Jak budujeme digitální výzkumnou infrastrukturu na MU

Michal Lorenz – Pavla Martinková

27. 2. 2020

LINDAT
CLARIAH-CZ

LINDAT/CLARIAH-CZ

- Co je digitální výzkumná infrastruktura?

- Digital Research Infrastructure for the Language Technologies, Arts and Humanities

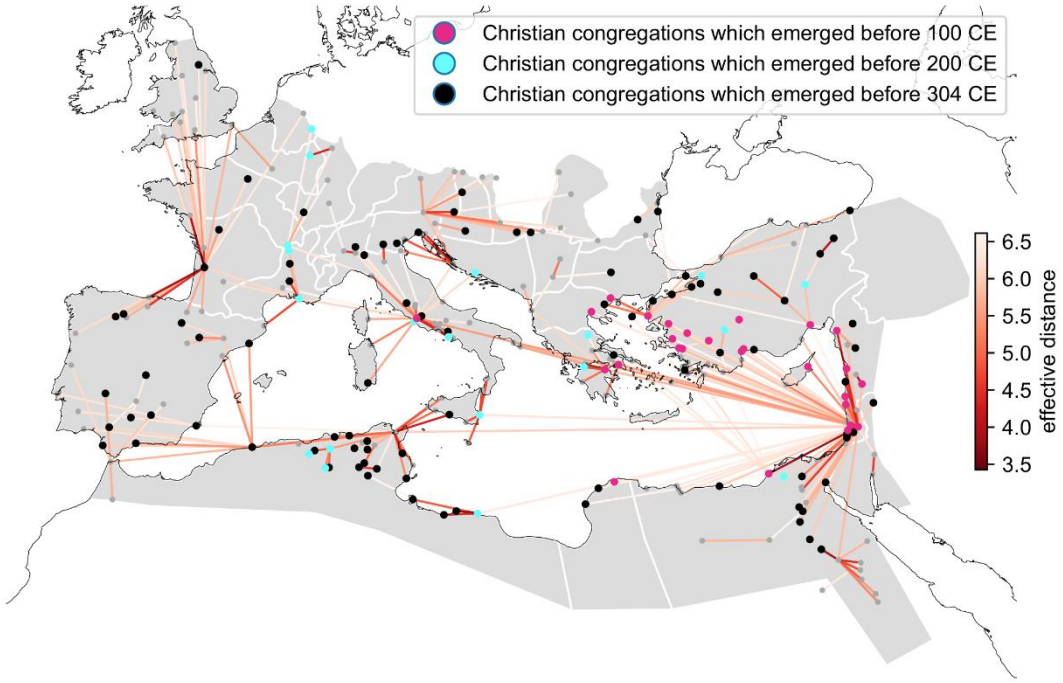
CLARIN – DARIAH – RIDICS

(https://www.clarin.eu) – (https://www.dariah.eu) – (https://vokabular.ujc.cas.cz)

