Tvoříme datovou platformu pro humanitní vědy

Doporučení pro tvůrce platforem

Zpracováno v projektu LINDAT/CLARIAH-CZ, Brno, březen 2022

Připravujete digitální sbírku, archiv, databázi nebo jinou platformu pro výzkum a přemýšlíte o jednotlivých krocích, přístupech a problémech? Pak je tento materiál coby základní uvedení do problematiky určen právě vám. Rádi se s vámi i setkáme a budeme se hlouběji věnovat vašim potřebám – ozvěte se nám prosím na e-mail lindat-clariah@phil.muni.cz *nebo si rezervujte osobní konzultaci.*

Každý originální výzkum začíná přípravou datové základny. Někdy lze využít už existujících informačních zdrojů, faktografických databází, datových sestav apod. Neexistují-li takové podklady pro zamýšlený výzkum nebo nemají-li odpovídající pokrytí, vhodnou strukturu či patřičnou validitu, dojde na tvorbu vlastní datové platformy. Tvůrce takového systému prochází komplexním procesem, v němž činí řadu rozhodnutí povahy koncepční, odborné, manažerské, technické. Tento materiál jsme připravili jako základní metodickou podporu tohoto procesu.

Následující doporučení jsou obecná, bez vazby na konkrétní technické řešení. První okruh doporučení – SEDMERO PRVOTNÍCH ÚVAH – se věnuje základním koncepčním rozhodnutím, druhý – PĚT KROKŮ K CÍLI – představuje doporučení jednotlivých postupů, metod a praktických řešení.

SEDMERO PRVOTNÍCH ÚVAH

Tato část obsahuje podněty pro úvodní rozvahu tvůrce datové platformy před tím, než se pustí do vlastní přípravy. Při plánování celého procesu je třeba myslet nejen na obsah, rozsah a strukturu zamýšlené databáze, ale i na dokumentaci jednotlivých kroků a výsledků, právní řešení přístupu veřejnosti k obsahu databáze či interakci uživatelů s informačním zdrojem a začlenění nově vzniklé platformy do datového univerza.

1) **Stanovení účelu a žádoucích funkcí**
   Základní otázky můžeme shrnout jako: Proč by měla datová platforma existovat? Co má být jejím obsahem? Kdo s ní bude pracovat a jakým způsobem?

   Důvod existence platformy je pravděpodobně dán výzkumným projektem a absencí jiného vyhovujícího zdroje. S platformou obvykle pracuje její správce (výzkumný tým) při přidávání a úpravě obsahu. Tento výzkumný tým ji dále využívá i po naplnění obsahem. Pokud je platforma veřejně přístupná, rozšiřuje se skupina potenciálních uživatelů a jejich potřeb. Potřeby uživatelů a žádoucí funkce je možné popsat více způsoby, od prostého textového popisu po různé typy modelů. Existuje také množství metod pro zjištění těchto potřeb (více bod 6). Pro funkčnost jsou zásadní také rozhodnutí týkající se popisu obsahu (více bod 2).

   Další požadavky se mohou týkat dat o využívání platformy. Například je vhodné určit, zda chceme sledovat počty stažení souborů nebo zobrazení jednotlivých záznamů a zda tato data chceme zobrazit veřejně přímo v systému.

2) **Popis obsahu**
   Aby bylo možné se sbírkovými objekty (naskenovanými materiály, daty, fotografiemi, texty aj.) dále pracovat, je potřeba je vhodně popsat. Komplexnost popisu se odvíjí od potřeb uživatelů, v první řadě samotného výzkumného týmu, který sbírku tvoří. Určujícím faktorem je také množství času, který je pro tvorbu k dispozici. Při formulování pravidel popisu je potřeba rozhodnout, jaké údaje zahrnout, odkud je zjišťovat a jakým způsobem je zapisovat. Všechna rozhodnutí o podobě popisu by měla být zdokumentována a dostupná všem,

kdo budou s obsahem pracovat. Tato rozhodnutí budou ovlivněna mimo jiné oborovými zvyklostmi a měla by být učiněna s ohledem na budoucí využitelnost a udržitelnost platformy a jejího obsahu. Lišit se bude také způsob, jakým budou popisné informace uloženy, a to v závislosti na zvoleném technickém řešení.

