Sem vložte zadanie Vašej práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra softwarového inženýrství

Diplomová práca

Spracovanie a vizualizácia chemických meraní v dátovom repozitári

Bc. Lukáš Koštenský

Vedúci práce: RNDr. David Antoš, Ph.D.

5. marca 2017

Poďakovanie

Doplňte, ak chcete niekomu za niečo poďakovať. V opačnom prípade úplne odstráňte tento príkaz.

Prehlásenie

Prehlasujem, že som predloženú prácu vypracoval(a) samostatne a že som uviedol(uviedla) všetky informačné zdroje v súlade s Metodickým pokynom o etickej príprave vysokoškolských záverečných prác.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona, v znení neskorších predpisov, a skutočnosť, že České vysoké učení technické v Praze má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe 5. marca 2017 .....

Odkaz na túto prácu

Koštenský, Lukáš. *Spracovanie a vizualizácia chemických meraní v dátovom repozitári.* Diplomová práca. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2017.

Abstrakt

V niekoľkých vetách zhrňte obsah a prínos tejto práce v slovenčine. Po prečítaní abstraktu by mal čitateľ mať dosť informácií pre rozhodnutie, či Vašu prácu chce čítať.

Klíčová slova Nahraďte zoznamom kľúčových slov v slovenčine oddelených čiarkou.

Abstract

Sem doplňte ekvivalent abstraktu Vašej práce v angličtině.

Keywords Nahraďte zoznamom kľúčových slov v angličtine oddelených čiarkou.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Kapitola 1

Popis problému

Repozitár slúži vo všeobecnosti ako centrálne miesto, ktoré sa stará o ukladanie a správu dát. Takúto službu môžu chcieť poskytovať rôzne inštitúcie (napríklad školy), knihovny,... Pod slovom repozitár si môžeme taktiež predstaviť konkrétny software, ktorý sa stará o ukladanie, archiváciu a sprístupnenie dát. V tejto diplomovej práci budeme pod slovom repozitár rozumieť práve software.

1.1 Zdieľanie dát v teame

Ústav organickej chémie VŠCHT Praha potrebuje vyriešiť ukladanie a sprístupnenie dát. Potrebujú ukladať, analyzovať a prezentovať infračervené vibračné, NMR a hmotnostné spektroskopické merania, chemické vzorce a reakcie.

Nad jedným datasetom môže pracovať viacero ľudí, ktorý môžu riešiť rôzne merania a pokusy alebo spoločne pracovať na jednom meraní. V oboch prípadoch, počas priebehu samotného výskumu, potrebujú prístup k dátam, ktoré vytvoril iný člen teamu. Taktiež musia mať možnosť dáta upravovať (napr. opakované merania a pokusy, keď je potrebné doplniť nové výsledky). Je teda potrebné vyriešiť zdieľanie dát v teame, rôzne oprávnenia pre osoby, ktoré majú mať k dátam prístup, verzovanie dát.

1.2 Open data/Open access

Taktiež je pri vývoji repozitára potrebné myslieť na možnosť zverejnenia (časti) dát pre širokú verejnosť s možnosťou ich ďalšieho využitia alebo odkazovania sa na ne. Takto zverejnené dáta označujeme pojmom Open data. V prípade zverejnených výzkumov hovoríme o Open access (OA).

1.3 Zálohovanie a archivácia

Zálohovaním dát rozumieme vytváranie kópie práve spracúvaných alebo v relatívne nedávnej dobe uložených dát. Archiváciou rozumieme uschovávanie dokumentačných materiálov.

Zálohované dáta môžu byť poškodené degradáciou média, fyzickým poškodením média alebo v súčasnosti rozšírenými cryptovírusmi. Zálohovať dáta na jedno médium nestačí. Je dobré sa riadiť pravidlom 3-2-1. Tri kópie všetkých dôležitých dát, na dvoch rôznych médiach, pričom jedna kópia by mala byť uložená off-site, teda niekde mimo pracovného prostredia. [1]

Pri archivácii dát je potrebné myslieť na čitateľnosť dát po dlhej dobe. Preto je potrebné myslieť nie len na zabezpečenie dát, ale aj na archiváciu programu potrebného pre prečítanie archivovaných dát.

