné zobrazit data o počtu nových insolvencí (po letech, po měsících), nejčastějších typech osoby dlužníka, věku dlužníka, délce řízení a počtu pohledávek.

- **F2.7** Statistiky zaměřené na průběh oddlužení. V kategorii statistik oddlužení bude kromě výstupů z požadavku F2.6 možné zobrazit navíc vizualizace těchto typů: konečná a předpokládaná míra uspokojení věřitelů v oddlužení, příjmy dlužníka a majetek dlužníka.

- **F2.8** U zobrazení statistik formou vizualizace dat z požadavků F2.5, F2.6, F2.7 bude možné zobrazit detailní zobrazení zahrnující možnosti

konfigurace filtrace dat vstupujících do vizualizace (typicky výběr časového období, typu osoby dlužníka, způsob řešení úpadku a případně další konfigurace specifické ke konkrétnímu typu grafu).

Nefunkční požadavky

- **N2.1** Statistiky budou prezentovány prostřednictvím webové aplikace. Aplikace bude responsivní, aby bylo možné ji bez problému zobrazit i v mobilním zařízení.
- **N2.2** Webová aplikace bude funkční ve většině v současnosti používaných verzích internetových prohlížečů s podporou javascriptu ve verzi standardu ES6 (Firefox 54+, Chrome 51+, Edge 15+). Nemusí být implementována zpětná kompatibilita se staršími prohlížeči jako je Internet Explorer.
- **N2.3** Snadné ovládání a srozumitelnost aplikace.

1.6 Použité technologie

V této sekci budou popsány technologie a knihovny, které jsem zvolil pro implementaci systému. Důvodem volby těchto technologií byly zejména moje předchozí zkušenosti s jejich použitím v průběhu studia.

Python 3.7 Tento jazyk bude použit pro implementaci sady nástrojů pro extrakci dat z insolvenčního rejstříku. Tyto nástroje budou implementovány s podporou pro souběžnost blokujících diskových a komunikačních operací, jako je komunikace s databází, stahování dokumentů, zápis na disk. K tomu bude využita knihovna asyncio.

PostgreSQL Datová vrstva bude využívat databázový systém PostgreSQL.

encode/databases Knihovna umožňující přístup k databázi prostřednictvím asynchronního rozhraní. Knihovna sjednocuje rozhraní pro práci jak s PostgreSQL, tak s MySQL [32]. Její použití tak umožní aplikaci používat i s databází MySQL.

pallets/click Knihovna pro tvorbu uživatelského rozhraní příkazové řádky.

PHP 8 Jazyk PHP bude použit pro implementaci webové aplikace pro zobrazení získaných dat. Použita bude verze 8, která je aktuálně nejnovější.

Laravel 8 Framework Laravel 8 bude použit pro implementaci webové sekce.

robmorgan/phinx Nástroj Phinx slouží k tvorbě databázových migrací [33]. Pomocí tohoto způsobu bude definován datový model aplikace. Změny datového modelu tak půjde snadno zaznamenávat v systému správy verzí a zároveň bude možné využití jiné databáze, jako je MySQL.

Kapitola 2

Návrh

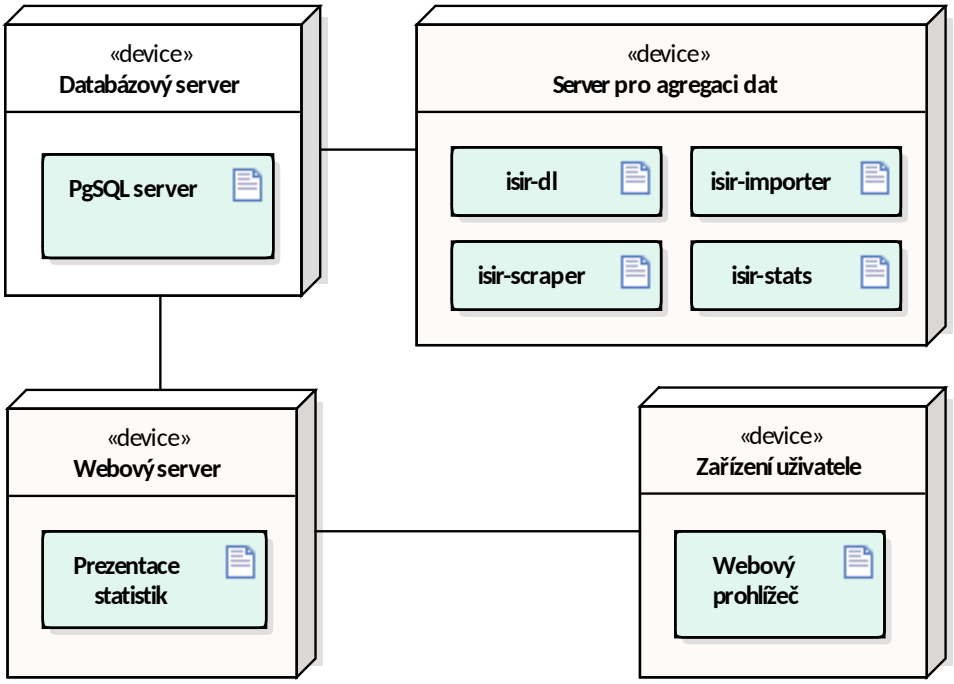
2.1 Stanovení hlavních částí aplikace

Celý systém se bude skládat ze dvou hlavních částí. První část bude tvořit sada aplikací pro extrakci dat z insolvenčního rejstříku a z PDF formulářů, následné zpracování dat, jejich import a statistické výpočty (dále pod souhrnným označením „nástroje pro agregaci dat“). Druhou částí bude webová aplikace, která bude výsledná data prezentovat uživateli. Obě části aplikace budou přistupovat ke sdílené databázi. Diagram možného nasazení systému s vyznačenými částmi aplikace je na obrázku 2.1.

2.1.1 Nástroje pro agregaci dat

Tato část aplikace bude implementována v jazyce Python. Bude rozdělena do samostatných nástrojů, viz diagram na obrázku 2.1. Tyto nástroje budou implementovány jako konzolové aplikace, jejichž návrh bude popsán dále v této kapitole. Jednotlivé nástroje budou označeny dle seznamu níže. Prefix „isir“ je zkrácené označení pro insolvenční rejstřík.

- **isir-ws** (webservice) – klient pro webovou službu insolvenčního rejstříku poskytovanou Ministerstvem spravedlnosti. Aplikace umožní stažení dostupných dat a jejich import do schématu relační databáze.
- **isir-scraper** – aplikace určená pro převod nejdůležitějších informací z PDF formulářů podporovaných typů do strukturované podoby (JSON soubor).
- **isir-importer** – aplikace pro import JSON souborů reprezentujících přečtené formuláře z nástroje isir-scraper do příslušných entit relační databáze.
- **isir-dl** (downloader) – aplikace pro hromadné stahování PDF formulářů z insolvenčního rejstříku.



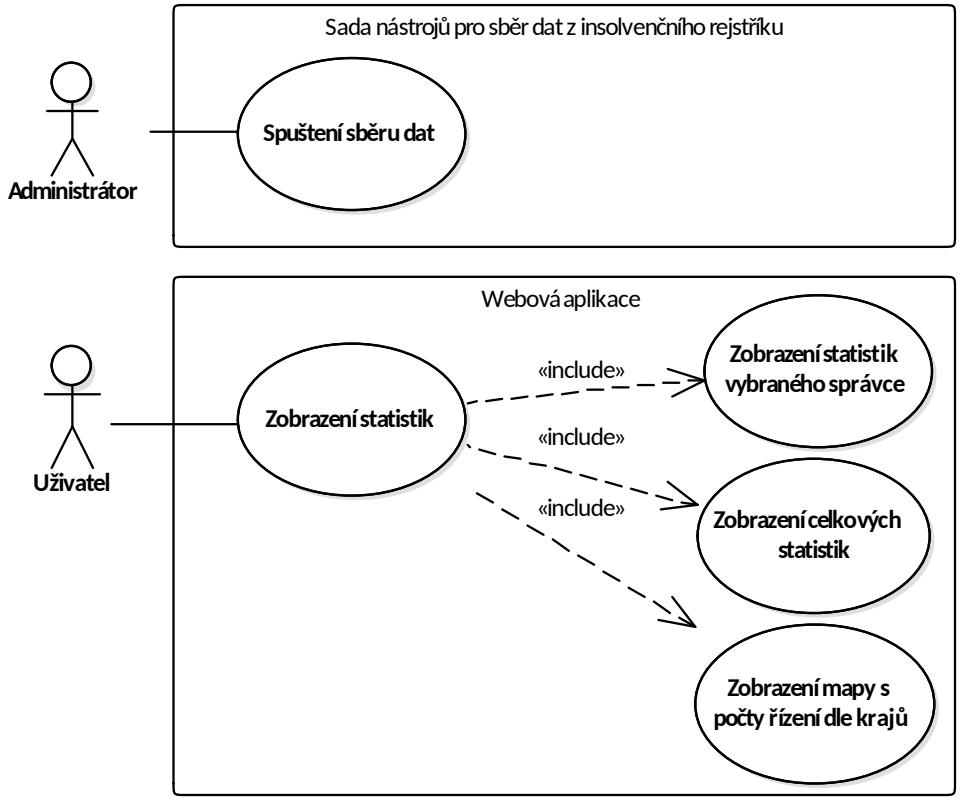
Obrázek 2.1: Diagram možného nasazení systému

- **isir-stats** – nástroj pro spuštění úloh pro zpracování dat a výpočet statistik nad databází s importovanými daty.

2.2 Uživatelé systému

Případy užití lze rozdělit dle jednotlivých částí aplikace. Diagram užití dle částí aplikace je na obrázku 2.2. Administrativní uživatel bude mít přístup k sadě nástrojů pro sběr a zpracování dat z insolvenčního rejstříku umístěných na serveru pro agregaci dat. Jeho hlavní úlohou je spuštění sběru dat za určité období, což bude zpravidla provedeno pouze při prvním nasazení aplikace. Průběžné aktualizace dat bude možné nastavit přidáním úlohy do automatického spouštěče úloh operačního systému.

Webová aplikace bude určena pro veřejnost a bude prezentovat získaná data o insolvencích. Mezi případy užití patří zobrazení detailů vybraného věřitele nebo insolvenčního správce, zobrazení celkových statistik nebo zobrazení statistik dle krajů. Kompletní seznam případů užití se odvíjí od funkčních požadavků na webovou sekci aplikace uvedených v sekci 1.5.2.



Obrázek 2.2: Diagram užití dle částí aplikace

2.3 Klient webové služby insolvenčního rejstříku

Klient webové služby insolvenčního rejstříku bude sloužit k vytvoření kopie databáze insolvenčního rejstříku v rozsahu, ve kterým jej služba zpřístupňuje. Databázové entity s daty získanými z webové služby budou mít v databázi prefix isir_ a jejich struktura je znázorněna na diagramu 2.3.

isir_osoba obsahuje údaje o všech osobách figurujících v insolvenčním řízení (dlužník, věřitelé, správce). Instance osob jsou unikátní vždy pouze v rámci konkrétního řízení – pokud se např. jeden věřitel vyskytuje u více řízení, bude evidován vícekrát.

isir_adresa je evidována zejména u osob dlužníků, což bude možné využít k tvorbě statistických map insolvencí dle sídla dlužníka. U osoby může být evidováno více adres různých typů (trvalá, přechodná, sídlo firmy, atd.) a různých časových platností.

isir_vec označuje insolvenční řízení identifikované spisovou značkou. U řízení je evidován jeho aktuální stav a podstatná data v jeho průběhu.

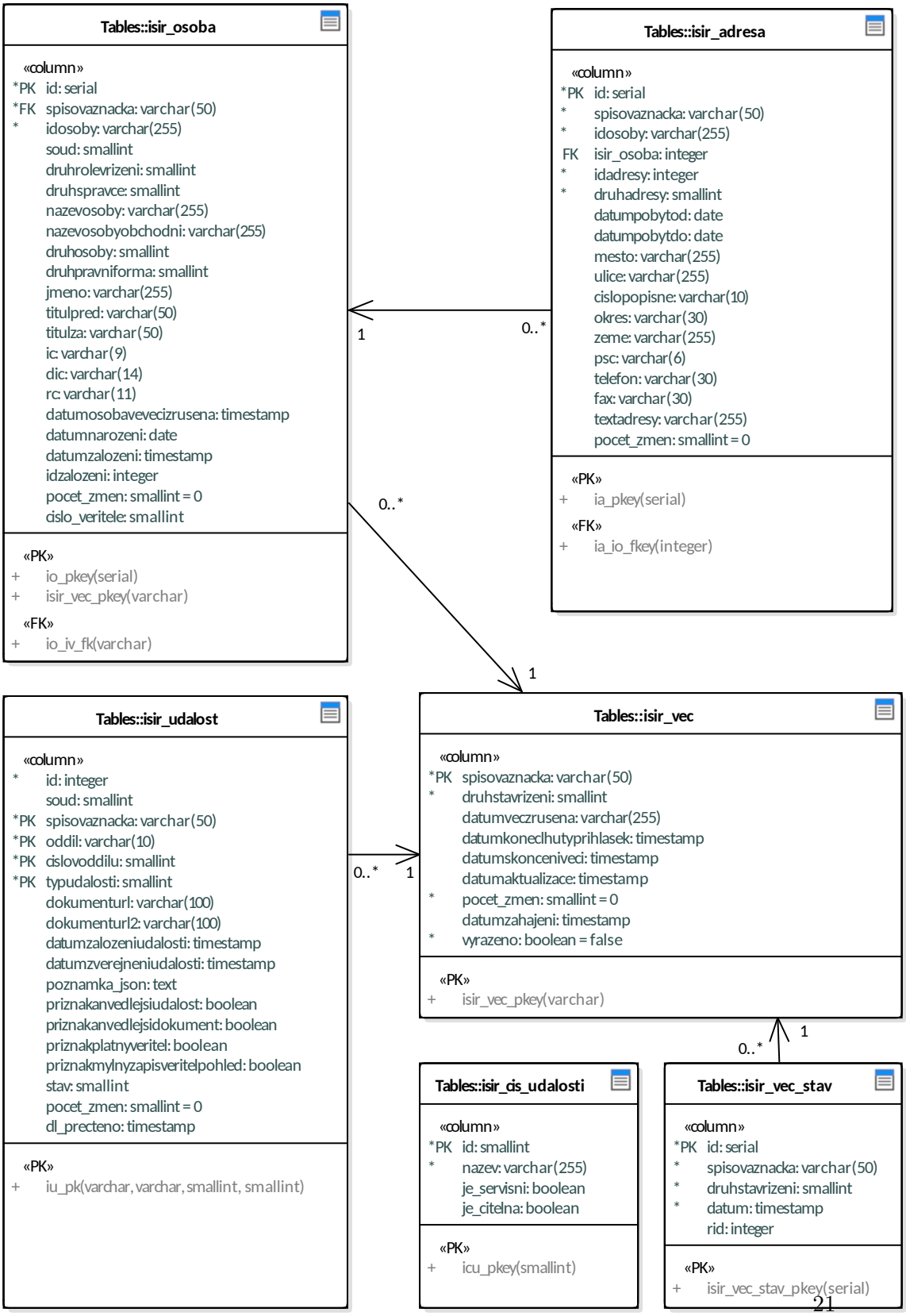
isir_vec_stav zachycuje změny stavů jednotlivých insolvenčních řízení v čase. Dokumentace webové služby definuje 14 možných stavů a specifikuje povolené přechody mezi jednotlivými stavy.

isir_udalost reprezentuje záznamy v rejstříku evidované u insolvenčních řízení. Události mohou být různých typů (např. Přihláška pohledávky, Insolvenční návrh, Zpráva, atd.). Ke každé události mohou být přiloženy až dva dokumenty (hlavní a vedlejší). Dokumenty jsou u událostí evidovány jako URL odkazy vedoucí ke stažení PDF souboru ze stránek insolvenčního rejstříku.

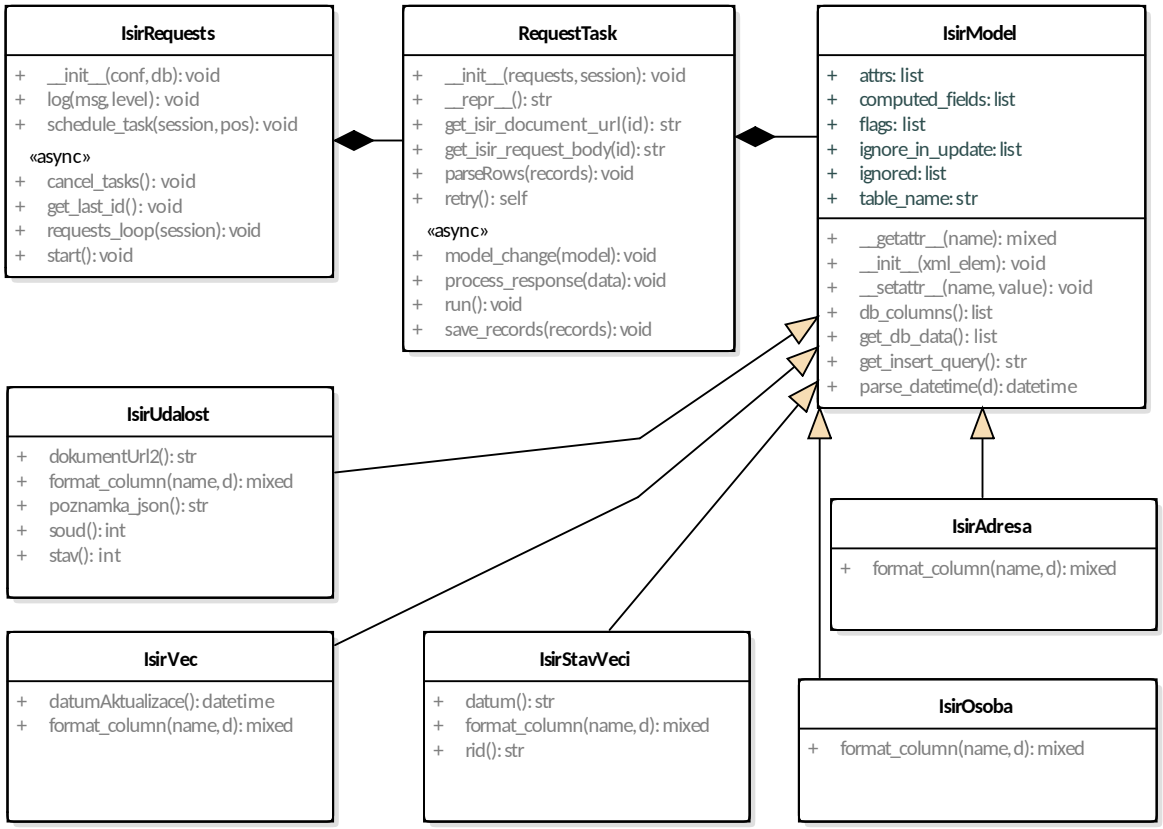
isir_cis_udalosti je číselník událostí pro převod čísla typu události na její název. Číselník je zveřejněn spolu s dokumentací webové služby a obsahuje více jak 1000 typů událostí. Do tabulky byl přidán příznak je_citelna, který bude pravdivý pro množinu událostí, pod kterými se v rejstříku zveřejňují dokumenty obsahující formuláře, které bude podporovat scraper.

Webová služba je k dipozici jako SOAP rozhraní poskytující jednu metodu s názvem `getIsirWsPublicPodnetId`, jejímž jediným parametrem je číslo události, od kterého služba vrátí 1000 následujících událostí v evidenci. Jednotlivé entity jako isir_vec, isir_osoba, isir_adresa jsou u některých událostí uloženy v atributu poznámka, kde jsou tyto entity reprezentovány ve formátu XML odpovídajícímu XSD struktuře zveřejněné spolu s dokumentací webové služby [26]. Cílem aplikace isir-ws bude volání této metody a import získaných dat do entit výše popsaného schématu. Jelikož jde o SOAP rozhraní, klienta služby by bylo možné vygenerovat ze zveřejněné WSDL specifikace, protože má však služba pouze jedinou metodu, komunikace bude zajištěna odesíláním parametrizovaného HTTP požadavku s obsahem a hlavičkami odpovídající specifikaci komunikace SOAP rozhraní, k čemuž bude využita asynchronní knihovna aiohttp.

Na obrázku 2.4 je zjednodušený třídní návrh programu isir-ws. Hlavní třídu bude představovat `IsirRequests`, jejíž instance bude vytvořena po spuštění skriptu. V rámci její inicializace dojde k připojení k databázi dle údajů v konfiguračním souboru a vytvoření aiohttp session, v rámci které budou odesílány požadavky na webovou službu. Po spuštění zároveň dojde k nalezení čísla poslední stažené události z databáze, od kterého bude zahájeno stahování událostí. Dotazy na službu budou odesílány a zpracovávány instancemi třídy `RequestTask` reprezentujícími požadavek s konkrétním číslem události. Odpovědí služby je XML obsahující zpravidla 1000 záznamů s událostmi, ze kterých budou vytvářeny instance třídy `IsirModel` v závislosti na obsahu XML struktury konkrétní události. Z každé události bude generována `IsirUdalost` a z vybraných událostí budou vytvářeny i doplňující entity jako `IsirOsoba`, `IsirAdresa` nebo `IsirVec`.



Obrázek 2.3: Databázový model pro data z webové služby ISIR



Obrázek 2.4: Zjednodušený třídní diagram návrhu modulu isir-ws

Instance `IsirModel` reprezentují databázové záznamy a realizují mapování mezi hodnotami entity v XML a hodnotami, které budou zapsány do databáze. Toto mapování zahrnuje validaci hodnot, jako je standardizace formátu časových údajů nebo převod textových konstant na číselné hodnoty dle interního číselníku pro efektivní uložení dat. Instance třídy `IsirModel` obsahují metody pro vytvoření parametrických SQL dotazů pro vložení entity daného typu. Instance třídy `RequestTask` po sestavení kolekce modelů z odpovědi webové služby iniciuje hromadné vložení této kolekce do databáze.

V rejstříku může být u událostí zveřejněn také tzv. vedlejší dokument. Vedlejší dokument může obsahovat další pomocné dokumenty, které se události týkají [26]. Tento typ dokumentů je ve webové službě evidován pomocí speciálních typů událostí majících příznak vedlejšího dokumentu a stejnou identifikaci v rámci řízení, jako hlavní událost, ke které takový dokument náleží. Vložení těchto typů událostí bude řešeno úpravou záznamu již dříve vložené hlavní události. V tabulce isir_udalost bude atribut dokumenturl reprezentovat odkaz na hlavní dokument a dokumenturl2 odkaz na vedlejší dokument.

Návrh umožňuje asynchronní zpracování více požadavků na webovou službu současně, což může být využito pro rychlejší import databáze. Toho

bude dosaženo vytvořením více instancí `RequestTask`, které jsou v prostředí asyncio zpracovávány paralelně díky využití asynchronní knihovny pro HTTP komunikaci aiohttp a přístupu k databázi databases. Instance `IsirRequest` si udržuje konstantní počet požadavků ve fázi zpracování. Počet souběžně spuštěných požadavků bude možné upravit v konfiguraci, přičemž ve výchozím nastavení bude tato funkce neaktivní, aby nedošlo k vytížení webové služby.

2.3.1 Možnosti konfigurace

Parametry klienta bude možné upravit pomocí záznamů v konfiguračním souboru a pomocí přepínačů příkazové řádky při spuštění programu. Bude umožněno nastavit tyto parametry:

- **last_id**, **min_id**, **max_id** – zadání přesného rozsahu identifikátorů událostí, které budou stahovány.
- **retry_times**, **request_timeout** – počet opakování požadavku v případě chyby komunikace a maximální doba trvání jednoho požadavku.
- **concurrency** – nastavení počtu souběžně zpracovávaných požadavků.
- **delay** – počet sekund zpoždění mezi požadavky pro snížení zatížení webové služby.

2.4 Scraper PDF formulářů

Scraper PDF formulářů bude nástroj, který bude sloužit pro převod PDF dokumentů vytvořených z oficiálních šablon do strukturovaného formátu, který umožní další zpracování dat z těchto dokumentů a jejich import do databáze. Pro výstupní formát bude použit JSON. Samotná funkcionalita importu do databáze bude oddělena do samostatného nástroje, aby scraper bylo možné využít i samostatně, např. pro analýzy dat, které nevyžadují import dokumentů do relační databáze, pro import do nerelačních databází, nebo do jiných indexovacích nástrojů. Scraper bude podporovat čtení typů formulářů zvolených v sekci 1.3.3, návrh však bude umožňovat snadné rozšíření o další typy.