3) **Dokumentace postupu i výsledku**
Na tvorbu dokumentace často nezbývá čas, přesto je nezbytná pro plné využití potenciálu pracně budované platformy. Technická dokumentace je zásadní pro případné strojové zpracování obsažených dat nebo jejich přesun do jiného systému (například pokud stávající řešení zastará a přestane správně fungovat). Pro inspiraci lze využít nabízenou Šablonu dokumentace (PDF) se základními okruhy, kterým je vhodné se věnovat.
Pro lepší využitelnost platformy je důležitá také dokumentace rozhodnutí, která vedla k jejímu vzniku a formovala její podobu. Základní informace o obsahu a cílech projektu by měly být dostupné rovněž na stránkách platformy, má-li veřejnou adresu.

4) **Zajištění autorských práv k vystavovanému obsahu**
Pokud je součástí platformy digitalizovaný obsah, je nutné vědět, zda a jak s ním lze nakládat. Je také nezbytné uvést vlastníky autorských práv. Speciální pozornost je pak potřeba věnovat osobním a citlivým datům, jsou-li ve sbírce obsažena.
Pokud uběhlo více než 70 let od úmrtí autora digitalizovaného obsahu, pak lze s dílem automaticky nakládat jako s volným dílem. Pokud však této ochranné lhůty nebylo dosaženo, pak je třeba dodržet ustanovení podle § 27b odst. 3 autorského zákona. V případě, že vlastníka autorských práv nedohledáme ve zmíněných zdrojích, pak postupujeme podle § 37a a zažádáme o patřičnou licenci.

5) **Volba licence pro poskytování dat a jejich citování**
Nejen u volně dostupných sbírek se nabízí otázka dalšího využití zahrnutých dat. Pokud jste sami tvůrci jejího obsahu nebo jeho části, je vhodné uvést na stránkách, pod jakou licencí je tento obsah poskytován. Je možné využít například licence Creative Commons. Jako nejméně restriktivní je doporučována varianta CC BY 4.0.
Licencovat lze také samotnou databázi jako celek. Ta může být chráněna autorským právem, pokud jde o tzv. originální databázi, nebo právem pořizovatele databáze. Více informací a názorný diagram lze najít na stránkách Otevřená data.
Pro další práci s obsahem je důležité jednoznačně určit, kdo je jeho tvůrcem a kdy byl obsah vytvořen a publikován. Dále je zvykem uvést, jak by měla být platforma správně citována, ať už jako celek, nebo jednotlivé obsažené položky.

6) **Péče o uživatele**
Uživateli platforem pro výzkum jsou obvykle samotní tvůrci. Pokud má platforma potenciál dalšího využití, je vhodné zapojit do jejího vzniku více osob a zjišťovat, jaké jsou jejich potřeby. Důležitou součástí tvorby webového obsahu a služeb je také přístupnost pro uživatele s různými druhy omezení.
Díky prototypům (průběžným, nehotovým verzím) je možné platformu otestovat a zjistit, zda je pro uživatele srozumitelná. Pro uživatelské testování je vhodné si připravit zadání úkolů, které mají testeři provést. Základní informace o realizaci uživatelského testování nabízí například web 100 metod, spolu s dalšími metodami pro zjišťování uživatelských potřeb.
Po spuštění lze zkoumat, jak uživatelé platformu používají. Pracovat lze s daty generovanými samotným systémem (např. s pomocí nástrojů Google Analytics nebo HotJar), nebo metodami jako pozorování či stínování uživatelů.

