Příloha č. 1 – Technická specifikace Aplikace

Popis

Digitální archiv Archeologické mapy ČR, je webová aplikace určená k prohlížení digitálních dokumentů uložených v Archeologických ústavech AVČR v Praze a Brně. Archivy těchto institucí obsahují dokumentaci terénních archeologických výzkumů na území ČR od roku 1919 do současnosti a vzhledem k zákonné roli ústavů na poli památkové péče se dále rychle rozrůstají. Jde o největší soubory dokumentů k archeologickým výzkumům a nálezům na území ČR a o významnou součást našeho kulturního dědictví. Portál je dostupný na adrese: http://digiarchiv.amapa.cz/

Digitální archiv AMČR obsahuje textové dokumenty (nálezové zprávy, expertní posudky a hlášení), fotografie z terénních výzkumů, letecké snímky, mapy a plány a digitální data (tabulky, databáze, vektorovou grafiku apod.), vždy včetně popisných údajů. Dokumenty a další informace jsou průběžně přebírány z AMČR (http://www.archeologickamapa.cz/), s níž je Digitální archiv propojen i uživatelskými účty. Dokumenty jsou v Digitálním archivu zveřejňovány v souladu s politikou otevřeného přístupu k informacím a se souhlasem jejich majitelů – příslušných odborných institucí. Většina dat a dokumentů je přístupná každému uživateli, menší část z nich až po zřízení uživatelského účtu, příp. pouze uživatelům z oprávněných archeologických organizací. Ke zveřejnění se připravují stovky tisíc dokumentů, o postupu jejich zahrnutí do aplikace informují údaje v základním menu a rubrika Novinky.

Digitální archiv AMČR je součástí Archeologického informačního systému ČR zapsaného do Cestovní mapy ČR velkých infrastruktur pro výzkum, experimentální vývoj a inovace pro léta 2016-2022 Ministerstva školství, mládeže a tělovýchovy, které budování infrastruktury v rámci stejnojmenného výzkumného projektu podporuje. Doplňkovou podporu získává infrastruktura i v rámci dalších výzkumných a rozvojových projektů.

Digitální archiv AMČR je skupina aplikací využívající technologie Java, Angular 2, bootstrap 3, leafletjs s OpenStreetMaps. Skupina aplikací se skládá z klientské a serverové části a ostatních samostatných podpůrných aplikací - Thumbs Generator a Apache Solr.

Zdrojové kódy aplikací

Zdrojové kódy aplikací jsou veřejně dostupné na adrese https://github.com/ARUP-CAS/arup-da-amcr. Součástí GIT repozitáře je i WiKi s popisem aplikací a jejich částí a popis konfiguračních parametrů. Aplikace jsou buildovány v nástroji Maven. Potřebné komponenty jsou Node.js 4 a vyšší, npm 3 a vyšší a angular-cli 1.0.0-beta.21. Apache Solr server je volně ke stažení ze stránek vlastního projektu http://lucene.apache.org/solr/.

Serverová část

O serverovou část se stará Tomcat server. Serverovou část reprezentují jednotlivé servlety pro komunikaci s API AMČR, SOLR, servlety pro komunikaci s klientskou aplikací a ostatní podpůrné servlety.

Aplikace provádí prostřednictvím servletů veškerou komunikaci se Solr pro získání dokumentu z indexu. Prohlížeč nikdy nepřistupuje přímo k Solr procesu. Náhledy obrázků jsou uložené v adresáři specifikovaném konfigurací. Prohlížeč k nim nemá přímý přístup. Aplikace je poskytuje na základě práv uživatelů opět prostřednictvím servletu.