2.4.1 Čtení dat z PDF formátu

Pro čtení PDF formátu bude použit program pdftotext z kolekce open source nástrojů Poppler utils, které využívají čtečku Xpdf verze 3.0 pro práci s PDF soubory [30]. Scraper bude tento nástroj využívat prostřednictvím spuštění podprocesu. Výstupem programu pdftotext je textový soubor, obsahující textové řetězce ze zadaného PDF souboru uspořádané tak, aby řádky textu ve

výstupním textovém souboru co nejpřesněji odpovídaly řádkům textu při zobrazení PDF souboru. Scraper bude využívat textovou reprezentaci formuláře k nalezení známých řetězců šablony formuláře, jako jsou nadpisy formulářových polí, což bude využito pro extrakci hodnot z příslušných formulářových polí. Program pdftotext dále umožňuje využití přepínače `-layout` pro co nejpřesnější reprodukci vizuálního rozložení textových prvků na stránce, čehož je dosaženo mnohonásobným vložením znaku mezera mezi textové řetězce reprezentující textové objekty v PDF, které jsou vizuálně na jedné řádce dále od sebe. Tato konfigurace bude využita, aby bylo možné v některých případech s větší jistotou rozlišit např. nadpisy formulářů, které jsou často od datových polí odděleny většími mezerami.

Pokud je v insolvenčním rejstříku k určité události kromě hlavního dokumentu zveřejněn i vedlejší dokument, je často publikován jako tzv. PDF portfolio. Jde o speciální funkci PDF formátu, která umožňuje do jednoho PDF souboru vložit přílohy ve formě dalších PDF souborů. V PDF standardu se tato funkce označuje jako *Portable collection* [28]. Čtení tohoto formátu často vyžaduje specializovaný nástroj, jako je Adobe Acrobat Reader, neboť open source implementace jej často nepodporují. Ani program pdftotext, resp. čtečka Xpdf, kterou využívá, jej nepodporuje. Pro podporu čtení PDF portfolií bude scraper využívat program PDFtk, který PDF portfolio rozdělí do samostatných PDF souborů, které již mohou být přečteny.

2.4.2 Dekódování textu po převodu z PDF

Při spuštění programu pdftotext na libovolný insolvenční formulář se ukazuje, že výstupem je nečitelný binární text. To souvisí i s tím, že z těchto PDF formulářů nelze při zobrazení v PDF prohlížeči kopírovat text nebo v nich vyhledávat.

Po prozkoumání interní struktury PDF souboru formuláře bylo zjištěno, že písmo, které je v souboru pro veškerý text použito, je v souboru integrováno ve formě CID písma. Vložené CID písmo (Embedded CID Font) je jeden ze způsobů reprezentace písma popsaný společností Adobe ve standardu formátu PDF [28]. Jednotlivé grafické znaky jsou v tomto způsobu uložení identifikovány pomocí jejich CID (Character ID) čísel. Vytvořením tohoto virtuálního kódování je zajištěno, že speciální znaky se zobrazí správně i v zařízeních se staršími operačními systémy, které nemají podporu pro Unicode kódování. Dalším důvodem pro tento způsob uložení písma je úspora velikosti souboru, neboť CID písmo často obsahuje pouze znaky, které jsou v dokumentu skutečně použity.

Aby bylo v PDF dokumentech používajících CID písma umožněno např. vyhledávání a kopírování textu, může interní struktura písma obsahovat ještě tzv. ToUnicode záznam, což je tabulka nebo mapování ve formátu CMap (Character map), která k CID kódům znaků mapuje odpovídající hodnoty v Unicode kódování [28]. Tato struktura však není povinná a žádný z insol-

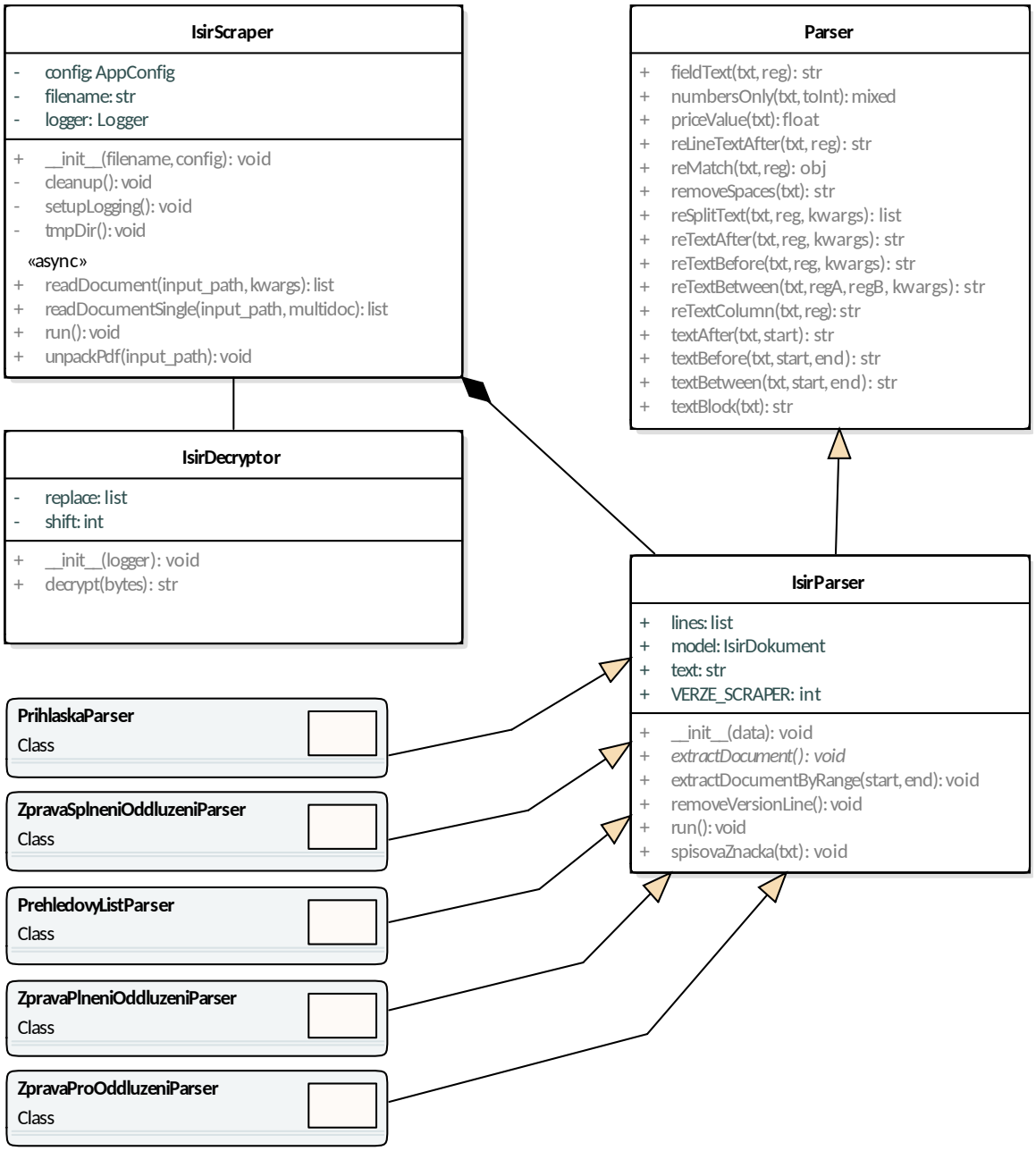
Obrázek 2.5: Tabulka CID kódování písma z PDF dokumentu formuláře

venčních formulářů generovaných z oficiálních šablon ji neobsahuje. To způsobí, že při extrakci textu z těchto souborů se místo Unicode kódů využijí CID kódy znaků ze souboru písma, a výsledný text je tak nečitelný.

Kódová tabulka CID hodnot písma získaného z jednoho formuláře pomocí nástroje pdfparser je na obrázku 2.5. Uspořádání základních znaků odpovídá uspořádání v tabulce ASCII, přičemž čísla znaků jsou o 31 nižší, než jejich kódy v ASCII. Speciální znaky české abecedy jsou v tomto příkladu v rozsahu 182-510. Zatím se nepodařilo ukázat, že sektor speciálních znaků z CID tabulky lze posuvně namapovat na některé běžně používané kódování podporující české znaky. Podstatné je, že CID čísla stejných znaků jsou stejná i mezi různými PDF formuláři a to i v případech formulářů z insolvenčních řízeních starých několik let. Scraper tak bude obsahovat modul pro dekódování textu, který bude využívat kódový posun pro základní znaky ASCII a tabulkové mapování pro speciální znaky. Protože každý dokument obsahuje pouze množinu znaků v něm se vyskytujících, mapovací tabulka speciálních znaků bude sestavena aplikací scraperu na velké množství dokumentů.

2.4.3 Návrh nástroje isir-scraper

Na obrázku 2.6 je zjednodušený třídní diagram návrhu nástroje isir-scraper. Hlavní třídou návrhu je `IsirScraper`, jejíž instance se stará o zpracování zadaného PDF souboru. Metoda `readDocument()` nejdříve zjistí, zda je zadaný soubor PDF portfolio a pokud ano, dojde k jeho rozbalení pomocí volání podprocesu nástroje pdftk. Na získané PDF dokumenty je následně aplikována metoda `readDocumentSingle()`, která voláním podprocesu pdftotext získá kódovaný textový obsah dokumentu, který je dekódován instancí třídy `IsirDecryptor` dle způsobu popsaného v sekci 2.4.2. Získaný textový obsah dokumentu je předán některé ze specializací třídy `IsirParser`, která je zodpovědná za konverzi textu do interní objektové reprezentace dle typu dokumentu. Výběr instance parseru pro konkrétní dokument je buď předem určen přepínačem při spuštění programu, nebo je využita automatická detekce dle charakteristik obsahu dokumentu. Při použití detekce se text dokumentu předá pro analýzu všem dostupným typům parseru a výsledek vrací pouze ty, prostřednictvím kterých se podaří přečíst kompletní obsah formuláře. Volání



Obrázek 2.6: Zjednodušený třídní diagram návrhu modulu isir-scraper

podprocesů budou mít asynchronní obsluhu, aby v případě souběžného zpracování více souborů najednou došlo k efektivní paralelizaci dalších blokujících činností.

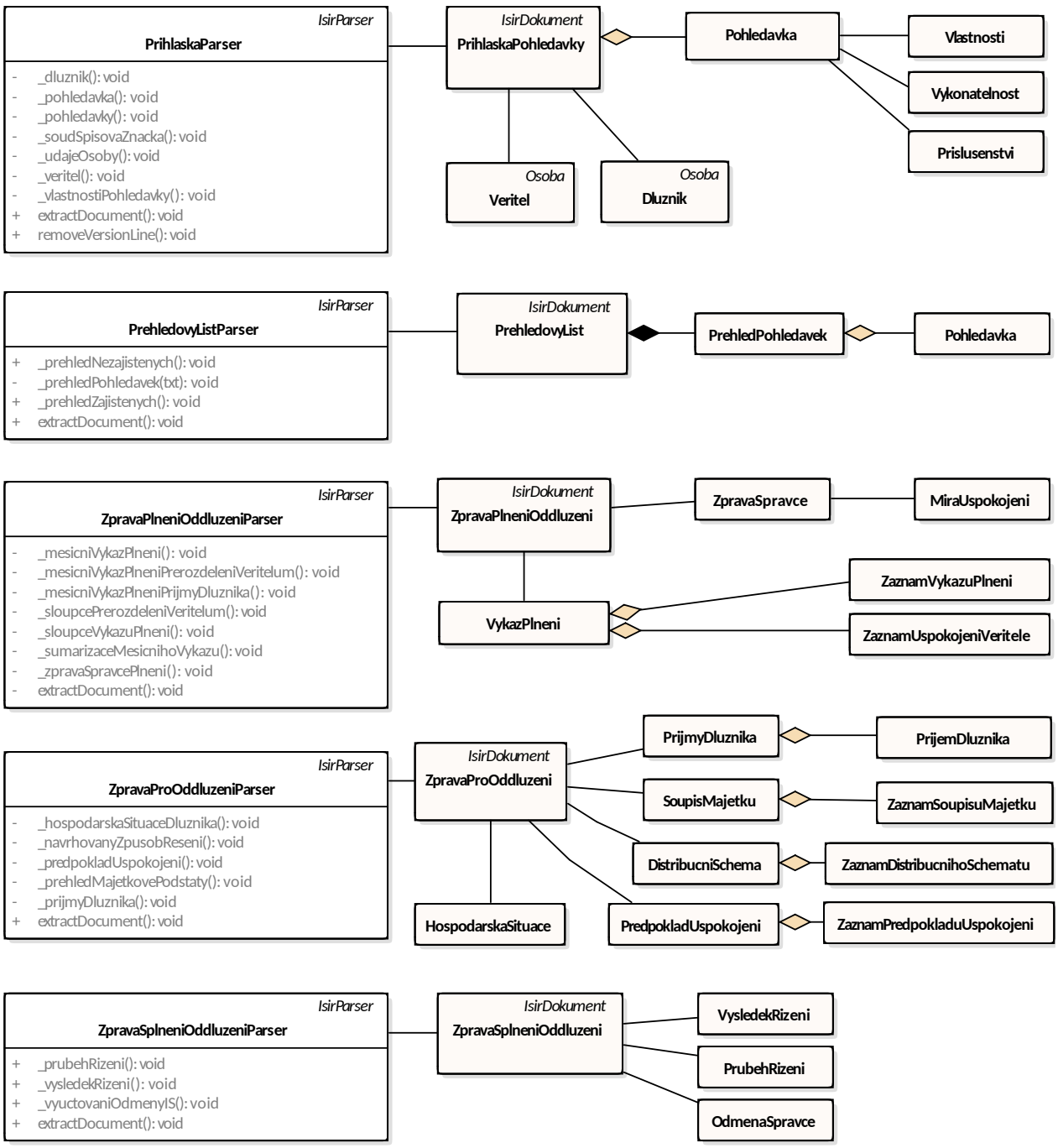
Návrh zohledňuje skutečnost, že v každém PDF dokumentu z insolvenčního rejstříku se může nacházet více formulářů. Nejedná se pouze o případy, kdy je použito PDF portfolio, jelikož i samostatné PDF soubory často obsahují více formulářů pod sebou a ty mohou být stejných i různých typů. Výsledkem metody `readDocumentSingle()` je proto pole s objektovými reprezentacemi obsahů všech formulářů, které byly v dokumentu nalezeny.

Specializace abstraktní třídy `IsirParser` budou implementovány pro každý typ insolvenčního formuláře, který scraper bude podporovat. Ve třídě `Parser` budou implementovány obecné metody pro extrakci dat z textu, které budou jednotlivé specializace využívat – např. metody pro dělení nebo ořezávání textu dle regulárních výrazů, pro extrakci číselných hodnot z formátů ve formulářových polích a jiné. Případné rozšíření scraperu o podporu dalších dokumentů by spočívalo v implementaci nové specializace třídy `IsirParser` a její abstraktní metody `extractDocument()`, která za využití metod pro zpracování textu z třídy `Parser` převede podstatný obsah ze zadaného textového souboru do objektové reprezentace. Detailně rozpracovaný návrh jednotlivých `Parser` tříd je na obrázku 2.7.

2.4.4 Výstupní datová struktura

Výstup scraperu bude prezentovat nejpodstatnější údaje z podporovaných formulářů v datové struktuře, jejíž pochopení by mělo být intuitivní a její členění by mělo usnadnit pozdější import dat do relační databáze. Parser dokumentu postupně naplní interní objektovou strukturu, která bude ve výsledku serializována na výstup. Použitý serializační formát bude JSON, podpora jiných formátů bude spočívat pouze ve vytvoření nástroje pro serializaci objektové struktury. Objektový návrh datových modelů jednotlivých parserů je zjednodušeně zachycen na diagramu 2.7. Pro datové třídy nejsou v diagramu vypsány čtené atributy, jeho cílem je zachycení vzájemných vazeb.

Součástí serializovaného výstupu s daty formuláře bude i sekce s metadaty obsahující identifikaci typu formuláře a verzi šablony formuláře, dle které byl původní PDF formulář vytvořen. Verze šablony se u většiny dokumentů nachází v záhlaví každé stránky dokumentu. Kromě verze šablony bude zaznamenána i verze třídy konkrétního parseru, který dokument přečetl. Tato informace umožní např. upravit proces importu v případě budoucí změny datové struktury výstupu nebo umožní evidovat počty dokumentů v databázi přečtených určitou verzí a případně iniciovat jejich opětovné přečtení v případě vydání nové verze parseru obsahující rozšíření rozpoznávaných polí formuláře.



Obrázek 2.7: Zjednodušený třídní diagram návrhu jednotlivých parserů a jejich datových modelů

2.4.5 Možnosti konfigurace

Nastavení scraperu bude možné upravit prostřednictvím konfiguračního souboru nebo pomocí přepínačů příkazové řádky při spuštění programu. Bude umožněno nastavit tyto parametry:

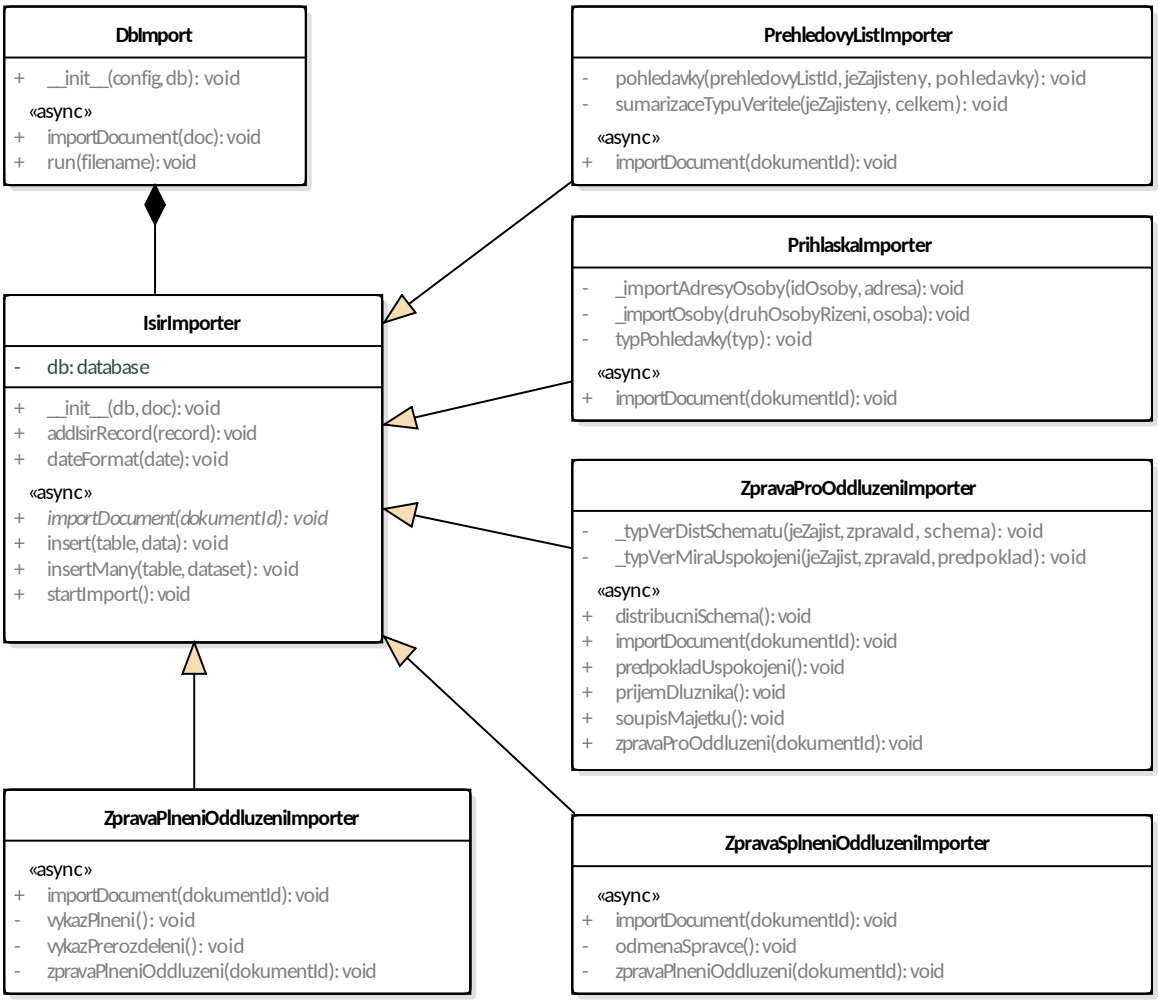
- **doctype** – určení typu formuláře pokud nemá být použita detekce typu.
- **unpack_filter** – regulární výraz pro názvy souborů z PDF portfolia, které mohou být pro čtení ignorovány. Ve vedlejších dokumentech jsou často zveřejňovány doručovací doložky, které mají jednotný prefix v názvu, pomocí něhož lze tyto PDF soubory při čtení ignorovat, neboť neobsahují žádné užitečné informace.
- **save_unreadable**, **save_text**, **save_unpacked** – možnosti aktivovat ukládání mezivýsledků procesů scraperu do datového adresáře programu. Jde o ukládání PDF souborů, ze kterých se nepodařil přečíst žádný formulář, ukládání dekódovaných textů dokumentů a ukládání výstupů z rozbalených PDF portfolií.
- **pdftotext**, **pdftk** – možnost pro nastavení jiné než výchozí cesty k umístění programů pdftotext a pdftk.

2.5 Nástroj pro import dokumentů

Strukturovaná data z formulářů získaná nástrojem isir-scraper bude možné importovat do schématu relační databáze pomocí nástroje isir-dbimport. Podstatnou částí návrhu tohoto nástroje je i návrh databázového schématu pro údaje z jednotlivých typů formulářů, který bude popsán dále v této sekci.

Na obrázku 2.8 je zjednodušený třídní diagram návrhu nástroje isir-dbimport. Hlavní třídou návrhu je `DbImport`, jejíž instance se stará o vytvoření databázového spojení, načtení vstupního souboru dat formuláře do interní objektové struktury a inicializace importní třídy dle importovaného typu dokumentu. Importní třídy budou implementovat rozhraní abstraktní třídy `DbImport` a jejich cílem bude transformace formuláře z interní objektové struktury do databázových relací. Databázové entity formuláře budou po importu obsahovat vazby pouze na úrovni entit tvořících strukturu konkrétního formuláře. Spojování na globální úrovni (např. propojení stejného věřitele vyskytujícího se v datech ve více formulářích) bude řešeno v další fázi zpracování dat.