7) **Sdílení informací a (meta)dat**
Aby se o platformě dozvěděli potenciální uživatelé, vyplatí se zjistit, zda existuje rejstřík platforem podobného zaměření, kam by bylo možné ji zaregistrovat. Platformy vzniklé na Filozofické fakultě MU shromažďujeme v katalogu na stránkách infrastruktury Digitalia MUNI ARTS. Dále existují společné vyhledávací služby, které sbírají metadata z více platforem. V takovém případě je obvykle potřeba, aby metadata splňovala stanovené požadavky.

PĚT KROKŮ K CÍLI

Tato část má být pomocníkem na cestě od prvotní výzkumné myšlenky až ke kvalitní datové platformě, jejímu zpřístupnění, prezentaci a využití. Je orientována na odborné otázky, každý krok ovšem vyžaduje i manažerská a technická řešení – k těm rovněž uvádíme vybrané podněty. Rozdělení do jednotlivých kroků je do jisté míry arbitrární, v praxi může probíhat plnění některých kroků (zčásti) souběžně.

Konkrétní příklady jsme se pokusili přiblížit prostřednictvím fiktivního badatele Alberta, který svými přístupy a pochybnostmi může podnítit další otázky.

Krok 1: CO CHCEME

Neboli: Základní pojetí, koncepce a účel připravované datové platformy.

Kde začít: Vyjasněním cílů chystaného výzkumu – co, jak, kde, kdy, jakým způsobem budeme zkoumat. Kdo a jak bude do výzkumu zapojen. Jak budou vypadat výsledky výzkumu. Jaké typy dat bude výzkum generovat, jak budou popsána (metadata), strukturována, formátována, provázána s dalšími datovými zdroji. Kdo a jakým způsobem bude tato data využívat. Jak bude zajištěna prezentace a propagace vzniklé datové platformy. Kdo, v jakých intervalech a jakými metodami bude datovou platformu udržovat, doplňovat a aktualizovat.

Možné překážky a problémy: Rozsah nebo strukturu dat nelze vždy s konečnou platností, přesností a úplností stanovit předem, protože se vyjevuje až v průběhu výzkumu. Výzkumem může být zjištěn parametr, vlastnost, charakteristika dat, která na počátku nebyla známa či se o ní neuvažovalo jako o podstatném hledisku analýzy.

Výstup tohoto kroku: Textový dokument shrnující všechna podstatná rozhodnutí uvedená výše, i s případnými nejasnostmi, pochybnostmi apod.

Manažerské otázky: Mám pro svůj výzkum dostatečné finanční zabezpečení? Je v něm rezervována položka na vznik datové platformy? Byl pro tuto proceduru vytvořen patřičný časový rámec? Zahrnuje můj tým všechny potřebné profese? Připravil/a jsem pro něj efektivní komunikační platformu?

Technické řešení: Programátoři jsou od počátku zapojeni do činnosti týmu a seznámeni s obsahem projektu. Navrhují variantní řešení komplexních otázek (volba SW, zálohování dat, uživatelské rozhraní atd.).

Tým LINDAT/CLARIAH-CZ pomůže mj.: formou konzultace, zaměřené zejména na to, zda nic podstatného v tomto kroku nebylo opomenuto, zda je dostatečně zřejmý rozsah odborné a časové investice do přípravy datové platformy.

Poznámky badatele Alberta: Zatím je to jasné. Problematice jazyka českých dětských komiksových časopisů se věnuji dlouhodobě, přehled o primární i sekundární literatuře mám dokonalý. Budeme zkoumat jazyk výpovědí hrdinů časopisu Čtyřlístek od jeho vzniku (1969) do roku 2020 a sledovat v něm intertextové a extratextové vazby. Můj tým zahrnuje lingvistu, literární vědkyni, pedagoga, specialistu na komiksový žánr a soukromého badatele orientujícího se výhradně na časopis Čtyřlístek. Potřebovat budeme ještě programátora. Výstupem bude strukturovaný korpus textů s vazbami na další informační objekty. Primárně bude sloužit lingvistům a literárním vědcům, užitečný bude i pro další badatele, laická skupina zájemců o Čtyřlístek je také velká. Přesná struktura dat a metadat ještě není určena, je to jeden z důležitých dalších kroků našeho týmu. Přesnou představu o prezentaci a aktualizaci vzniklé databáze ještě nemáme. Určitě bych využil konzultaci s týmem LINDAT/CLARIAH-CZ, abych věděl, že nic podstatného neuteklo.

