Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Diplomová práce

Odborné fórum, ÚISK FF UK, 4. 3. 2024

Filip Kersch, FilipKersch@gmail.com, kersch@knav.cz

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Cíl práce:

Na základě strojového analýzy tištěných publikací zjistit, jaká témata byla předmětem bádání České akademie císaře Františka Josefa pro slovesnost, vědy a umění v prvních dvaceti letech jejího fungování.

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Cíl práce:

Na základě strojového analýzy tištěných publikací zjistit, jaká témata byla předmětem bádání České akademie císaře Františka Josefa pro slovesnost, vědy a umění v prvních dvaceti letech jejího fungování.

Proč?

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Cíl práce:

Na základě strojového analýzy tištěných publikací zjistit, jaká témata byla předmětem bádání České akademie císaře Františka Josefa pro slovesnost, vědy a umění
v prvních dvaceti letech jejího fungování.

Proč?

Ucelený a zároveň dostatečně podrobný pohled na publikační činnost dovolující identifikovat konkrétní oblasti zájmu České akademie zejména ze začátku jejího působení dosud chybí.
A existují nové možnosti, jak tento pohled získat.

Struktura práce

Teoretická část

- Co o tématech víme?----------> 1. Česká akademie věd a umění
- Co nového máme k dispozici?------> 2. Digitalizace fondů
- Jak to lze využít?----------> 3. Makroanalýza digitálních dat

Výzkumná část

4. Modelování témat

5. Diskuze

1. Co o tématech víme?

- Česká akademie císaře Františka Josefa pro vědy, slovesnost a umění vznikla 23. ledna 1890
- **Organizována byla do 4 tříd:**

I. třída

byla určena filozofii,

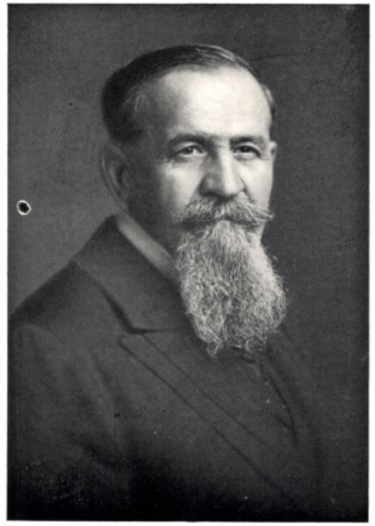
právu a historii.

Antonín Randa

II. třída

byla věnována matematickým,

přírodním a lékařským vědám.



Bedřich Procházka

III. třída

měla pěstovat filologické obory

a zejména češtinu a českou literaturu.

Václav Sládek

1. Co o tématech víme?

- **Na publikační činnost mělo významný vliv členění oborů do tříd:**
  - společné publikace (*Věstník, Almanach)*,
  - každá třída vydávala ***Rozpravy***,
  - monografie,
  - podpora vydávání časopisů.

1. Co o tématech víme?

- **Na publikační činnost mělo významný vliv členění oborů do tříd:**
  - společné publikace (*Věstník, Almanach),*
  - každá třída vydávala ***Rozpravy****,*
  - monografie,
  - podpora vydávání časopisů.

- **Témata (dosavadní stav poznání):**
  - I. třída: historie, právní problematika a filozofie,
  - II. třída: matematika, fyzika, chemii, biologie (bryologie), lékařství, technika,
  - III. filologie, čeština, česká literatura, lexikografie, dialektologie, orientální jazyky,…

2. Co máme k dispozici?

- Fond obsahující publikační činnost ČAVU převzala Základní knihovna založená v roce 1952 spolu s ČSAV. Dnes jej spravuje KNAV.
- Knihovny v České republice včetně KNAV digitalizují své fondy:

2. Co máme k dispozici?

- Fond obsahující publikační činnost ČAVU převzala Základní knihovna založená v roce 1952 spolu s ČSAV. Dnes jej spravuje KNAV.
- Knihovny v České republice včetně KNAV digitalizují své fondy:
  - **Ochrana fyzických dokumentů:**

    Z počátku pozornost napříč knihovnami věnována zejména periodikům z 19. století, která jsou významně ohrožena degradací kyselého papíru.

    Postupné rozšíření na širokou paletu dokumentů.

2. Co máme k dispozici?

- Fond obsahující publikační činnost ČAVU převzala Základní knihovna založená v roce 1952 spolu s ČSAV. Dnes jej spravuje KNAV.
- Knihovny v České republice včetně KNAV digitalizují své fondy:
  - Ochrana fyzických dokumentů.
  - **Překonání času a prostoru:**

    Digitální dokumenty mají mnoho výhod: jsou snadno šiřitelné, mohou být dostupné 24 hodin denně, umožňují prohledávání a porovnávání plných textů digitalizovaných dokumentů.

    Tyto vlastnosti a nástroje činí z digitálních knihoven nový a mimořádně bohatý zdroj poznání

2. Co máme k dispozici?

- Fond obsahující publikační činnost ČAVU převzala Základní knihovna založená v roce 1952 spolu s ČSAV. Dnes jej spravuje KNAV.
- Knihovny v České republice včetně KNAV digitalizují své fondy:
  - Ochrana fyzických dokumentů.
  - **Překonání času a prostoru:**

    Digitální dokumenty mají mnoho výhod: jsou snadno šiřitelné, mohou být dostupné 24 hodin denně, umožňují prohledávání a porovnávání plných textů digitalizovaných dokumentů.

    Tyto vlastnosti a nástroje činí z digitálních knihoven nový a mimořádně bohatý zdroj poznání a **dat pro výzkum.**

2. Digitální knihovna pro digitální humanitní vědy

- Obecně jsou digitální knihovny koncipovány především pro vyhledávání, procházení a prohlížení jednotlivých dokumentů:
  - přímo **neumožňují práci s agregovanými daty** o publikacích
  - **ani jednoduché stažení** potřebných dat k analýze.
    - --> Tímto se digitální knihovny výrazně odlišují od jazykových korpusů, které představují významný textový zdroj sloužící k výzkumu v oblasti digitálních humanitních věd.
    - --> Na rozšíření možností vytěžování dat z českých digitálních knihoven se zaměřil projekt DL4DH (KNAV, NK ČR, MZK)
      - Stažení dat a metadat titulů uložených v digitální knihovně a obohacení externími nástroji a dalšími interními moduly.

3. Jak lze digitalizaci dokumentů využít?

- Snadná dostupnost velkého množství digitalizovaných textů zavazuje badatele, aby nespoléhali pouze na pečlivé čtení úzké skupiny textů, ale věnovali se mimo to také analýze širších textových korpusů pomocí softwarových nástrojů.*

* JOCKERS, Matthew Lee. *Macroanalysis : Digital Methods and Literary History*, Urbana, Chicago, Springfield: University of Illinois Press, 2013. ISBN: 978-0252079078. s. 3 – 10.

3. Jak lze digitalizaci dokumentů využít?

- Snadná dostupnost velkého množství digitalizovaných textů zavazuje badatele, aby nespoléhali pouze na pečlivé čtení úzké skupiny textů, ale věnovali se mimo to také analýze širších textových korpusů pomocí softwarových nástrojů.*
  - **Makroanalýza / Distant reading****
    - Stěžejním principem metody je zpracování textových informací pomocí počítačů – ze vzdálenosti
    - rozpoznání těžko identifikovatelných vzorců a souvislostí
    - makro-mikro

* JOCKERS, Matthew Lee. *Macroanalysis : Digital Methods and Literary History*, Urbana, Chicago, Springfield: University of Illinois Press, 2013. ISBN: 978-0252079078. s. 3 – 10.

** MORETTI, Franco. *CONJECTURES ON WORLD LITERATURE*. Online. New Left Review, 2000, vol. 1. Dostupné z: https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature. [cit. 2023-11-10].

3. Jak lze digitalizaci dokumentů využít?

- **Makroanalýza je jednou z metod digital humanities.**

- Podle České asociace pro digitální humanitní vědy (CzADH):

  *„odborné činnosti jako využívání **digitálních metod** a **nástrojů** k vědeckému bádání o humanitních a společenskovědních otázkách nebo k prezentaci jeho výsledků; získávání, uchovávání, propojování a zpřístupňování **dat v digitální podobě** s důrazem na jejich maximální otevřenost […]“.**

* ČESKÁ ASOCIACE PRO DIGITÁLNÍ HUMANITNÍ VĚDY. *O nás*. Online. CzADH: Česká asociace pro digitální humanitní vědy. Dostupné z: https://www.czadh.cz/o-nas/. [cit. 2023-11-10]. Zvýraznění vlastní.