Repozitár by mal byť pre užívateľov možnosťou ako dáta zálohovať. Zároveň jeho napojenie na služby CESNETu umožní ochranu dát, akú by bolo na pracovisku VŠCHT Praha ťažké dosiahnúť.

V budúcnosti bude možné repozitár rozšíriť o nástroje, ktoré by umožnili aj dlhodobú archiváciu dát.

Kapitola 2

Súčasné riešenia

2.1 Repozitáre

Existujú rôzne repozitáre, ktoré sa od seba líšia použitou technológiou, možnosťou rozšírenia, používajú rôzne metadátové schémy. Niektoré sú voľné dostupné ako open source iné ako proprietárny software alebo hosťované aplikácie. V tejto časti je prehľad dostupných aplikácií. Zameriavam sa najmä na vlastnosti, ktoré boli pre ďalší vývoj repozitára kľúčové a to: open source (aby bolo možné software ďalej upravovať), použitie metadátovej schémy, modulárnosť softwaru (jednoduchá možnosť rozšírenia o ďalšie nástroje) a verzovanie (najmä kvôli zdieľaniu a zálohovaniu dát).

2.1.1 Metadáta

Na popis uložených dokumentov slúžia metadáta. Metadáta sú štrukturované dáta nesúce informáciu o primárnych dátach.[2] Kvôli vzájomnej prepojenosti repozitárov, vyhľadávaní dát a správnej interpretácii informácií je snaha o vyvinutie celosvetovo používaného štandardu pre popis dát.

O to sa snažia rôzne metadátove schémy, pomocou ktorých je možné zdroje popísať. Medzi najznámejšie schémy patrí Dublin Core [`http://dublincore.org/`] a MARC [`http://www.loc.gov/marc/`].

2.1.1.1 Dublin Core

Dublin Core (skrátene DC) vznikol s cieľom jednoducho a všeobecne popísať webové zdroje. Táto schéma obsahuje 15 prvkov. To sú: názov (title), autor (creator), predmet (subject), popis (description), vydavateľ (publisher), prispievateľ (contributor), dátum (date), typ (type), formát (format), identifikátor (identifier), zdroj (source), jazyk (language), vzťah (relation), pokrytie (coverage) a práva (rights). Tieto prvky nie sú povinné a môžu sa opakovať. Jednotlivé vlastnosti sú teda pomenované. Ako sa World Wide Web menil, v snahe o vytvorenie sémantického webu, sa vyvinul aj štandard Dublin Core.

Tabuľka 2.1: Typ informácie v kóde MARC

0XX	Kontrolná informácia, identifikačné a klasifikačné čísla,...
1XX	Hlavné údaje
2XX	Názvy a kapitoly (názov, edícia, vydanie)
3XX	Fyzický popis,...
4XX	Informácie o dieloch/sériách
5XX	Poznámky
6XX	Kontaktné informácie na subjekty
7XX	Pridané informácie (iné než o subjektoch, dieloch/sériách); linkovacie polia
8XX	Rada pridaných informácií, informácie o holdingoch
9XX	Vyhradené pre lokálnu implementáciu

Od roku 2008 obsahuje formálne domény a rozsahy v definiciách vlastností. Táto aktualizovaná varianta vlastností sa nazýva dcterms. Jednotlivé prvky môžu byť ďalej rozšírené o kvalifikátor. Ten môže lepšie určiť, čo daná položka popisuje. Napríklad namiesto všeobecného autora tak môžeme upresniť, či išlo o ilustrátora (dc:creator.ilustrator), editora (dc:creator.editor),... Pre systémy, ktoré kvalifikátory nepoužívajú ale musí zostať význam zachovaný.

2.1.1.2 MARC

MARC využívajú najmä knihovníci. Bol navrhnutý pre popis bibliografických údajov v strojovo čitateľnej podobe. Schéma obsahuje vlastnosti, ktoré sú očíslované. Kým názov v dcterms je označený ako title, v MARCu je označený číslom 245 (title proper statement). Na rozdiel od dcterms obsahuje niekoľko pomocných polí (ako napríklad 222 kľúčový názov, 240 unifikovaný názov,...). Takéto označenie je ľahko čitateľné pre stroje, knihovníci si pri každodennej práci s týmito číslami, ich významy zapamätajú. Človek, ktorý ich vidí prvýkrát ale významu nerozumie.