Krok 2: KDE VEZMEME DATA

Neboli: Metody sběru a tvorby dat, jejich strukturace.

Kde začít: Identifikací všech dostupných pramenů dat nejrůznějších typů (klasických i elektronických), které se mohou stát dílčími zdroji dat. Připravit metody excerpce těchto dat, stanovit strukturu jejich uložení a zajistit kontrolu validity. Totéž určit pro data získaná terénním výzkumem. Provést procesní analýzu zpracovávaných dat – tedy, co se s daty při zpracování děje (od jejich identifikace až po jejich konečné uložení v databázi) ve smyslu jejich formalizace, unifikace, normalizace atp., a stanovit odpovědnost zpracovatelů za jednotlivé procedury. Zajistit požadovanou retrospektivu, aktuálnost, granularitu, úplnost, konzistenci dat.

Možné překážky a problémy: Data, jejich část nebo jejich určitý parametr se ve stanovém rozsahu nebo kvalitě nepodaří získat pramenným ani terénním průzkumem, nebo je jejich získání spojeno s neúměrnými náklady.

Výstup tohoto kroku: A) Strukturovaný textový dokument, který popisuje výše uvedené parametry – prameny a jejich výtěžnost, metody terénního výzkumu, vlastnosti a parametry dat a s nimi spojených procedur. B) Na základě stanovených kritérií strukturovaný soubor dat.

Manažerské otázky: Byla dostatečně vyřešena otázka autorských práv u dat, která pocházejí z jiných zdrojů? Nepřekračují nároky stanovené na rozsah a kvalitu dat možnosti pracovního týmu?

Technické řešení: Příprava databáze, jejíž struktura reflektuje požadavky na formát a strukturu dat a možnosti práce s nimi (editace, import/export, prohlížení, filtrování, vyhledávání, statistické přehledy, vizualizace, analytické zpracování).

Tým LINDAT/CLARIAH-CZ pomůže mj.: zejména při definici vlastností a parametrů dat, metod jejich prvotního získávání a následného zpracování.

Poznámky badatele Alberta: S některými pojmy se setkávám poprvé (třeba granularita dat). Jinak našim pramenem jednoznačně jsou jednotlivá čísla Čtyřlístku. Teprve rozhodneme, jestli využijeme už digitalizovaných podkladů (vyžaduje spolupráci s redakcí časopisu), nebo přistoupíme k vlastní digitalizaci. Zvažujeme možnost OCR textů, což je vzhledem ke komiksovému formátu nevyzkoušená metoda, takže možná bude jednodušší použít „hlavoruční" přepis. Data zatím máme strukturována takto: výpověď (vždy jedna „bublina"), mluvčí, identifikace výpovědi (ročník, číslo, stránka, sekvence). Výpovědi budou zapisovány přesně tak, jak jsou uvedeny v dokumentu, dořešit musíme co s překlepy, tiskovými chybami a dalšími zjevnými formálními nedostatky, jak naložit s meta- a polyvýpověďmi apod. Uvažujeme o třístupňovém modelu zpracovatelů: „plniči" (prvotní zpracovatelé dat), editoři (kontrolují kvalitu a úplnost dat), administrátor (řeší systémové otázky). S týmem LINDAT/CLARIAH-CZ bych chtěl probrat zejména systémové řešení nepravidelností a inkonzistence vstupních dat.

Krok 3: PŘIPRAVUJEME METADATA

Neboli: Definice a parametrizace metadat.

Kde začít: Stanovením rozsahu popisných, strukturálních a administrativních metadat. Určit obsah, strukturu a formát jednotlivých typů metadat, pokud možno s ohledem na existující standardy. Zvážit využití řízených slovníků pro stanovený okruh metadatových polí. Naplánovat využití nástrojů a postupů zajišťujících úplnost a konzistenci metadat. Použít modelování (např. pomocí orientovaných grafů) pro vyjádření vztahu dat a metadat a jejich implementaci ve vyhledávacím rozhraní.