- Ústřední portál LINDAT/CLARIAH-CZ: https://lindat.cz

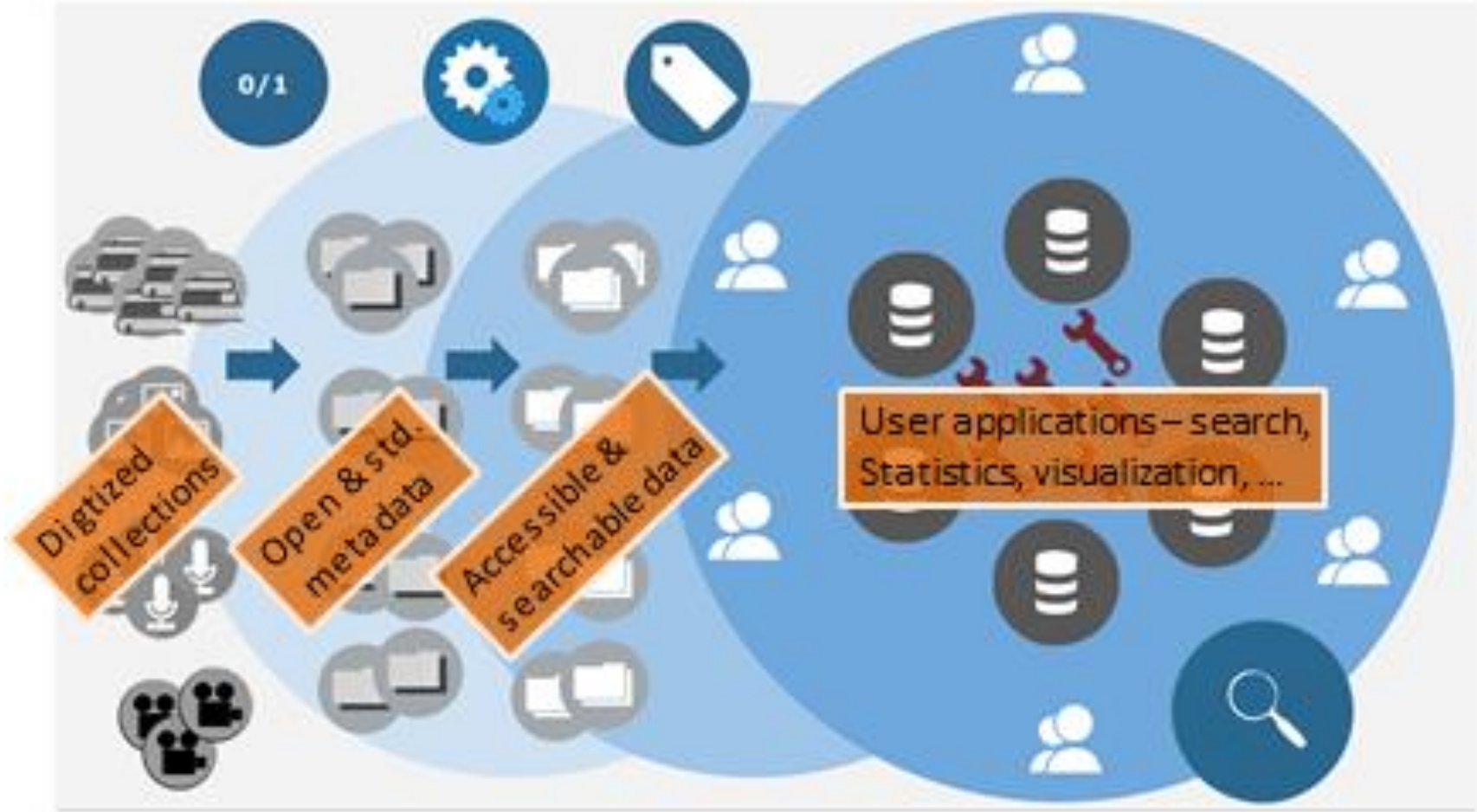
Digitální věda a MU

- technická a datová základna pro digitální výzkum
- výpočetní síla - servery a důvěryhodné datové úložiště
- vedoucí role lingvistiky (výzkum zpracování textu a řeči) a religionistiky na MU (Centrum pro digitální výzkum náboženství (CEDRR)



- využití strojového učení, umělé inteligence, lingvistických korpusů a digitálních fondů a sbírek paměťových institucí

DARIAH-CZ


0/1
Digtized collections
Open & std. metadata
Accessible & searchable data
User applications – search, Statistics, visualization, ...

Příprava projektu

- financování – OP VaVpl, Operační program Výzkum a vývoj pro inovace: konkurenceschopnost státu a podpora znalostní ekonomiky
- **Ministerstvo školství, mládeže a tělovýchovy**

- zapojení MU do přípravy 2016
- 1. kolo – žádost 50 stran v angl., 1. průzkum infrastruktur na MU
- 2. kolo – žádost 100 stran v angl., celkem 14 dní na zpracování
- 3. kolo – obhajoba před mezinárodní komisí expertů na ministerstvu
- zahájení projektu 1. 1. 2019

Co jsme museli řešit na začátku projektu?