Pro komunikaci s databází bude použita knihovna encode/databases [32], která poskytuje sjednocující rozhraní pro databáze PostgreSQL a MySQL (za využití knihoven MagicStack/asyncpg a aio-libs/aiomysql). Použitou databázi bude možné zvolit v konfiguračním souboru při zadání údajů k databázi formou URL. Jedná se o asynchronní rozhraní s funkcionalitou udržování více

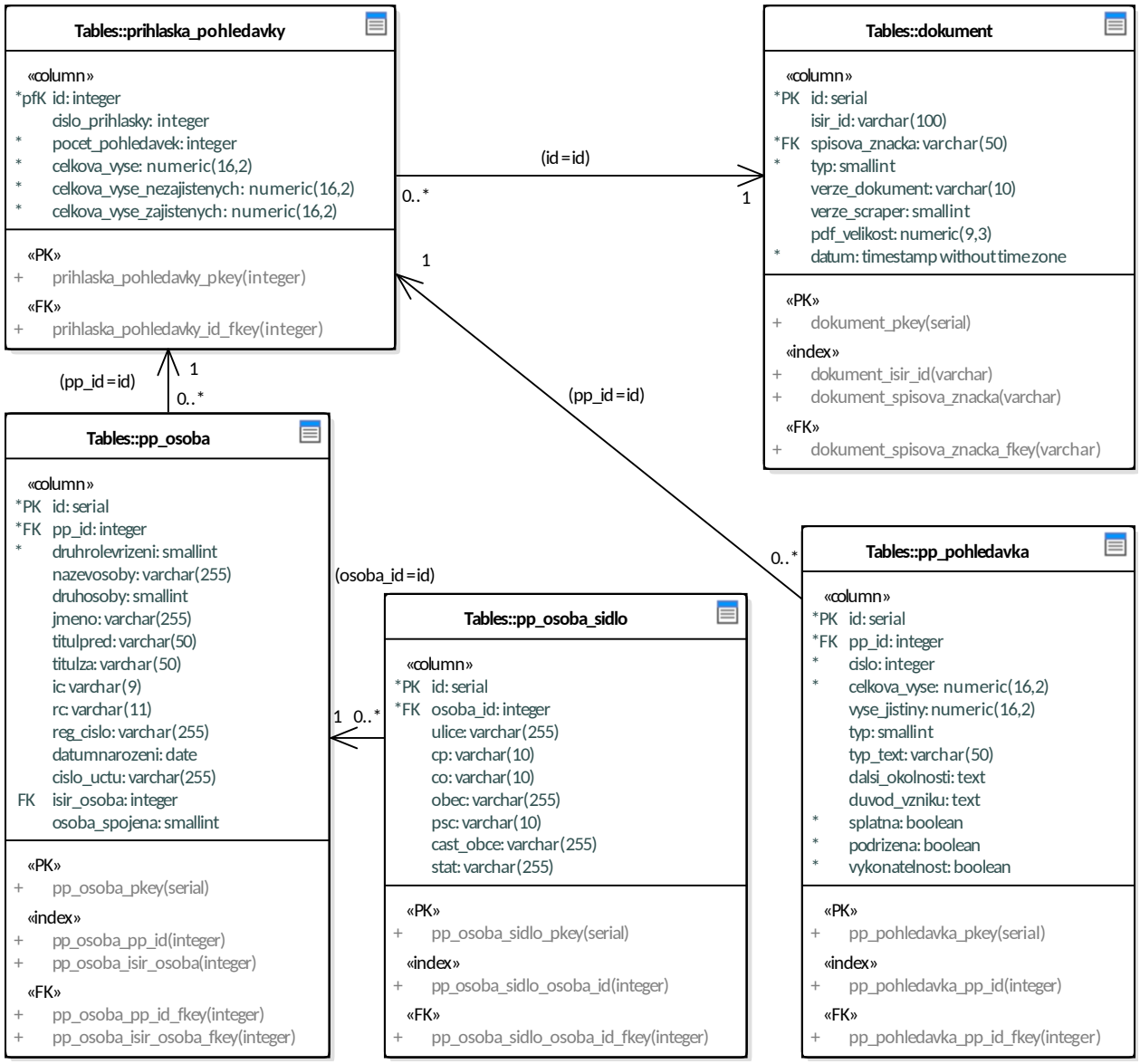


Obrázek 2.8: Zjednodušený třídní diagram návrhu modulu isir-dbimport

aktivních spojení s databází (connection pooling), která bude využita pro efektivní paralelizaci nástroje isir-dl využívající třídu `DbImport` v rámci stahování a okamžitého importu formulářů, což bude popsáno v sekci 2.6.

2.5.1 Návrh databázového schématu

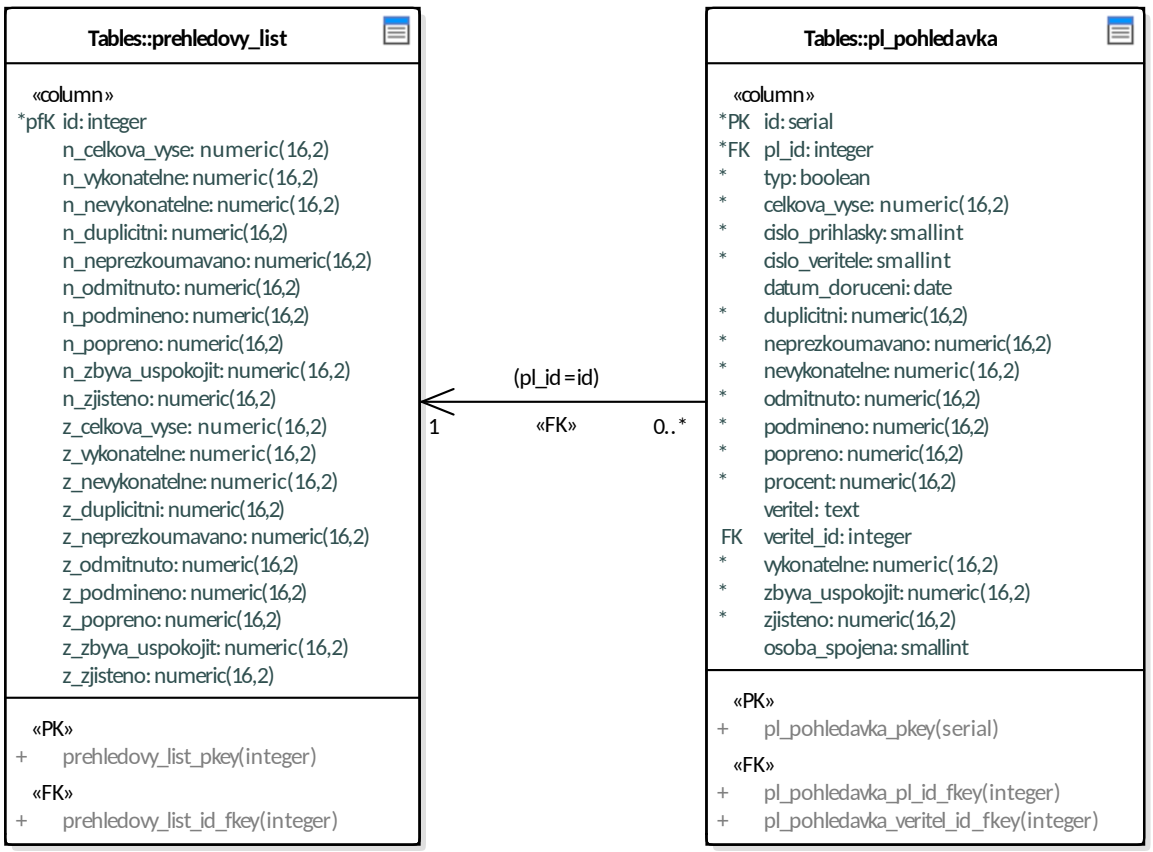
Data formulářů budou ukládána do databázových entit dle typů formuláře: `prihlaska_pohledavky`, `prehledovy_list`, `zprava_pro_oddluzeni`, `zprava_plneni_oddluzeni` a `zprava_splneni_oddluzeni`. Každá z těchto entit bude propojena s vedlejšími entitami pro zachycení datové struktury příslušného formuláře. U každého formuláře bude pomocí vazby na tabulku `dokument` evidováno, z jakého dokumentu byl přečten. U dokumentu bude specifikováno k jakému insolvenčnímu řízení (`isir_vec`) náleží, což je znázorněno na diagramu návrhu schématu přihlášky pohledávky na obr. 2.9.



Obrázek 2.9: Databázový model dokumentu typu Přihláška pohledávky

2.5.1.1 Přihláška pohledávky

Pro formulář přihlášky pohledávky budou v první řadě evidovány informace o věřiteli. Pro to budou určeny tabulky `pp_osoba` a `pp_sidlo`. Webová služba sice poskytuje výčet věřitelů figurujících v řízení, ale již neudává informace o tom, který věřitel přihlásil které pohledávky. V rámci linkování dat budou záznamy věřitelů z `pp_osoba` spojeny s existujícím záznamem věřitele v `isir_osoba`.



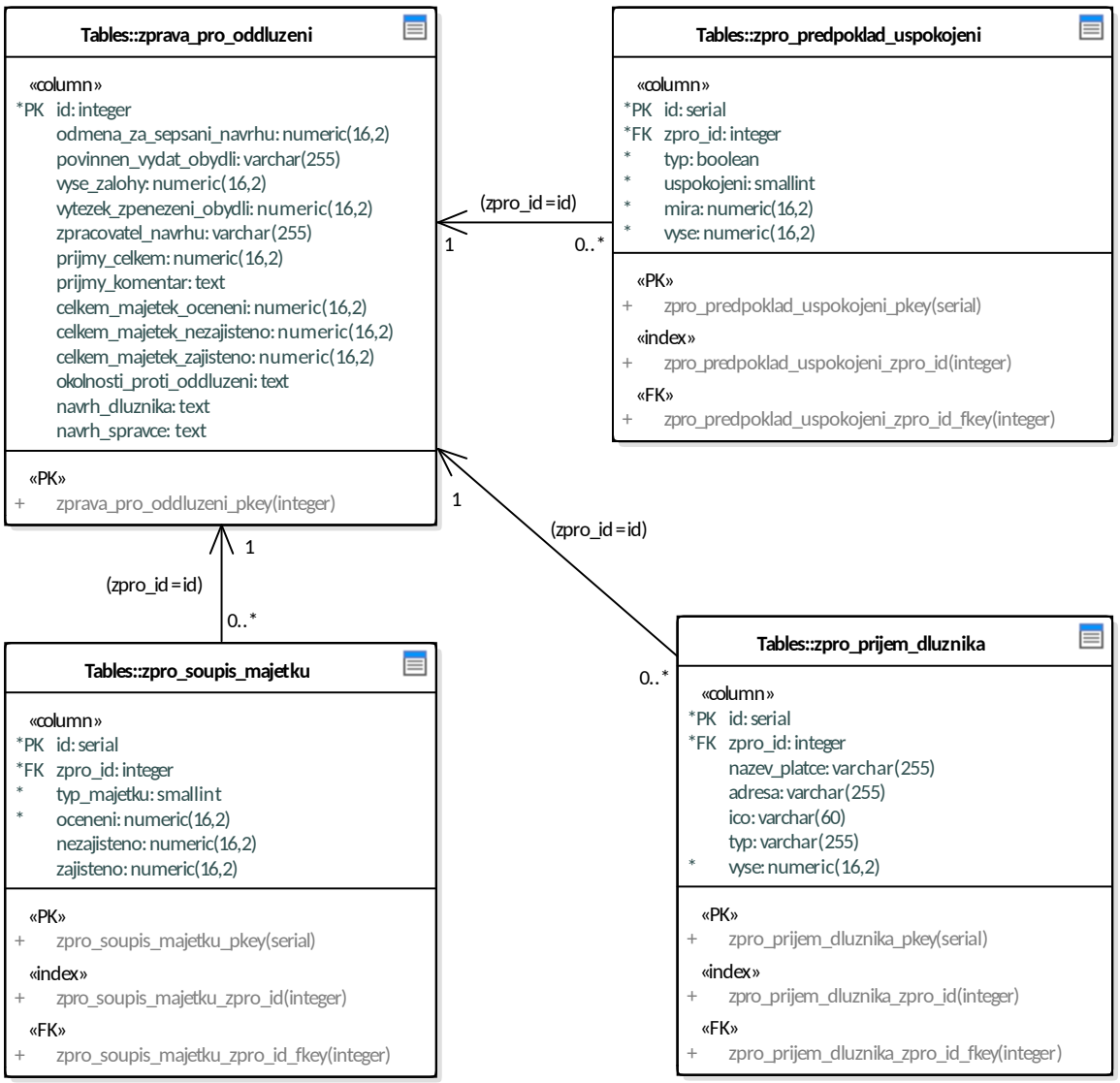
Obrázek 2.10: Databázový model dokumentu typu Přehledový list

Jedna přihláška může obsahovat větší množství pohledávek. Detaily pohledávek budou evidovány v `pp_pohledavka`. Mezi podstatné údaje patří výše pohledávky a její typ (zajištěná nebo nezajištěná), textový popis předmětu pohledávky a její vlastnosti (splatná, podřízená, vykonatelná). U přihlášky pohledávky budou dále evidovány celkové výše pohledávek a tento údaj bude dále rozdělen na výši pro zajištěné a nezajištěné věřitele. Jedná se o redundantní informaci vzhledem k samostatným údajům v `pp_pohledavka`, mohou ale být použity pro kontrolu správného přečtení jednotlivých částek z dokumentu pohledávky. Detail návrhu znázorňuje diagram na obrázku 2.9.

2.5.1.2 Přehledový list

Přehledový list obsahuje sumarizaci přihlášených pohledávek po jejich přezkumu insolvenčním správcem. Hodnoty sumarizující celkový stav přihlášených pohledávek rozdělené dle toho, zda jde o zajištěné nebo nezajištěné věřitele, budou ukládány do tabulky `prehledovy_list`.

Detaily o jednotlivých pohledávkách z přehledového listu budou evido-



Obrázek 2.11: Databázový model dokumentu typu Zpráva pro oddlužení

vány v `pl_pohledavka`. U každé pohledávky budou k dispozici např. informace o tom, zda byla popřena, odmítnuta nebo duplicitní (příp. do jaké výše) a další údaje, které se zjistí po přezkumu pohledávky. V přehledovém listu jsou jednotlivé pohledávky identifikovány pořadovým číslem přihlášky a věřitele. Pomocí těchto údajů bude možné tyto záznamy spojit se záznamem věřitele v `isir_osoba` klíčem na atributu `veritel_id`. Detail návrhu znázorňuje diagram na obrázku 2.10.

2.5.1.3 Zpráva pro oddlužení

Zpráva pro oddlužení se zveřejňuje před zahájením oddlužení. Insolvenční správce v ní shrnuje aktuální hospodářskou situaci dlužníka a uvede svoje stanovisko k tomu, zda má být oddlužení povoleno, případně jakou formou. Shrnující informace ze zprávy budou evidovány v `zprava_pro_oddluzeni`. Jde např. o hodnotu měsíčních příjmů dlužníka, hodnotu jeho majetku po ocenění a textové vyjádření správce i dlužníka.

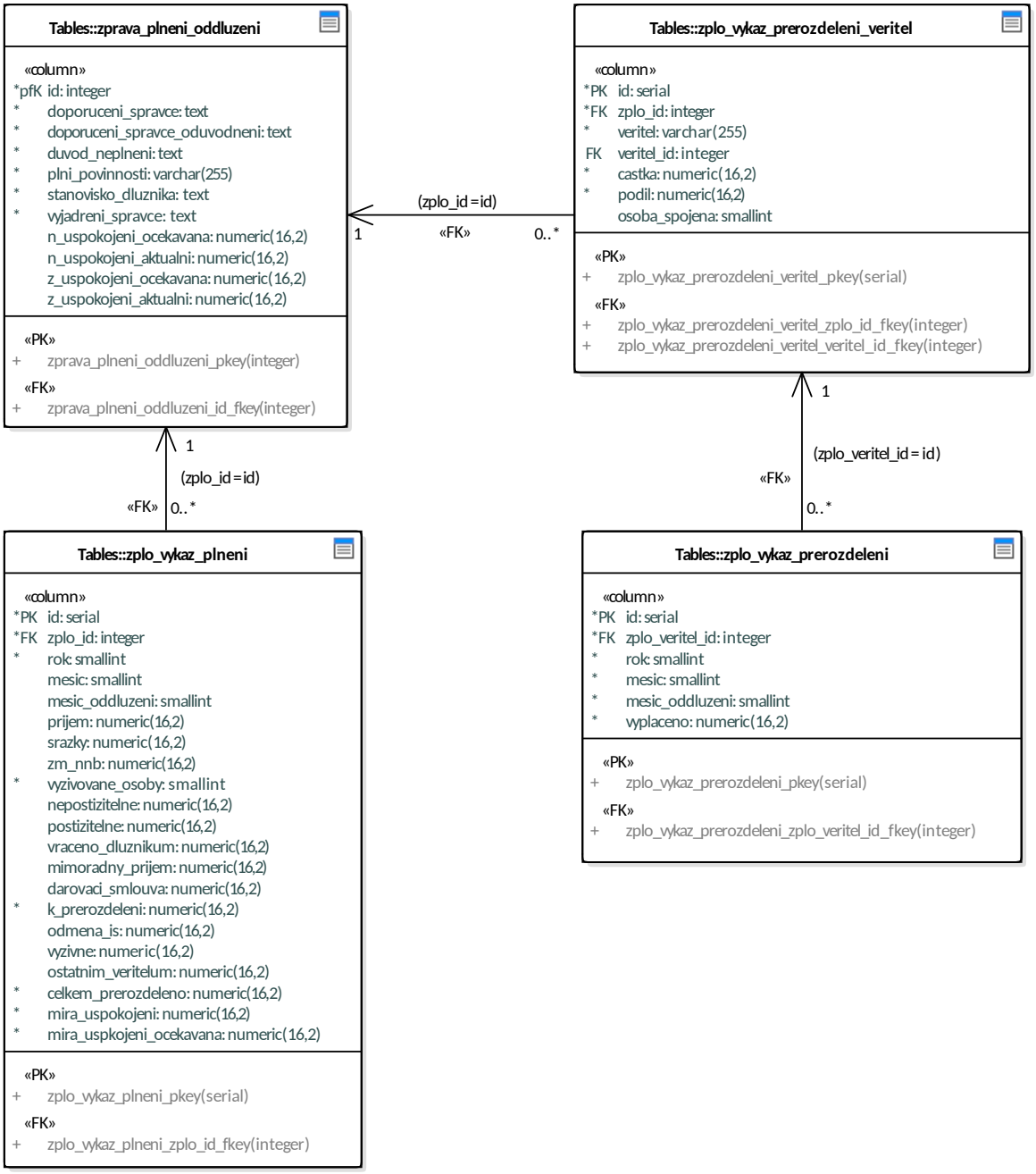
Detailní informace o příjmech dlužníka budou evidovány v tabulce `zpro_prijem_dluznika`. Ke každému příjmu bude evidována jeho výše, jeho typ (např. pracovní smlouva, renta, darovací smlouva) a identifikace subjektu poskytující tento příjem (pokud je ve formuláři uvedena). Detail dlužníkova majetku bude v tabulce `zpro_soupis_majetku`, kde bude uvedena hodnota majetku a jeho typ (např. finanční prostředky, movitý nebo nemovitý majetek). Další sekcí zprávy pro oddlužení je sekce s předpokladem uspokojení věřitelů, kde je na základě zjištěných okolností a aktuální hospodářské situace dlužníka vypočítána očekávaná míra uspokojení v průběhu oddlužení a to samostatně pro zajištěné a nezajištěné věřitele. Tyto údaje budou evidovány v `zpro_predpoklad_uspokojeni`. Detail návrhu znázorňuje diagram na obrázku 2.11.

2.5.1.4 Zpráva o plnění oddlužení

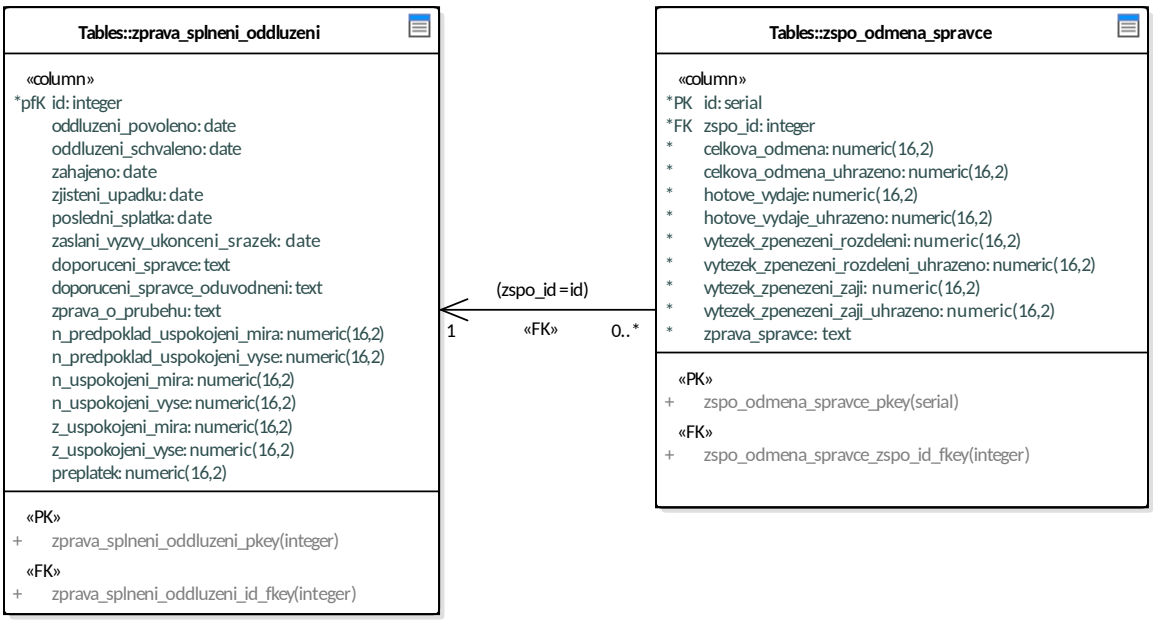
Ve zprávě o plnění oddlužení insolvenční správce shrnuje aktuální průběh oddlužení a poskytuje přehled o tom, jaké částky byly přerozděleny věřitelům a jaký podíl k uspokojení ještě zbývá. Zpráva obsahuje detailní údaje rozepsané po měsících a může obsahovat údaje za období až 6 měsíců. Shrnující informace ze zprávy budou evidovány v tabulce `zprava_plneni_oddluzeni`. Jedná se o informace jako aktuální a očekávaná míra uspokojení věřitelů, stanovisko správce, zda dlužník plní povinnosti oddlužení a doporučení správce, zda má oddlužení pokračovat.

Ve zprávě následuje výkaz plnění oddlužení popisující plnění povinností dlužníka po měsících popisovaných zprávou. Údaje pro každý z těchto měsíců budou ukládány do `zplo_vykaz_plneni`. Mezi tyto údaje patří příjem dlužníka ve zkoumaném měsíci, kolik z tohoto příjmu bylo vyhrazeno dlužníkovi v rámci životního minima, kolik bylo vyplaceno insolvenčnímu správci, kolik bylo určeno k přerozdělení věřitelům a kolik bylo skutečně přerozděleno a jak to ovlivnilo celkovou míru uspokojení.

Další sekcí ve zprávě je výkaz přerozdělení věřitelům, který pro každého věřitele vyčísluje vývoj uspokojení jeho pohledávek ve všech měsících popisovaných zprávou. Ve formuláři jsou v tabulce tohoto výkazu u věřitelů uvedeny i informace, které na měsíci nezávisí, proto bude tento výkaz reprezentován dvěmi tabulkami `zplo_vykaz_prerozdeleni_veritel` (jednotný záznam věřitele) a `zplo_vykaz_prerozdeleni` (záznamy uspokojení pohledávek věřitele



Obrázek 2.12: Databázový model dokumentu typu Zpráva o plnění oddlužení



Obrázek 2.13: Databázový model dokumentu typu Zpráva o splnění oddlužení

za určitý měsíc). U záznamů věřitelů bude opět možné spojení se záznamem v `isir_osoba` přes atribut `veritel_id`. Detail návrhu znázorňuje diagram na obrázku 2.12.

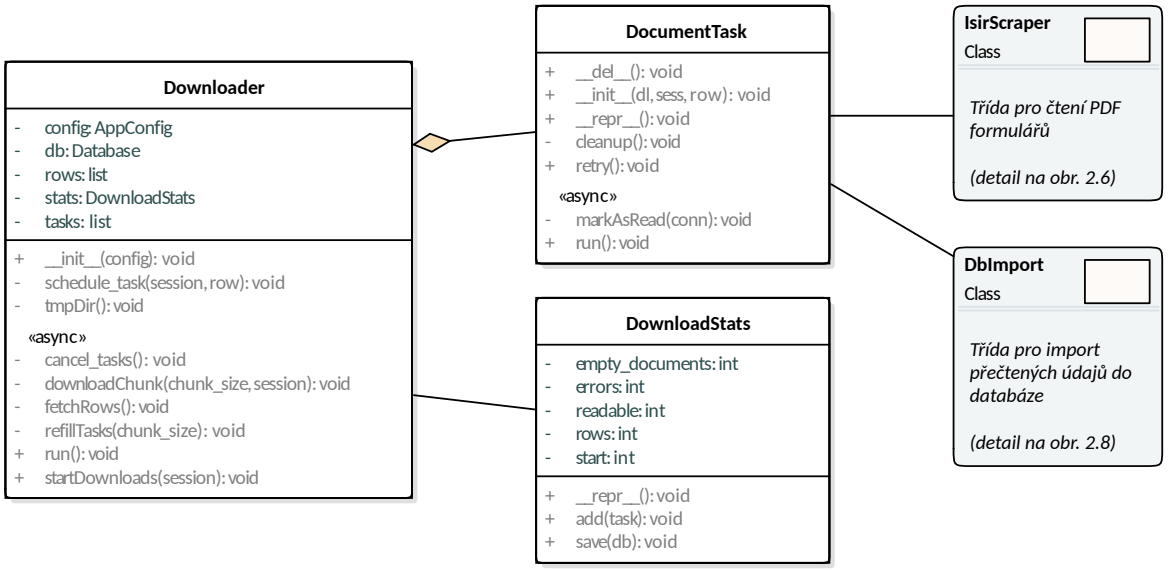
2.5.1.5 Zpráva o splnění oddlužení

Zpráva o splnění oddlužení se zveřejňuje po skončení oddlužení a insolvenční správce v ní shrnuje výsledky oddlužení. Údaje z tohoto typu formuláře budou ukládány do `zprava_splneni_oddluzeni`. Budou evidovány informace jako předpokládaná a konečná míra uspokojení věřitelů, textové shrnutí správce o průběhu oddlužení a doporučení správce o tom, zda má být dlužník zbaven zbývajících dluhů v případě, že dluh nebyl uhrazen v plné výši.