Možné překážky a problémy: Formálně, úzce nebo naopak široce koncipovaná popisná metadata, která neumožňují s daty pracovat předpokládaným způsobem nebo dostatečně přesně. Přílišná akcentace administrativních metadat, která může ubírat síly na vlastní obsahová (popisná a strukturální) metadata.

Výstup tohoto kroku: A) Strukturovaný textový dokument, který popisuje výše uvedené parametry – jednotlivé typy metadat (popisná, strukturální, administrativní), jejich formát a strukturu, metody jejich zpracování. B) Na základě stanovených kritérií strukturovaný soubor metadat připojený k dříve zpracovaným datům.

Manažerské otázky: Splňují metadata všechny nároky technické i právní standardizace?

Technické řešení: Připravit uložení a zobrazení metadat v databázi, jejich propojení s daty.

Tým LINDAT/CLARIAH-CZ pomůže mj.: s rozlišením jednotlivých typů metadat, stanovením jejich konkrétní struktury a formátu, doporučeními vhodných standardů a efektivních metod tvorby metadat.

Poznámky badatele Alberta: Naše popisná metadata se týkají vlastních textů (výpovědí), subjektu těchto výpovědí a kontextu výpovědi. Každá vrstva metadat zahrnuje další kritéria a parametry, např. u textu složky lingvistické analýzy, u subjektu pohlaví, emoční/fyzický stav, u kontextu denní doba, místo výpovědi (interiér/exteriér) aj. U vrstvy textu bychom pro segmentaci textu rádi nasadili standardní korpusové nástroje, u dalších vrstev řízené slovníky, pokud možno standardizované a parametrizovatelné. S týmem LINDAT/CLARIAH-CZ bych rád podrobně probral všechny aspekty zvolených metadat, abych si byl jist, že jsme neopomněli nějaká hlediska a že nás nepostihla „profesní slepota". Budeme také rádi za doporučení, jakou minimální úroveň by měla mít administrativní metadata.

Krok 4: ZPŘÍSTUPŇUJEME VÝSLEDKY

Neboli: Uživatelské rozhraní a možnosti vyhledávání.

Kde začít: Určením dat a metadat, která budou používána nejčastěji a ovlivní tak i základní strukturu uživatelského rozhraní. Zvolit strukturu hlavní stránky uživatelského rozhraní a parametry dílčích stránek spojených se zobrazením, prohlížením, filtrací a vyhledáváním údajů. Promyšleně rozlišit možnosti a rozsah plnotextového, nestrukturovaného a strukturovaného vyhledávání. Zapojit do testování systému koncové uživatele.

Možné překážky a problémy: Strukturované vyhledávání dostatečně nezužitkovává všechny parametry metadat a omezuje tak analytický přístup uživatele k obsahu databáze. Výsledky vyhledávání jsou z hlediska koncového uživatele zobrazovány nepřehledně nebo nesrozumitelně.

Výstup tohoto kroku: A) Strukturovaný textový dokument, který popisuje výše uvedené parametry – jednotlivá rozhraní, prohledávatelná pole a jejich obsah, formát zobrazení údajů. Stane se základem uživatelské dokumentace – nápovědy. B) Konkrétní uživatelské rozhraní s funkčními a otestovanými možnostmi vyhledávání. C) Kontextová nápověda vycházející z A a implementovaná do B.

Manažerské otázky: Bude se lišit přístup k datové platformě podle typu uživatele? Jaké licence budou pro konkrétní obsah a uživatele nastaveny? Respektují tyto licence publikační politiku naší instituce (open access)? Jakým způsobem lze výsledky projektu uplatnit při hodnocení výzkumné a vědecké činnosti?

Technické řešení: Na základě stanovené struktury připravit prototyp uživatelského rozhraní se zohledněním pravidel přístupnosti a dalších standardů. Po opakovaném testování zpřístupnit výslednou podobu datové platformy.