Používané servlety a jejich popis

- **InitServlet** - interní servlet, nastavuje cesty
- **IndexerServlet** - indexer se stará o získávání dat z AMČR API a o indexaci do SOLR
- **ImageServlet** a **PdfServlet** - poskytují náhledy souborů. Soubory nejsou přímo přístupné pro prohlížeč. Servlety testují, zda uživatel má práva získat soubor a pokud ano, tak ho poskytnou
- **StaticServlet** - interní servlet mapuje adresy URL. Zachytává adresy URL
- **SearchServlet** - dává výsledky hledání ze SOLR, SOLR není dostupný zvenčí
- **ConfigServlet** - mergovanání konfigurací, klient nemá přístup do konfigurace. Konfigurace pro serverovou i klientskou část.
- **I18nServlet** - servlet starající se o lokalizace
- **LoginServlet** - provádí přihlášení proti API AMČR a ukládá tato do session
- **FavoritesServlet** - ukládá a vrací obsah oblíbených. Dostupné pro registrované
- **TextsServlet** - servlet slouží k ukládání a vracení editovatelných součástí homepage

Struktura aplikace

Pro Tomcat se očekává vybuildovaná aplikace ve formátu balíku Web Archive. Tento balík se ukládá do adresáře webapps Tomcat serveru. Balík Web Archive obsahuje adresářové struktury:

- / - kromě podadresářů je v balíku statický soubor index.html a klientská JavaScript aplikace
- */assets* - zde jsou vlastnosti, styly, fonty a výchozí konfigurace aplikace
- */META-INF* - adresář metadat aplikace
- */WEB-INF/classes* - obsahuje třídy pro komunikaci se SOLR, generování náhledů, poskytování souborů a ostatní třídy
- */WEB-INF/lib* - obsahuje knihovny aplikace i knihovny používané tomcatem
- */WEB-INF/web.xml* - obsahuje specifikaci servletů

Konfigurace

O načítání konfigurace se stará ConfigServlet. Tento načte výchozí konfiguraci, jak je uchystána po buildu aplikace a následně do této výchozí konfigurace promítne uživatelskou konfiguraci, která leží mimo strukturu aplikace. Toto servlet dělá jak pro klientskou, tak pro serverovou část.

Klientská část

Klientská část obsahující výchozí hodnoty je v souboru *assets/config.json*. Zde jsou zkonfigurovány parametry, který používá JavaScript pro zobrazení, např. překlady, facety, počty řádků. Aplikace se z prohlížeče dotáže na tento soubor. Tato konfigurace je výchozí a mění se s buildem aplikace. Konfigurační soubor v repozitáři https://github.com/incad/arup-da-amcr je zde:

searchapp/src/main/ngClient/src/assets/config.json

Serverová část

V souboru *WEB-INF/classes/cz/incad/arup/searchapp/server_config.json* je serverová část konfigurace. Zde jsou zkonfigurovány parametry připojení do API AMČR, parametry spojení se Apache Solr, překlady, parametry pro zpracování zdrojových dokumentů, parametry uživatelů, indexace, cesty a ostatní. Konfigurační soubor v repozitáři https://github.com/incad/arup-da-amcr je zde:

searchapp/src/main/resources/cz/incad/arup/searchapp/server_config.json

Společná uživatelská část

Veškeré uživatelské změny konfigurace je potřeba dělat mimo root aplikace v souboru *%HOME%/.amcr/config.json*. Tento obsahuje uživatelskou konfiguraci pro klientskou (sekce "client":{ ... }) i serverovou (sekce "server":{ ... }) část aplikace. Soubor obsahuje i přístupy a různá ostatní nastavení. Hodnoty zde uvedené přepíší výchozí hodnoty z výchozích konfiguračních souborů. Umístění této externí konfigurace lze změnit:

- pomocí systemové proměnné
- při startu tomcat přidáním JAVA proměnné *-Damcr_app_dir = adresarkdebudeconfig.json*
- v nastavení context.xml aplikace

Struktura konfiguračního souboru

Viz https://github.com/ARUP-CAS/arup-da-amcr/wiki/Konfigurace

Indexace

Aplikace indexuje existující databázi AMČR. Indexace získává data z exportu do csv formátu. Indexace probíhá dvoufázově:

- v první fázi se indexují všechny pomocné tabulky do dvou core, relations a export
- následně se indexuje core dokument, kde jsou dokumenty uložené s celou vazebnou strukturou

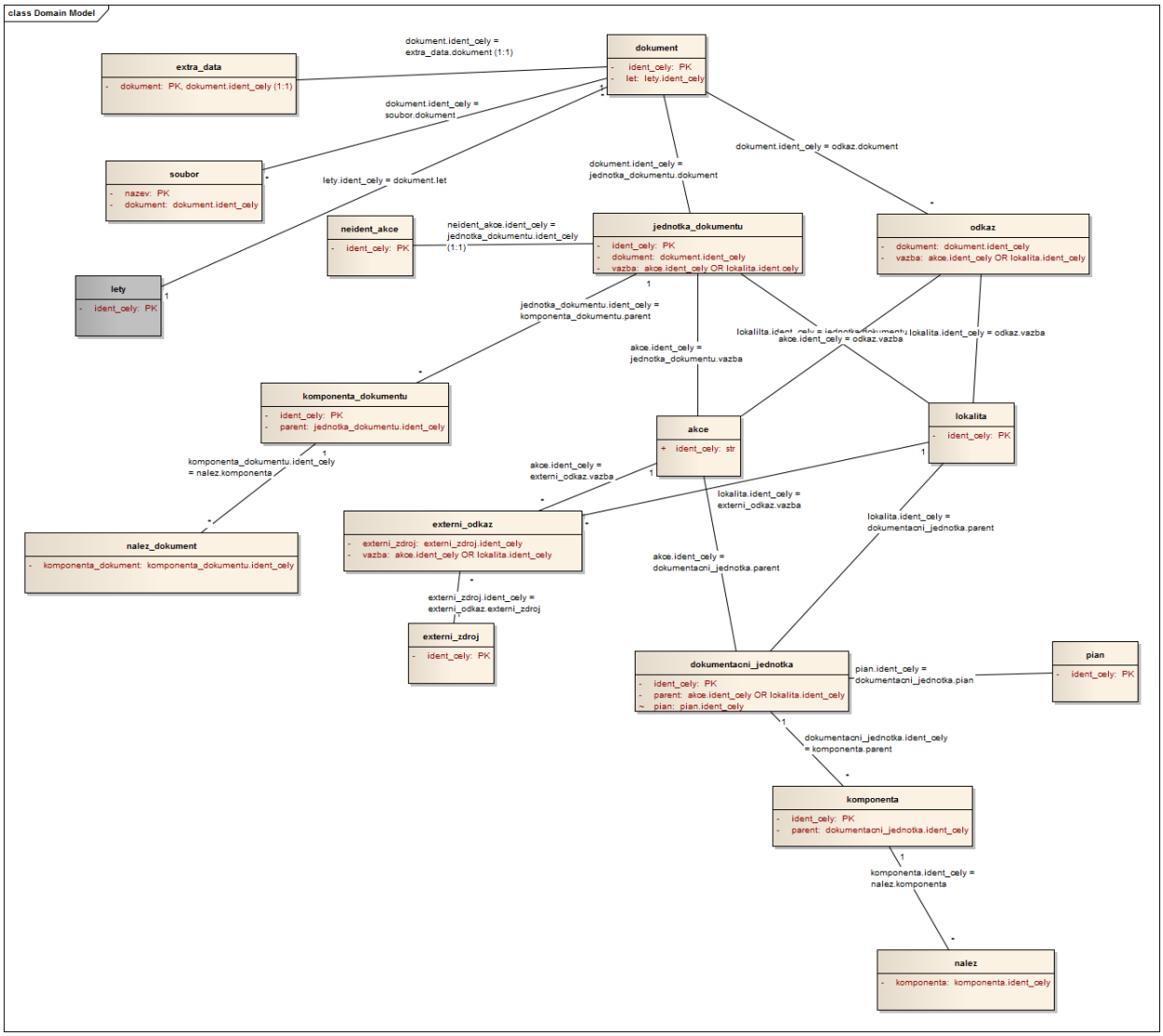
Pro spuštění indexace slouží servlet IndexerServlet. Kompletní reindexaci dat vyvoláme zavoláním URL url_aplikace/indexer?action=FULL. Parametrem clean můžeme předem smazat index vyhledávacího systému, URL: url_aplikace/indexer?action=FULL&clean=true

Pro provedení indexace je třeba být přihlášen do aplikace s oprávněním administrátora.