Správce dále ve zprávě vyčíslí své výdaje a vypočítá, jaká odměna mu má být přiznána soudem. Informace o vyčíslení odměny a nákladů správce budou ukládány do tabulky `zspo_odmena_spravce`. Zde budou také údaje o případném výtěžku zpeněžení dlužníkova majetku a komentář správce popisující shrnutí této sekce. Detail návrhu znázorňuje diagram na obrázku 2.13.

2.6 Nástroj pro stahování dokumentů

Program isir-dl bude umožňovat hromadné stažení PDF formulářů z insolvenčního rejstříku a jejich okamžité přečtení a import do databáze. Tímto způsobem bude možné automatizovaně naplnit databázi daty z formulářů za zadané



Obrázek 2.14: Zjednodušený třídní diagram návrhu modulu isir-dl

období. Program bude využívat návrh tříd scraperu a nástroje pro import přečtených formulářů. Využití tohoto nástroje předpokládá, že v databázi již existuje kopie databáze insolvenčního rejstříku získaná z webové služby pomocí nástroje isir-ws, neboť tabulka `isir_udalost` bude sloužit jako rejstřík existujících dokumentů, které bude program postupně stahovat a importovat.

Návrh tříd tohoto modulu je na diagramu 2.14. Hlavní třídou návrhu je `Downloader`, která si bude průběžným dotazováním databáze udržovat seznam událostí v rejstříku s dosud nepřečteným dokumentem. Dokumenty budou stahovány pouze k událostem, jejichž typ má v číselníku událostí nastaven příznak is_readable, který určuje typy, pod kterými jsou typicky zveřejňovány dokumenty s podporovanými formuláři. Pro každý dokument budou vytvářeny úlohy importu reprezentované pomocí instancí `DocumentTask`. Tato třída bude zajišťovat stažení konkrétního PDF souboru z insolvenčního rejstříku, přečtení všech formulářů v něm obsažených (za využití `IsirScraper`) a jejich import do databáze (za využití `DbImport`). Návrh umožňuje paralelní zpracování více importních úloh současně. To je zajištěno využitím asynchronních knihoven aiohttp a aiofiles pro HTTP komunikaci a pro práci se souborovým systémem v rámci návrhu nástroje isir-dl, v kombinaci s asynchronním návrhem využitých tříd `IsirScraper` a `DbImport`.

2.6.1 Možnosti konfigurace

Nastavení isir-dl bude možné upravit prostřednictvím konfiguračního souboru nebo pomocí přepínačů příkazové řádky při spuštění programu. Bude umožněno nastavit tyto parametry:

- **delay**, **delay_after** – parametry pro konfiguraci časových intervalů mezi požadavky na jednotlivé dokumenty v zájmu snížení zátěže serverů insolvenčního rejstříku rozdělením stahovaného objemu dat do delšího časového úseku. Stahování je pozastaveno na *delay* sekund po stažení *delay_after* dokumentů.
- **limit**, **start** – možnosti pro specifikaci rozsahu událostí, pro které se budou stahovat dokumenty. Nastavení *start* udává číslo počáteční události a *limit* počet událostí ke stažení.
- **keep_pdf** – možnost pro zachování stažených PDF souborů na disku. Ve výchozím stavu jsou dokumenty po jejich přečtení a importu odstraněny pro úsporu diskového prostoru.
- **request_timeout**, **retry_times** – konfigurace časového limitu na požadavek a maximální počet opakování v případě neúspěšného stahování.
- **concurrency** – stanovení počtu souběžně zpracovávaných dokumentů.

2.7 Zpracování dat a výpočet statistik

Po importu dat z přečtených formulářů do databáze bude nutné mít možnost nad databází spouštět určité úlohy pro zvýšení kvality dat. Obsaženy budou funkce pro kontrolu konzistence dat, filtry chybných údajů, funkce pro propojení entit z různých formulářů a úlohy pro výpočet statistických údajů pro prezentaci ve webové sekci aplikace. Pro tento účel bude sloužit nástroj isir-stats. Nástroj bude koncipován jako spouštěč pojmenovaných úloh, kde název úlohy bude programu předán prostřednictvím argumentu.

2.7.1 Implementované operace

V následujících sekcích jsou popsány operace nad databází, které bude nutné v rámci zpracovávání dat programem isir-stats implementovat.

2.7.1.1 Doplnění čísla přihlášky

Přihlášky pohledávek jsou v rejstříku identifikovány svým pořadovým číslem dle toho, v jakém pořadí byly doručeny na insolvenční soud. Operace pro doplnění čísla přihlášky zapíše k záznamům v tabulce `prihlaska_pohledavky` jejich číslo dle názvu oddílu události, u které byl zdrojový dokument evidován.

2.7.1.2 Propojení věřitelů v přihláškách pohledávky

Pokud jsou ve formulářích údaje vztažené k věřitelům figurujícím v řízení, bude nutné je asociovat k referenční entitě v tabulce `isir_osoba`. Propojení

věřitelů z přihlášek pohledávek bude probíhat dle informací vyplněných o konkrétním věřiteli v pořadí od nejjednoznačnějších údajů (IČ, rodné číslo) až po kombinace jména a příjmení nebo obchodního názvu.

2.7.1.3 Doplnění čísla věřitele

Své pořadové číslo mají i věřitelé. Číslo věřitele bude užitečné k asociaci záznamů distribučního schématu přerozdělení věřitelům ve zprávě o plnění oddlužení. Pro jednoznačné doplnění čísla věřitele k záznamům věřitelů v `isir_osoba` půjdou využít záznamy přehledového listu. Každý řádek v přehledovém listu obsahuje číslo věřitele a číslo přihlášky. Dle čísla přihlášky bude vyhledán záznam `prihlaska_pohledavky` a k věřiteli v `isir_osoba` přiřazenému u této přihlášky zapsáno číslo věřitele z přehledového listu.

2.7.1.4 Propojení věřitelů ze zprávy o plnění oddlužení

V tabulce distribučního schématu přerozdělení věřitelům ve zprávě o plnění oddlužení jsou věřitelé kvůli malému prostoru v tabulce často označovány zkratkovitým názvem, což by způsobovalo nepřesnosti při jejich spojování. Pro spojení proto bude použito číslo věřitele, které je v tabulce také uvedeno.

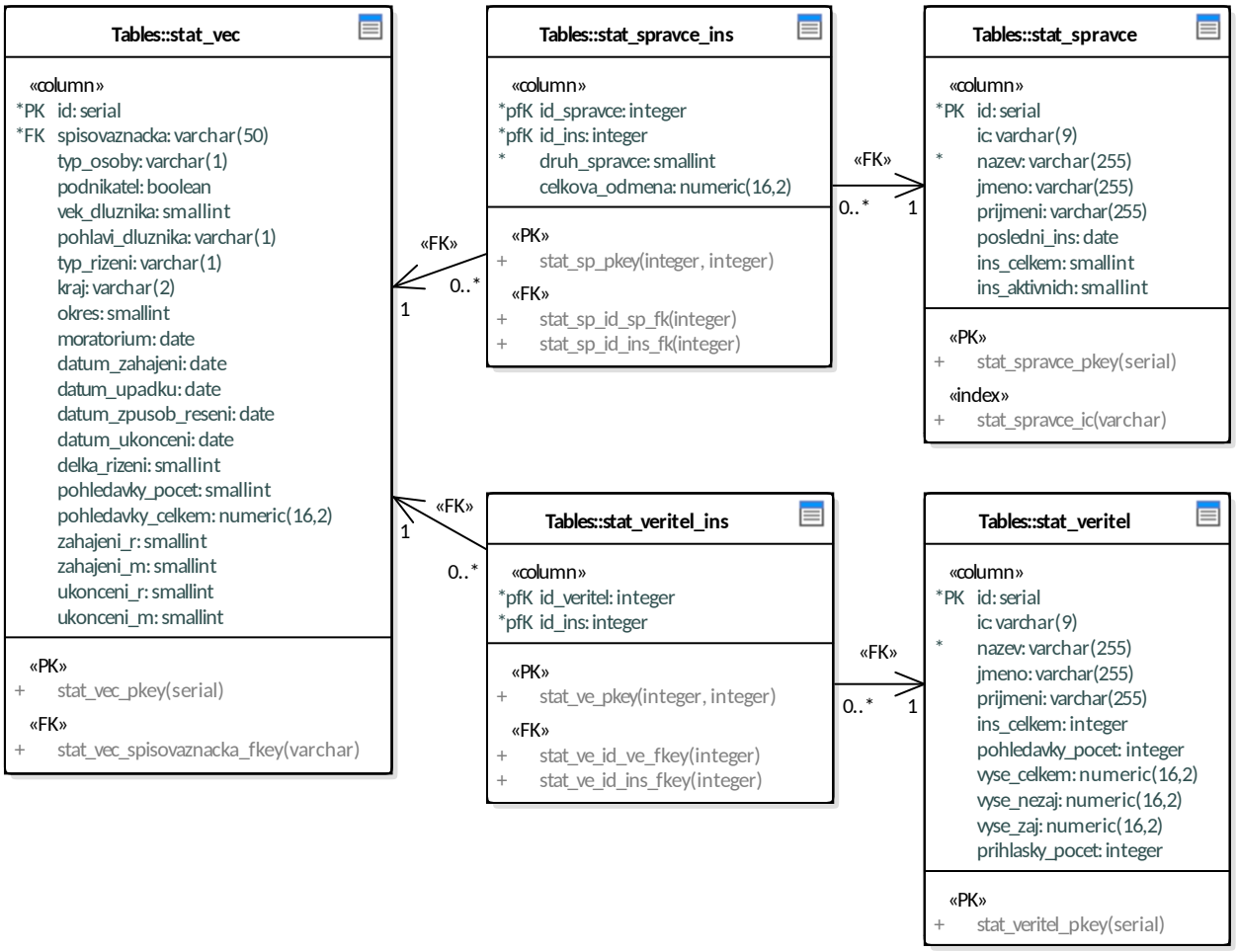
V záznamech distribučního schématu zprávy pro oddlužení jsou věřitelé identifikování také číslem věřitele, a proto bude spojení realizováno stejným způsobem.

2.7.1.5 Určení kraje a okresu řízení

Pro statistiky dle krajů ČR bude nutné každé insolvenční řízení přiřadit ke kraji. K tomu nelze využít prefix spisové značky řízení, který obsahuje signaturu krajského soudu, ve kterém je řízení projednáváno, protože platí, že ne všechny kraje mají svůj vlastní krajský soud. Pro zařazení do kraje tak bude využita trvalá adresa dlužníka nebo adresa jeho sídla nebo místa podnikání, dle toho, která informace je evidována. Pro převod adresy na kraj bude použito PSČ z adresy, které bude asociováno na okres dle veřejně dostupného číselníku poskytovaného Českou poštou [34]. Číslo okresu bude následně převedeno na kraj dle veřejně dostupného číselníku okresů ČR [35]. V případě, že PSČ nebude v adrese evidováno, bude použit alternativní způsob za využití veřejného číselníku všech obcí ČR [36].

2.7.1.6 Určení typu insolvenčního řízení

Ve statistikách bude nutné insolvenční řízení rozdělit dle způsobu řešení úpadku (zda jde o konkurz, reorganizaci nebo oddlužení). Tento údaj bude možné získat z tabulky `isir_stav`, která obsahuje seznam změn stavů u každého řízení. Možných stavů řízení je definováno 14 a dokumentace webové služby insolvenčního rejstříku také specifikuje graf povolených přechodů mezi

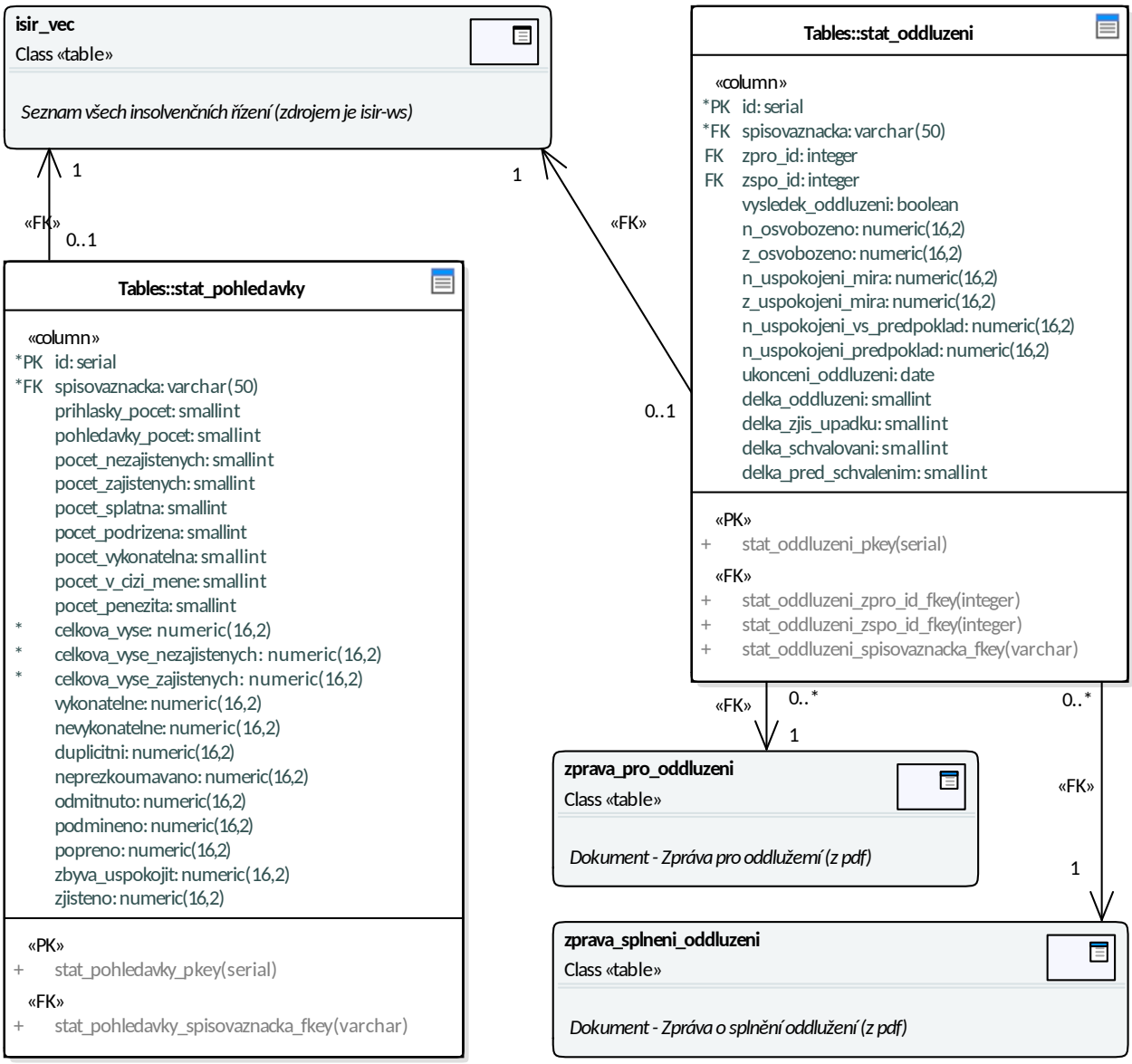


Obrázek 2.15: Databázový model pro uložení souhrnných informací o insolvenčních řízeních, správcích a věřitelích

jednotlivými stavy [26]. Insolvenční řízení může nabývat během svého trvání více forem řešení úpadku (např. v případech, kdy je zrušeno oddlužení a je vyhlášen konkurz). Pro účely kategorizace řízení dle způsobů řešení úpadku bude jako typ řešení zvolena forma, která ve stavové historii řízení nastane jako první.

2.7.2 Datový model pro uložení statistik

Datový model popsaný v předchozích kapitolách je navržen pro uložení dat z webové služby insolvenčního rejstříku a dat přečtených formulářů v originální podobě, není však vhodný pro efektivní dotazovaní statistických údajů pro webovou sekci aplikace. V tabulce `isir_osoba` jsou např. stejní věřitelé či správci evidování jako nové osoby pokaždé, co se vyskytnou v novém řízení. Dokumenty jako Zpráva pro oddlužení nebo Zpráva o splnění oddlužení



Obrázek 2.16: Databázový model pro uložení souhrnných informací o pohledávkách insolvenčního řízení a detailů řízení mající formu oddlužení

mohou být u jednoho řízení evidovány ve více verzích, ačkoliv pro dotazování informací o konkrétním řízení stačí pouze nejaktuálnější dokument od každého typu. Pro uložení statistik jednotlivých řízení a informací o unikátních věřitelích a správcích bude v datovém modelu určena množina entit s prefixem `stat_`. Data do těchto entit budou vložena nástrojem isir-stats v poslední fázi přípravy statistik.

Tabulka `stat_vec` bude obsahovat základní informace ke každému insolvenčnímu řízení. Dle adresy sídla dlužníka zde bude evidován kód kraje

a okresu řízení, dle stavových změn řízení v `isir_stav` bude určen typ řízení (způsob řešení úpadku) a podstatné časové údaje o průběhu řízení (datum zahájení, ukončení, zjištění úpadku). Detail návrhu znázorňuje diagram 2.15.

Tabulka `stat_spravce` bude obsahovat základní údaje o všech insolvenčních správcích a `stat_spravce_ins` jejich vazby na insolvenční řízení, ve kterých figurují. Program isir-stats bude sjednocovat jednotlivé záznamy správců z `isir_osoba` do jednotného záznamu v `stat_spravce` a to primárně dle IČ subjektu správce. Pokud správce nemá IČ přiděleno nebo není u záznamu v `isir_osoba` evidováno, bude pro asociaci použita kombinace jména a příjmení správce, případně obchodní název správce.

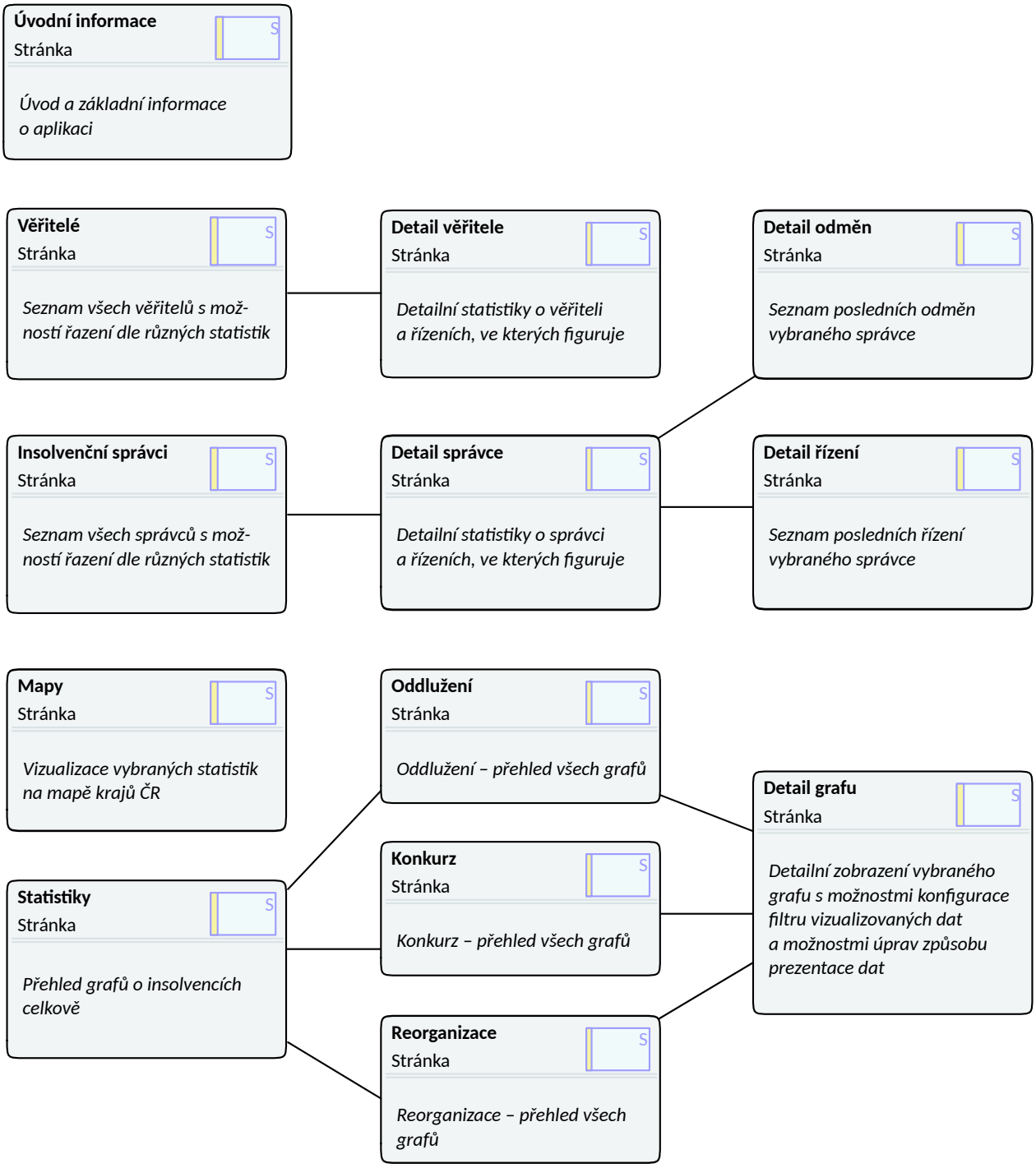
Tabulka `stat_veritel` bude obsahovat základní údaje o všech věřitelích a `stat_veritel_ins` jejich vazby na insolvenční řízení, do kterých mají přihlášené pohledávky. Při sjednocování záznamů věřitelů z `isir_osoba` do jednotných záznamů věřitelů v `stat_veritel` bude použito pouze IČ, neboť dle požadavku F2.1 nebude webová sekce zobrazovat údaje o věřitelích typu nepodnikající fyzická osoba.

Tabulka `stat_pohledavky` bude obsahovat statistiky pohledávek konkrétního insolvenčního řízení. Jde o hodnoty jako počty pohledávek dle jednotlivých typů a vlastností a celková výše pohledávek pro toto řízení. Nástroj isir-stats tyto hodnoty získá agregací údajů z přečtených dokumentů pohledávek a přehledových listů pro každé řízení. Tabulka `stat_oddluzeni` bude obsahovat detailní údaje k řízením vedeným formou oddlužení. Atributy `zpro_id` a `zspo_id` budou obsahovat odkazy na nejnovější verzi dokumentů Zpráva pro oddlužení a Zpráva o splnění oddlužení, jsou-li u daného řízení k dispozici. Detail návrhu těchto entit znázorňuje diagram na obrázku 2.16. Již dříve popsané entity jsou na tomto diagramu znázorněny v zjednodušeném zobrazení.