Tým LINDAT/CLARIAH-CZ pomůže mj.: s nastavením vyhledávacích filtrů, definicí vyhledávacího rozhraní, volbou vhodného zobrazení výsledků vyhledávání a testováním uživatelského rozhraní.

Poznámky badatele Alberta: Za hlavní filtry našich údajů považujeme předměty (objekty) výpovědí, subjekty výpovědí a časové hledisko. Ty by měly být součástí hlavní stránky stejně jako možnost plnotextového vyhledávání ve výpovědích. Zároveň bychom rádi umožnili badatelům maximálně zužitkovat obsah databáze kladením analytických, kombinovaných, parametrizovatelných dotazů, byť i z našeho pohledu irelevantních nebo nesmyslných (třeba jak často subjekt typu *muž* v emocionálním stavu *rozčilen* používá extratextové odkazy k objektům typu *stavba*). Od týmu LINDAT/CLARIAH-CZ očekávám v tomto kroku zejména konzultaci k možnostem analytického vyhledávání a intenzivní testování uživatelského rozhraní s akcentem na pokročilé vyhledávání.

Krok 5: ZŮSTÁVÁME TU NADLOUHO

Neboli: Správa a údržba datové platformy, dlouhodobá udržitelnost.

Kde začít: Stanovením koncepce dalšího rozvoje datové platformy. Určení okruhů dat a metadat, která budou ve stanovených intervalech aktualizována a doplňována. Identifikace datových i funkčních nedostatků, které vyvstaly z prvotní projektové fáze. Určení dlouhodobé potřeby rozšíření datové základny, metadatového popisu a vylepšených nebo nových funkcí.

Možné překážky a problémy: Nedostatek motivace nebo zdrojů pro dlouhodobou správu datové platformy. Nesoustavná aktualizace údajů ústící do zastaralosti nebo neúplnosti dat.

Výstup tohoto kroku: A) Strukturovaný textový dokument, který stanovuje pravidla a procedury správy a aktualizace datové platformy. B) Průběžně aktualizovaný a doplňovaný obsah datové platformy.

Manažerské otázky: Kdo bude pověřen dlouhodobou správou datové platformy? Bude nutné pro údržbu databáze získávat i mimorozpočtové finanční zdroje?

Technické řešení: Implementace nástrojů umožňujících efektivní sledování obsahových i formálních změn datové platformy.

Tým LINDAT/CLARIAH-CZ pomůže mj.: s definicí parametrů krátkodobého, střednědobého a dlouhodobého rozvoje datové platformy, modelováním potenciálních uživatelů databáze, úpravou struktury, obsahu a uživatelského rozhraní.

Poznámky badatele Alberta: Z této fáze mám největší obavy. Datovou platformu vytváříme v rámci projektu a po jeho skončení hrozí, že vše poběží „na volnoběh" nebo se databáze zakonzervuje ve stavu ke konci projektu, což bychom neradi, protože podnětů k jejímu dalšímu rozvoji máme řadu. Rady týmu LINDAT/CLARIAH-CZ by proto pro nás mohly být přínosné s ohledem na stanovení priorit rozvoje a zajištění jejich realizace včetně financování.

ZPĚTNÁ VAZBA

Budete-li to považovat za vhodné a užitečné, dejte nám prosím vědět při našem příštím setkání (nebo na e-mail lindat-clariah@phil.muni.cz), nakolik byl pro vás tento dokument přínosný:

A – Podstatně mi ozřejmil záležitosti spojené s tvorbou datové platformy.

B – Vysvětlil mi některé nejasnosti, přinesl dílčí podněty, ale celkově můj vhled do problematiky nezměnil.

C – Byl pro mě dost obecný, celkový přehled o problematice mám a potřebuji spíše řešit konkrétní kroky.

D – Není mi jasné, o co tu jde, potřebuji asi problematiku vysvětlit jinou formou.

Poznámky: _____

Děkujeme za váš názor. Zohledníme jej v naší další práci.