MARC je od DC komplikovanejší, dokáže ale presnejšie popísať zdroj. Prvé číslo v číselných kódoch určuje o aký typ informácie ide, jednotlivé kódy sú popísané v tabuľke 2.1. V prípade kódov 1XX, 4XX, 6XX, 7XX a 8XX sa obsah upresňuje doplnením dvojice čísel. Zvyčajne sa dodržiavajú následujúce dvojice: X00 - Mená osôb, X40 - Bibliografické názvy, X10 - Názvy firiem, X50 - Tématické pojmy, X11 - Názvy stretnutí/konferencií, X51 - Názvy miest, X30 - Jednotné názvy.

Použitie navrhovaného repozitára by malo byť jednoduché aj pre užívateľov, ktorí s metadátami nemajú veľké skúsenosti a nepotrebujú komplikovaný popis dát. Pre skúsenejších užívateľov by však bolo dobré zachovať možnosť použitia zložitejších, prípadne vlastných metadátových schém.

2.1.2 Software

Výpis vlastností najrozšírenejších repozitárov:

2.1.2.1 Digital Commons

[http://digitalcommons.bepress.com/]

Hostovaná platforma inštitucionálneho repozitára. Zameraný na školy a školské dokumenty.

Používa Dublin Core schému, v používateľskom rozhraní podporuje aj iné vlastnosti než len DC, aj keď nepodporuje iné schémy (vrátane MARC).

Autori vedia prispôsobiť repozitár požiadavkám klienta.

Nepodporuje verzovanie.

2.1.2.2 LIBSYS

[http://www.libsys.co.in/]

Proprietárny software. Repozitár funguje ako webová aplikácia.

Používa MARC ako schému metadát.

2.1.2.3 SimpleDL

[http://www.simpledl.com/]

Proprietárny software.

Metadáta na základe Dublin Core. Môžu byť rozšírené o iné schémy.

2.1.2.4 Greenstone

[http://www.greenstone.org/]

Repozitár vyvinutý na Univerzite Waikato.

Používa MARC schému.

Modulárna architektúra, napísaný v jazyku Java. Pluginy v jazyku Perl.

Nepodporuje verzovanie.

Open source

2.1.2.5 Invenio

[http://inveniosoftware.org/]

Software bol pôvodne vyvinutý pre CERN. Umožňuje vytvoriť digitálnu knihovňu alebo repozitár dokumentov dostupný cez web.

Používa špecifikáciu MARC pre metadáta.

Má modulárnu architektúru. Napísaný v jazyku Python.

Podporuje verzovanie uložených dát.

Open Source

2.1.2.6 EPrints

[http://www.eprints.org/]
Vyvinutý na Univerzite Southampton.
Používa rôzne typy metadátových polí, ktoré je možné nastavovať (upraviť zobrazovanie, indexovanie, vyhľadávanie).
Modulárny software napísaný v jazyku Perl.
Podporuje verzovanie dát.
Open Source

2.1.2.7 DSpace

[http://www.dspace.org/]
Software pôvodne vyvinutý MIT a Hewlett-Packard. Od vzniku má viac ako 2000 inštalácií po celom svete.
Ako východziu schému pre popis dát používa Dublin Core, je však možné použiť aj iné schémy.
Ide o súbor spolupracujúcich Java webových aplikácií. K dispozícií je RESTful webové užívateľské rozhranie.
Neumožňuje verzovanie uložených dát.
Open source

2.1.2.8 Fedora

[http://www.fedora-commons.org/]
Je možné použiť rôzne schémy pre popis dát.
Flexibilný, jednoducho rozšíriteľný, modulárny repozitár. Napísaný v programovacom jazyku Java.
Umožňuje verzovanie uložených dát.
Open Source