- Jaké digitální kolekce a platformy na FF MU existují?
  - vlastnictví, otevřenost, dlouhodobé uložení, ochota poskytnout zdroje
- Jaké je vhodná technologie pro repozitář?
  - vývojářská komunita, rychlost, škálovatelnost, kurátorství, interoperabilita, customizace rozhraní, heterodiverzita dat
- Jaké jsou uživatelské potřeby a požadavky?
  - akademici, digitální vědci, studující, vyučující digitálních humanitních věd
- Pod jakou certifikací chceme provozovat dlouhodobé uchování dat?
  - CoreTrustSeal, Nestor Seal, ISO 16363, Drambora
- Jaká metadata potřebujeme zpracovávat?
  - centrální portál – standard, platformy - obohacení, ontologie?

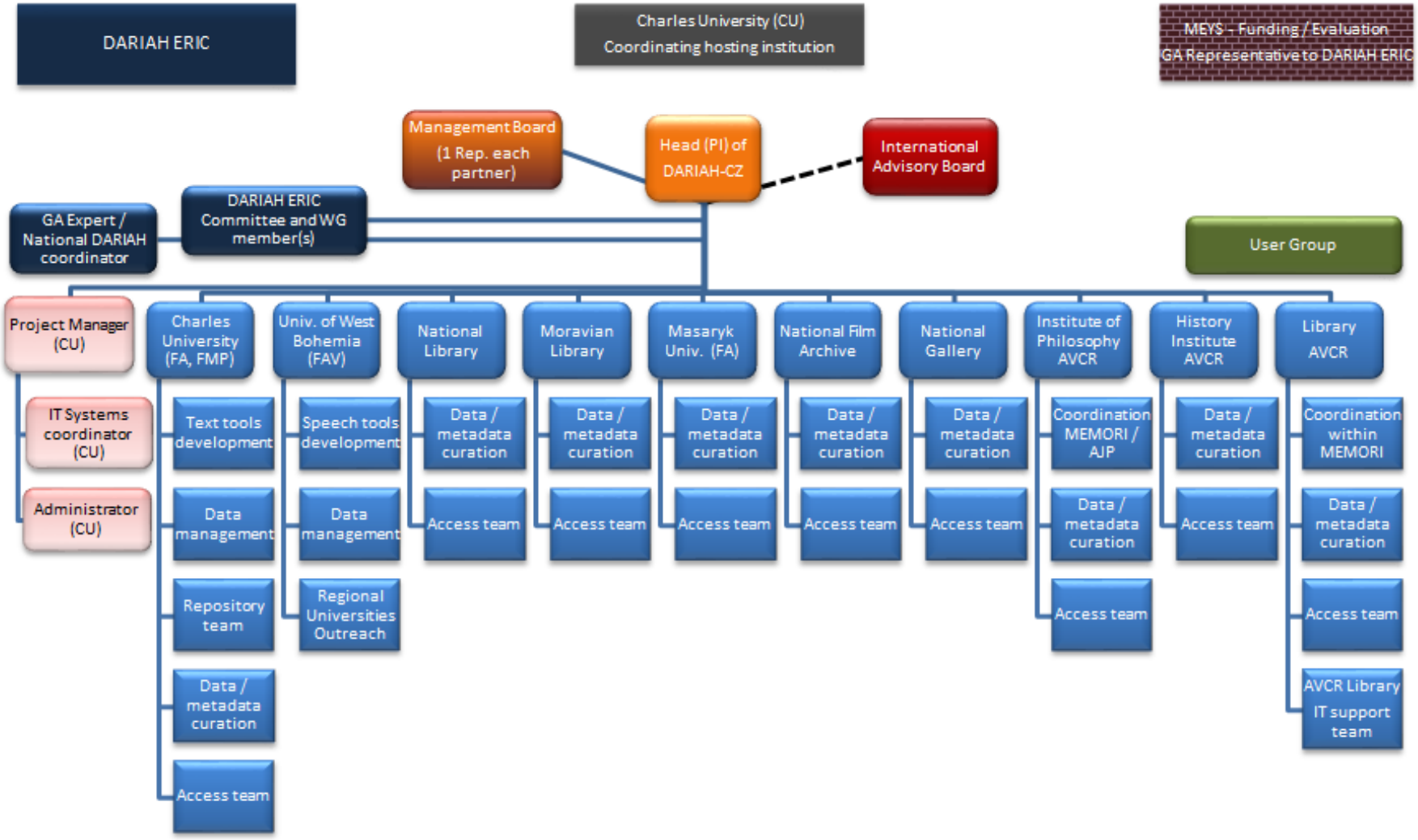
Tým CLARIAH na MU

KISK	CIT	Knihovna FF	ÚVT	Externí expert
				
Michal Lorenz	Hana Kubelková	Pavla Martinková	Vlastimil Krejčíř + Nicole Benčíková + Alžbeta Strakošová	Michal Konečný