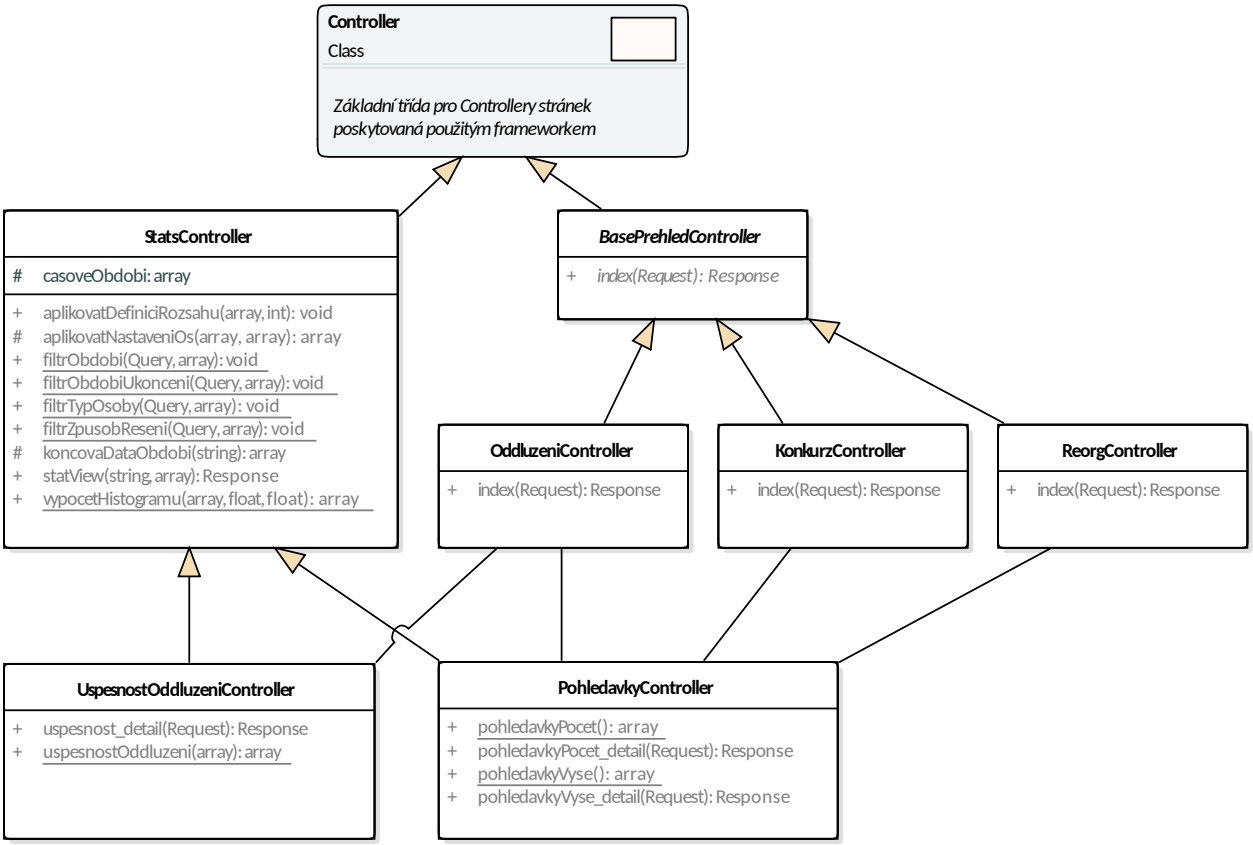
2.8 Webová aplikace

Webová aplikace bude získaná data prezentovat formou statistických přehledů a grafů. Aplikace bude přistupovat k databázi dle diagramu nasazení na obrázku 2.1 a data bude získávat primárně z entit s prefixem `stat_` obsahujících předzpracovaná data, která budou indexována pro filtraci a agregační dotazy. Aplikace bude navržena pro implementaci v jazyce PHP 8.0 za využití frameworku Laravel 8.

Diagram na obrázku 2.17 znázorňuje návrh mapy stránek webové aplikace. Stránky znázorněné na diagramu v levém sloupci reprezentují primární kategorie, které budou přístupné z hlavní nabídky. Propojení mezi stránkami v diagramu symbolizuje, že na stránce je hypertextový odkaz mezi příslušnými stránkami, který není součástí hlavní nabídky. Webová aplikace bude členěna do 4 hlavních sekcí – statistiky o nejčastějších věřitelích, statistiky insolvenčních správců, sekce s mapami s grafickým rozlišením rozdílů ve statistikách mezi kraji ČR a sekce s grafy pro jednotlivé způsoby řešení úpadku.



Obrázek 2.17: Mapa stránek webové aplikace



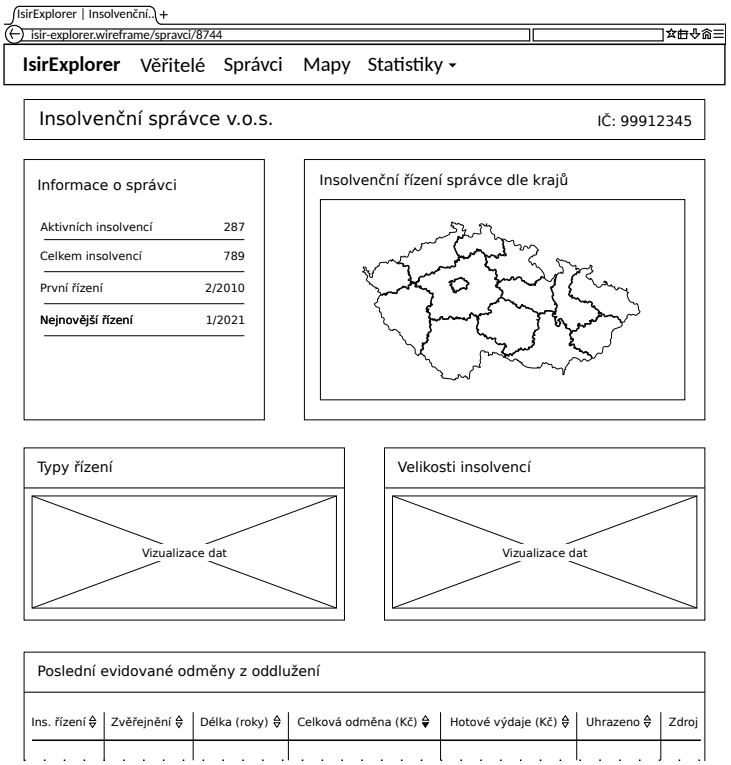
Obrázek 2.18: Zjednodušený návrh tříd pro sekci Statistiky ve webové aplikaci

2.8.1 Návrh tříd webové aplikace

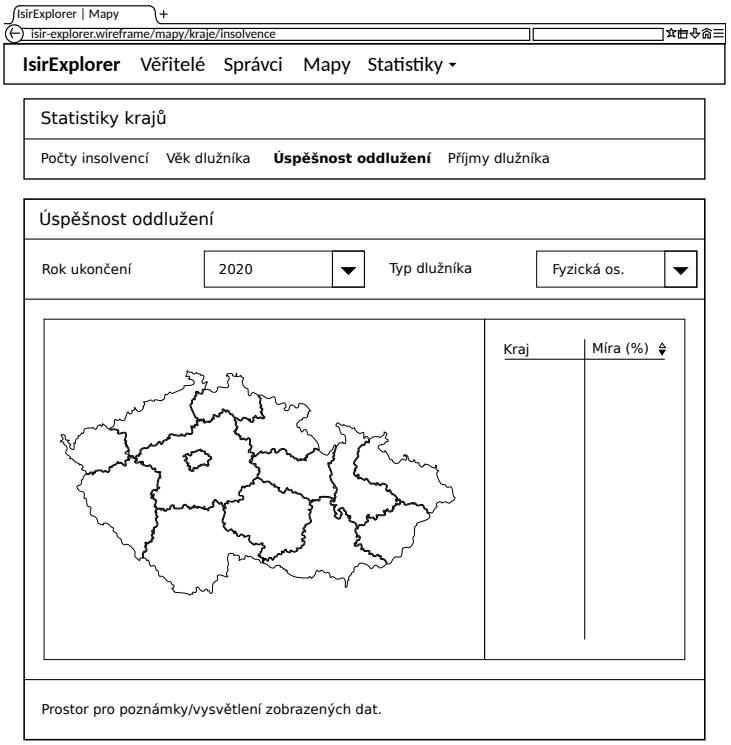
Hlavní úlohou tříd webové aplikace bude sestavení databázových dotazů dle aktuálně zobrazované statistiky a odeslání získaných dat uživateli v HTTP odpovědi společně s obsahem stránky. Návrh tříd určených pro obsluhu požadavků na stránky v sekci Statistiky znázorňuje diagram 2.18. Třída `OddluzeniController` slouží k zobrazení přehledu všech grafů o oddlužení, k čemuž využívá třídy pro jednotlivé statistické výstupy (v diagramu jen úspěšnost oddlužení a statistiky pohledávek). Tyto třídy zajišťují i zobrazení detailu, kde jsou na rozdíl od zobrazení v přehledu aplikovány i zadané filtry od uživatele.

2.8.2 Návrh uživatelského rozhraní

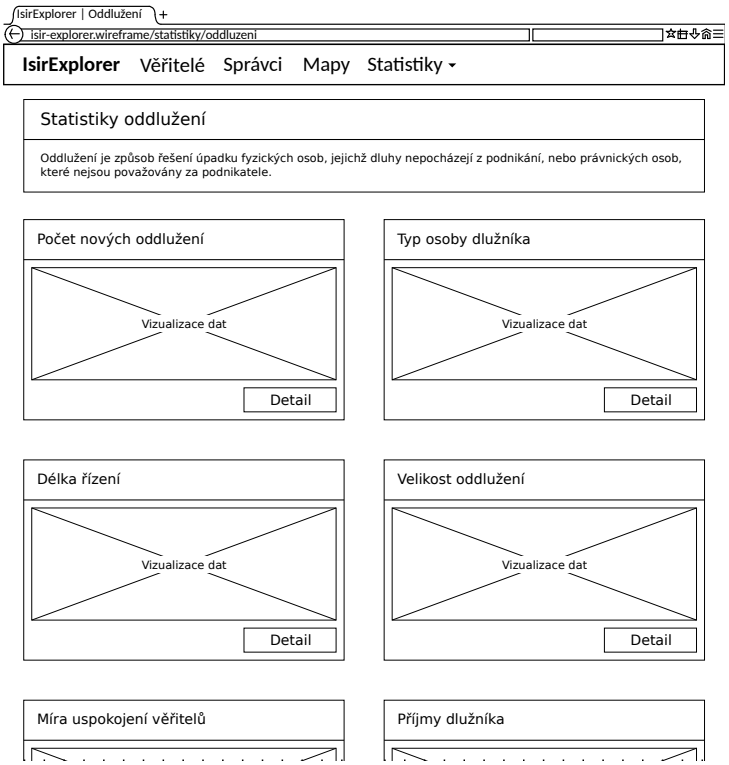
Návrh nejdůležitějších stránek webové aplikace znázorňují wireframy na obrázcích 2.19 – 2.22. Znázorněn je návrh stránky s detailem vybraného insolvenčního správce (obr. 2.19), sekce se statistikami krajů ČR (obr. 2.20), přehledová stránka se statistikami oddlužení (obr. 2.21) a detailní zobrazení vybrané vizualizace dat (obr. 2.22).



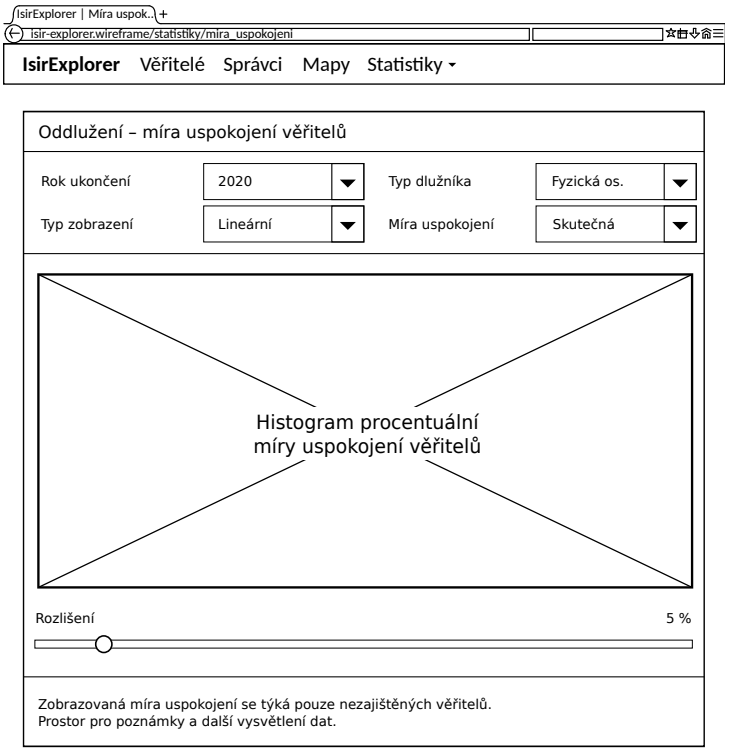
Obrázek 2.19: Wireframe – Detail správce



Obrázek 2.20: Wireframe – Mapy



Obrázek 2.21: Wireframe – Statistiky oddlužení



Obrázek 2.22: Wireframe – Detail grafu

Kapitola 3

Realizace

Implementaci aplikace jsem provedl dle návrhu popsaného v předchozí kapitole. Při implementaci jsem dbal na to, aby výsledné řešení splňovalo všechny funkční požadavky popsané v sekci 1.5. Zdrojový kód aplikace a uživatelská dokumentace jsou v příloze této práce. Vybrané aspekty realizace budou popsány v této kapitole.

3.1 Modifikace nástroje pdftotext

Typ písma použitý v PDF dokumentech insolvenčních formulářů postrádá mapování CID kódů znaků na kódy Unicode. To způsobuje, že nástroj pdftotext není schopen z takových souborů extrahovat čitelný text. Výstupem konverze je binární soubor s CID kódy znaků, který je implementovaným modulem `IsirDecryptor` programu isir-scraper následně převáděn na čitelný text. K tomu jsou využity poznatky o způsobu číslování znaků v používaném typu písma popsané v sekci 2.4.2.

Při implementaci se ukázalo, že program pdftotext není uzpůsoben pro použití na tento typ dokumentů, neboť s výstupem pracuje jako by se jednalo o úspěšně extrahovaný text v Unicode kódování. To má za následek, že některé znaky jsou z výstupu vyřazeny. Jde o znaky, které by v Unicode textu normálně odpovídaly prázdným znakům, ale jakožto CID kódy jsou to znaky, které by bylo možné ve fázi konverze výstupu převést na neprázdné Unicode znaky v textu formuláře. V tomto ohledu bylo nutné program pdftotext modifikovat, aby bylo možné z formulářů získat kompletní text.

3.1.1 Oprava chybějících znaků

Nástroj pdftotext je implementován v jazyce C++ a má otevřený zdrojový kód. Bylo tak možné vytvořit jeho upravenou verzi pro použití specificky na insolvenční formuláře. Velká část tohoto nástroje je implementována v souboru TextOutputDev.cc, což je sada tříd implementující rozhraní pro vykreslování

```
bool UnicodeIsWhitespaceSimple(Unicode ucs4) {
  static Unicode const spaces[] = {
    // 0x0009, // horizontal tab   => 0x28 = (
    // 0x000A, // NL new line      => 0x29 = )
    // 0x000B, // VT vertical tab  => 0x2A = *
    // 0x000C, // NP new page      => 0x2B = +
    // 0x000D, // CR carriage ret. => 0x2C = ,
    // 0x0020, // space            => 0x3F = ?
    0x0085, 0x00A0, 0x2000, 0x2001, 0x2002, 0x2003, 0x2004,
    0x2005, 0x2006, 0x2007, 0x2008, 0x2009, 0x200A, 0x2028,
    0x2029, 0x202F, 0x205F, 0x3000 };
  Unicode const *end = spaces + sizeof(spaces) / sizeof(spaces[0]);
  Unicode const *i = std::lower_bound(spaces, end, ucs4);
  return (i != end && *i == ucs4);
}
```

Ukázka 3.1: Úprava funkce určující prázdné Unicode znaky

PDF v textovém režimu. Ve funkci `TextPage::addChar()` dochází ke členění znaků z textových objektů PDF souboru do slov, která jsou později dle jejich pozice zapisována na výstup. Tato funkce používá pro rozlišení oddělovačů mezi slovy funkci `UnicodeIsWhitespace(Unicode ucs4)` ze souboru UTF.cc. Tato funkce je příčinou ztráty informace při použití na insolvenční formuláře, protože některé významově užitečné CID hodnoty považuje za prázdné znaky. Tuto funkci jsem proto nahradil za `UnicodeIsWhitespaceSimple(Unicode ucs4)`, která vybrané znaky ponechá v textu. Tato funkce je na ukázce kódu 3.1. V komentáři je seznam kódů původních prázdných znaků, jejich název dle ASCII a také, jakým znakům tyto kódy odpovídají po převodu CID kódu modulem `IsirDecryptor`.

3.1.2 Eliminace znaků s duplicitním významem

Další problém nastal v případě znaku pro nový řádek (kód znaku 0x0A). Program pdftotext vkládá do výstupního textu odřádkování tak, aby řádky ve výsledném textovém souboru co nejpřesněji odpovídaly řádkům textu při zobrazení PDF. V případě insolvenčních formulářů odpovídá CID kód 0x0A znaku reprezentující uzavírací závorku. Na výstupu programu pdftotext měl tedy kód 0x0A duplicitní význam a z pohledu modulu `IsirDecryptor` nebylo možné rozlišit, kdy který význam použít.

Tato duplicita byla řešena další úpravou ve funkci `TextPage::addChar()`, kde je znak 0x0A obsažený v textových objektech formuláře (tj. znak závorky) nahrazen novým kódem 0x7F (znak DEL, který se v běžném textu nevysky-

tuje). Při zpracování textu v modulu `IsirDecryptor` se tento znak nahradí za závorku, a eliminuje se tak duplicita významu se znakem odřádkování.

3.1.3 Filtrace dle typu písma

Program pdftotext byl dále upraven tak, aby při extrakci textu z PDF byly ignorovány všechny znaky, jejichž písmo nenáleží do rodiny písem Myriad, která je použita v šablonách insolvenčních formulářů. Cílem této změny je zajistit, aby se na výstupu nevyskytovaly části textu v Unicode kódování kombinované s textem kódovaným pomocí CID kódů. Části textu v jiných písmech neobsahují žádné informace užitečné pro isir-scraper, protože se vyskytují mimo formulář, a nebyly by tak přečteny do výsledné datové struktury.

V některých insolvenčních formulářích je do záhlaví stránek vkládána insolvenční značka řízení nebo jiná signatura spisu v nestandardním formátu. Toto záhlaví není součástí šablony formuláře, a je tak pravděpodobné, že ji tam vkládá některý z podpůrných programů určených pro usnadnění agendy insolvenčních správců. Záhlaví je vkládáno rozdílným písmem od obsahu formuláře, a bylo jej tak možné tímto způsobem odstranit. To přispělo k nápravě chyb programu isir-scraper při detekci formulářových polí přesahující rozhraní dvou stránek, kde neočekávaný formát záhlaví mohl způsobovat chybnou kategorizaci obsahu polí.

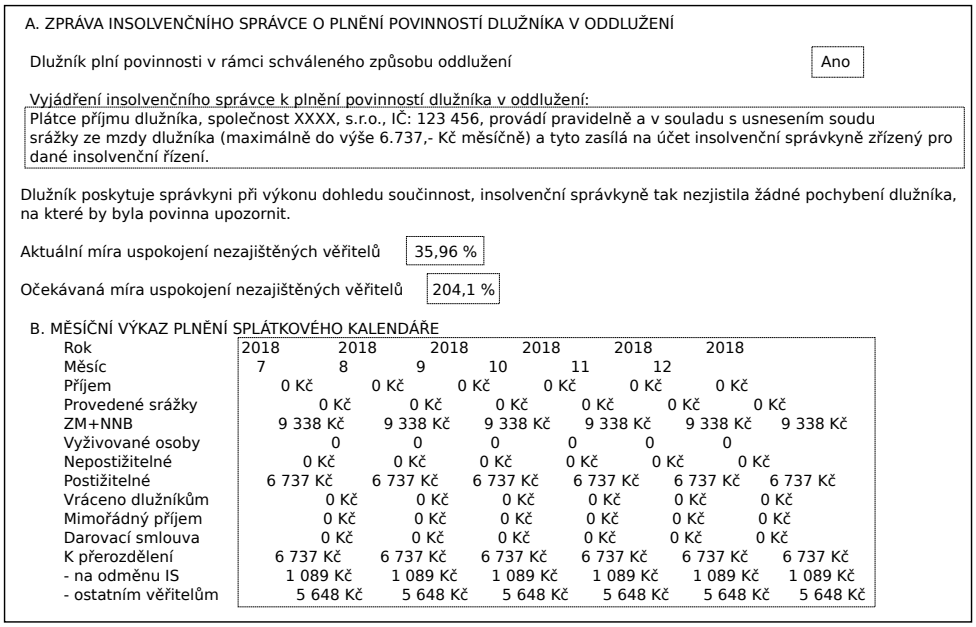
3.1.4 Poznámky k úpravám

Modifikovaná verze programu pdftotext byla zveřejněna spolu s kódem projektu na GitHubu. Pro správné použití programu isir-scraper si uživatel bude muset tuto verzi zkompilovat a nastavit cestu ke zkompilované verzi v konfiguračním souboru programu isir-scraper.

Funkcionalitu modulu `IsirDecryptor` by bylo možné v budoucnu implementovat přímo v modifikované verzi programu pdftotext. Tento způsob zatím nebyl zvolen, neboť převodní tabulka CID znaků na Unicode není známa pro všechny speciální znaky, a tak může být potřeba ji průběžně rozšiřovat, když se v dokumentech objeví speciální znaky, jejichž kódy dosud nejsou známy. Pro uživatele bude přívětivější, když budou případné aktualizace vydávány spolu s aktualizacemi kódu programu isir-scraper, a uživatel tak nebude muset při každé aktualizaci znovu spouštět kompilaci modifikovaného nástroje pdftotext.

3.2 Čtení údajů z PDF formulářů

Pro extrakci údajů formulářových polí z výstupu programu pdftotext byly využívány zejména konstantní textové řetězce obsažené ve formulářové šabloně, jako jsou názvy textových polí, vysvětlující texty formuláře nebo názvy sekcí. Text je při zpracování nejdříve rozdělen na menší celky dle sekcí v závislosti



A. ZPRÁVA INSOLVENČNÍHO SPRÁVCE O PLNĚNÍ POVINNOSTÍ DLUŽNÍKA V ODDLUŽENÍ

Dlužník plní povinnosti v rámci schváleného způsobu oddlužení: Ano

Vyjádření insolvenčního správce k plnění povinností dlužníka v oddlužení:

Plátce příjmu dlužníka, společnost XXXX, s.r.o., IČ: 123 456, provádí pravidelně a v souladu s usnesením soudu srážky ze mzdy dlužníka (maximálně do výše 6.737,- Kč měsíčně) a tyto zasílá na účet insolvenční správkyně zřízený pro dané insolvenční řízení.

Dlužník poskytuje správkyni při výkonu dohledu součinnost, insolvenční správkyně tak nezjistila žádné pochybení dlužníka, na které by byla povinna upozornit.

Aktuální míra uspokojení nezajištěných věřitelů: 35,96 %

Očekávaná míra uspokojení nezajištěných věřitelů: 204,1 %

B. MĚSÍČNÍ VÝKAZ PLNĚNÍ SPLÁTKOVÉHO KALENDÁŘE

Rok	2018	2018	2018	2018	2018	2018
Měsíc	7	8	9	10	11	12
Příjem	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Provedené srážky	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
ZM+NNB	9 338 Kč	9 338 Kč	9 338 Kč	9 338 Kč	9 338 Kč	9 338 Kč
Vyživované osoby	0	0	0	0	0	0
Nepostižitelné	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Postižitelné	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč
Vráceno dlužníkům	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Mimořádný příjem	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
Darovací smlouva	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč	0 Kč
K přerozdělení	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč	6 737 Kč
- na odměnu IS	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč	1 089 Kč
- ostatním věřitelům	5 648 Kč	5 648 Kč	5 648 Kč	5 648 Kč	5 648 Kč	5 648 Kč

Obrázek 3.1: Příklad výstupu pdftotext s vyznačením čtených údajů

na typu formuláře (např. kapitoly či údaje o jednotlivých pohledávkách). Ze sekcí jsou následně dle jejich obsahu extrahovány údaje pomocí běžných metod pro zpracování textu (vyhledávání v textu, aplikace regulárního výrazu, dělení řádku dle textu nebo regulárního výrazu atd.). Příklad dekódovaného výstupu programu pdftotext pro část formuláře typu Zpráva o plnění oddlužení je na obrázku 3.1. Ohraničením jsou na obrázku vyznačeny čtené formulářové údaje různých typů (jednořádková hodnota, víceřádkový text, tabulka).

3.2.1 Typy formulářových polí

Přístup k extrakci údajů se liší dle typu a grafického rozložení konkrétního formulářového pole. Základní typy formulářových polí dle způsobu implementace jejich čtení je možné rozdělit do těchto skupin:

- Jednořádkové textové pole.
- Víceřádkové textové pole zaujímající celou šíři dokumentu.
- Víceřádkové textové pole s horizontálním omezením šíře obsahu.
- Tabulka konstantní velikosti.
- Tabulka s předem neurčeným počtem sloupců a řádků.
- Zaškrtávací pole s pravdivostní hodnotou ano/ne.