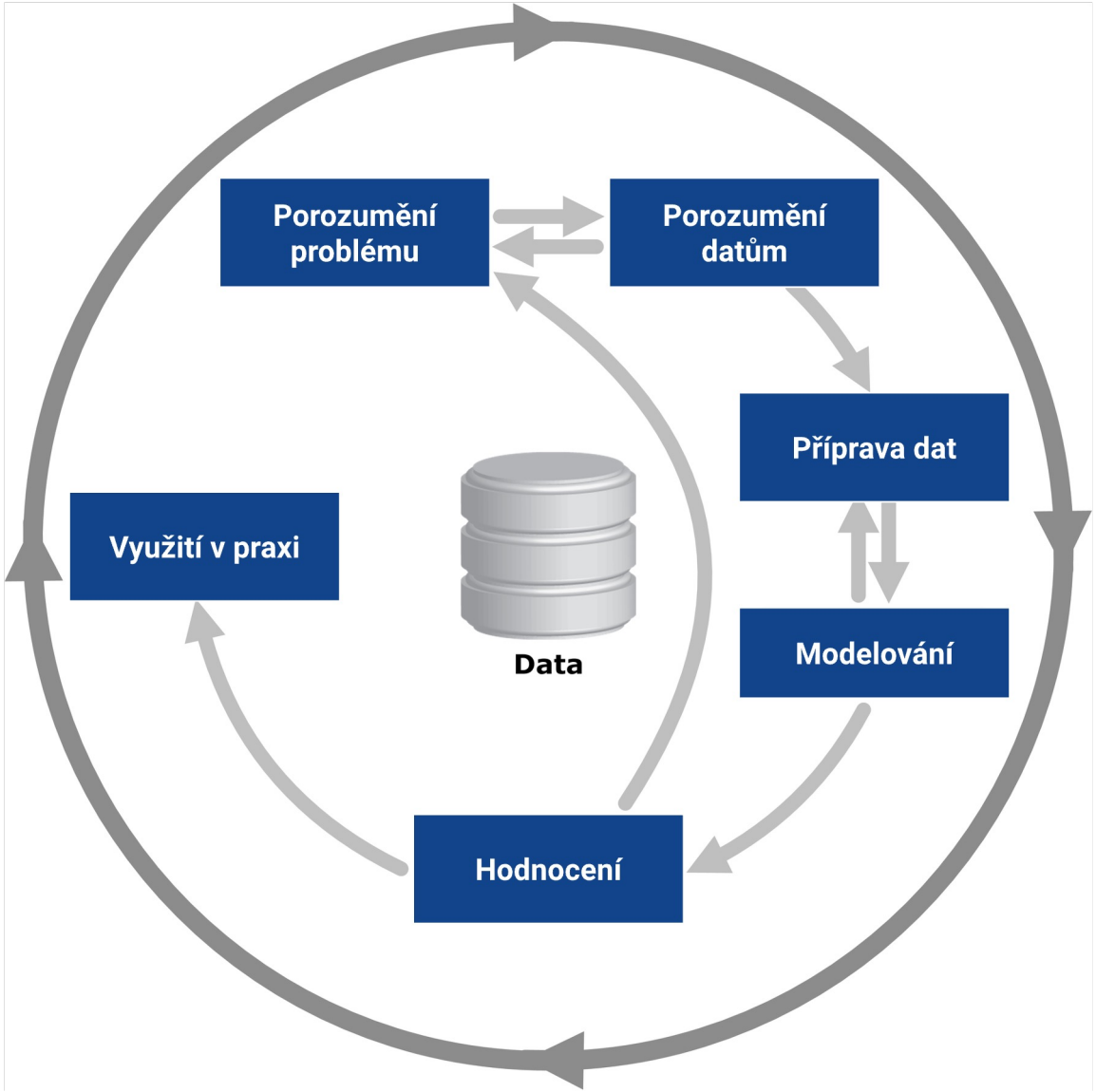
3. Makroanalýza digitálních dat

- **digitální metody ----------> ** makroanalýza (tematické modelování)
- **data v digitální podobě ----------> ** digitalizované publikace ČAVU
- **digitální nástroje ----------> ** Digitální knihovna AV ČR (Kramerius)

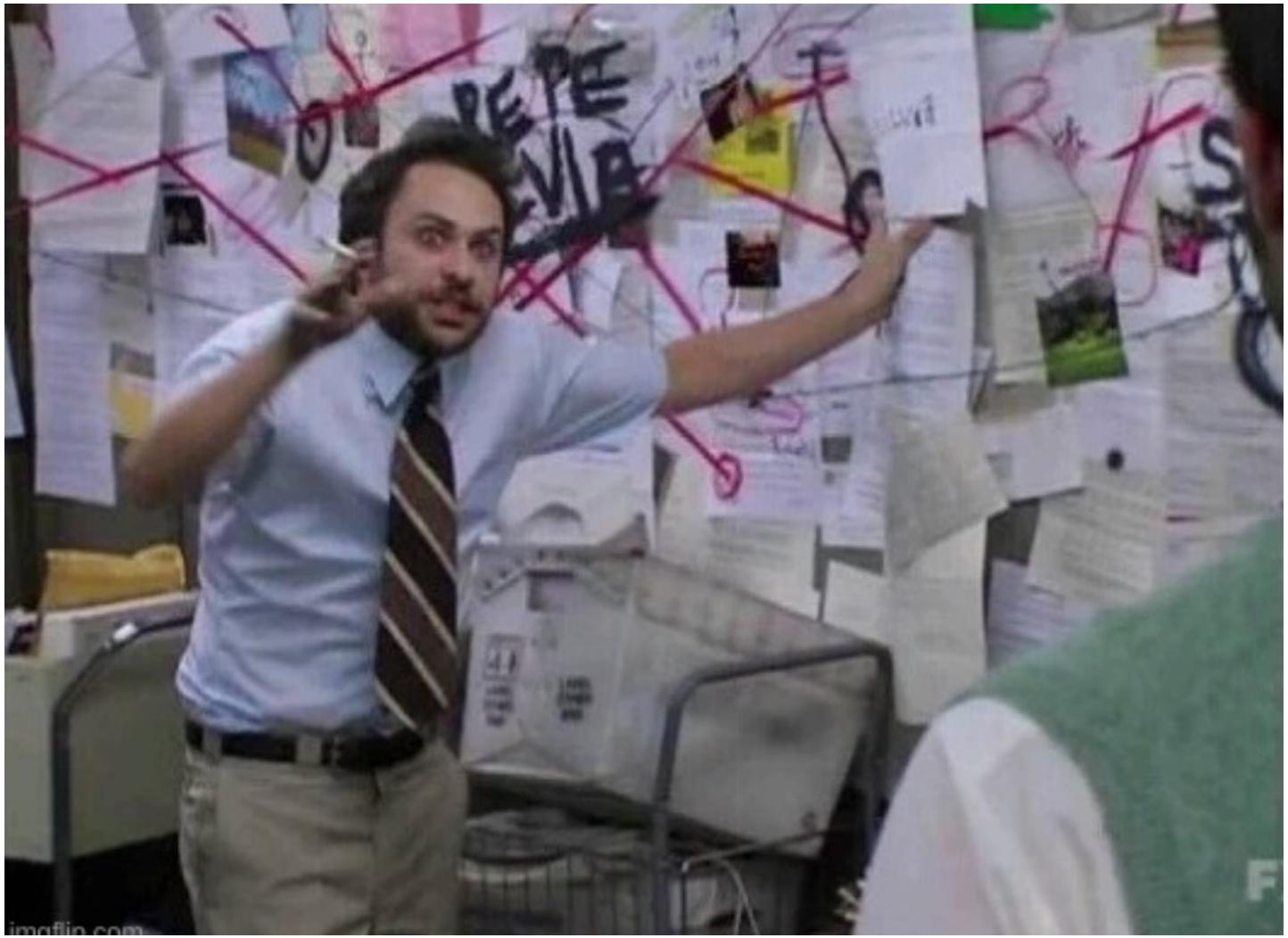
  **----------> ** softwarové prostředí R (RStudio)