Management infrastruktury

- stránky projektu – https://it.muni.cz/phil/lindat-clariah
- tutoriály pro širokou veřejnost – Praha
- tutoriály pro MU
- neformální setkávání
- DARIAH ANNUAL EVENT 2020: SCHOLARLY PRIMITIVES (https://dariah-ae-2020.sciencesconf.org/)

Management infrastruktury



Průzkum platforem

- Jaké digitální kolekce a platformy na FF MU existují?
  - vlastnictví, otevřenost, dlouhodobé uložení, ochota poskytnout zdroje

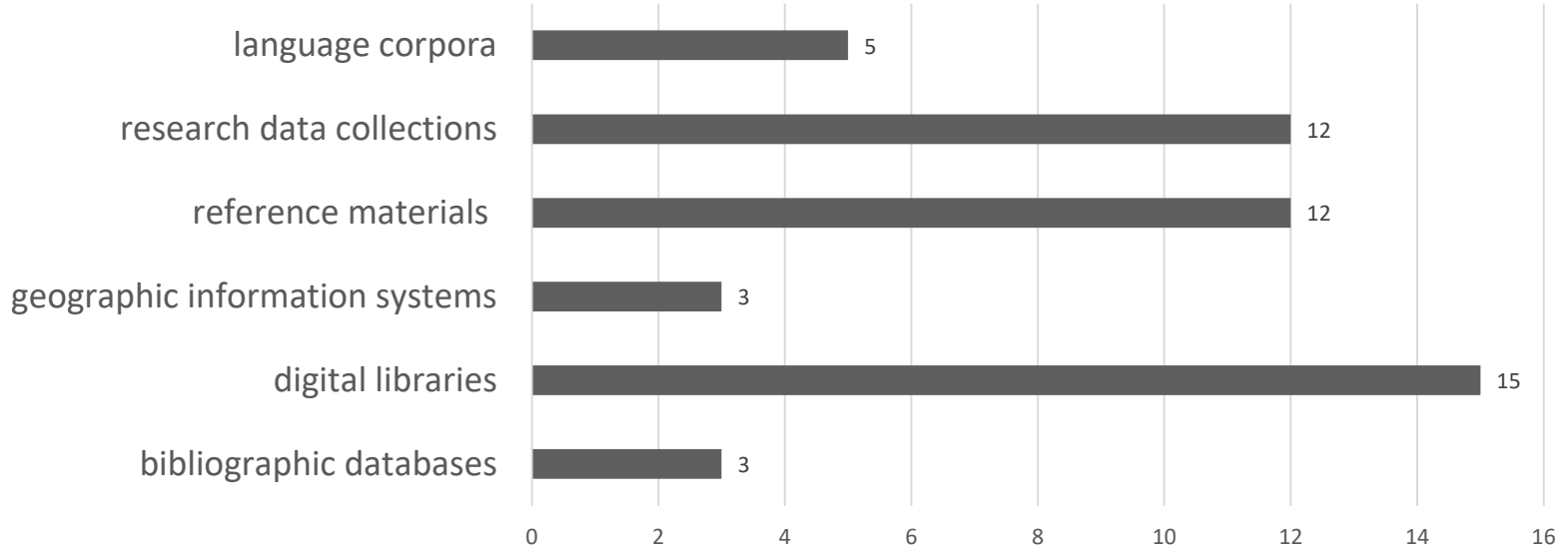
1. Před-výzkum před zahájením projektu
2. Výzkum od stolu
   - Projekty pracovišť
3. Dotazníkové šetření
   - Pro získání přehledu
   - 26 pracovišť, následné telefonické a osobní rozhovory
4. Polostrukturované rozhovory
   - Pro získání podrobnějších informací
   - 13 rozhovorů
5. Průběžná aktualizace
   - Aktivně i pasivně