Typ formuláře	Počet importovaných	Verze formulářových šablon
Přihláška pohledávky	1 162 171	5-a (47,91 %), 3-h (11,66 %), 4-a (11,16 %), 3-g (11,05 %), 3-e (6,90 %), 3-f (5,11 %), 3-d (4,47 %), 3-c (1,72 %), 3-b (0,01 %)
Přehledový list	60 997	2-b (98,51 %), 2-a (1,16 %), 1-b (0,28 %), 1-a (0,04 %), 2-d (0,01 %), 2-c (0,01 %)
Zpráva pro oddlužení	65 257	2-b (57,66 %), 3-a (39,44 %), 2-a (2,21 %), 1-a (0,37 %), 1-b (0,30 %), 2-d (0,06 %)
Zpráva o plnění oddlužení	133 519	2-b (98,86 %), 2-a (0,81 %), 1-b (0,26 %), 1-a (0,06 %)
Zpráva o splnění oddlužení	58 270	2-b (94,17 %), 2-a (2,39 %), 1-b (1,90 %), 1-a (1,53 %)

Tabulka 3.1: Počet úspěšně importovaných formulářů za období od 1. 1. 2016 do 31. 12. 2020 a podíl výskytu verzí formulářových šablon

Při implementaci bylo dbáno na ošetření i méně častých situací, jako např. že jeden PDF dokument může obsahovat více formulářů stejných typů, nebo že formulářové údaje přesahující rozhraní dvou stran v PDF mohou být odděleny více řádky či mohou být proloženy záhlavím stránky obsahujícím verzi formuláře.

3.2.2 Různé verze formulářů

Šablony formulářů se s postupem času vyvíjely, a scraper proto musí být připraven i na různé verze formulářů. Pozitivní zjištění bylo, že jednotlivé verze neobsahují zásadní změny v rozložení již existujících formulářových prvků. Rozdíly mezi verzemi ve většině případů spočívají v přidání nových polí, opravě chyb v popisných textech nebo změně uspořádání obsahu některých sekcí, což někdy může vést k přečíslování již očíslovaných polí formuláře.

Chybějící pole ve starších verzích formuláře nepředstavují pro scraper problém, neboť implementovaná metodika čtení polí je taková, že pokud se hledané pole nepodaří najít, nebo jeho obsah neodpovídá očekávanému formátu, je toto pole ze čtení vyřazeno a na výstupu je jeho hodnota NULL. Přizpůsobení různým popisným textům mezi verzemi většinou spočívalo v přidání alternativy do vyhledávacího regulárního výrazu pro dané formulářové pole. Vyhledávací regulární výrazy byly tvořeny v co nejobecnějším tvaru, aby např. nezáviselo na počtu mezer oddělujících detekované popisné texty, nebo aby nezáviselo na konkrétním čísle hledaného formulářového pole.

Verze šablony formuláře je uvedena vždy v dolní části každé stránky. Scraper tuto hodnotu přečte v první fázi zpracování dokumentu, aby bylo případně možné chování scraperu měnit v závislosti na verzi formuláře. Verze formuláře je v rámci importu ukládána jako součást metadat přečteného dokumentu. Po

úspěšném importu přibližně 1,5 milionu dokumentů bylo rozlišeno celkem 29 různých verzí formulářů a jejich podíl výskytů je pro jednotlivé typy formulářů uveden v tabulce 3.1. Z tabulky vyplývá, že pro každý typ formuláře existují většinou 1 až 2 verze, které dohromady tvoří většinové zastoupení. Právě na tyto verze bylo zaměřeno testování správného čtení dokumentů, které bude dále popsáno v sekci 4.3 kapitoly Testování.

3.3 Výsledky importu dokumentů

Nejdříve byla implementovaným nástrojem isir-ws získána kopie databáze insolvenčního rejstříku v rozsahu poskytovaném webovou službou. Po dokončení importu má tabulka `isir_udalost` přibližně 17 milionů záznamů. Následně byl zahájen import dokumentů nástrojem isir-dl, který z tabulky `isir_udalost` vybírá odkazy na podporované typy dokumentů, provádí jejich stahování, čtení a import do databáze.

3.3.1 Průběh stahování a importu

Proces stahování a importu dokumentů byl spuštěn na serveru vyhrazeném pro tento účel, kde tato operace mohla probíhat nepřetržitě. Stahovány byly formuláře zveřejněné v období od 1. 1. 2016 do 31. 12. 2020. Rok 2016 byl zvolen jako počátek tohoto období, protože dokumenty zveřejňované v dřívějších letech ještě nevyužívaly aktuálních formulářových šablon.

Stahování a import probíhal celkem 282 hodin (přibližně 12 dní). Tato doba zahrnuje i časové intervaly mezi požadavky pro snížení zátěže serveru s dokumenty (popsáno v sekci 2.6). Celkový objem stažených dat PDF formulářů byl 2 TB. Medián velikosti jednoho PDF souboru byl 0,21 MB a největší stažený PDF dokument dosahoval velikosti 138 MB (obsahem byly naskenované stránky). Celková velikost databáze s extrahovanými textovými a číselnými údaji je po importu přibližně 14 GB. Celkem bylo importováno 1 480 214 formulářů. V tabulce 3.1 jsou rozepsány počty importovaných formulářů dle typů. Nejčastějším dokumentem byla Přihláška pohledávky.

3.3.2 Vyhodnocení úspěšnosti importu

Pro vyhodnocení úspěšnosti importu bude srovnáván počet stažených dokumentů s počtem úspěšně importovaných dokumentů. Z těchto dvou hodnot bude počítán podíl vyjádřený procenty, který bude označen jako míra přečtení. Tato hodnota určuje podíl dokumentů v rejstříku, který lze přečíst nástrojem isir-scraper, a lze ji tedy použít k vyhodnocení použitelnosti implementovaného způsobu automatizovaného čtení dokumentů. Míra přečtení nemůže u implementovaného řešení dosáhnout 100 %, protože mezi stahovanými dokumenty se vyskytují i naskenované formuláře a formuláře neznámých

	2016	2017	2018	2019	2020
Stažených dokumentů	149 865	266 882	253 301	355 442	457 062
Úspěšně importovaných	77 151	169 903	197 791	305 547	410 442
Míra přečtení (%)	51,48	63,66	78,09	85,96	89,80

Tabulka 3.2: Počet importovaných formulářů typu Přihláška pohledávky

	2016	2017	2018	2019	2020
Stažených dokumentů	85 195	161 837	173 170	170 672	188 041
Úspěšně importovaných	0	28 543	95 349	94 135	99 428
Míra přečtení (%)	0,00	17,64	55,06	55,16	52,88

Tabulka 3.3: Počet importovaných formulářů typu Přehledový list, Zpráva pro oddlužení, Zpráva o plnění oddlužení, Zpráva o splnění oddlužení

formátů, které nevyužívají oficiálních šablon. Cílem definovaným v požadavku F1.2 bylo přečtení alespoň 50 % formulářů.

V tabulce 3.2 je uvedena míra přečtení formuláře typu Přihláška pohledávky. Hodnota je v tabulce rozepsána po letech zkoumaného období 2016 až 2020. Je patrné, že míra přečtení se s každým rokem zkoumaného období zvyšuje a v posledních dvou letech dosahuje více jak 85 %. Zvyšující se míru přečtení lze vysvětlit např. zvyšující se adaptací šablon formulářů mezi insolvenčními správci nebo zvyšujícím se podílem elektronických podání na insolvenční soud, a tedy nižším podílem naskenovaných dokumentů.

Míru přečtení pro ostatní podporované typy formulářů nelze vyjádřit samostatně pro jednotlivé typy kvůli jejich nejednoznačné kategorizaci v insolvenčním rejstříku. U přihlášky pohledávky je tomu tak, že tento typ dokumentu se v rejstříku vždy zveřejňuje pod typem události Přihláška pohledávky. Ostatní čtené dokumenty však v číselníku událostí nemají svůj vlastní typ a mohou být zveřejňovány pod několika různými typy událostí, které mají širší význam a negarantují tak výskyt hledaného formuláře v dokumentu zveřejněném pod takovým typem události. Například formulář typu Zpráva o splnění oddlužení se může vyskytovat pod událostmi evidovanými dle číselníku událostí jako Sdělení správce o splnění oddlužení nebo také Usnesení o nesplnění oddlužení. Tyto dvě události však nemusí vždy obsahovat formulář Zpráva o splnění oddlužení. Nelze tak stanovit počet stažených dokumentů daného typu nutný pro vyčíslení míry přečtení.

Program isir-dl stahuje všechny dokumenty zveřejňované u událostí takových typů, které mohou některý z podporovaných formulářů obsahovat, a ve staženém dokumentu se hledají všechny podporované typy pomocí automatické detekce. Pro výpočet míry přečtení zbývajících typů formulářů bude použit počet stažených dokumentů zveřejněných u událostí všech stahova-

ných typů kromě typu Přihláška pohledávky. Kvůli okolnostem popsaným výše bude takto spočítaná míra přečtení nižší než skutečnost, ale měla by poskytovat dobrý dolní odhad skutečné míry přečtení. Výsledné hodnoty jsou zaznamenány v tabulce 3.3.

Je vidět, že v roce 2016 se nepodařilo přečíst ani jeden formulář vybraných 4 typů. Pravděpodobně je to způsobeno tím, že aktuálně používané šablony formulářů těchto typů byly zavedeny až v průběhu roku 2017. Pro roky 2019 a 2020 se však podařilo přečíst více jak polovinu dokumentů. Celkově se tedy podařilo importovat údaje z nadpoloviční většiny všech stažených dokumentů za období 2019 až 2020, a cíl vytyčený požadavkem F1.2 byl tak splněn.

3.4 Zpracování dat

Po dokončení importu formulářů došlo na fázi zpracování dat a jejich přípravu pro prezentaci ve webové sekci. K tomu byl dle návrhu implementován nástroj isir-stats. Nástroj isir-stats je realizován jako spouštěč pojmenovaných úloh nad databází. Celkem bylo pro tento nástroj implementováno 13 úloh, z toho 8 slouží ke zvýšení kvality dat a propojení entit a 5 úloh se podílí na vytvoření záznamů do statistických entit optimalizovaných pro dotazovaní z webové sekce aplikace.

3.4.1 Implementované operace

Jako součást nástroje isir-stats byly implementovány nad daty tyto operace:

doplnit_cislo_prihlasky Operace pro doplnění čísla přihlášky k entitám z přečtených dokumentů přihlášek. Číslo přihlášky je určeno číslem v oddílu u události, ze které byl dokument přečten.

doplnit_cislo_veritele Doplnění čísla věřitele k věřitelům evidovaných v tabulce `isir_osoba`. Číslo věřitele je získáno z přečteného formuláře typu Přehledový list, k jehož záznamům je věřitel asociován dle čísla jím podané přihlášky.

doplnit_datum_zahajeni Doplnění data zahájení řízení do tabulky `isir_vec`. Jako datum zahájení se považuje datum zveřejnění první události pod spisovou značkou řízení.

doplnit_datum_zverejneni_dokument Doplnění data zveřejnění dokumentu k záznamům přečtených formulářů v tabulce `dokument`.

link_distribucni_schema_veritel Operace pro přiřazení osob věřitelů z `isir_osoba` k záznamům distribučního schématu splátkového kalendáře ve zprávě pro oddlužení. K asociaci jsou použita čísla věřitelů.

link_prihlaska_osoba Úloha pro přiřazení osoby z přihlášky pohledávky k záznamu osoby v `isir_osoba` náležící věřiteli, který tuto přihlášku odeslal. Pro spojení jsou použity údaje o věřiteli uvedené v přihlášce a to v pořadí od nejvyšší specificity (IČ, rodné číslo, jméno nebo název subjektu). Po spuštění této operace na 1,1 milionu přihlášek bylo změřeno, že spojení se podaří nalézt v 99,12 % případů. Ve zbylých případech často nebylo o věřiteli uvedeno dostatečně údajů, aby bylo jednoznačné spojení možné. Byly nalezeny ale i případy přihlášek, kde jsou údaje o věřiteli vyplněny chybně (IČ neodpovídá názvu subjektu, překlepy v názvu aj.).

link_vykaz_prerozdeleni_veritel Úloha pro přiřazení osob věřitelů z `isir_osoba` k záznamům měsíčního výkazu uspokojení věřitelů ze zprávy o plnění oddlužení. V této tabulce nejsou věřitelé identifikováni číslem věřitele, ale pouze svým názvem nebo jménem. Chybí zde tedy možnost přesného spojení podle IČ nebo rodného čísla. Dlouhé názvy obchodních společností zde bývají navíc často různým způsobem zkracovány, protože pole pro název subjektu má ve formulářové tabulce omezenou šířku.

Implementovaný algoritmus pro spojování se nejdříve na začátku názvu pokusí dle regulárního výrazu najít číslo věřitele. Tento údaj není ve formuláři povinný, ale někteří správci jej před název subjektu pro přehlednost přesto zapisují. Obdobným způsobem je proveden pokus o dohledání IČ v názvu, neboť jej někteří správci zapisují za název subjektu do závorky. V poslední fázi je použito spojování dle nejvyšší podobnosti množin slov názvu subjektu, které si klade za cíl nalezení shody i v případě, že je některé slovo delšího názvu zkráceno nebo vynecháno.

Po spuštění této operace na 2,3 miliony záznamů výkazu přerozdělení bylo změřeno, že tato implementace nalezla spojení v 83,3 % případů.

odstranit_duplicitni_zmeny_stavu Tato operace slouží pro smazání duplicitních záznamů o stavových změnách řízení v tabulce `isir_vec_stav`. Protože jsou při čtení dat z webové služby insolvenčního rejstříku stavové záznamy evidovány i události, které změnu stavu řízení nezpůsobily, je možné stavový graf následně zredukovat. Po aplikaci této operace bylo odstraněno 96,2 % stavových záznamů.

stats_ins_spravci Tato operace slouží k sestavení seznamu insolvenčních správců v `stat_spravce` a jejich přiřazení k řízením vytvořením asociací v `stat_spravce_ins`. Správci jsou získáváni z tabulky `isir_osoba`, kde jsou evidováni samostatně pro každé řízení, ve kterém figurují. Pro vytvoření unikátního seznamu je použito IČ správce, případně kombinace jména a příjmení u správců, kteří IČ nemají.

stats_ins_spravci_pocty_rizeni Doplňující úloha k stats_ins_spravci, která ke správcům doplní počty jejich řízení.

stats_ins_vec Operace pro vytvoření statistických záznamů o jednotlivých řízeních do tabulky `stat_vec`.

stats_ins_vec_datum Doplňující úloha k stats_ins_vec pro doplnění měsíce a roku začátku a ukončení řízení. Tyto údaje jsou pro každé řízení doplněny z tabulky `stat_vec`, která zaznamenává změny stavu řízení v čase.

stats_ins_veritele Úloha pro vytvoření seznamu unikátních věřitelů `stat_veritel`. Pro seskupení věřitelů je použito IČ subjektu.

3.4.2 Možnost aktualizace dat

Implementované řešení je připraveno na průběžnou aktualizaci dat. Aktualizaci dat je možné provést spuštěním sekvence příkazů: isir-ws (pro stažení nových událostí v insolvenčním rejstříku), isir-dl (pro import dokumentů zveřejněných u těchto nových událostí) a isir-stats (pro zpracování dat a jejich přípravu pro prezentaci ve webové sekci). Uživatel má možnost z těchto příkazů vytvořit skript a spouštět jej periodicky pomocí automatického spouštěče úloh v operačním systému.

3.5 Implementace webové sekce

Webová sekce pro prezentaci získaných dat byla implementována za využití frameworku Laravel 8 v jazyce PHP 8. Vybrané detaily implementace budou popsány v této sekci.

3.5.1 Implementace uživatelského rozhraní

Pro základ grafického vzhledu stránky byla použita šablona Bootswatch: Lumen veřejně dostupná pod licencí MIT [37]. Pro zajištění responzivního zobrazení byl použit grid systém CSS frameworku Boostrap 4. Pro přehlednější zápis CSS kódu byl použit preprocesor Sass. Pro hromadnou kompilaci Sass kódu společně s minifikací Javascript kódu pro nasazení byl použit nástroj Laravel Mix, který slouží ke kompilaci všech frontendových souborů za využití nástroje Webpack [38].

3.5.2 Způsoby prezentace dat

Při implementaci webové aplikace byly použity tyto 4 způsoby zobrazení dat uživateli:

Tabulky Zobrazení dat v tabulkách bylo použito např. v sekci se seznamem insolvenčních správců nebo pro seznam nejčastějších věřitelů. V sekci Mapy jsou data krajů kromě mapy zobrazena i v tabulce pro lepší odečtení konkrétních hodnot. U všech tabulek byla přidána možnost seřazení dle zvoleného sloupce.

Mapy krajů Zobrazení údajů na mapě krajů slouží pro znázornění rozdílů vybraných statistik insolvenčních řízení mezi kraji ČR. Každý kraj je v mapě zbarven odstínem odpovídajícím hodnotě statistiky ve srovnání s ostatními kraji.

Pro zobrazení mapové komponenty byla použita javascriptová knihovna Leaflet. Mapový podklad poskytuje služba Mapbox za využití dat OpenStreetMap. Polygony s hranicemi krajů ČR zobrazované na mapě byly získány z OpenStreetMap pomocí filtračního nástroje Overpass a následně zjednodušeny nástrojem simplify-geojson, aby byla velikost souboru vhodná pro zobrazení na webu.

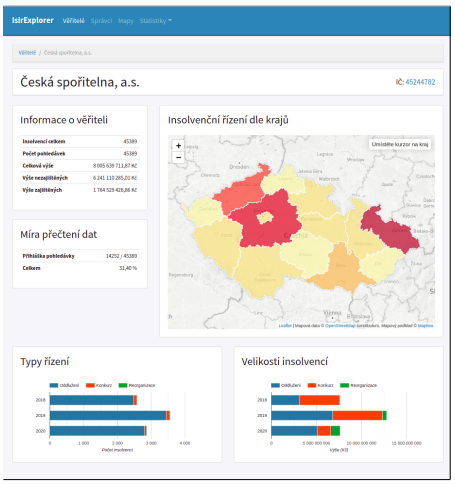
Při zobrazení stránky se geojson stáhne pomocí AJAX požadavku a v rámci následného zpracování jsou k jednotlivým obrysům krajů doplněny hodnoty zkoumané statistiky. Zobrazení krajů na mapě a jejich obarvení je realizováno pomocí možností Leaflet komponenty.

Histogramy Histogramy jsou použity u většiny vizualizací v sekci Statistiky. Tento způsob byl použit, aby uživatel mohl získat přehled o přibližném statistickém rozdělení zkoumané veličiny. Pro zobrazení histogramů je použita knihovna Plotly.

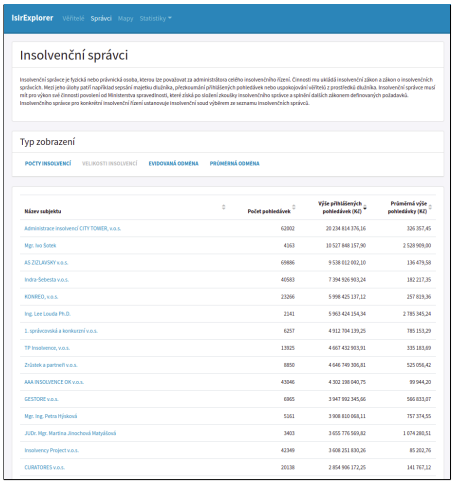
U dat zobrazených formou histogramu byla implementována možnost změny šířky intervalů (tříd) použitých pro vytvoření histogramu. Uživatel má možnost interaktivně měnit rozlišení histogramu pomocí posuvníku pod grafem. Tato funkcionalita umožní nastavit buď velmi jemné rozlišení histogramu (vhodné pro přesné znázornění rozdělení hodnot nebo pro hledání anomálií v datech), nebo naopak hrubé rozlišení, při kterém je celý histogram tvořen jen malým počtem sloupců (vhodné pro odečtení četností na větších intervalech).

Histogram je spočítán na serveru z dat vyhovujícím filtrům, které uživatel nastavil. Data histogramu jsou vložena do stránky v relativně vysoké přesnosti s četnostmi v řádu až pro stovky intervalů. Před zobrazením dat v grafu dojde ke zjednodušení dat histogramu pomocí součtu sousedních intervalů tak, aby výsledné zobrazení histogramu odpovídalo aktuálně zvolené přesnosti.

Sloupcové grafy Zobrazení sloupcových a pruhových grafů je realizováno za využití javascriptové knihovny Google Charts.



(a) Stránka s detailem věřitele



(b) Seznam insolvenčních správců podle velikosti insolvencí

Obrázek 3.2: Ukázky obsahu ze sekcí Věřitelé a Správci

3.5.3 Implementované pohledy na data

V této sekci budou popsány implementované datové výstupy, aby bylo možné vytvořit si představu o tom, jaké informace se uživatel webové aplikace může při jejím prohlížení dozvědět.

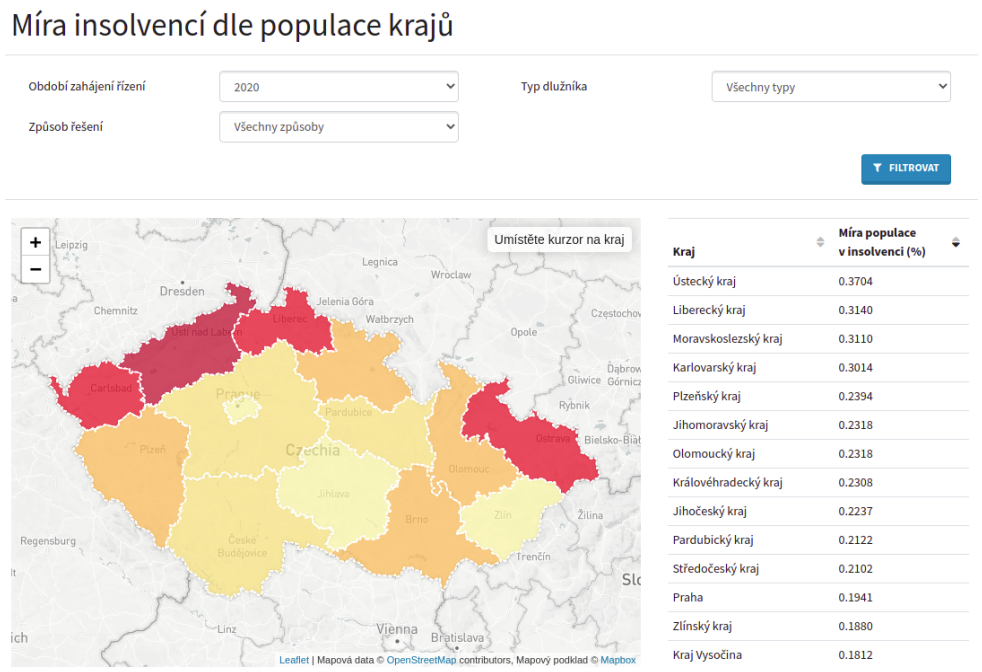
3.5.3.1 Sekce Věřitelé

Seznam věřitelů je možné přepínat mezi různými způsoby zobrazení, které se liší zobrazovanými údaji v tabulce. Implementovány jsou zobrazení dle počtu insolvencí, celkové výše přihlášených pohledávek, průměrné výše pohledávky a dle průměrného počtu pohledávek v jedné přihlášce. U každého věřitele je možný přechod na stránku s detailními informacemi.

Detailní zobrazení věřitele obsahuje základní informace o věřiteli a mapu krajů, ze kterých nejčastěji pocházejí jeho dlužníci. V dolní části stránky se nachází pruhové grafy nejčastějších typů řízení a velikostí insolvencí, do kterých tento věřitel přihlásil své pohledávky. Ukázka stránky s detailem věřitele je na obrázku 3.2a.

3.5.3.2 Sekce Správci

Stejně jako seznam věřitelů je možné i seznam správců přepínat mezi různými způsoby zobrazení. Implementovány jsou zobrazení dle počtu insolvencí, velikosti insolvencí, celkové odměny správce a průměrné odměny správce. Údaje o odměnách správce jsou rozděleny na celkovou odměnu a hotové výdaje



Obrázek 3.3: Ukázka zobrazení mapy počtu insolvencí na obyvatele dle krajů

správce. U každého správce je možný přechod na stránku s detailními informacemi. Ukázka stránky se seznamem správců je na obrázku 3.2b.

Detail insolvenčního správce zobrazuje základní informace o subjektu, mapu krajů nejčastějších řízení správce a pruhové grafy se statistikami insolvenčních řízení správce vždy za poslední 3 roky: typy řízení správce, velikosti insolvencí dle přihlášené částky nebo dle počtu pohledávek, výše popřených pohledávek. Následuje tabulka s posledními evidovanými odměnami z oddlužení a tabulka posledních skončených oddlužení. Tabulka s odměnami umožňuje rozšířené zobrazení s výpisem více záznamů. U všech záznamů v obou tabulkách je přidán odkaz na zdrojový PDF dokument na stránkách insolvenčního rejstříku. V dolní části stránky se nachází časová osa doby trvání insolvenčních řízení správce s barevným rozlišením záznamů dle typů řízení. Ukázka stránky s detailem insolvenčního správce je v obrazové příloze na obrázku D.3.