Další pomocná cores:

- heslář pro zobrazení seznamu hodnot z různých heslářů.
- translations - pro překlad heslářů.
- favorites - uložené oblíbené stránky uživatelů.

Datový model je na schématu níže:



Dotazování

Aplikace odesílá dotazy na index uložený v core "dokument". Při fulltextovém hledáni jsou nalezeny dokumenty obsahující hledaný výraz ve všech polích tabulek. Některé tabulky mají omezený seznam polí:

```
"akce": [
  "ident_cely",
  "stav",
  "pristupnost",
  "poznamka",
  "okres",
  "katastr",
  "dalsi_katastry",
  "vedouci_akce",
```

```
  "organizace",
  "organizace_ostatni",
  "vedouci_akce_ostatni",
  "hlavni_typ",
  "vedlejsi_typ",
  "datum_zahajeni_v",
  "datum_ukonceni_v",
  "lokalizace",
  "ulozeni_nalezu"
],

"dokumentacni_jednotka":[
  "ident_cely",
  "parent",
  "nazev",
  "typ",
  "pian"
],

"externi_zdroj": [
  "nazev",
  "podnazev",
  "typ_dokumentu",
  "oznaceni",
  "organizace",
  "sbornik_editor",
  "sbornik_nazev",
  "edice_rada",
  "casopis_denik_nazev",
  "casopis_rocnik",
  "datum_rd",
  "misto",
  "vydavatel",
  "isbn",
  "issn",
  "ident_cely",
  "sysno",
  "paginace_titulu"
],

"extra_data": [
  "cislo_objektu",
  "pas",
  "easting",
  "northing",
  "zachovalost",
  "nahrada",
  "pocet_variant_originalu",
  "odkaz",
  "format",
  "vyska",
  "sirka",
  "meritko",
```

```
  "zeme",
  "region",
  "udalost",
  "udalost_typ",
  "rok_od",
  "rok_do",
  "osoby"
],

"komponenta": [
  "ident_cely",
  "obdobi",
  "obdobi_poradi",
  "jistota",
  "presna_datace",
  "areal",
  "aktivita_sidlistni",
  "aktivita_pohrebni",
  "aktivita_vyrobni",
  "aktivita_tezebni",
  "aktivita_kultovni",
  "aktivita_komunikace",
  "aktivita_deponovani",
  "aktivita_boj",
  "aktivita_jina",
  "aktivita_intruze",
  "poznamka"
],

"komponenta_dokumentu":[
  "ident_cely",
  "parent",
  "poradi",
  "obdobi",
  "obdobi_poradi",
  "jistota",
  "presna_datace",
  "areal",
  "aktivita_sidlistni",
  "aktivita_pohrebni",
  "aktivita_vyrobni",
  "aktivita_tezebni",
  "aktivita_kultovni",
  "aktivita_komunikace",
  "aktivita_deponovani",
  "aktivita_boj",
  "aktivita_jina",
  "aktivita_intruze",
  "poznamka"
],

"lokalita": [
  "ident_cely",
```

```
  "stav",
  "pristupnost",
  "okres",
  "katastr",
  "dalsi_katastry",
  "typ_lokality",
  "vedouci_akce",
  "vedlejsi_typ",
  "vedouci_akce_ostatni",
  "organizace_ostatni",
  "nazev",
  "druh",
  "popis",
  "poznamka"
],

"nalez": [
  "druh_nalezu",
  "specifikace",
  "pocet",
  "poznamka"
],

"neident_akce":[
  "ident_cely",
  "stav",
  "pristupnost",
  "okres",
  "katastr",
  "vedouci",
  "rok_zahajeni",
  "rok_ukonceni",
  "lokalizace",
  "popis",
  "pian",
  "poznamka"
],

"soubor": [
  "nazev",
  "rozsah",
  "mimetype",
  "size_bytes",
  "filepath"
],

"pian": [
  "ident_cely",
  "geom",
  "presnost",
  "typ",
  "centroid_e",
  "centroid_n"
```