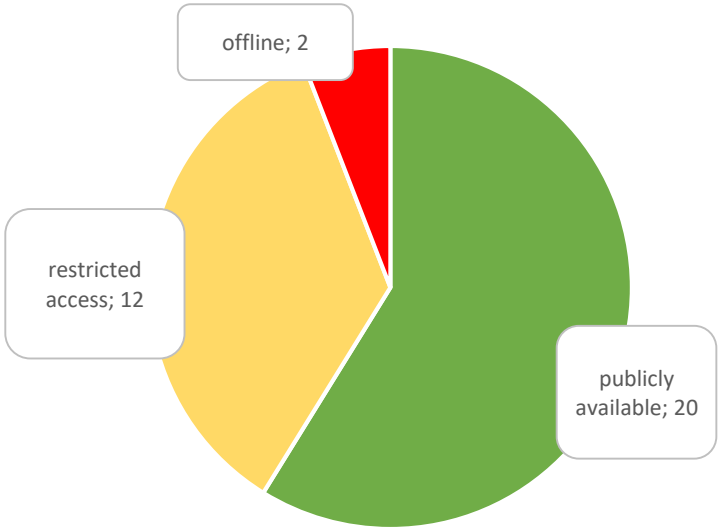
Photo by Helloquence on Unsplash

Zjištění a výstupy

- Celkem 50 platforem (léto 2019)
  - 34 v provozní fázi
  - Základní informace o 50, podrobný popis u 11
- Potřeby a obavy akademiků
  - Základy komunity
- Online katalog veřejně přístupných

Název	URL	Popis	Pracoviště	Fáze	Kategorie	Přístup k obsahu	Předávání metadat	Přesun do CR	Kontaktní osoba	Rozhovor	Katalog platforem	Další informace
Allgemeines historisches Künstler-Lexikon für Böhmen und zum Theile auch für Mähren und Schlesien	http://www.ceskyh	Neveřejný slovník Centra hudební lexikografie.	Ústav hudební vědy	provozní	výkladový slovník	dostupné online s omezením	ne	ne	dr. Macek	proběhl	CHL	
Archiv Centra hudební lexikografie	http://www.ceskyh	Neveřejný archiv Centra hudební lexikografie.	Ústav hudební vědy	provozní	výzkumná data	dostupné online s omezením	ne	ne	dr. Macek	proběhl	CHL	
Bibliografická databáze oborového časopisu Filmový kurýr	https://www.muni.c	2019 | Databáze rozpracovaná ve spolupráci s Markem Stehlíkem z FI MU, po dokončení umožní fulltextové vyhledávání v anotacích jednotlivých textů Filmového kurýra, vyhledávání podle autorů	Ústav filmu a audiovizuální kultury	realizační	bibliografická databáze		nezjištěno	nezjištěno	doc. Skopal		zatím ne	doplňující informace od KE
Bibliografie Aleše Tichého	http://www.phil.mu	Bibliografický přehled knižních překladů z britské literatury. Digitalizována po roce 2000, je dodnes funkční v původní podobě, ale neaktualizuje se, protože její funkci převzal jiný projekt, do nějž byla začleněna. Bylo by možné vypnout, ale chtějí	Katedra anglistiky a amerikanistiky	ukončeno	bibliografická databáze		ne	ne	dr. Rambousek		ne	soubor doplněk