  **----------> ** nástroje LINDAT/CLARIAH-CZ

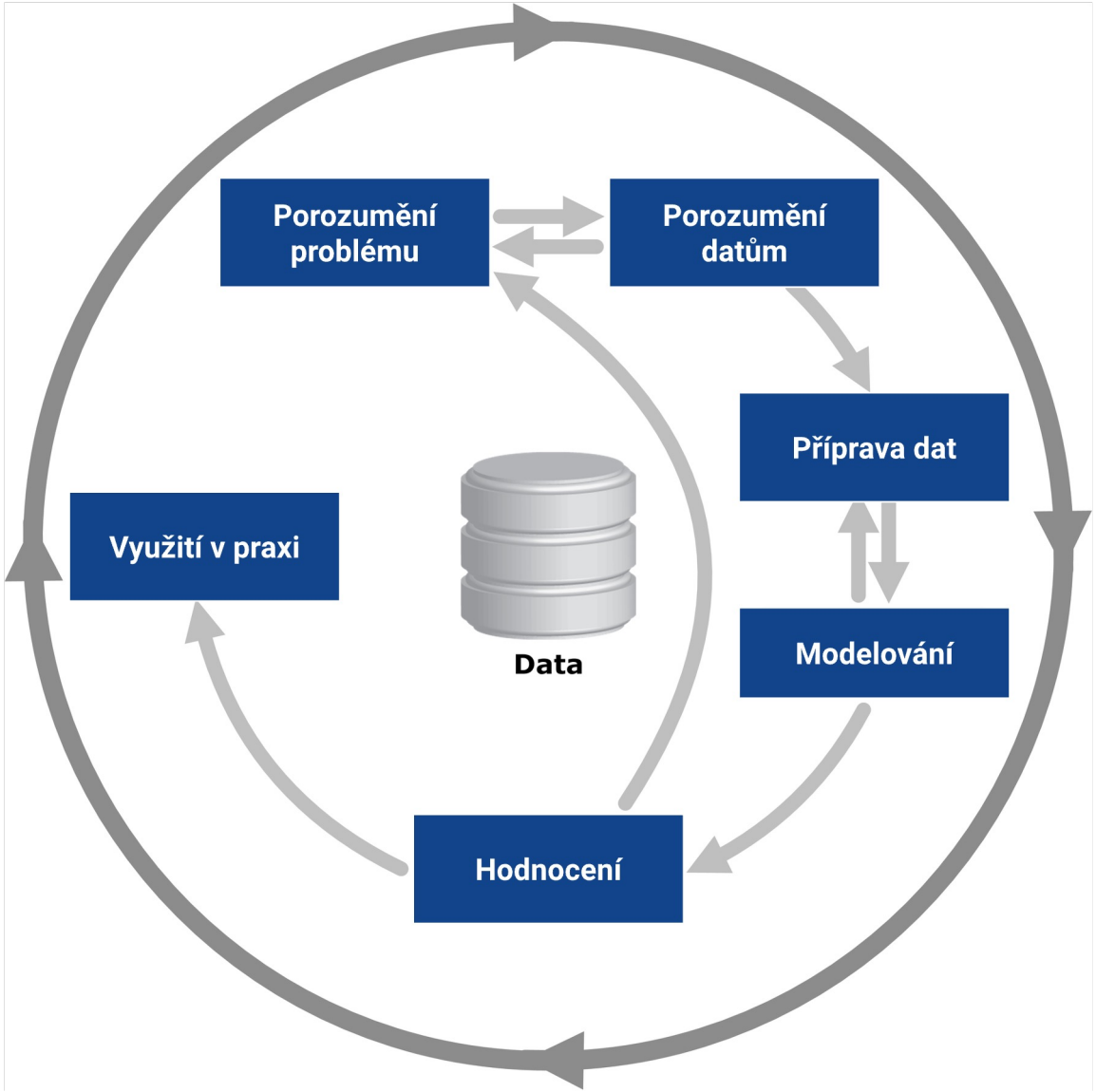
4. Výzkumná část



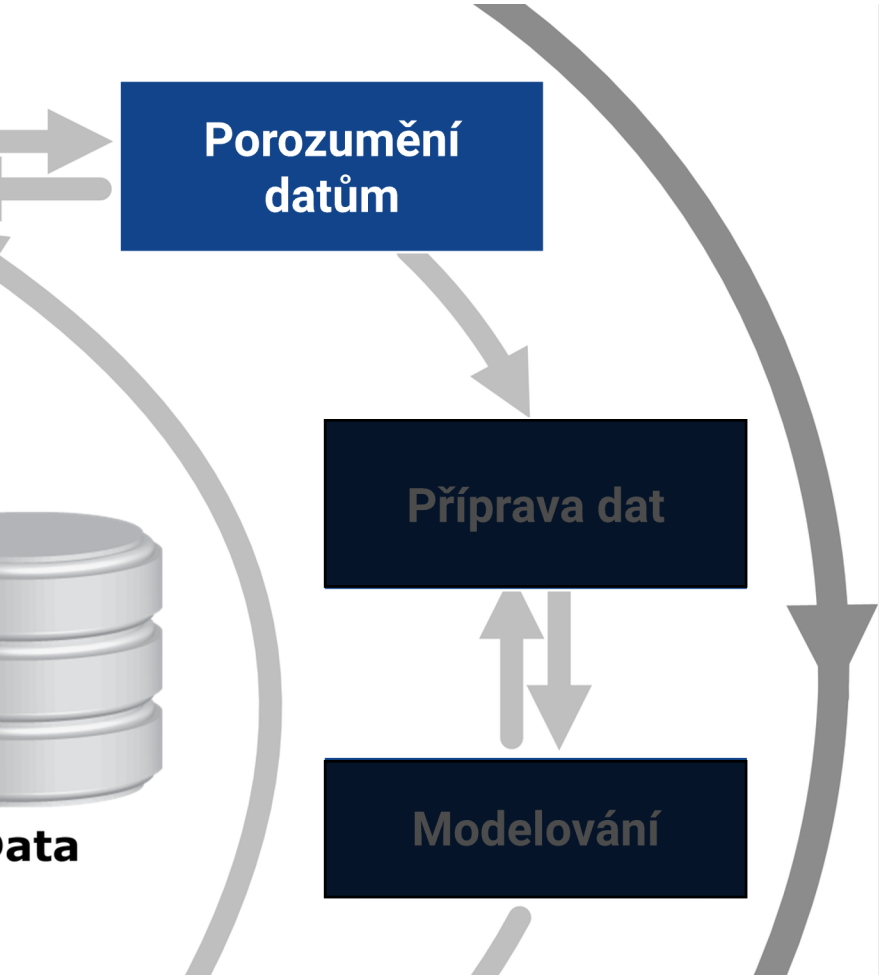
4.1 Porozumění problému



4. Výzkumná část



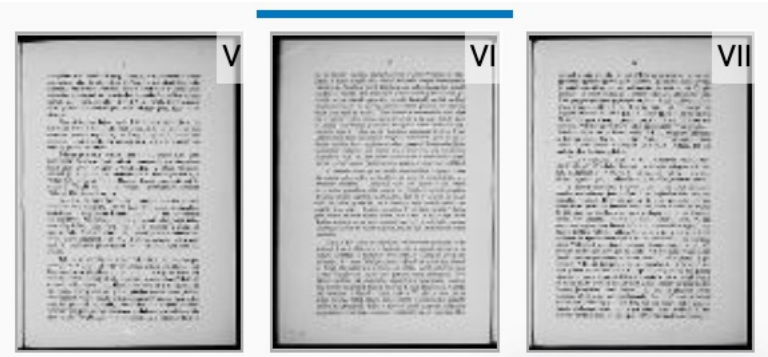
4.2 Porozumění datům



1892

Číslo 1

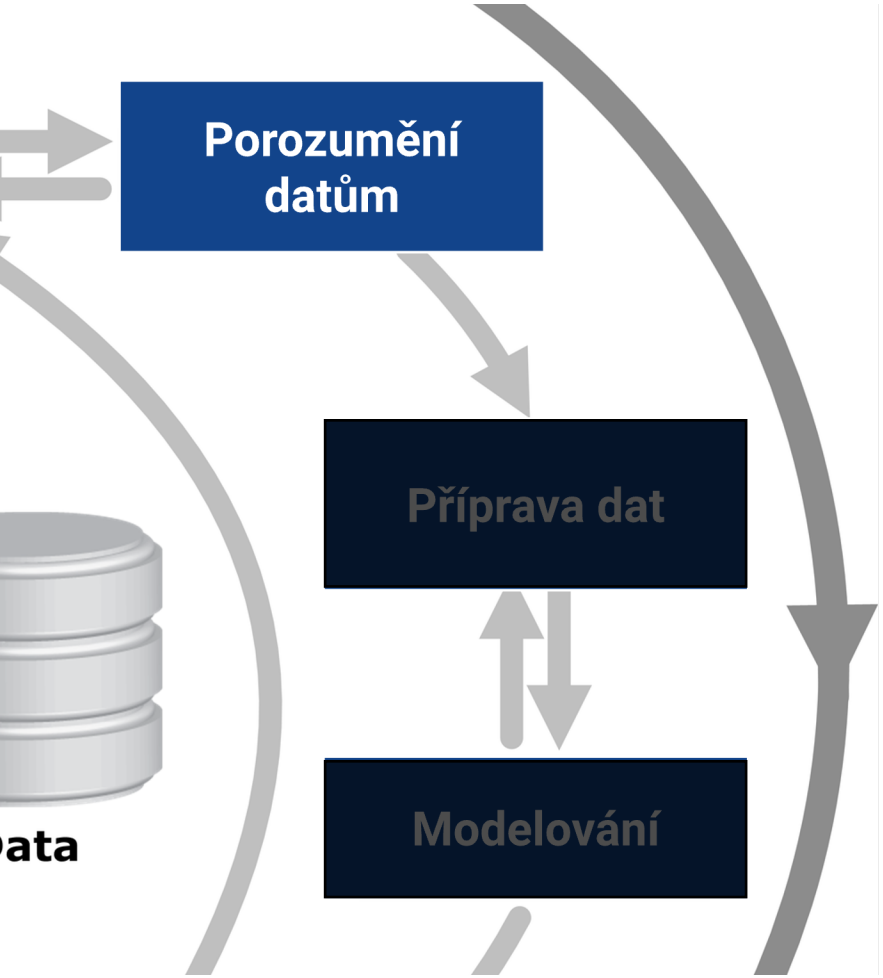
Digitální knihovna Akademie věd ČR

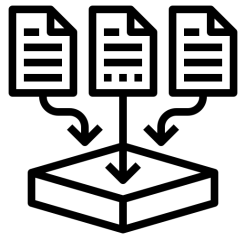


ocr | stránka | číslo periodika | ročník periodika | noviny | URL

PŘEDMLUVA.
Roku 1878 vydal jsem »Vergilstudien nebst einer Collati
V mnohých recensích vysloveno přání a očekávání, že pod
Toto pokračování podávám nyní, kdy zřízením České Akade
V práci této podány jsou k velikému počtu míst Vergilio