]

Lokalizace

Lokalizační soubory mají JSON formát. Tyto soubory jsou umístěné v adresáři i18n. Pro každý jazyk musí být samostatný soubor s názvem *xx.json*, kde *xx* je jazykový kód. Například *cs.json*.

Výchozí soubory se nacházejí ve složce */assets/i18n applikace*. Aplikace dohledá ostatní soubory nacházející se v adresáři *%APP_DIR%/i18n/* a sloučí nalezené soubory s těmito výchozími v aplikaci následujícím způsobem:

- Pokud daný klíč existuje ve výchozím souboru, bude přepsán.
- Pokud klíč neexistuje, bude přidán.

Pro překlad heslářů je používán jiný přístup. V adresáři s uživatelskou konfigurací je složka "*thesauri*", která obsahuje seznam přeložených heslářů jako .csv soubory. Tyto soubory se indexují do Solr core translations. Každý záznam má následující pole:

- id
- heslar
- pole
- heslo
- cs
- en

Při volání */assets/i18n/cs.json*, servlet I18n vrátí json, kde jsou sloučené výchozí json soubory a záznamy v Solr core translations. Například uvidíme:

- "zeme_Česká republika": "Česká republika",
- "areal_druha_polní opevnění": "polní opevnění",

které odpovídají obsahu heslářů:

- heslar: zeme, heslo: Česká republika, a obsah pole cs: Česká republika
- heslar: areal_druha, heslo: polní opevnění, a obsah pole cs: polní opevnění

Pokud nějaký klíč v lokalizačním souboru chybí, aplikace zobrazí hodnotu klíče.

Klientská část

Klientská část je JavaScript aplikace Tomcatem předávána prohlížeči klienta, kde se spouští. Klientská část byla uchystána v Angular 2 frameworku s podporou bootstrap a leafletjs s OpenStreetMaps. Klientská část komunikuje s vybranými servlety pod Tomcatem.

Konfigurace

Konfiguraci aplikace získává prostřednictvím ConfigServletu, ten se postará o vydání platné konfigurace, jak je popsáno v předchozí kapitole.

Popis aplikace

Popis vyhledávání v Digitálním archivu AMČR:

- Hledání - provádí se fulltextové vyhledávání v datech na základě dotazu zadaného uživatelem. Dotaz je zadáván do pole "Zadejte dotaz pro vyhledávání v dokumentech".
- Rozšířené hledání - Tlačítko otevírá formulář s opakovaným polem. Dotaz omezuje hledaní výrazu na konkrétní zvolené pole. V dotazu lze kombinovat více polí opakováním pomocí "přidat další kritérium". Výčet polí vychází ze struktury databáze.
- Mapa - zobrazuje PIANy nalezených dokumentů na mapě. Při větším počtu nálezů se mapa zobrazuje jako grafy hustoty výskytu bodů formou heatmapy.