2.2 Nástroje pre zber a organizáciu chemických dát

Výskumníci v oblasti chémie si vedú laboratórne denníky so záznamami hypotéz, experimentov, analýz alebo interpretáciou experimentov. V súčasnosti sa denníky vedú v elektronickej forme s využitím elektronických laboratórnych denníkov (často sa pre tento software používa skratka ELN). Ústav organickej chémie VŠCHT Praha využíva nasledujúce programy:

2.2.1 E-Notebook

[http://www.cambridgesoft.com/Ensemble/E-notebook/] Software od firmy Perkin Elmer. V súčasnosti je k dispozícii len ako Enterprise verzia s inštaláciou na serveroch Oracle priamo pre koncového zákazníka alebo ako súčasť

cloudových aplikácií Elements `https://elements.perkinelmer.com/` a plánovaného ChemDraw E-notebook `http://chemdrawenotebook.perkinelmer.cloud/`.

2.2.2 Open Enventory

[`https://www.chemie.uni-kl.de/goossen/open-enventory/`] Webová open source aplikácia napísaná v jazyku PHP. Využíva MySQL databázu.

Kapitola 3

Analýza a návrh riešenia

3.1 Analýza požiadavkov

3.1.1 Funkčné požiadavky

Repozitár musí umožniť:

- Uložiť nové dáta.
- Upraviť existujúce dáta.
- Zobraziť existujúce dáta.
- Uložiť históriu zmien dát.
- Vyhľadávať v metadátach.
- Import dát z aplikácie Open Eventory.

Prístup k jednotlivým objektom a poliam ale môže byť limitovaný. Vytváranie a úprava jednotlivých objektov je umožnená len konkrétnym užívateľom.

Aplikácia Open Eventory musí byť upravená tak, aby umožnila export dát vo formáte vhodnom pre import do repozitára.

3.1.2 Požiadavky na vlastnosti repozitára

- Repozitár bude pre užívateľov dostupný ako webová aplikácia.
- Repozitár musí umožniť ďalšie rozšírenie pre iné typy dát.
- Repozitár bude napojený na služby CESNETu.

3.2 Výber technologií

3.2.1 Fedora

Keďže ani jeden existujúci repozitár nespĺňa všetky požiadavky alebo nevie uložiť/zobraziť dáta pre Ústav organickej chémie VŠCHT Praha, bolo potrebné vytvoriť nový alebo upraviť stávajúci software. Vytvorenie nového softwaru od základov by bolo neefektívne. Vyššie zmienené repozitáre fungujú, niektoré ich časti by teda boli programované nanovo. Zvolená bola možnosť doplniť/upraviť funkčnosť existujúceho repozitára. Výber vhodného repozitára bol možný z open source repozitárov, ktoré umožňujú úpravu kódu.

Okrem toho boli, pri výbere vhodného repozitára, do úvahy brané ďalšie kritéria. A to možnosť použitia viacerých metadátových schém, programovací jazyk, podpora komunity. Do finálneho výberu sa dostali DSpace a Fedora. Oba repozitáre sú podobne výkonné (zvládajú milióny záznamov).

Vďaka návrhu Fedory je do tohto softwaru jednoduchšie pridávanie rozšírení, taktiež už má vyriešené verzovanie uložených dát. DSpace má k dispozícii webové užívateľské rozhranie, ktoré je možné upravovať a ďalej rozširovať. Fedora používa jednoduché webové rozhranie, ktoré umožňuje len základnú prácu s dátami. Vytvorenie samostatného, nového webového užívateľského rozhrania s využitím RESTapi je ale jednoduchšie než úprava jadra DSpace, aby zvládal verzovanie uložených dát.

Z existujúcich možností bola zvolená Fedora, ako najvhodnejší software pre možnosť ďalších, potrebných, úprav pre použitie vrámci služieb CESNET z.s.p.o. a splnenie požiadavkov Ústavu organickej chémie VŠCHT Praha.

