Projekt INDIHU – vývoj nástrojů a infrastruktury pro digital humanities

V roce 2016 byl zahájen pětiletý projekt INDIHU – vývoj nástrojů a infrastruktury pro digital humanities. Koordinátorem je Knihovna AV ČR a na projektu s ní spolupracuje také Národní knihovna ČR a šest ústavů Akademie věd ČR. Hlavním cílem projektu je vývoj nástrojů, které usnadní práci s digitálními daty a propojení dat v humanitních oborech. V rámci projektu vznikne zastřešující platforma, která bude sloužit k prohledávání rozmanitých databází z oblasti humanitních věd. Dále bude vyvinut prototyp virtuální výstavy, virtuální znalostní báze a bude integrován open source OCR mechanismus. Projekt bude probíhat do konce roku 2020.

Projekt INDIHU

Knihovna AV ČR, Národní knihovna ČR a šest dalších ústavů Akademie věd ČR ve společném projektu vytváří nástroje a infrastrukturu pro vědeckou práci v humanitních oborech. Projekt s názvem INDIHU – vývoj nástrojů a infrastruktury pro digital humanities byl odstartován v roce 2016 a bude probíhat ještě následující čtyři roky. Informace o projektu jsou zpřístupněny na internetových stránkách *www.indihu.cz.*

Mezi výstupy projektu budou patřit nástroje, jež umožní práci s digitálním obsahem, jeho efektivní využití k výzkumu a použití nových metod a přístupů. Bude připraveno softwarové řešení pro virtuální výstavy umožňující využití různých zdrojů dat. Vznikne softwarové řešení pro virtuální znalostní bázi, jež bude přehledným informačním systémem a zdrojem pro jednotlivé badatele, a zároveň bude integrován open source mechanismus, který umožní vytvoření strojově čitelných plných textů, které budou sloužit k efektivnějšímu výzkumu a vyhledávání v požadovaných dokumentech a publikacích.

Do řešení projektu je zapojeno těchto osm institucí:

- Knihovna AV ČR
- Národní knihovna ČR
- Archeologický ústav AV ČR, Brno
- Archeologický ústav AV ČR, Praha
- Ústav pro českou literaturu AV ČR
- Filosofický ústav AV ČR
- Etnologický ústav AV ČR
- Ústav dějin umění AV ČR

Zapojené instituce spojuje zájem systematicky propojit své digitální kolekce dat a svá výzkumná data a společně je využívat. Velký důraz je v rámci projektu kladen také na napojení na mezinárodní infrastrukturu v oblasti digital humanities, jako jsou například DARIAH a Europeana.

Cíle projektu

Hlavním cílem projektu je vytvořit potřebné nástroje a infrastrukturu pro vědeckou práci v humanitních oborech, jejímž základem nebudou dle stávajících zvyklostí analogové dokumenty, nýbrž zdroje dostupné v elektronické podobě – například digitalizované dokumenty, elektronické databáze či různé separátní kolekce elektronických dat vědecké povahy, jež vznikají na akademických pracovištích. Tato data dosud nejsou používána v uspokojivém množství, nemluvě o absenci jejich systematického propojení.

Projekt má tři hlavní cíle, v rámci jednoho z nich vzniknou tři softwarové nástroje:

1. vytvoření zastřešující platformy INDIHU,
2. vývoj nástrojů pro vědeckou práci a prezentaci kulturního dědictví v digitálním světě,
   a) prototyp virtuální výstavy,
   b) virtuální znalostní báze,
   c) integrace open source OCR mechanismu,
3. vzdělávání vysokoškolských studentů v oblasti použití digitálního obsahu v rámci digital humanities.

Vytvoření zastřešující platformy INDIHU

Centrální řešení INDIHU bude sloužit pro integraci dat z oblasti humanitních věd. Jednotlivé zapojené ústavy AV ČR, Knihovna AV ČR a Národní knihovna ČR budou v rámci projektu propojovat své zdroje, které doposud fungují separátně, mnohdy o nich navzájem není velké povědomí a je prakticky nemožné je používat společně pro tvorbu konkrétních analýz a specifických výzkumů. Předpokladem vzniku jednotné platformy INDIHU tak bude vytvoření funkčního prostředí pro mapování metadat, jež jsou k těmto zdrojům dostupné, a rovněž vytvoření modelu pro napojení dalších obdobných zdrojů. Scelení doposud separátních zdrojů a jejich propojení do společného prostředí tak významně přispěje k větší efektivitě komparativních výzkumů historické a kulturní kontinuity a k jejich zpřístupňování širšímu okruhu odborné veřejnosti.