•'—1 Seznal jsem ovšem také tu i tam, ale celkem ve vel
l*

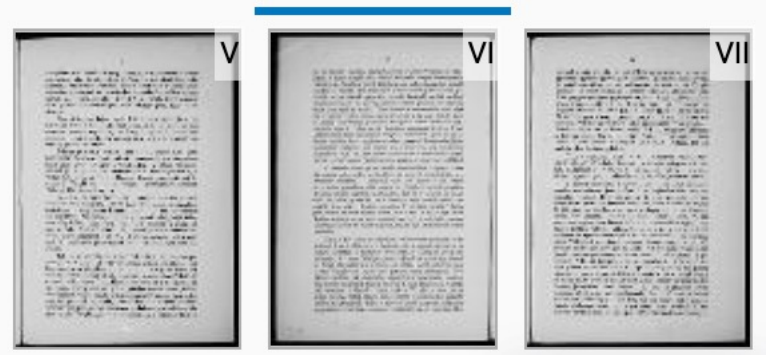
4.2 Porozumění datům





.txt soubory:
887 svazků,
31 147 stran

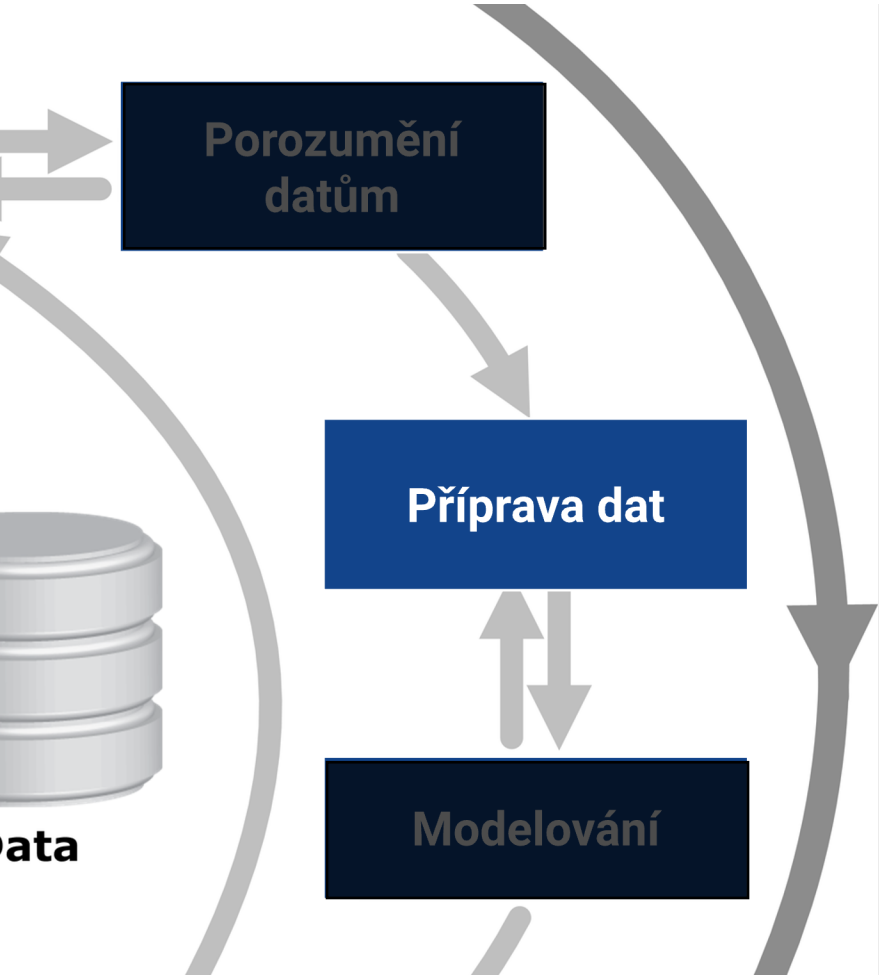
Digitální knihovna Akademie věd ČR



ocr | stránka | číslo periodika | ročník periodika | noviny | URL

```
PŘEDMLUVA.
Roku 1878 vydal jsem »Vergilstudien nebst einer Collati
V mnohých recensích vysloveno přání a očekávání, že pod
Toto pokračování podávám nyní, kdy zřízením České Akade
V práci této podány jsou k velikému počtu míst Vergilio
•'—1 Seznal jsem ovšem také tu i tam, ale celkem ve vel
l*
```

4.3 Příprava dat

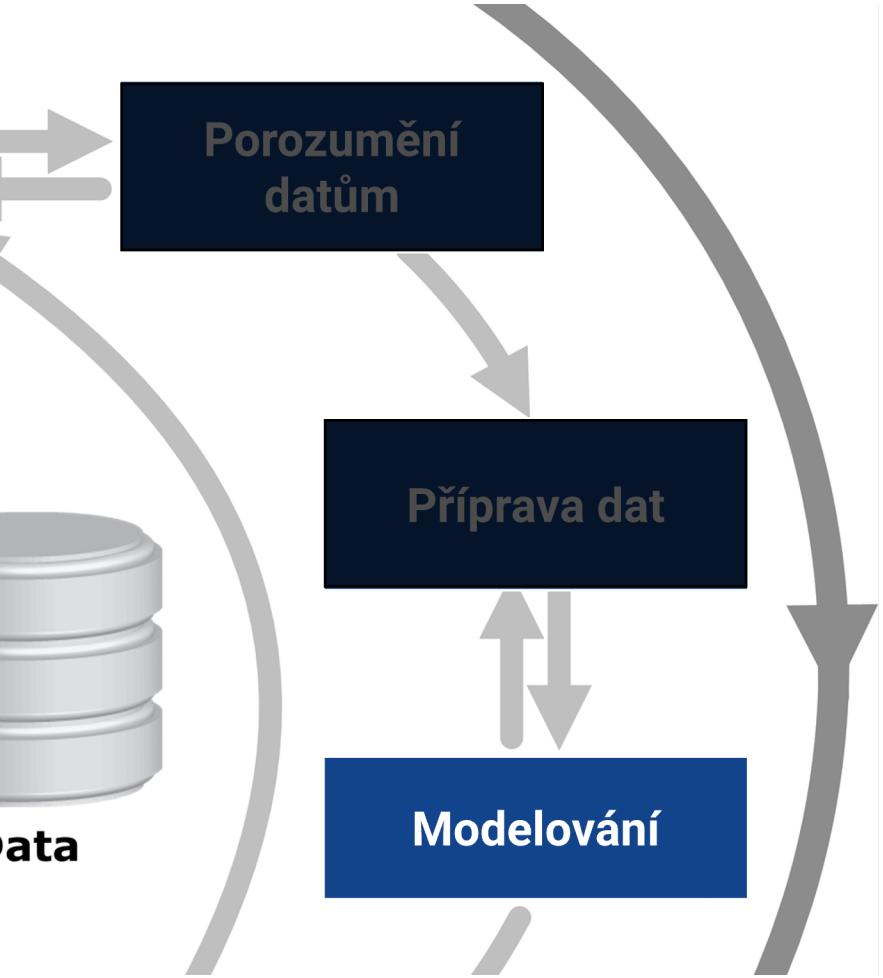


Transformace, úpravy, obohacení:

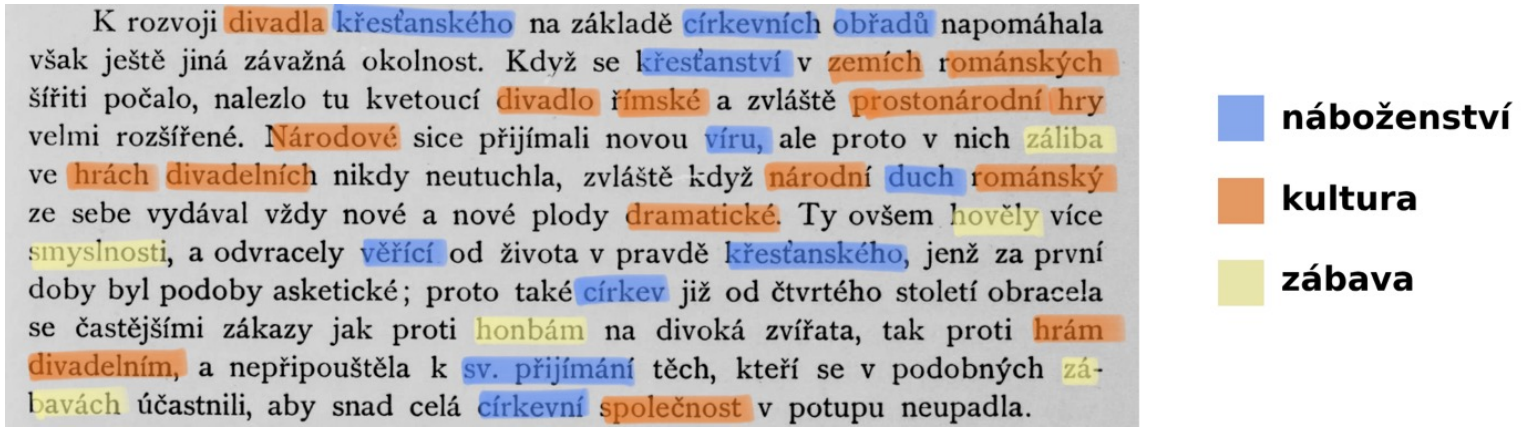
- odstraněna slova obsahující dva nebo méně znaků,
- obohacení pomocí UDPipe 2 (tokenizace, lemmatizace a identifikaci slovního druhu),
- vyfiltrována jen podstatná jména.

Pro každý svazek *Rozprav* vznikl jeden .txt soubor obsahující pouze podstatná jména v základním tvaru.

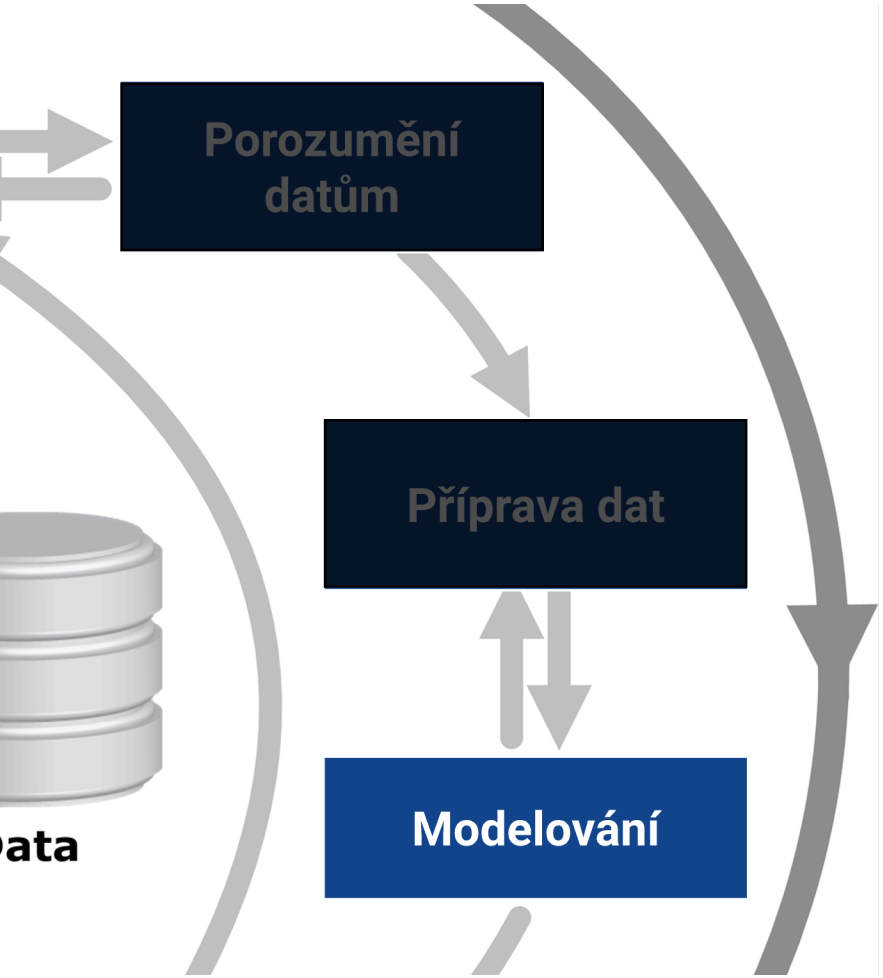
4.4 Modelování – Latentní Dirichletova Alokace (LDA)



- LDA: v každém dokumentu je možné identifikovat více různých témat (téma = seznam slov, která se spolu vyskytují statisticky významným způsobem).
- Každému významovému slovu v textu můžeme intuitivně přiřadit určité téma.
- Výsledkem takového postupu může být přehled témat, která jsou v dokumentu obsažena.



4.4 Modelování – Latentní Dirichletova Alokace (LDA)



- LDA předpokládá, že dokumenty lze reprezentovat jako náhodné směsi témat,
  - každé téma je možné charakterizovat skrze specifikovanou množinu slov.

- Pro své fungování pracuje model s předpokladem imaginárního procesu, kterým dokumenty ve zkoumané kolekci vznikly.

4.3 Odhalování témat

- Způsob, kterým se hledají neznámá témata v korpusu, je do jisté míry převrácením generativního postupu.

Co víme:

- Z dokumentů získáme dataset slov a frekvencí, se kterými se v jednotlivých dokumentech vyskytují.
- Máme, respektive vytvoříme, seznam unikátních slov vyskytujících se v korpusu.
- Parametr α, který představuje prvotní odhad rozdělení témat v dokumentech.
- Parametr β, který představuje prvotní odhad ohledně rozdělení slov v tématech.

4.3 Odhalování témat

- Způsob, kterým se hledají neznámá témata v korpusu, je do jisté míry převrácením generativního postupu.

Co musíme zjistit:

- Počet témat *K* vyskytujících se v korpusu.
- Rozdělení témat v jednotlivých dokumentech θ.
- Rozdělení slov v jednotlivých tématech ϕ.
- Přidělení slov k jednotlivým tématům Z.

4.3 Odhalování témat

- Způsob, kterým se hledají neznámá témata v korpusu, je do jisté míry převrácením generativního postupu.