3.5.3.3 Sekce Mapy

V sekci Mapy je možné zkoumat vybrané statistiky insolvencí na mapě krajů České republiky. Typ zobrazovaných dat lze volit v horní části stránky. Každé z těchto zobrazení umožňuje další filtrace dat vstupujících do vizualizace prostřednictvím filtračních polí bezprostředně nad mapou. U většiny zobrazení je možné data filtrovat dle období zahájení nebo ukončení řízení, dle typu osoby

dlužníka a dle způsobu řešení úpadku. Vedle mapy je zobrazena tabulka s přesnými hodnotami zkoumané statistiky pro jednotlivé kraje. Přesnou hodnotu je však možné odečíst i z informačního prvku zobrazeného po umístění kurzoru na vybraný kraj.

Pro zobrazení na mapě krajů byly implementovány tyto statistiky: počet insolvencí, míra populace v insolvenci (dle počtu obyvatel daného kraje), celková výše přihlášených pohledávek, průměrný věk dlužníka, úspěšnost oddlužení, počet zrušených oddlužení, průměrná výše osvobození od dluhů po splnění oddlužení a průměrné příjmy dlužníka v době zahájení oddlužení. Ukázka stránky s mapou míry populace krajů v insolvenci je na obrázku 3.3.

3.5.3.4 Sekce Statistiky

Sekce Statistiky obsahuje 4 přehledové stránky: celkové statistiky o insolvencích a statistiky dle 3 možných způsobů řešení úpadku. Na každé z přehledových stránek jsou zobrazeny všechny dostupné vizualizace ve výchozím zobrazení a uživatel má možnost přejít na detailní zobrazení vybraného grafu, kde může konfigurovat upřesňující filtry pro zobrazená data nebo upravovat možnosti zobrazení.

U většiny následujících vizualizací je v detailním zobrazení možné nastavovat filtr na rok zahájení řízení, způsob řešení úpadku a typ osoby dlužníka (fyzická osoba – podnikatel, fyzická osoba – nepodnikatel nebo právnická osoba). U vybraných histogramů je navíc k dispozici možnost přepínání mezi lineárním a logaritmickým zobrazením os.

Počet nových insolvencí Sloupcový graf se zobrazením počtu nových insolvencí. Graf je možné zobrazit buď po letech (pro rozmezí od roku 2008 do roku 2020), nebo po měsících pro zvolený rok. Je možná filtrace dle způsobu řešení úpadku a dle typu dlužníka (fyzická nebo právnická osoba).

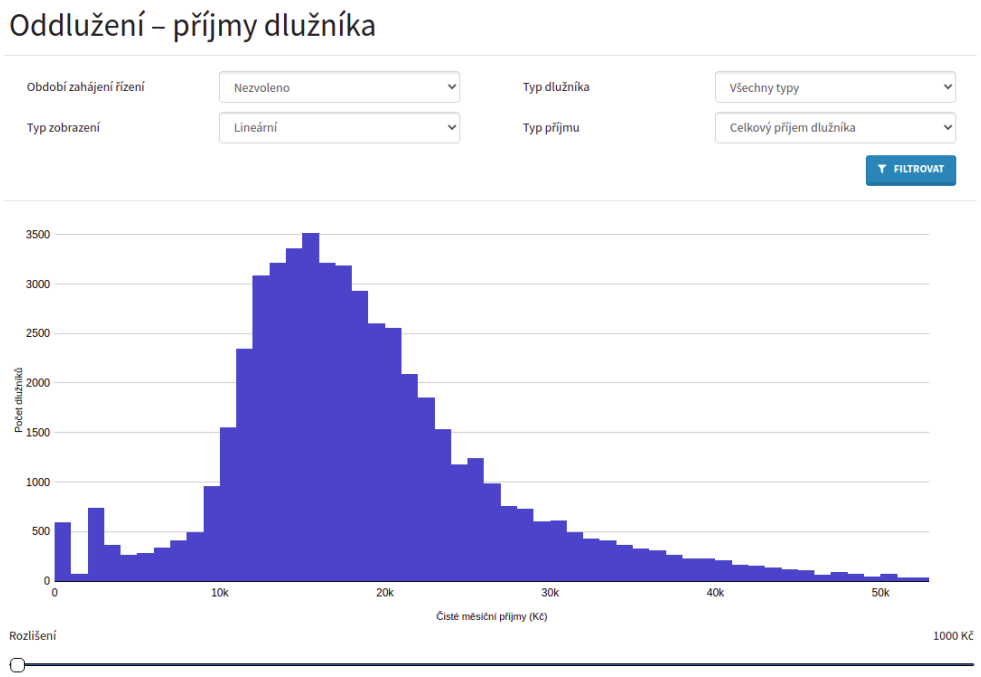
Typ osoby dlužníka Pruhový graf zobrazující nejčastější typ osoby dlužníka (fyzická nebo právnická osoba).

Věk dlužníka Histogram věku dlužníka. Pro kalkulaci histogramu je použit věk dlužníka v době zahájení řízení.

Délka řízení Histogram délky trvání řízení. V kalkulaci jsou zahrnuty délky pouze již skončených řízení.

Počet pohledávek Histogram zobrazující četnosti počtů přihlášených pohledávek do insolvenčního řízení.

Velikosti insolvencí Histogram celkové výše přihlášených pohledávek do řízení. U grafu je kromě společných filtrů navíc umožněna možnost výběru typu pohledávky (zajištěná nebo nezajištěná).



Obrázek 3.4: Ukázka detailního zobrazení grafu Příjmy dlužníka

Následují implementované vizualizace specifické pro řešení úpadku oddlužením. Tyto výstupy byly přidány jako součást plnění požadavku F2.7 (zaměření statistických výstupů na průběh oddlužení).

Forma oddlužení Pruhový graf s nejčastěji navrhovanými formami oddlužení (Splátkový kalendář, Zpeněžení majetku nebo Splátkový kalendář se zpeněžením majetku). U grafu je umožněna volba navrhující strany – je možné srovnávat formy oddlužení navrhované dlužníkem nebo insolvenčním správcem.

Míra uspokojení věřitelů Histogram procentuální míry uspokojení věřitelů po skončení oddlužení. Je možné filtrovat mezi skutečnou mírou uspokojení a mírou uspokojení odhadovanou na počátku oddlužení. Mezi možnostmi je i zobrazení rozdílu skutečné a předpokládané míry, což umožňuje zkoumat, jak často má skutečná míra uspokojení tendenci přesahovat počáteční odhad nebo na něj vůbec nedosahovat.

Příjmy dlužníka Histogram měsíčních příjmů dlužníka. Je umožněna filtrace dle kategorie příjmu (např. mzda a plat, starobní a jiný důchod, výsluhový příspěvek atd.).

Majetek dlužníka Histogram ocenění majetku dlužníka na počátku oddlužení. Je umožněna filtrace dle typu majetku (finanční prostředky, movitý majetek, nemovitý majetek, pohledávky).

```
$filtr = InsRizeni::query();
self::filtrObdobi($filtr, $conf);
self::filtrZpusobReseni($filtr, $conf);
self::filtrTypOsoby($filtr, $conf);
$rows = $filtr
 ->join('stat_oddluzeni', 'stat_oddluzeni.spisovaznacka',
        '=', 'stat_vec.spisovaznacka')
 ->join('zprava_pro_oddluzeni', 'zprava_pro_oddluzeni.id',
        '=', 'stat_oddluzeni.zpro_id');
 ->join('zpro_soupis_majetku', 'zpro_soupis_majetku.zpro_id',
        '=', 'stat_oddluzeni.zpro_id')
 ->select('oceneni AS majetek')
 ->where('typ_majetku', '=', $conf['typMajetku'])
 ->where('oceneni', '>=', 0)
 ->get();
```

Ukázka 3.2: Příklad způsobu sestavení SQL dotazu dle filtrů z formuláře

3.5.4 Způsob dotazování nad daty

Pro sestavování dotazů respektujících variabilní počet zadaných filtrů od uživatele byl použit nástroj Query Builder dostupný v použitém frameworku. V ukázce kódu 3.2 je příklad syntaxe sestavení dotazu pro histogram ocenění dlužníkova majetku dle zadaného typu majetku. Příklad je převzat z třídy `OddlMajetekController`, ve které jsou dostupné funkce jako `filtrObdobi()` nebo `filtrZpusobReseni()`, které na filtrační objekt aplikují příslušné podmínky, pokud byly nastaveny ve filtračním formuláři.

3.5.5 Optimalizace rychlosti zobrazení

V případě stránek, u kterých na serveru dochází k výpočtu dat histogramu, může odpověď serveru trvat až několik sekund. To by mohlo odradit netrpělivé uživatele, a proto bylo ve webové sekci implementováno cachování odpovědí. Webová prezentace obsahuje konečný počet statických stránek s konečným počtem kombinací filtračních možností dat u datových vizualizací. Proto je možné využít optimalizaci ve formě uložení všech stránek na disk a následný provoz webové sekce bez přístupu do databáze.

Cachování bylo implementováno tak, že odpověď se uloží na disk, pouze pokud odpověď na požadované URL ještě není součástí cache. Tímto způsobem dojde ke generování obsahu nejvýše jednou pro každou stránku. Administrátor webové sekce má možnost spustit v adresáři umístění webové aplikace připravený skript, který cache vygeneruje předem pro vybrané nejkomplikovanější stránky. Tento skript se také spustí automaticky po nasazení aplikace.

3.6 Dokumentace

Pro dokumentaci zdrojového kódu byl použit nástroj Sphinx, který využívá popisy tříd, metod a konstant, uvedených v komentářích zdrojového kódu, a generuje přehlednou dokumentaci ve formátu HTML. Pro úplné pokrytí dokumentačními řetězci byl zvolen modul Parser, který obsahuje definici metod využívaných pro extrakci údajů z formulářů. Bylo nezbytné, aby tento modul byl dostatečně dokumentován, neboť jeho správné pochopení je důležité pro případné budoucí úpravy parserů jednotlivých formulářů, které z tohoto modulu vycházejí.

Uživatelská dokumentace pro jednotlivé nástroje pro agregaci dat z insolvenčního rejstříku je obsažena přímo v těchto nástrojích a uživatel si ji může zobrazit při spuštění programu v režimu nápovědy použitím přepínače `--help`. Výstup se vytváří jako součást uživatelského rozhraní pro příkazovou řádku, k čemuž je v implementovaných nástrojích použita knihovna Click. Tento způsob uživatelské dokumentace je doplněn obecnými informacemi o použití jednotlivých nástrojů v informačním souboru na stránkách projektu na GitHubu.

Dokumentace databázového schématu je realizována prostřednictvím komentářů všech databázových tabulek a většiny jejich atributů. To usnadní orientaci ve schématu uživatelům, kteří budou mít zájem nad databází spouštět vlastní analytické dotazy.

3.7 Způsob instalace a nasazení

Sadu nástrojů pro agregaci dat z insolvenčního rejstříku (isir-ws, isir-sraper, isir-stats, isir-dl) je možné nainstalovat pomocí nástroje pip, správce balíčků pro moduly programovacího jazyka Python. Instalátor využívá definici projektu v souboru setup.py k instalaci potřebných knihoven a k nastavení aliasů pro spuštění jednotlivých nástrojů.

Pro instalaci webové aplikaci je kromě standardního způsobu možné použít nástroje Docker. Potřebné technologie (webserver Nginx, PHP v potřebné konfiguraci, nástroje npm a composer) jsou při použití tohoto způsobu instalovány jako samostatné Docker kontejnery definované v souboru docker-compose.yml. Tento způsob instalace vyžaduje od uživatele pouze poskytnutí konfiguračního souboru s údaji k databázi a spuštění instalačního skriptu.

Instalační příručka ke všem částem projektu se nachází v informačním souboru na stránkách projektu na GitHubu (opendatalabcz/isir-explorer). Webová sekce projektu byla nasazena a je dostupná na adrese **isir-explorer.eu** nebo také na adrese **isir-explorer.opendatalab.cz**.

Kapitola 4

Testování

4.1 Uživatelské testování

Pro ověření míry splnění požadavku N2.3 (snadné ovládání a srozumitelnost) bylo provedeno uživatelské testování výsledné webové aplikace. Cílem bylo ověřit, zda uživatelé, kteří aplikaci uvidí poprvé, se budou schopni rychle zorientovat v jejím uživatelském rozhraní a rychle nalézt hledané informace.

4.1.1 Proces testování

Pro testování byl vytvořen testovací scénář sestavený z 9 jednoduchých úkolů. Úkoly byly navrženy tak, aby představovaly dostatečně realistické případy skutečného použití aplikace a zároveň aby jejich splnění vyžadovalo práci s co nejvíce částmi uživatelského rozhraní aplikace. Všechny úkoly spočívají v nalezení určité informace týkající se dat prezentovaných v aplikaci. Při testování byl seznam úkolů předložen uživateli a bylo zkoumáno, jak uživatel s aplikací pracuje a co mu při plnění úkolů činí největší problémy. Seznam úkolů testovacího scénáře byl následující:

- **Úkol 1** Určete kraj ČR s nejvyšším podílem míry nových oddlužení na obyvatele v období 1. 1. 2019 do 31. 12. 2019.
- **Úkol 2** Určete kraj ČR s nejnižším průměrným příjmem dlužníka typu Fyzická osoba – podnikatel v oddlužení pro řízení ukončená v období 1. 1. 2019 a 31. 12. 2019.
- **Úkol 3** U věřitele figurujícího celkově v nejvíce insolvenčních řízeních nalezněte, ze kterého kraje ČR pochází nejvíce jeho dlužníků.
- **Úkol 4** Nalezněte insolvenčního správce s nejvyšší průměrnou odměnou na jedno řízení a zjistěte, jaká byla jeho odměna za oddlužení, které skončilo 4. 11. 2020.

- **Úkol 5** Nalezněte insolvenčního správce figurujícího v největších insolvenčních řízeních (měřeno dle průměrné výše přihlášené pohledávky).
- **Úkol 6** Zjistěte, kolik nových insolvencí bylo zahájeno v únoru 2017.
- **Úkol 7** Zjistěte nejčastější rozsah příjmu typu Starobní a jiný důchod u fyzických osob – nepodnikatelů v oddlužení. Velikost hledaného rozsahu je 5 000 Kč.
- **Úkol 8** Nalezněte nejčastější formu oddlužení (dle počtu návrhů ze strany insolvenčního správce).
- **Úkol 9** Určete kraj ČR s nejnižší průměrnou mírou uspokojení věřitelů v oddlužení (pro splněná oddlužení ukončená v průběhu roku 2019).

Testování bylo provedeno celkem s 5 uživateli, z toho 1 uživatel použil pro testování mobilní zařízení. Věková skupina uživatelů byla 20 – 30 let u 3 uživatelů a 60 – 70 u 2 uživatelů. Dosažené vzdělání uživatelů bylo vysokoškolské – magisterský studijní program u 4 uživatelů a vysokoškolské – bakalářský studijní program u 1 uživatele. Většina uživatelů měla již před testováním alespoň základní povědomí o fungování insolvenčního procesu a o jeho účastnících. V průběhu testování mohli uživatelé klást testerovi dotazy ohledně problematiky insolvenčních řízení do míry nutné pro pochopení úkolů z testovacího scénáře. Nebylo však možné pokládat dotazy ohledně postupu pro řešení jednotlivých úkolů nebo ohledně uživatelského rozhraní aplikace.

4.1.2 Výsledky testování

Pro jednotlivé úkoly budou popsány akce, které byly od uživatelů očekávány, a shrnutí akcí, které uživatelé při testování vykonali.

1. **Úkol 1** Určete kraj ČR s nejvyšším podílem míry nových oddlužení na obyvatele v období 1. 1. 2019 do 31. 12. 2019.

   Očekávaný postup: Volba kategorie *Mapy* v hlavní nabídce, přepnutí zobrazení mapy do režimu *Míra populace v insolvenci*, nastavení časového období a přepnutí způsobu řešení na *Oddlužení*.

   Postup uživatelů: Většina uživatelů nejdříve informaci hledala v sekci *Statistiky – Oddlužení*. Po zjištění, že tato sekce neobsahuje výstupy členěné dle krajů, uživatelé vstoupili do sekce *Mapy*. Následné nastavení filtru a zjištění hledané hodnoty jim již nečinilo problém.

   Další zjištění: Uživatel s mobilním zařízením zmínil, že zobrazení mapy krajů je po načtení stránky ve výchozím stavu příliš přiblíženo a pro pohled na celou republiku je nutné mapu o jeden stupeň oddálit. Poznámka byla zaznamenána pro opravu. Nedostatek však nepředstavoval problém pro splnění testovacího úkolu.

2. **Úkol 2** Určete kraj ČR s nejnižším průměrným příjmem dlužníka typu Fyzická osoba – podnikatel v oddlužení pro řízení ukončená v období 1. 1. 2019 a 31. 12. 2019.

   Očekávaný postup: Přepnutí zobrazení mapy do režimu *Oddlužení – příjmy dlužníka* a nastavení typu osoby na *Fyzická osoba – podnikatel*.

   Postup uživatelů: Tento úkol opět vyžaduje práci se sekcí *Mapy*, se kterou se již uživatelé seznámili během předchozího úkolu. Tento úkol proto všichni splnili bez problému.

3. **Úkol 3** U věřitele figurujícího celkově v nejvíce insolvenčních řízeních nalezněte, ze kterého kraje ČR pochází nejvíce jeho dlužníků.

   Očekávaný postup: Volba kategorie *Věřitelé* a zobrazení detailu věřitele na prvním místě v seznamu seřazeném ve výchozím stavu dle počtu řízení. Určení kraje s největším počtem dlužníků dle mapy zobrazené v detailu věřitele.

   Postup uživatelů: Jeden uživatel úspěšně dohledal hledaného věřitele v tabulce, ale pak si nevšiml, že je možné kliknout na název věřitele pro zobrazení více informací. To vedlo k tomu, že opustil sekci *Věřitelé* a neúspěšně údaj hledal v sekci *Mapy*. Později se však vrátil a detail nalezl. Ostatní uživatelé údaj dohledali relativně rychle.

4. **Úkol 4** Nalezněte insolvenčního správce s nejvyšší průměrnou odměnou na jedno řízení a zjistěte, jaká byla jeho odměna za oddlužení, které skončilo 4. 11. 2020.

   Očekávaný postup: Volba kategorie *Správci* a přepnutí zobrazení seznamu do režimu *Průměrná odměna*, zobrazení detailu správce zobrazeného na prvním místě v seřazeném seznamu. Nalezení odměny za hledané řízení v tabulce posledních odměn správce.

   Postup uživatelů: Uživatelská rozhraní stránek *Věřitelé* a *Správci* jsou si velice podobná, a uživatelé tak po předchozím úkolu již věděli, jak rychle dohledat a zobrazit detail hledaného správce.

   Jeden z uživatelů nalezl problém s řazením tabulky *Poslední skončená oddlužení* v detailu správce. Řazení tabulky dle data skončení řízení fungovalo v abecedním režimu a seřazené řádky tak nebyly řazeny chronologicky dle data. Tato chyba byla opravena.

   Další zjištění: Uživatel s mobilním zařízení zmínil, že tabulka se seznamem insolvenčních správců se v mobilním zobrazení nevejde na obrazovku a část posledního sloupce přetéká vpravo mimo viditelnou oblast stránky. Všechny tabulky ve webové aplikaci jsou tvořeny maximálně 4 sloupci. V tomto případě se však na stránku vejdou pouze 3. Problém byl vyřešen tak, že s obsahem tabulky je nyní možno horizontálně posouvat, pokud se nevejde na stránku.

5. **Úkol 5** Nalezněte insolvenčního správce figurujícího v největších insolvenčních řízeních (měřeno dle průměrné výše přihlášené pohledávky).

   Očekávaný postup: Přepnutí zobrazení seznamu správců do režimu *Velikost insolvencí* a přepnutí řazení sloupce na *Průměrná výše pohledávky*.

   Postup uživatelů: Všichni uživatelé úkol splnili bez problémů.

   Další zjištění: Několik uživatelů upozornilo na matoucí umístění šipek pro řazení tabulky. Šipky se zobrazovaly vždy na opačné straně sloupce, než na jakou byl zarovnán text jeho obsahu. První sloupec je zarovnán doleva a další sloupce doprava, což vyvolávalo nejasnosti ohledně toho, které šipky náleží kterému sloupci. Tato chyba již byla opravena a šipky pro řazení se zobrazují vždy v pravé části sloupce.

6. **Úkol 6** Zjistěte, kolik nových insolvencí bylo zahájeno v únoru 2017.

   Očekávaný postup: Volba sekce *Statistiky – Všechny insolvence* a přechod na detail grafu *Počet nových insolvencí*. Přepnutí grafu z výchozího zobrazení po letech do režimu zobrazení po měsících. Přečtení hodnoty u sloupce 2017/02.

   Postup uživatelů: Jeden uživatel měl menší potíže s nalezením sekce *Statistiky*, ostatní úkol splnili bez problémů.

   Další zjištění: Jeden z uživatelů, který uvedl, že má zkušenosti s prací se statistickými dashboardy business intelligence aplikací od poskytovatelů SAP a IBM, prohlásil, že chování grafu neodpovídá jeho očekávání. Uživatel předpokládal, že při kliknutí na sloupec v grafu s hodnotami pro rok 2017 se graf přepne do zobrazení měsíců pro zvolený rok.

7. **Úkol 7** Zjistěte nejčastější rozsah příjmu typu Starobní a jiný důchod u fyzických osob – nepodnikatelů v oddlužení. Velikost hledaného rozsahu je 5 000 Kč.

   Očekávaný postup: Volba sekce *Statistiky – Oddlužení* a přechod na detail grafu *Příjem dlužníka*. Přidání filtrační podmínky omezující zobrazovaná data pouze na typ příjmu *Starobní a jiný důchod*. Změna rozlišení histogramu na interval délky 5 000 Kč.

   Postup uživatelů: Všem uživatelům se podařilo nalézt graf s příjmy dlužníka. Histogram je však ve výchozím stavu zobrazen pro intervaly délky 1 000 Kč, a zadání úkolu tedy vyžaduje změnu rozlišení histogramu pomocí posuvníku pod grafem. Většina uživatelů tuto funkcionalitu přehlédla, a proto pro ně tento úkol představoval velký problém.

8. **Úkol 8** Nalezněte nejčastější formu oddlužení (dle počtu návrhů ze strany insolvenčního správce).

   Očekávaný postup: Volba sekce *Statistiky – Oddlužení* a přečtení hledaného údaje z grafu *Správcem navrhovaná forma oddlužení*.

   Postup uživatelů: Všichni uživatelé úkol splnili bez problémů.

9. **Úkol 9** Určete kraj ČR s nejnižší průměrnou mírou uspokojení věřitelů v oddlužení (pro splněná oddlužení ukončená v průběhu roku 2019).

   Očekávaný postup: Volba kategorie *Mapy*, přepnutí zobrazení mapy do režimu *Úspěšnost oddlužení*, nastavení časového období na rok 2019 a zjištění kraje s nejnižší hodnotou.

   Postup uživatelů: Všichni uživatelé úkol splnili bez problémů.

Testování považuji za úspěšné, neboť vedlo k nalezení několika chyb uživatelského rozhraní a uživatelé poskytli několik návrhů na zlepšení. Většina nalezených chyb byla opravena a návrhy na zlepšení mohou být použity pro budoucí rozvoj aplikace.

4.2 Jednotkové testy

Jednotkové testy byly použity pro testování funkčnosti částí aplikace isir-scraper používané pro čtení dat z PDF formulářů. Pro realizaci jednotkových testů byl použit nástroj Pytest. Pro automatické spouštění testů byl použit nástroj pro průběžnou integraci GitHub Actions dostupný na GitHubu, kde je kód aplikace zveřejněn. Tento nástroj při každé změně kódu provede v izolovaném systému stažení závislostí programu nástrojem pip a následné spuštění testů.

Aktuální testovací sada obsahuje 51 testů, které pokrývají všechny metody modulů `IsirParser` a `IsirDecryptor`. Tyto moduly byly zvoleny pro úplné pokrytí jednotkovými testy, neboť jejich bezchybná implementace je nutností pro správné čtení všech podporovaných typů dokumentů.

4.3 Testování správného čtení dokumentů

Pro testování správného čtení údajů z PDF formulářů byla požita metoda srovnávání výstupu programu s referenčním výstupem pro zadaný vstupní formulář. Jde o způsob testování klasifikovaný jako tzv. *black box testing*, který se vyznačuje zkoumáním výstupů aplikace v závislosti na změně vstupů. Tento způsob je možný i bez znalostí vnitřního fungování aplikace [39].