Vývoj nástrojů pro vědeckou práci a prezentaci kulturního dědictví v digitálním světě

Práce s digitálním obsahem do jisté míry kopíruje způsob vědeckého použití analogových zdrojů. Digitální data lze stejně jako originální předměty prezentovat v rámci specificky zaměřených výstav, lze s nimi pracovat v rámci budování odborných kartoték nebo je excerpovat pro tvorbu seznamů literatury či poznámkového aparátu. Zdroje převedené do elektronické formy však mají svou přidanou hodnotu, neboť se nad ně dají nasadit efektivní mechanismy umožňující přesné vyhledávání potřebného obsahu, strojové rychlé zpracování či automatické vyčítání statistických údajů apod. To může urychlit dosud zdlouhavé výzkumy. Konkrétní nástroje nahrazující ruční práci, jež operují i s digitálním obsahem a jsou volně dostupné badatelské obci, však až na výjimky chybějí. Mezi výstupy předkládaného projektu tak bude patřit i vývoj specifických nástrojů, jež by mohly tuto mezeru zacelit. Všechny výstupy, které v rámci projektu vzniknou, budou šířené pod open source licencí a bude je tak moci využít jakákoliv instituce. Jedná se konkrétně o prototyp virtuální výstavy, virtuální znalostní bázi a integrovaný open source mechanismus.

Prototyp virtuální výstavy

Prototyp virtuální výstavy by vznikl propojením izolovaných řešení k prezentaci kulturního dědictví (interaktivní časová osa, referencování osob či fotografií, zpřístupnění digitálních 3D objektů, interaktivní myšlenková mapa, virtuální prohlídka místnosti na základě snímání QR kódů atd.); taková prezentace by pro menší paměťové instituce sloužila jako základní open source nástroj pro vyhotovení virtuální výstavy, která by se dala publikovat v prostředí internetu nebo uzavřené sítě. Další využití tohoto prototypu/softwaru by bylo určeno k didaktickým účelům, kdy by se univerzitní studenti mohli v rámci výuky učit sestavovat takovéto výstavy.

Virtuální znalostní báze

Virtuální znalostní báze by měla nahradit dosud používané tzv. excerpční karty, jež se standardně používají v různých vědních oborech (například archeologické nálezové zprávy, slovní excerpce jazykových ústavů atd.). V klasické podobě je práce s takovouto znalostní bází obtížná, zejména vyhledání potřebné informace pořízené před několika lety, někdy v důsledku špatné čitelnosti starších výpisků, chybného založení karet, jejich ztráty atd. V současné době se samozřejmě nabízí budování osobní znalostní báze v elektronické podobě, která umožní překonat většinu uvedených problémů a může nabídnout řadu dalších funkcí, služeb, nástrojů, přímého propojení na informační zdroje a způsobů využití. Součástí budovaného prostředí budou i interaktivní rešeršní služby a RSS kanály, komunikační nástroje pro práci v týmu (tracking listy, sdílení dat, virtuální konference aj.) či trend tracking tool management pro sledování trendů odborné literatury i trendů ve vlastní znalostní bázi.

Integrace open source OCR mechanismu

Integrace open source OCR mechanismu by měla badatelům umožnit transformaci osobního digitálního obsahu do formy, která by skýtala možnost full-textového vyhledávání a propojení do osobní virtuální kartotéky, jež by napomohlo automatickému zhotovování citací a jejich vkládání do vědeckých prací.

Vzdělávání vysokoškolských studentů v oblasti použití digitálního obsahu v rámci digital humanities

Součástí tohoto projektu bude příprava série workshopů pro studenty a pedagogy vysokých škol humanitního zaměření za účelem seznámení se s výstupy řešení projektu a v obecné rovině i se specifiky digital humanities a digital curation.