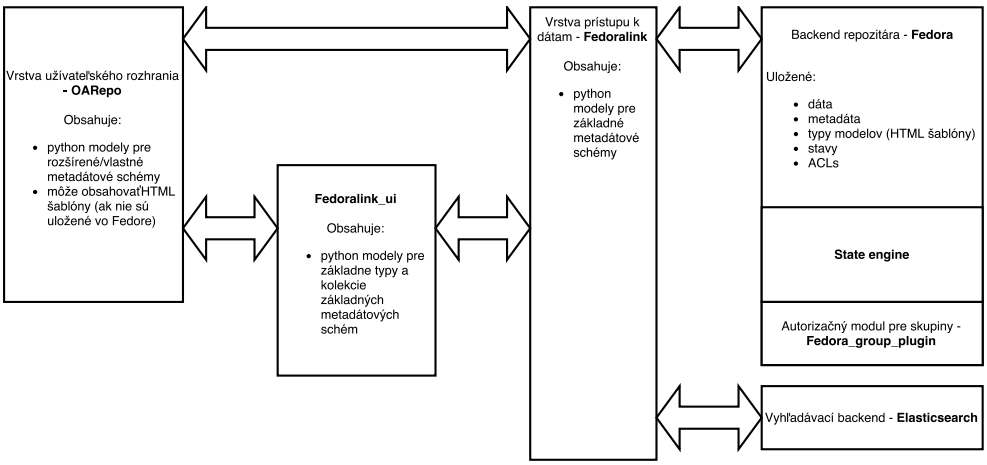
3.2.2 Elasticsearch

Na vyhľadávanie v repozitáry bude použitá samostatná aplikácia Elasticsearch `https://www.elastic.co/products/elasticsearch`. Aplikácia umožňuje veľmi rýchle vyhľadávanie v indexovaných dátach. Dotazi je možné posielať s využitím RESTful api a JSONu.

Kapitola 4

Implementácia

4.1 Design aplikácie



4.1.1 Fedora

Ako už bolo zmienené v predchádzajúcej kapitole, ako backend pre repozitár bola zvolená Fedora. V nej sú uložené dáta, metadáta, typy modelov spolu s HTML šablónami, stavy a oprávnenia (ACL).

4.1.2 State engine

Pre možnosť využívania stavov bude nutné rozšíriť Fedoru o tento modul. Modul rieši prechody medzi stavmi, zmenu stavov, zmenu kontroléru stavov, konkrétnu operáciu povolí len oprávneným osobám. Oprávnené osoby sú určené pomocou ACL.

4.1.3 Fedora_group_plugin

Doplňujúci modul do Fedory, ktorý umožňuje overenie oprávnení aj na základe členstva v django skupinách. Samotná Fedora umožňuje overenie autorizácie na základe On-Behalf-Of hlavičky, toto rozšírenie umožňuje autorizáciu na základe On-Behalf-Of-Django-Groups hlavičky. Autorom pluginu je Mgr. Miroslav Šimek. Plugin je súčasťou git repozitára fedoralinku.

4.1.4 Elasticsearch

Aplikácia, ktorá umožňuje rýchle vyhľadavánie v metadátach.

4.1.5 Fedoralink

Aplikácia napísaná pre potreby repozitára záverečných prác VŠCHT Praha, v programovacom jazyku Python s využitím frameworku Django, stará sa o komunikáciu s Fedorou a Elasticsearch. Autorom fedoralinku je Mgr. Miroslav Šimek. Fedoralink bol počas vývoja repozitára, vrámci diplomovej práce, ďalej upravovaný. Aktuálnu verziu je možné nájsť na `https://github.com/mesemus/fedoralink`

4.1.6 Fedoralink_ui

Súčasťou git repozitára fedoralinku. Modul sa stará o užívateľské rozhranie aplikácie. Pre komunikáciu s Fedorou využíva fedoralink.

generic_urls.py Funkcie vrámci tohto súboru mapujú URL adresy vrámci aplikácie na správne časti kódu pre zobrazenie, editáciu alebo vyhľadávanie. Z URL adresy zistíme ID objektu alebo kolekcie vo Fedore.