Analýza technického řešení

- Jaké je vhodná technologie pro repozitář?
  - vývojářská komunita, rychlost, škálovatelnost, kurátorství, interoperabilita, customizace rozhraní, heterodiverzita dat

1. Výzkum od stolu a předvýběr
2. Testování systémů
   - Samvera (Hyku/Hyrax), DSpace a Islandora
   - Testovací scénáře na základě požadavků
3. Závěrečná zpráva a rozhodnutí
   - Podle stanovených kritérií

Photo by Pakata Goh on Unsplash

Testovací scénář

	název scénáře:	Autentizace		
	popis:	Systém pro autentizaci (i externí SSO typu Shibboleth). Osoby z MUNI sa prihlasujú cez Shibboleth, pre externých sa vytvorí účet v systéme.		
	projekt:			
	manažer:			
	tester:		**datum testování:**	**22.09.2019**

Krok #	Popis kroku a vstupních dat.	Očekávaný výsledek	Hyku: Skutečný výsledek	Hyku: PASS/FAIL
1	Systém umožní využitie externého autorizačného systému (napr. Shibboleth).			
2	Systém umožní užívateľovi vytvoriť účet.			
3	Tester si na stránke vyberie možnosť prihlásenia cez Shibboleth.	Zobrazí sa mu formulár pre prihlásenie na MUNI.	Zobrazí sa mu formulár pre prihlásenie na MUNI.	PASS
4	Tester zadá svoje univerzitné prihlasovacie údaje.	užívateľa.	užívateľa.	PASS
5	Tester sa odhlási zo systému.			
6	Tester si na stránke vyberie možnosť registrácie nového užívateľa.	Zobrazí sa mu registračný formulár.	Zobrazí sa mu registračný formulár.	PASS
7	Tester vyplní údaje a potvrdí ich.	Vytvorí sa užívateľský účet a užívateľ je prihlásený.	Vytvorí sa užívateľský účet a užívateľ je prihlásený.	PASS

Závěrečná zpráva

- Společné shrnutí a závěr
  - **Islandora**
- Podrobné srovnání systémů
  - Podle požadavků
  - Plusy a mínusy

Kompletní shrnutí a závěr

Tabulka uvádí přehled zkoumaných vlastností jednotlivých systémů spolu s hodnocením (1 - nejlepší, 2- prostřední, 3 - nejslabší).

	Islandora	*DSpace*	*Hyku/Hyrax*
Procházení, tvorba hierarchických kolekcí	*1*	*2*	*2*
Úprava prostředí, GUI	*1*	*1*	*3*
Autentizace	*1*	*1*	*1*
Autorizace (a embargo)	*2*	*1*	*?*
Ukládání různých typů dat (typy objektů)	*1*	*2*	*?*
Vkládání a editace metadat a dat (přes UI, kooperativně, strojově)	*1*	*3*	*2*
Metadatová schémata	*1*	*1-2*	*?*
Škálovatelnost	*?*	*2*	
Možnosti propojení s externími aplikacemi	*1*	*1*	*1*
Možnost uživatelské customizace systému (vlastní kolekce apod.)	*1*	*3*	*1*
Prezervace dat (LTP)	*3*	*2*	*3*
Statistiky	*1*	*3*	*2*
Stažení kolekce	*2*	*1*	*3*

A co dál?

- Výběr platforem podle kritérií
  - **Nahraditelnost**: Dokáže nové řešení suplovat funkce původního a naplnit potřeby uživatelů?
  - **Faktická proveditelnost**: Bude možné přenést původní obsah, vzhledem k závazkům původních projektů a/nebo autorskému právu?
  - **Technická proveditelnost**: Je k dispozici dokumentace a/nebo vývojář původního systému?
  - **Otevřenost obsahu**: Nakolik je obsah možno veřejně zpřístupnit? Ideální možností je udělení CC licence s minimem omezení.
  - **Kvalita metadat**: Jaká metadata jsou poskytnuta, v jakém formátu a nakolik jsou úplná nebo mohou být doplněna?
  - **Využívanost**: Nakolik je obsah využíván – například zda je zapojován do výuky nebo využíván mezinárodně.
- Tvorba prototypu a testování
  - Digitální knihovna Arna Nováka, Digitální knihovna FF MU, GISTRALIK a Filmové Brno