Výsledky vyhledávání je možné zobrazit formou seznamu nebo formou mapy. V obou případech poskytuje aplikace možnost přepnout zobrazení

- Filtry - umožňují počet vyhledaných dokumentů zúžit prostřednictvím hodnot filtrů v levém sloupci. Hodnoty jednotlivých filtrů se rozbalují. Výsledky lze také filtrovat posuvným prvkem časového období nad seznamem výsledků.
- Mapové zobrazení funguje jako jeden z filtrů.
- Řazení výsledků - uspořádá nalezený seznam.
- Možnosti zobrazení - řádky s náhledy, řádky bez náhledů, sloupce.
- Zobrazení záznamu - umožňuje uživateli prohlížet zvětšené náhledy dokumentů, a také stáhnout připojené soubory, má-li uživatel oprávnění. Detailní zobrazení také vypisuje další dostupná metadata, umožňuje tisk a kopírování persitentního identifikátoru.
- Export – vyhledané výsledky je možné zobrazit jako tabulku uzpůsobenou pro tisk (počet zobrazených záznamů je pomocí konfiguračního souboru limitován na 1000)

Uživatelské účty

Přihlášení do aplikace probíhá prostřednictvím AMČR API. Existuji následující typy uživatelů:

- A anonym
- B badatel
- C archeolog
- D archivář
- E administrátor

Thumbs Generator

Samostatnou aplikací je Thumbs Generator pro vytváření náhledů. Thumbs Generator je java aplikace, která běží jako systémová služba pod neprivilegovaným uživatelem a má přístup k digitálním dokumentům systému AMČR a k tabulce v databázi s odkazy na tyto dokumenty. Aplikace generuje náhledy, které ukládá do struktury, ke které mají přístup vybrané servlety pod Tomcatem.

Soubory - náhledy jsou uloženy na disku dle názvu tak, aby rovnoměrně vytvářely adresářovou strukturu. Ta se skládá ze dvou písmen názvu odzadu, a má tři úrovně. Například soubor "MTX200900065.pdf" bude uložen do adresáře:

- malý náhled: .../thumbs/65/00/90/MTX200900065.pdf.jpg
- střední náhledy stránek: .../thumbs/65/00/90/MTX200900065.pdf/0.jpg .../thumbs/65/00/90/MTX200900065.pdf/1.jpg

Jakmile Thumbs Generator začne zpracovávat PDF soubor, tak jeho jméno uloží do souboru **processing.txt**. Pokud dojde k pádu procesu, tak při příštím spuštění procesu dojde ke kontrole obsahu processing.txt a případně je jeho obsah zapsán do souboru **unprocessables.txt**. Proces přeskakuje generování náhledů uvedených v tomto seznamu. Je na lidské obsluze zjistit, co je za problém s konkrétním PDF souborem a vyřešit jeho vnitřní formátování.

Konfigurace

Thumbs Generator sdílí serverovou část konfigurace s ostatními servlety. Významným parametrem je položka *maxpixels*. Při generování náhledu může dojít k OutOfMemory chybě. To lze omezit právě tímto konfiguračním parametrem. Jeho hodnota znamená maximální počet pixelů (šířka x výška) obrázku, který bude zpracován. Obrázky s větším počtem pixelů budou přeskočeny a toto přeskočení bude zalogováno jako "skipping page 3 in file xxx"

Po dokončení aplikace vypíše: "Generate thumbs finished. Files processed ..." a další doplňující informace.

Apache Solr

Je použita standardní instalace Solr (viz Solr Tutorial). Do *%solrhome%/server/solr* umístíme konfigurační soubory, které jsou v repozitáři https://github.com/ARUP-CAS/arup-da-amcr/ v adresáři *solr/*.

Apache Solr běží jako samostatný lokální proces na výchozím portu TCP/8983. Na tento port se obrací vybrané servlety. Pro správné fungování by měl mít proces k dispozici paměť o stejné

velikosti, jako je velikost indexu. V současné době to je 1GB. S růstem počtu dokumentů by se paměť měla navyšovat. Pamět Solru přidělíme pomoci startovacího skriptu parametrem -m (viz Solr Control Script Reference). Velikost paměti lze nastavit také v souboru s konfigurací, např. v *bin/solr.in.sh*, proměnná SOLR_HEAP.

Další informace k související aplikaci AMČR lze nalézt na adrese
http://www.archeologickamapa.cz/help/