Co musíme zjistit:

- **<u>Počet témat *K* vyskytujících se v korpusu.</u>**
- Rozdělení témat v jednotlivých dokumentech θ.
- Rozdělení slov v jednotlivých tématech ϕ.
- Přidělení slov k jednotlivým tématům Z.

$$p(\theta, \phi, z|w, \alpha\beta,) = \frac{p(\theta, \phi, z, w|\alpha\beta,)}{p(w|\alpha\beta,)}$$

DALRING, W.M. A *Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling*. Online výukový materiál. University of Guelph, 2011. Dostupné z: https://coli-saar.github.io/cl19/materials/darling-lda.pdf. [cit. 2023-01-11].

4.3 Odhalování témat

- Způsob, kterým se hledají neznámá témata v korpusu, je do jisté míry převrácením generativního postupu.

Co musíme zjistit:

- **Počet témat *K* vyskytujících se v korpusu.**
  - Volba tohoto parametru je pro celý proces zcela zásadní a umožňuje badateli vstoupit do procesu modelování témat s vlastní představou o tematické granularitě korpusu.*

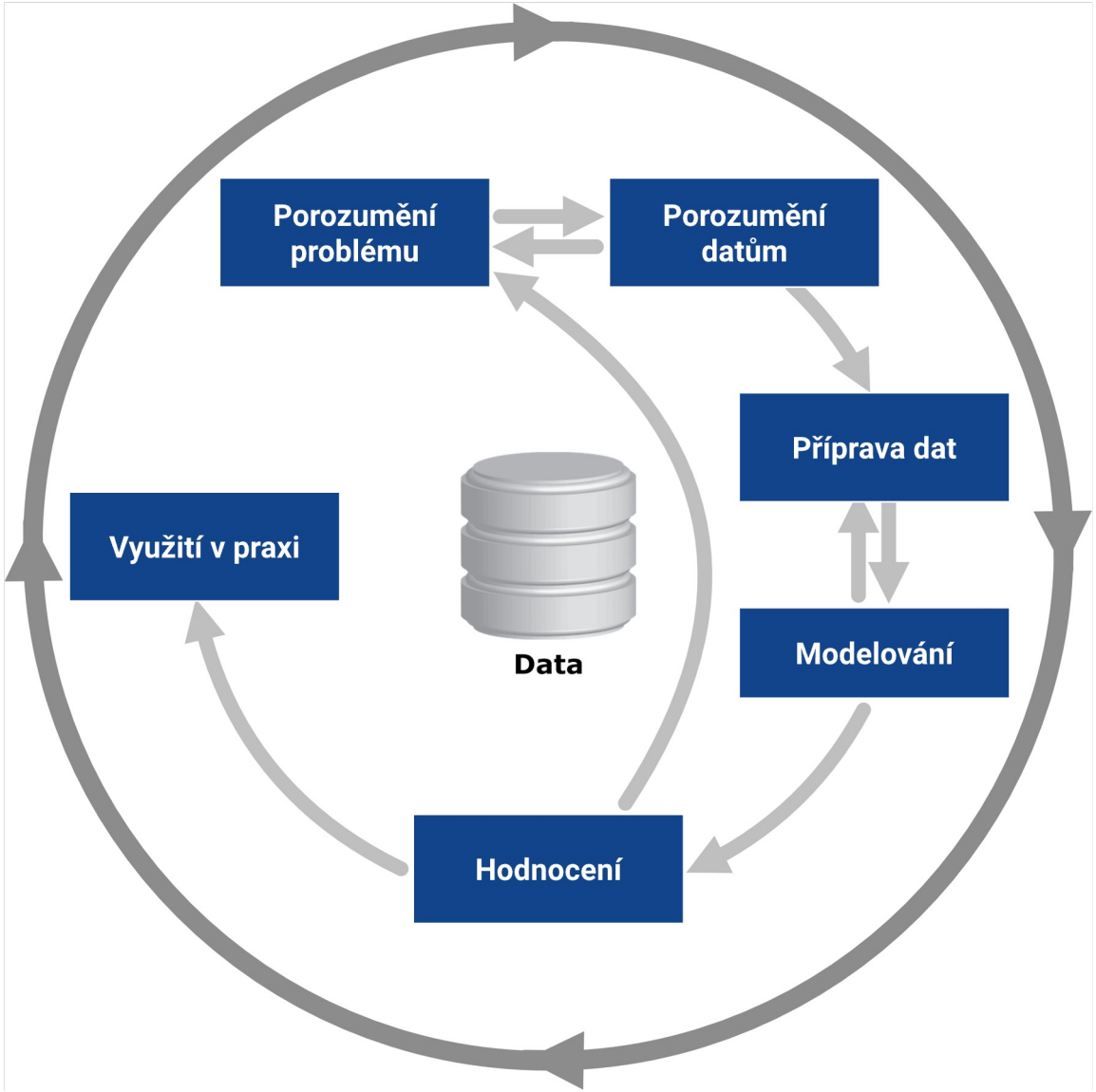
* HLADÍK, Radim. Modelování témat v české sociologii: typy autorství a citační ohlas v odborných textech. In: HLADÍK, Radim (ed.). *Digitální obrat v českých humanitních a sociálních vědách.* Studia nových médií. Praha: Univerzita Karlova, nakladatelství Karolinum, 2022. ISBN: 978-80-246-5193-4.

4.3 Stanovení cílů a počtu témat

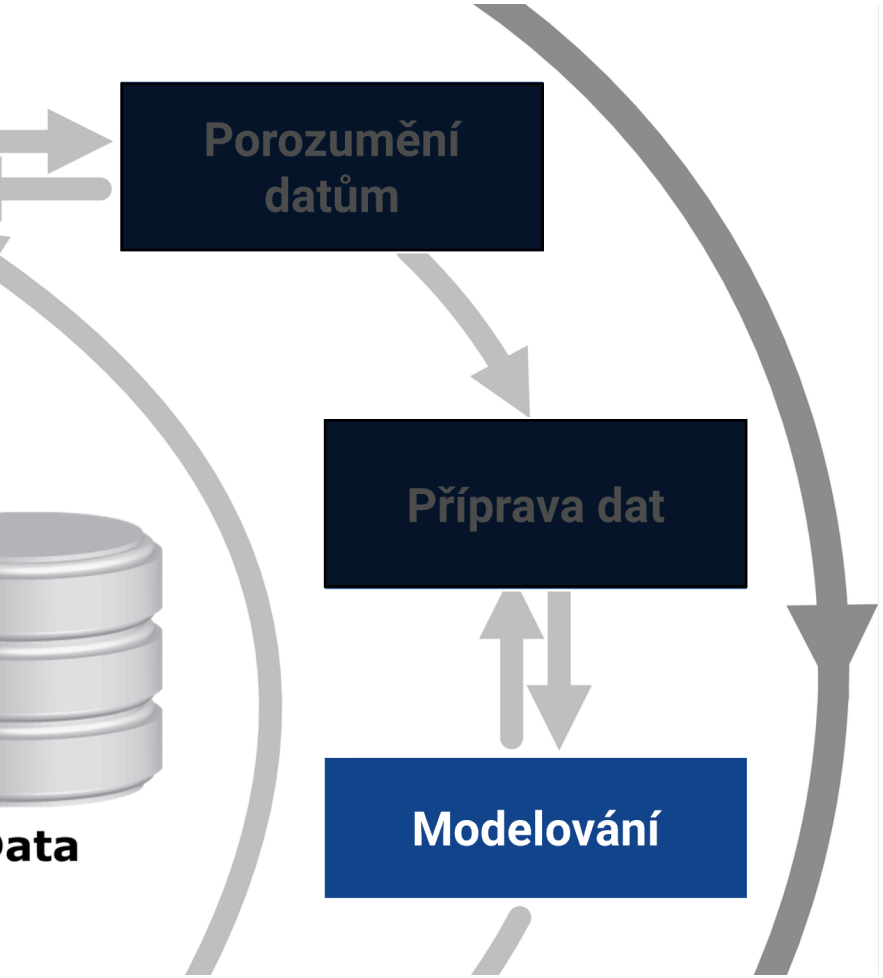
- **Cílem modelování je identifikovat témata s alespoň o úroveň jemnější granularitou, než se to podařilo doposud.**
  - Pokud se tedy publikace věnovaly například chemii, cílem je zjistit, kterým konkrétním chemickým oborům (organická chemie, analytická chemie, anorganická chemie, …)

- Celkem by modelování mělo identifikovat **35 témat.**
  - Toto číslo bylo zvoleno na základě kvalifikovaného odhadu vycházejícího z počtu sekcí vědních oblastí současné AV ČR (9) a počtem jejích jednotlivých pracovišť (54).

Výzkumná část



Modelování



- Vstupem pro model LDA je **matice dokument–slovo.**
- K samotnému modelování témat byla využita knihovna `topicmodels`*.