Harmonogram projektu

Pětiletý projekt INDIHU probíhá od 1. 3. 2016 a bude ukončen 31. 12. 2020. Projekt je rozdělen do tří fází. První fáze, analytická příprava softwaru, proběhla v roce 2016. Druhá fáze, vývoj softwarových komponent a prezentace výstupů projektu odborné veřejnosti, bude probíhat v letech 2017–2018. Závěrečná třetí fáze, zprovoznění zastřešující platformy INDIHU a představení práce s propojeným obsahem studentům VŠ, je naplánována na roky 2019–2020.

Pro vytvoření jednotlivých nástrojů, které patří mezi hlavní výstupy projektu, byla v prvním roce řešení projektu zpracována detailní analýza. V následujících letech budou řešitelé z jednotlivých ústavů AV ČR průběžně spolupracovat při zapojení jimi provozovaných zdrojů do platformy INDIHU i při návrhu a testování nově vyvinutých řešení. Po schválení vstupní analýzy budou pokračovat programovací práce na jednotlivých softwarových výstupech včetně průběžného testování. V poslední fázi bude zpracováno a otestováno uživatelské prostředí. Tyto činnosti Použitá literatura:
LHOTÁK, Martin. Souhrnná roční zpráva o průběhu a dílčích výsledcích řešení projektu za rok 2016. Praha, 2016.
by měly být uzavřeny do roku 2018. Průběžně bude vznikat analýza funkčních vlastností zastřešující platformy INDIHU. Ta bude v roce 2018 předložena dodavateli, který zabezpečí programovací práce. Následovat bude opět testování a v roce 2020 po schválení řešiteli bude platforma spuštěna pro veřejnost.

Vývoj projektu v prvním roce

V roce 2016 proběhla 1. etapa projektu – Analytická příprava softwarových řešení, v jejímž rámci byly řešeny tyto dílčí cíle:

- zpracování vstupních analýz propojovaných zdrojů a
- zpracování definicí funkčních vlastností jednotlivých aplikací jakožto podkladů pro programování.

Zpracování vstupních analýz propojovaných zdrojů

Do projektu INDIHU bylo přihlášeno 22 databází, provozovaných v sedmi institucích. Analýzou dat z těchto databází bylo zjištěno, že pro indexování v rámci společného vyhledávání je momentálně vhodných třináct databází. Další tři databáze byly navrženy pro pozdější zapojení do společného vyhledávání, protože vyžadují větší úpravy dat nebo případně jejich doplnění. Kritéria pro zařazení do společného vyhledávacího systému byla následující:

- databáze je přístupná pro uživatele,
- metadata v databázi jsou ve formátu vhodném pro indexaci vyhledávacím nástrojem,
- metadata obsahují alespoň základní pole, nutná pro indexaci,
- metadata obsahují trvalé odkazy na metadatový záznam a/nebo na digitální objekt v původní databázi,
- databáze obsahuje náhledy digitálních objektů (týká se pouze obrazových dat).

Zbylých osm databází nesplňuje většinu kritérií pro zařazení do společného vyhledávání nebo již nejsou funkční, proto budou využity pouze jako odkaz na existující zdroje dat.

Příprava podkladů pro vytvoření softwarového nástroje pro virtuální výstavy

Řešení projektu v této oblasti započalo zpracováním rešerše aktuálních trendů, zejména zahraniční provenience. Cílem bylo zjistit, jak k tématu virtuálních výstav přistupují v jiných institucích a jaké funkcionality virtuální výstavy obvykle nabízejí. V dalším kroku byla provedena analýza poskytovaných forem obsahu od zapojených institucí. Jednalo se o tyto formy obsahu – text, obraz, časová osa, mapa, 3D objekty, video, audio, vizualizace/infografika. Přitom první tři jmenované se vyskytují u všech subjektů.

V kombinaci s rešerší byly definovány funkcionality, které bude virtuální výstava nabízet; jednotlivé prvky jsou se základním popisem fungování uvedeny na *http://indihu.dipozitiv.cz/.* Ukázalo se, že jedním z koncepčních přístupů je chronologické pojetí virtuální výstavy, které bude vycházet z časové osy. Druhým je topologické uspořádání, skrze nějaký, byť imaginární prostor, tedy mapu. Posledním způsobem je uspořádání tematické. Tematický přístup by měl pracovat s různými kapitolami, které se k tématu váží.