Z insolvenčního rejstříku byl stažen náhodný vzorek dokumentů, které byly použity pro testování. Jedná se o 50 dokumentů obsahující formuláře všech podporovaných typů v nejčastějších verzích. Na tuto množinu dokumentů byl

aplikován nástroj isir-scraper a výsledné JSON soubory byly manuálně zkontrolovány, zda obsahují správně přečtené údaje z příslušných vstupních formulářů. Pro testování byl vytvořen skript v jazyce Bash, který spustí program isir-scraper postupně na všechny testovací PDF a testuje výstup programu s referenčními JSON soubory. Tento způsob testování byl užitečný zejména při vývoji parserů jednotlivých typů dokumentů pro ověření, že změna určité části nezpůsobí nefunkčnost již implementovaných částí. Nevýhodou tohoto způsobu je, že v případě rozšíření množiny čtených polí jistého typu formuláře bylo nutné aktualizovat i množinu referenčních výstupů.

4.4 Statická analýza kódu

Statická analýza kódu slouží k auditu zdrojového kódu bez jeho spuštění. Na zdrojový kód implementované aplikace byly aplikovány statické analyzátory s flake8 a Bandit. Jejich automatické spuštění při změně kódu bylo realizováno jejich přidáním do GitHub Actions v repositáři aplikace.

Nástroj flake8 kontroluje, že formátování kódu je konzistentní v celé aplikaci a odpovídá konvencím dle standardu PEP 8 pro formátování Python kódu [40]. Nástroj Bandit je určen pro hledání častých bezpečnostních zranitelností v Python kódu [41]. Nedostatky v kódu zjištěné oběma programy byly prověřeny a opraveny. Některá bezpečnostní upozornění programu Bandit nebyla relevantní pro konkrétní výskyt v kódu aplikace, a tak byly tyto části kódu vyřazeny z kontroly (šlo např. o upozornění na použití potenciálně nebezpečných modulů subprocess pro volání podprogramu, nebo modulu xml.etree pro čtení obsahu XML dokumentů).

Závěr

Cílem práce bylo navrhnout, implementovat a otestovat systém pro zpracování a přehledné zobrazení agregovaných dat o insolvencích. Implementoval jsem sadu nástrojů pro extrakci dat z insolvenčního rejstříku jak za využití oficiálního programového rozhraní rejstříku, tak za využití extrakce dat z vybraných typů PDF formulářů zveřejňovaných v insolvenčním rejstříku. Čtením údajů z formulářů se podařilo získat data, která dle provedené analýzy nejsou součástí žádných v současné době zveřejňovaných statistik o insolvencích. Mezi tyto údaje patří např. informace o příjmech a hodnotě majetku dlužníků, o odměnách insolvenčních správců a do určité míry i údaje o úspěšnosti oddlužení. Implementoval jsem také webovou prezentaci zobrazující získaná data pomocí grafů a jiných vizualizací. Implementované řešení bylo otestováno jak z hlediska ověření správnosti extrakce údajů, tak z pohledu návrhu uživatelského rozhraní. Cíl práce byl proto splněn.

Metoda pro extrakci dat pomocí čtení obsahu PDF formulářů se ukázala jako poměrně efektivní, neboť např. pro dokument typu Přihláška pohledávky se podařilo přečíst a importovat údaje z více jak 85 % dokumentů zveřejněných mezi roky 2019 a 2020. U ostatních 4 podporovaných typů formulářů se za stejné období podařilo importovat údaje z více jak 50 % dokumentů. Dokumenty, které se importovat nepodařilo, většinou obsahují formuláře v naskenované podobě nebo k jejich vytvoření nebyly použity oficiální elektronické šablony.

Aplikaci jsem navrhl s ohledem na možnosti budoucího rozšíření. Možná rozšíření lze rozdělit do dvou směrů. Jeden představuje zlepšování kvality extrakce dat z insolvenčního rejstříku jako např. přidání podpory pro další typy formulářů či implementace metod extrakce údajů i z dokumentů, které nevyužívají známé šablony. Druhým směrem rozvoje je zlepšování kvality výstupů ve webové sekci aplikace prezentující získaná data. Jedním z možných vylepšení webové prezentace je přidání dalších nebo podrobnějších statistických výstupů, neboť možnosti analýzy údajů ve vytvořeném datovém modelu zdaleka nejsou vyčerpány. Další možností rozvoje webové sekce jsou detailnější popisy

dat a přidání analytických textů, které by data shrnovaly, vyvozovaly z nich závěry či pokládaly jejich význam do kontextu souvislostí aktuální insolvenční legislativy.

Výslednou webovou prezentaci může využít jak odborná, tak neodborná veřejnost k získání přehledu o statistikách insolvenčních řízení v České republice. I samotný datový model a sada nástrojů pro extrakci dat z insolvenčního rejstříku může být užitečným zdrojem při tvorbách analýz ze strany odborné veřejnosti. Zdrojový kód implementovaného řešení byl zveřejněn na GitHubu pod licencí GNU GPL v3, kde jej plánuji dále udržovat. Při zpracování této práce jsem získal mnoho praktických zkušeností, stejně jako poměrně rozsáhlý pohled na fungování insolvenčního procesu v České republice.

Bibliografie

1. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvence - Slovníček insolvenčních pojmů* [online]. 2018 [cit. 2020-12-13]. Dostupné z: `https://insolvence.justice.cz/slovnik-insolvencnich-pojmu/`.
2. *Zákon č. 182/2006 Sb., o úpadku a způsobech jeho řešení (insolvenční zákon)*. 2006.
3. BANKY.CZ. Realitní slovník [online]. 2020 [cit. 2020-12-13]. ISSN 2464-4579. Dostupné z: `https://www.banky.cz/realitni-slovnik/`.
4. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvence - Insolvenční návrh* [online]. 2018 [cit. 2021-01-07]. Dostupné z: `https://insolvence.justice.cz/jak-ven-z-dluhove-pasti/insolvencni-navrh/`.
5. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvence - Co Vás čeká po podání návrhu* [online]. 2018 [cit. 2021-01-07]. Dostupné z: `https://insolvence.justice.cz/jak-ven-z-dluhove-pasti/co-vas-ceka-po-podani-navrhu/`.
6. DOLEČEK, Marek. *Insolvence – úpadek a způsoby jeho řešení* [online]. BusinessInfo.cz, CzechTrade, 2020 [cit. 2020-11-22]. Dostupné z: `https://www.businessinfo.cz/navody/insolvence-upadek-a-zpusoby-jeho-reseni-ppbi/`.
7. HOVORKA, Jiří. *Osobní bankrot se zkrátí. Oddlužení má být na tři roky* [online]. Peníze.cz, 2020-11-02 [cit. 2021-01-07]. Dostupné z: `https://www.penize.cz/osobni-bankrot/421565-osobni-bankrot-se-zkrati-oddluzeni-ma-byt-na-tri-roky`.
8. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvence - Oddlužení* [online]. 2018 [cit. 2021-01-07]. Dostupné z: `https://insolvence.justice.cz/jak-ven-z-dluhove-pasti/oddluzeni/`.
9. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvenční rejstřík* [online]. 2020 [cit. 2020-12-13]. Dostupné z: `https://isir.justice.cz`.

10. ADVOKÁTNÍ DENÍK. *Mapa insolvence: obraz Česka jako rizikové země neplatičů je falešný* [online]. 2019-10-30 [cit. 2021-01-19]. ISSN 2571-3558. Dostupné z: https://advokatnidenik.cz/2019/10/30/mapa-insolvence-obraz-ceska-jako-rizikove-zeme-neplaticu-je-falesny/.

11. INSOLCENTRUM, S.R.O. *Vše o insolvencích v České republice* [online]. 2020 [cit. 2020-12-03]. Dostupné z: https://www.insolcentrum.cz/insolvence-cr/.

12. SURVEILLIGENCE, S.R.O. *Vzdělávací centrum - Surveilligence* [online]. 2020 [cit. 2021-01-19]. Dostupné z: http://www.surveilligence.com/cs/vzdelavaci-centrum.

13. SURVEILLIGENCE, S.R.O. *Insolvency report 12/2020* [online]. 2021-01-03 [cit. 2021-01-19]. Dostupné z: http://www.surveilligence.com/content/3-vzdelavacie-centrum/5-insolvency-report/20210103-insolvency-report-12-2020/insolvency-report-2020-12-by-surveilligence_s.pdf.

14. CRIF - CZECH CREDIT BUREAU, A. S. *CRIBIS – Informace o firmách* [online]. 2021 [cit. 2021-01-20]. Dostupné z: https://www.informaceofirmach.cz/.

15. CRIF - CZECH CREDIT BUREAU, A. S. *CRIF: Počet návrhů na firemní bankrot zůstal v listopadu velmi vysoký* [online]. 2020-12-04 [cit. 2021-01-20]. Dostupné z: https://www.informaceofirmach.cz/crif-pocet-navrhu-na-firemni-bankrot-zustal-v-listopadu-velmi-vysoky/.

16. BISNODE ČESKÁ REPUBLIKA, A.S. *Seznam produktů - Bisnode Česká republika* [online]. 2021 [cit. 2021-01-20]. Dostupné z: https://www.bisnode.cz/seznam-produktu/.

17. CREDIT CHECK, S.R.O. *Prověřování zákazníků - CreditCheck.cz* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://www.creditcheck.cz/.

18. CREDIT CHECK, S.R.O. *Insolvenční rejstřík - Creditcheck.cz* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://www.creditcheck.cz/InfoSourceDetail.aspx?id=01.

19. INSOLVENCE 2008, A.S. *Monitoring rejstříku* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://monitoringrejstriku.cz/.

20. ZELIUS, S.R.O. *Monitor Justice - Sledování insolvenčního rejstříku zdarma* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://monitorjustice.cz/.

21. I4B S.R.O. *Sledování insolvence a hlídání insolvenčního rejstříku* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://sledovani-insolvence.cz/.

22. GRIT, S.R.O. *Monitoring insolvence - GRiT* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://www.grit.eu/cs/monitoring-insolvencniho-rejstriku/.

23. GREGOR, Jiří; HEJLOVÁ, Hana. *Tematický článek o finanční stabilitě – 4/2020* [online]. Česká národní banka, 2020 [cit. 2021-01-21]. Dostupné z: https://www.cnb.cz/export/sites/cnb/cs/financni-stabilita/.galleries/tematicke-clanky-o-financni-stabilite/tcfs_2020_04_cz.pdf.

24. ČESKÝ STATISTICKÝ ÚŘAD. *Příjmová chudoba ohrožuje necelou desetinu obyvatel* [online]. 2019-03-21 [cit. 2021-01-21]. Dostupné z: https://www.czso.cz/csu/czso/prijmova-chudoba-ohrozuje-necelou-desetinu-obyvatel.

25. HÖFFEROVÁ, Markéta; KOVANDA, Lukáš. *Mezi Čechy je v rámci EU druhý nejnižší podíl dlužníků se závazky po splatnosti. Hned po Lucembursku* [online]. Kurzy.cz, 2018-11-08 [cit. 2021-01-21]. Dostupné z: https://www.kurzy.cz/zpravy/472814-mezi-cechy-je-v-ramci-eu-druhy-nejnizsi-podil-dluzniku-se-zavazky-po-splatnosti-hned-po-lucembursku/.

26. CCA GROUP A.S. *Webová služba aplikace ISIR - Popis způsobu používání webové služby* [online]. Ministerstvo spravedlnosti ČR, 2015-07-16 [cit. 2021-01-21]. Dostupné z: https://isir.justice.cz/isir/help/Popis_WS_1_v2_7.pdf.

27. MINISTERSTVO SPRAVEDLNOSTI ČR. *Insolvenční rejstřík - Formuláře* [online]. 2020 [cit. 2020-12-13]. Dostupné z: https://isir.justice.cz/isir/common/stat.do?kodStranky=FORMULAR.

28. ADOBE SYSTEMS INCORPORATED. *Document management — Portable document format* [online]. 2008-07-01 [cit. 2021-01-20]. Dostupné z: https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdfs/PDF32000_2008.pdf.

29. HAUSENBLAS, Michael. *5-hvězdičková otevřená data* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://5stardata.info/cs/.

30. FREE SOFTWARE FOUNDATION, INC. *Poppler* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://poppler.freedesktop.org/.

31. ARTIFEX SOFTWARE, INC. *ghostscript - High Level Output Devices* [online]. 2020 [cit. 2021-01-20]. Dostupné z: https://www.ghostscript.com/doc/current/VectorDevices.htm#TXT.

32. THE ENCODE COMMUNITY. *Databases* [online]. 2021-01-12 [cit. 2021-01-12]. Dostupné z: https://github.com/encode/databases.

33. THE CAKEPHP COMMUNITY. *Phinx Documentation* [online]. 2017 [cit. 2021-01-23]. Dostupné z: https://phinx.readthedocs.io/en/latest/index.html.

34. ČESKÁ POŠTA, S.P. *Seznam PSČ částí obcí a obcí bez částí České pošty* [online]. 2020 [cit. 2021-02-23]. Dostupné z: `https://www.ceskaposta.cz/documents/10180/3738087/xls_pcobc.zip/50617e56-6e9a-4335-9608-96fec214e6ef`.

35. MINISTERSTVO ZEMĚDĚLSTVÍ ČR. *Číselník okresů České republiky* [online]. 2019 [cit. 2021-02-23]. Dostupné z: `https://eagri.cz/ssl/nosso-app/DataKeStazeni/Okresy`.

36. MINISTERSTVO VNITRA ČR. *Seznam obcí České republiky* [online]. 2015-04-10 [cit. 2021-02-23]. Dostupné z: `https://seznam.gov.cz/otevrena-data/datove-sady/2015-04-10/8302`.

37. PARK, Thomas. *Bootswatch – Lumen* [online]. 2021 [cit. 2021-03-22]. Dostupné z: `https://bootswatch.com/lumen/`.

38. WAY, Jeffrey. *Laravel Mix* [online]. 2021 [cit. 2021-04-14]. Dostupné z: `https://laravel-mix.com/`.

39. PATTON, Ron. *Software Testing*. Sams Publishing, 2000. ISBN 0-672-31983-7.

40. PYTHON CODE QUALITY AUTHORITY. *Flake8* [online]. 2021-04-18 [cit. 2021-04-18]. Dostupné z: `https://github.com/pycqa/flake8`.

41. PYTHON CODE QUALITY AUTHORITY. *Bandit* [online]. 2021-04-18 [cit. 2021-04-18]. Dostupné z: `https://github.com/PyCQA/bandit`.

Příloha A

Seznam použitých zkratek

AJAX Asynchronous JavaScript And XML

ASCII American Standard Code for Information Interchange

CID Character ID

CMap Character map

CSS Cascading Style Sheets

ČR Česká republika

ES6 ECMAScript 6

GNU GNU's Not Unix

GPL General Public License

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

ISIR Insolvenční rejstřík

JSON JavaScript Object Notation

MIT Massachusetts Institute of Technology

PDF Portable Document Format

PEP Python Enhancement Proposals

PSČ Poštovní směrovací číslo

SOAP Simple Object Access Protocol

SQL Structured Query Language

URL Uniform Resource Locator

UTF Unicode Transformation Format

WSDL Web Services Description Language

XML Extensible markup language

XSD XML Schema Definition

Příloha B

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `docs` ..... programová dokumentace
  - `build` ..... dokumentace zdrojového kódu v HTML
  - `schema` ..... kompletní diagram databázového modelu
- `src`
  - `pdf-scraper` ..... nástroj pro extrakci dat
  - `isir-explorer` ..... webová sekce pro prezentaci dat
  - `schema` ..... definice databázového schématu
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF

Příloha C

Instalační příručka

Nástroje pro extrakci dat

Nástroje pro extrakci dat z insolvenčního rejstříku se nachází v adresáři pdf-scraper. Pro jejich instalaci lze postupovat takto:

1. Je doporučeno nejdříve vytvořit a aktivovat nové Python prostředí. K tomu lze použít následující příkaz:
   `python3.7 -m venv venv && . venv/bin/activate`
2. Instalace se spustí příkazem `pip install .`
3. Po dokončení budou v prostředí dostupné nástroje isir-scraper, isir-ws, isir-dl, isir-importer a isir-stats. Nápovědu lze zobrazit spuštěním vybraného nástroje s přepínačem `--help`.

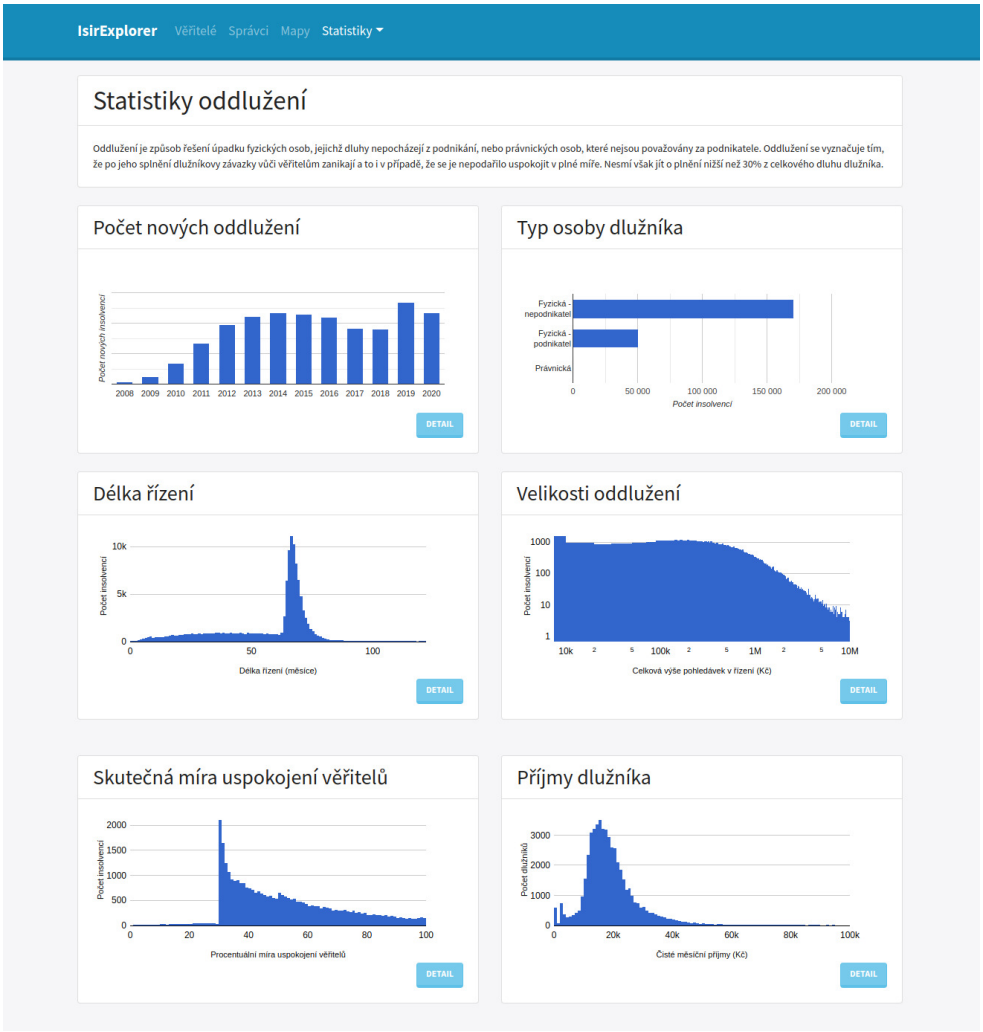
Webová aplikace

Aplikaci je možné nainstalovat pomocí nástroje docker-compose. Po naklonování repositáře lze pro instalaci postupovat takto:

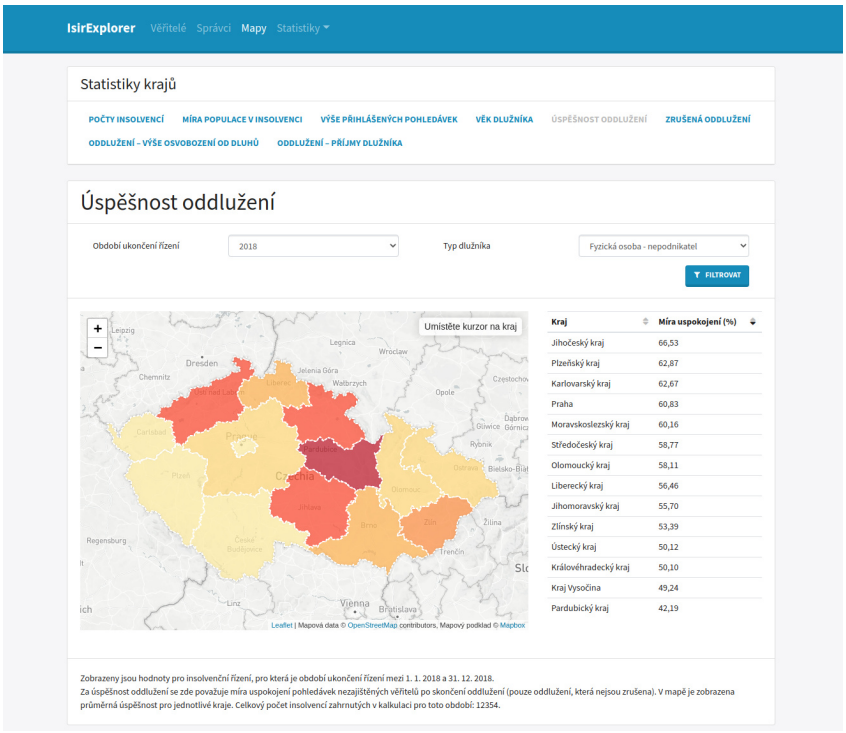
1. Vytvořit konfigurační soubor .env. Je možné použít šablonu v souboru .env.example. V tomto souboru je nutné vyplnit údaje pro připojení k databázi, URL, kde bude aplikace provozována (APP_URL), a API klíč ke službě Mapbox pro zobrazení mapového podkladu (MAPBOX_KEY).
2. Spustit skript `./docker-build.sh`, který obsahuje příkazy pro sestavení potřebných Docker kontejnerů dle docker-compose.yml.
3. Spustit `docker-compose up -d` v adresáři projektu. Tím se spustí vytvořené kontejnery a služba bude dostupná portu specifikovaném v docker-compose.yml (ve výchozím stavu 8080).

Příloha D

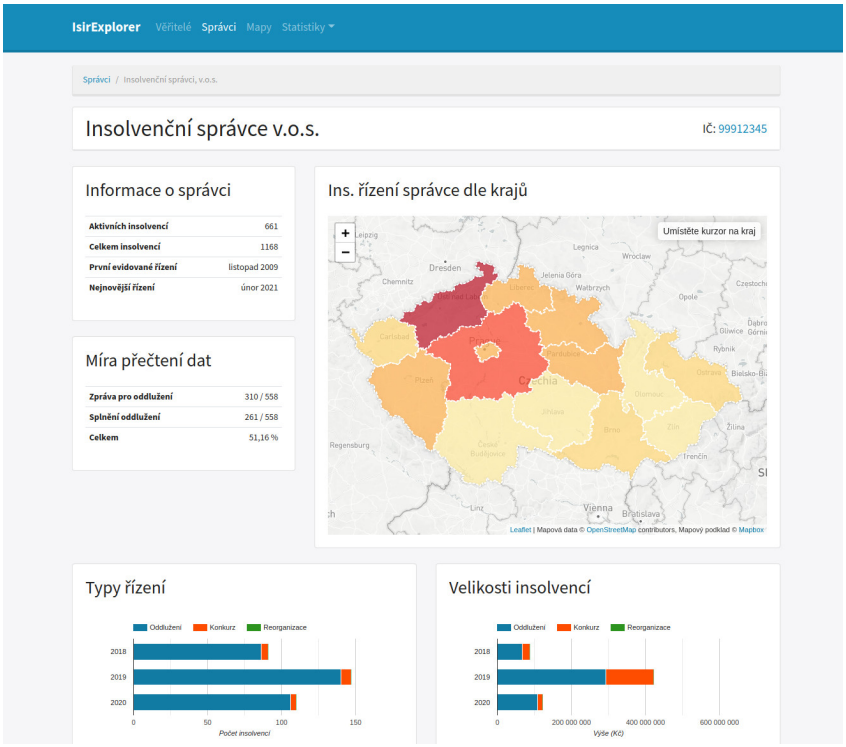
Obrazová příloha



Obrázek D.1: Obsah části sekce Statistiky – Oddlužení



Obrázek D.2: Obsah sekce Mapy



Obrázek D.3: Ukázka horní části stránky s detailem správce