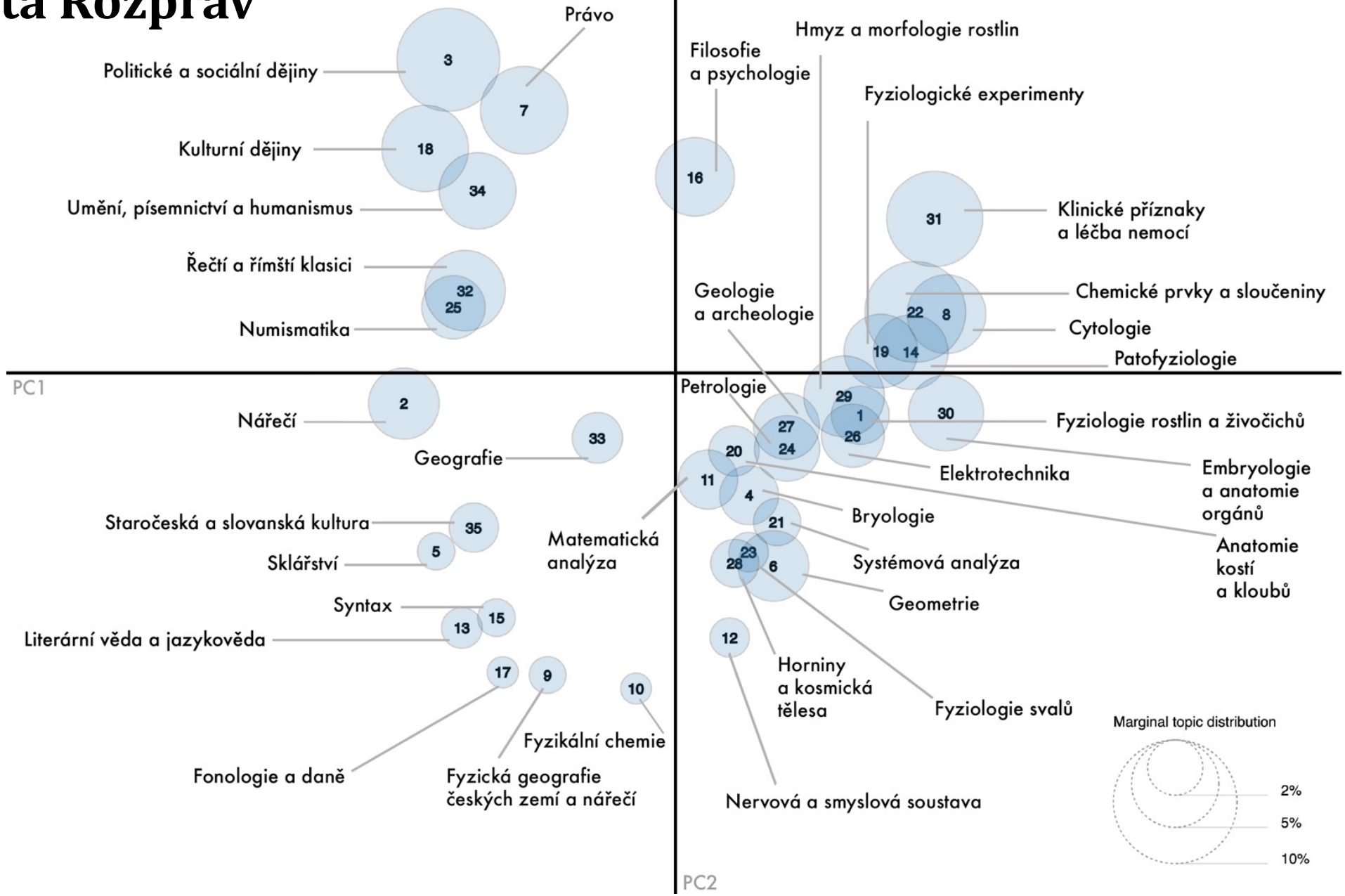
* GRÜN, Bettina. *Package 'topicmodels'* Online. Verze 0.2-15, 27. 11. 2023. Dostupné z: https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf. [cit. 2023-12-07].

Modelování

- Výstupem modelu jsou dvě matice:
  - **slovo–téma**, ve které je u každého slova z korpusu uvedeno,
    s jakou pravděpodobností „patří“ k jakému tématu.
  - **dokument–téma** popisující pravděpodobnost příslušnosti dokumentu
    ke konkrétnímu tématu.

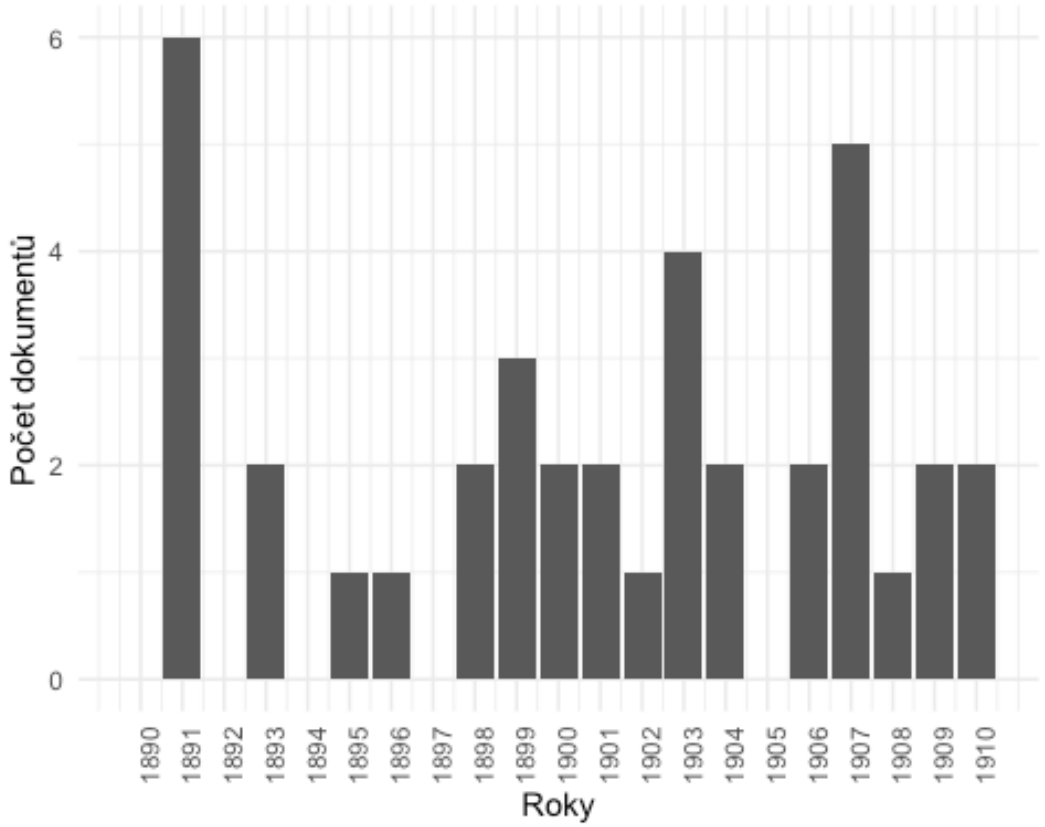
- **Model rozpoznaná témata v korpusu označuje pouze číslem**,
  pro pojmenování byly využity tři techniky:
  - nejfrekventovanější slova, která se v každém tématu vyskytují (17 případů),
  - analýza začátků dokumentů patřících k danému tématu s nejvyšší pravděpodobností,
  - analýza zdrojových digitalizátů.