V závěru této fáze projektu bylo rozhodnuto, že virtuální výstava bude nabízet přibližně pět výstavních šablon, které bude moci uživatel využít a vytvořit vlastní virtuální výstavu.

Vytvoření softwarového nástroje pro virtuální znalostní báze

V této části se řešitelský tým věnuje vytvoření nástroje zaměřeného na podporu individuální vědecké práce využívající excerpce textových a jiných informací z různých typů dokumentů pro budování znalostní báze.

Virtuální znalostní báze bude databázovou aplikací zajišťující různé funkce podporující několik činností při práci s textovými (vč. grafů a tabulek), obrazovými a dalšími typy informací při zpracování a vytěžování vědeckých a zdrojových dokumentů. Jedná se zejména o pořizování excerpcí, jejich třídění, ukládání a vyhledávání, pořizování citací, pohotové získávání informací z knihovních katalogů a plných textů z digitálních knihoven. Virtuální znalostní báze bude mít následující funkční komponenty:

- virtuální excerpční kartotéka
- automaticky generovaný rejstřík literatury
- osobní báze dokumentů
- napojení na katalogy největších knihoven
- napojení na digitální knihovny
- napojení na generátor citací
- trend tracking systém
- aplikační rozhraní

Většina atributů bude sloužit pro vyhledání excerpované informace. Ostatní funkční komponenty budou sloužit pro pohotové získávání informací z externích zdrojů, pro jejich další zpracování a uchovávání prostřednictvím aplikace, pro automatické generování citací, případně pro další operace a funkce. Jedná se v ČR o nově řešenou problematiku. Hlavní přínos v této oblasti lze spatřovat v racionalizaci individuální vědecké a odborné práce, v poskytnutí nástroje pro systematické třídění poměrně heterogenních typů informací a v automatizaci některých časově náročných operací.

Integrace open source OCR mechanismu

V rámci výzkumných aktivit uskutečněných v roce 2016 byl v oblasti integrace open source nástroje pro OCR podniknut rozsáhlý průzkum stávajících systémů, které byly navzájem porovnány. Uvedené srovnání umožnilo zjistit výhody nejen open source řešení, jež jsou klíčová pro naplnění cílů projektu, ale došlo i k jejich poměření ke komerčním produktům společnosti ABBYY. Jakožto možná open source řešení (nejen samotné OCR mechanismy, ale i návazné softwarové nástroje) byly zvoleny Aletheia, OCRopus a Tesseract. Jakožto nejvhodnější možnost pro využití v rámci aktivit projektu byl dle stávajícího stavu poznání zvolen OCR mechanismus Tesseract, a to zejména z důvodů stabilní komunity uživatelů a vývojářů, otevřenosti řešení pro různé operační systémy (Unix, Mac, Windows) a dobré dokumentace na GitHubu. Limity do budoucna, na kterých bude řešitelský tým dále pracovat s externím dodavatelem programátorských prací, jsou spojené zejména s absencí GUI rozhraní – pro ovládání je používán pouze příkazový řádek. Největší výhodou systému je pak možnost budovat vlastní vzdělávací databáze, které systému umožňují postupně vylepšovat výsledky OCR.

Plánovaný vývoj projektu v roce 2017

V roce 2017 bude zahájena další etapa řešení projektu, která bude končit v prosinci 2018. V jejím rámci budou vytvořeny softwarové nástroje pro podporu oblasti Digital Humanities – Virtuální znalostní báze, Virtuální výstava, Integrace open source OCR nástroje, které budou odevzdány v roce 2018, a dále bude uspořádán tematický workshop pro pracovníky AV ČR.

Použitá literatura:

LHOTÁK, Martin. Souhrnná roční zpráva o průběhu a dílčích výsledcích řešení projektu za rok 2016. Praha, 2016.

Mgr. Ivana Šlapáková, slapakova@lib.cas.cz
Ing. Martin Lhoták lhotak@lib.cas.cz
(Knihovna Akademie věd ČR)