Průzkum uživatelů

- Jaké jsou uživatelské potřeby a požadavky?
  - akademici, digitální vědci, studující, vyučující digitálních humanitních věd

1. World Café
2. Dotazník
   - Online, malá návratnost
3. Micro Timeline Interview
4. Uživatelské testování

World Café

- Společná diskuze na téma digitální humanitní vědy
  - Informování o projektu a zjišťování potřeb
- Diskuzní metoda
  - Jak probíhá (váš) výzkum v DH?
  - Jaké bariéry vnímáte ve výzkumu v DH?
  - Jak připravit studující na výzkumv DH?
- Shrnutí na webu

Micro Timeline Interview

- = rozhovor formou mikro časové osy
- Vědci využívající jednotlivé platformy
- Analýza bariér a jejich překonání
- Analýza potřeb

Photo by Maranda Vandergriff on Unsplash

Proces vyjednávání

- rektor – dopadová zpráva, podpora vedení univerzity
- označení vyvíjeného řešení:
  - **digitální knihovna** – spravovaná sbírka s odpovídajícími službami, organizace poskytující sbírky digitálních zdrojů
  - **digitální archiv** – poskytuje digitalizované informační objekty jako evidenci minulosti
  - **datový repozitář** – syn. datová knihovna, datový archiv: infrastruktura tvořená federací několika databází, poskytující přístup k množině dat a službami spojenými s údržbou a využitím dat
  - **datový sklad** (*data warehouse*) – syn. datové skladiště, pouhá agregace dat z více zdrojů
  - **datové jezero** (*data lake*) – repozitář nestrukturovaných dat s klasifikačními metadaty a taggy, usnadňuje spojení formátů a norem dat
  - **digitální tržiště** (*data marts*) – repozitář zaměřený na uživatelské potřeby, použitelnost, zabezpečení různých množin dat, všechna data nejsou uživateli přístupná
- citace dat – Altmetrie?
- digitální stopy
- datové kultury

Komunita

- Online Fórum pro digitální humanitní vědy
  - Konzultanti
  - Zdroje
- Nabídka konzultací a materiálů
  - Doporučení
- Pozvánky na akce
- Neformální setkávání

Photo by Vonecia Carswell on Unsplash

Vzdělávání

- Workshopy
- Kurz Digital humanities
  - **Vznik**: Analýza a komparace téměř 50 sylabů kurzů na českých i zahraničních univerzitách. Rozhovory s českými a zahraničními vyučujícími a vědci.
  - **Cíl**: Seznámit se s oblastí Digital Humanities, co přinesly digitální technologie jednotlivým oblastem humanitních věd, jak se změnily vědecké postupy v dané oblasti, jaké otázky mohou být zodpovězeny za pomocí digitálních technologií. Seznámit se s technologiemi, osobnostmi a projekty, které něco znamenaly pro oblast Digitálních humanitních věd.

Photo by Mikael Kristenson on Unsplash

Témata kurzu

- Co jsou DH, milníky ve vývoji DH, projekty z oblasti DH
- Typy a zdroje dat v DH, digitální a digitalizovaná data
- Získávání a čištění dat
- Obrazová data a jejich zpracování
- Textová data a jejich zpracování
- Prostorová data a jejich zpracování
- Sítě a síťová analýza
- Metadata
- Vizualizace dat (teorie)
- Vizualizace dat (praxe)
- Publikování dat
- Archivace a uchovávání dat

Nástroje pro výuku:

- Otevřené
- Jednodušší
- Bez znalosti kódu
- Zdarma
- Pro základní pochopení tématu
- Pro pokročilejší práci představení složitějších nástrojů (odkazy na kurzy na MU i online)

- *Google MyMaps*
- *Voyant Tools*
- *Gephi*
- *Omeka*
- *Neatline*
- *Raw graphs*
- *Image plot*
- *Xpath*
- *Open refine*