Vzory využívajúce regulárne výrazy pre URL adresy:

- '^$' - index
- r'^(?P<collection_id>[a-fA-F0-9_/-]*)extended_search(?P<parameters>.*)$' - vyhľadávanie vrámci kolekcie
- '^(?P<id>.*)/addSubcollection$' - pridanie novej subkolekcie
- '^(?P<id>.*)/add$' - vytvorenie nového objektu ako potomka objektu s daným ID
- '^(?P<id>.*)/edit$' - upravenie objektu s daným ID
- '^(?P<id>.*)$' - zobrazenie objektu s daným ID

O zobrazenie správnych údajov v správnej šablóne, prípadne o vytvorenie nového objektu/kolekcie so správnymi údajmi sa ďalej stará kód v súbore views.py.

V samotnej aplikácií (vo fedoralinku alebo v koncovej aplikácii) musí byť model objektu - trieda v Pythone. Ostatné potrebné veci sú uložené priamo vo Fedore. Jednotlivé typy objektov, pri kolekciách je uložený typ subkolekcií a potomkov.Tieto objekty môžu mať naviac uložené šablóny pre zobrazenie, úpravu a vytvorenie potomkov. Taktiež je možné do Fedory uložiť typy jednotlivých polí a k nim šablóny pre ich zobrazenie/editáciu.

Kód vo views.py teda z ID získa objekt, ku ktorému nájde vo Fedore uložený správny typ. Z neho následne získa šablónu, ktorú zobrazí. Ak sa správna šablóna pre objekt alebo pole nenachádza vo Fedore skúsi ju nájsť v aplikácii alebo použije šablóny uložené vo fedoralink_ui, ktoré umožňujú aspoň základné zobrazenie informácií.

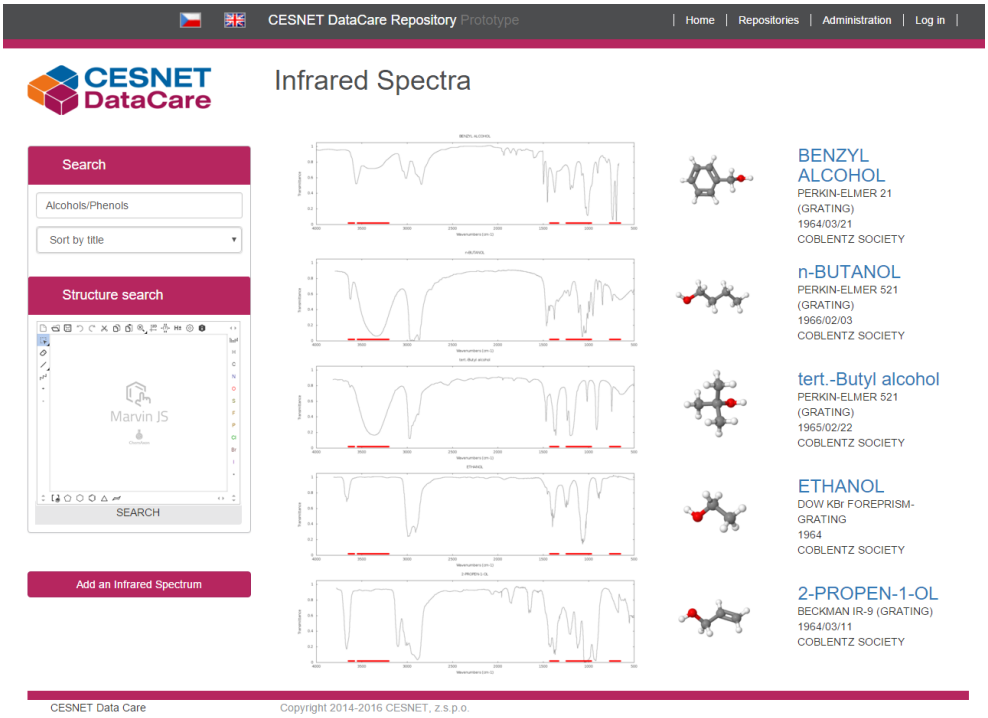
Fedoralink_ui taktiež obashuje kód potrebný pre cachovanie výsledných šablón zložených zo šablón typu objektu a jednotlivých polí, keďže získanie týchto údajov z Fedory je časovo náročné. Pre získanie výslednej šablóny je potrebné množstvo dotazov na Elasticsearch a následne Fedoru, počet dotazov záleží hlavne na komplikovanosti modelu objektu.

Repozitár záverečných prác VŠCHT Praha pôvodne využíval aplikáciu s vlastnými šablónami. Po vzniku fedoralink_ui ale aj tento repozitár začal využívať fedoralink_ui.

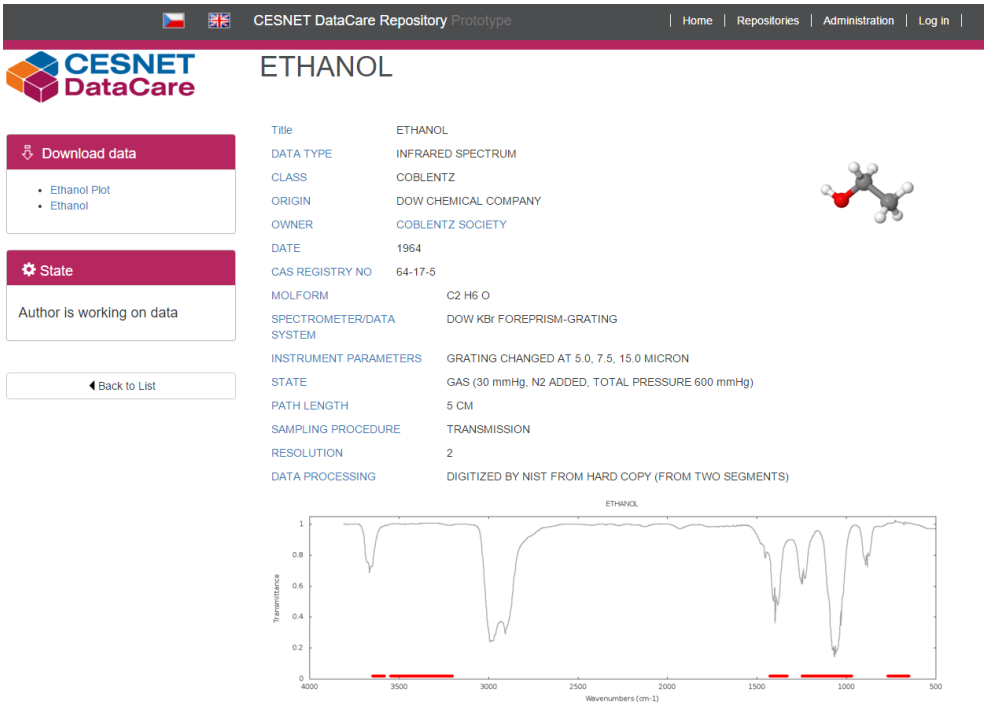
4.1.7 Návrh grafického rozhrania

Repozitár bude nasadený ako jedna zo služieb diskového úložiska CESNET, z.s.p.o. `https://du.cesnet.cz/`, preto navrhnuté grafické rozhranie vychádza z už existujúcich služieb.

Návrh zobrazenia pre dáta organickej chémie:



Obr. 4.1: Výpis dát v kolekcii Infrared Spectra



Obr. 4.2: Zobrazenie detailu konkrétnej položky (Ethanolu)

Záver

Literatúra

[1] Strnad, M.: Svěřte svá data vhodnému médiu – díl 1. *V: LinuxEXPRES [online]*, október 2013, [cit. 2016-11-07]. Dostupné z: `https://www.linuxexpres.cz/praxe/sverte-sva-data-vhodnemu-mediu-dil-1`

[2] Český normalizační institut: *ČSN ISO 8459-5 (01 0175) Informace a dokumentace - sborník bibliografických datových prvků. Část 5, Datové prvky pro výměnu katalogizačních dat a metadata.* 2004.

Dodatok **A**

Zoznam použitých skratiek

ELN Electronic lab notebook

GUI Graphical user interface

XML Extensible markup language

Dodatok **B**

Obsah priloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresár so spustiteľnou formou implementácie
- `src`
  - `impl` ..... zdrojové kódy implementácie
  - `thesis` ..... zdrojová forma práce vo formáte LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce vo formáte PDF
  - `thesis.ps` ..... text práce vo formáte PS