Témata Rozprav



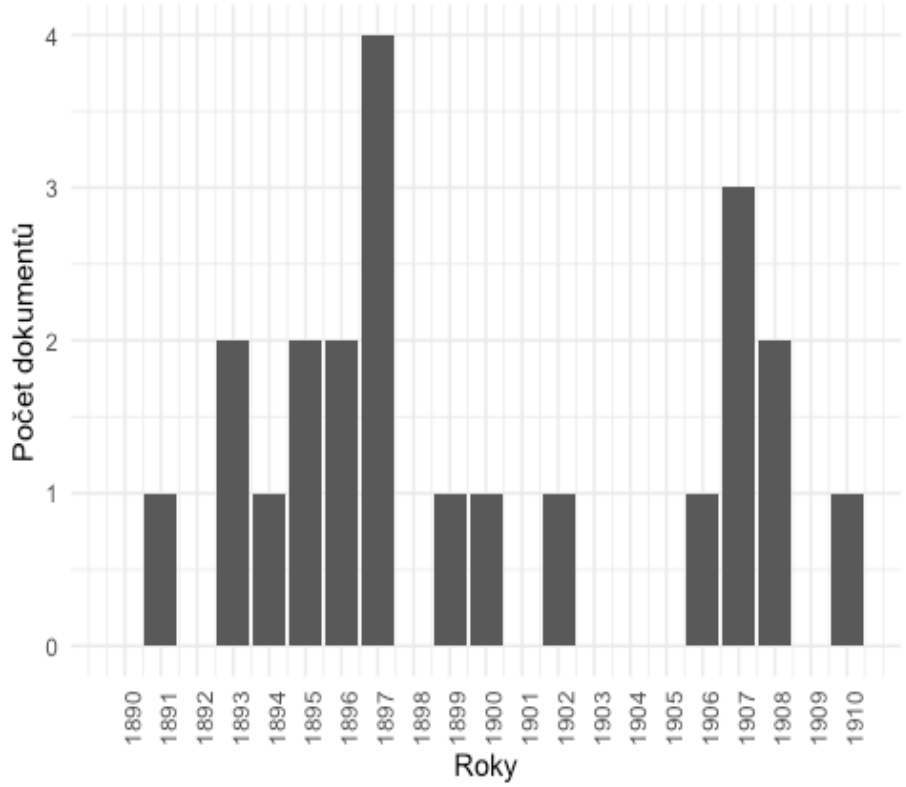
Příklad: **Téma č. 29: Hmyz a morfologie rostlin**

- Celkem 38 svazků *Rozprav*.
- Spojuje různé obory biologie.

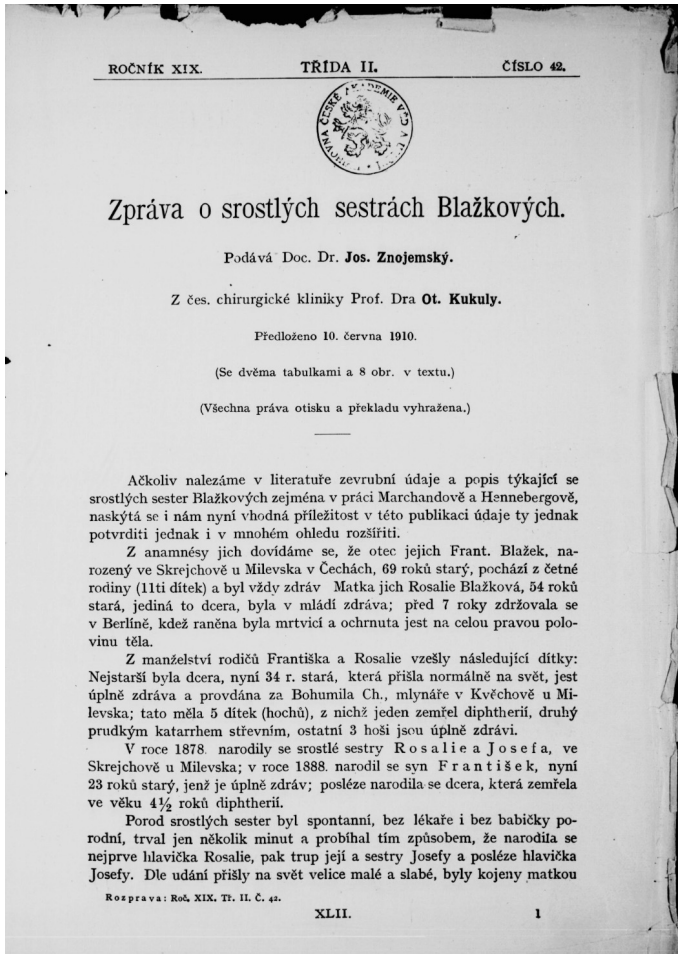


Příklad: **Téma č. 20: Anatomie kostí a kloubů**

- Celkem 22 svazků *Rozprav*.
- svazky zařazené pod toto téma spojuje zejména makroskopický zájem o pojivové tkáně těla, neobsahují ale pouze lékařskou teorii.



Příklad: **Téma č. 20: Anatomie kostí a kloubů**



ROČNÍK XIX. TŘÍDA II. ČÍSLO 42.

Zpráva o srostlých sestrách Blažkových.

Podává Doc. Dr. **Jos. Znojemský.**

Z čes. chirurgické kliniky Prof. Dra **Ot. Kukuly.**

Předloženo 10. června 1910.

(Se dvěma tabulkami a 8 obr. v textu.)

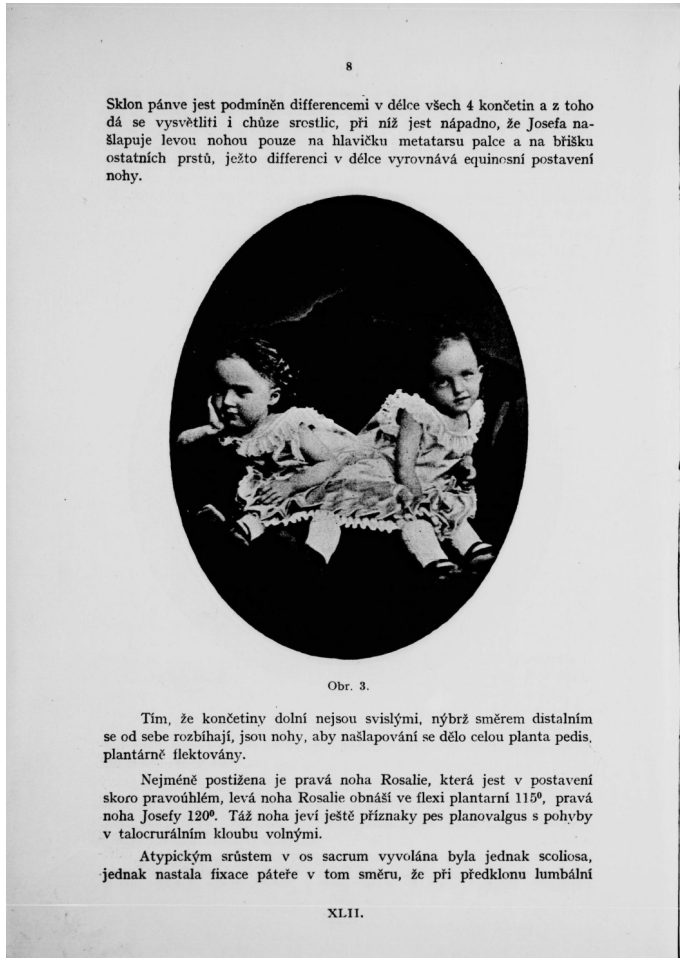
Ačkoliv nalezáme v literatuře zevrubní údaje a popis týkající se srostlých sester Blažkových zejména v práci Marchandově a Hennebergově, naskýtá se i nám nyní vhodná příležitost v této publikaci údaje ty jednak potvrditi jednak i v mnohém ohledu rozšířiti.

Z anamnésy jich dovídáme se, že otec jejich Frant. Blažek, narozený ve Skrejchově u Milevska v Čechách, 69 roků starý, pochází z četné rodiny (11ti dítek) a byl vždy zdráv Matka jich Rosalie Blažková, 54 roků stará, jediná to dcera, byla v mládí zdráva; před 7 roky zdržovala se v Berlíně, kdež raněna byla mrtvicí a ochrnuta jest na celou pravou polovinu těla.

Z manželství rodičů Františka a Rosalie vzešly následující dítky: Nejstarší byla dcera, nyní 34 r. stará, která přišla normálně na svět, jest úplně zdráva a provdána za Bohumila Ch., mlynáře v Kvěchově u Milevska; tato měla 5 dítek (hochů), z nichž jeden zemřel diphtherií, druhý prudkým katarrhem střevním, ostatní 3 hoši jsou úplně zdrávi.

V roce 1878. narodily se srostlé sestry Rosalie a Josefa, ve Skrejchově u Milevska; v roce 1888. narodil se syn František, nyní 23 roků starý, jenž je úplně zdráv; posléze narodila se dcera, která zemřela ve věku 4½ roků diphtherií.

Porod srostlých sester byl spontanní, bez lékaře i bez babičky porodní, trval jen několik minut a probíhal tím způsobem, že narodila se nejprve hlavička Rosalie, pak trup její a sestry Josefy a posléze hlavička Josefy. Dle udání přišly na svět velice malé a slabé, byly kojeny matkou



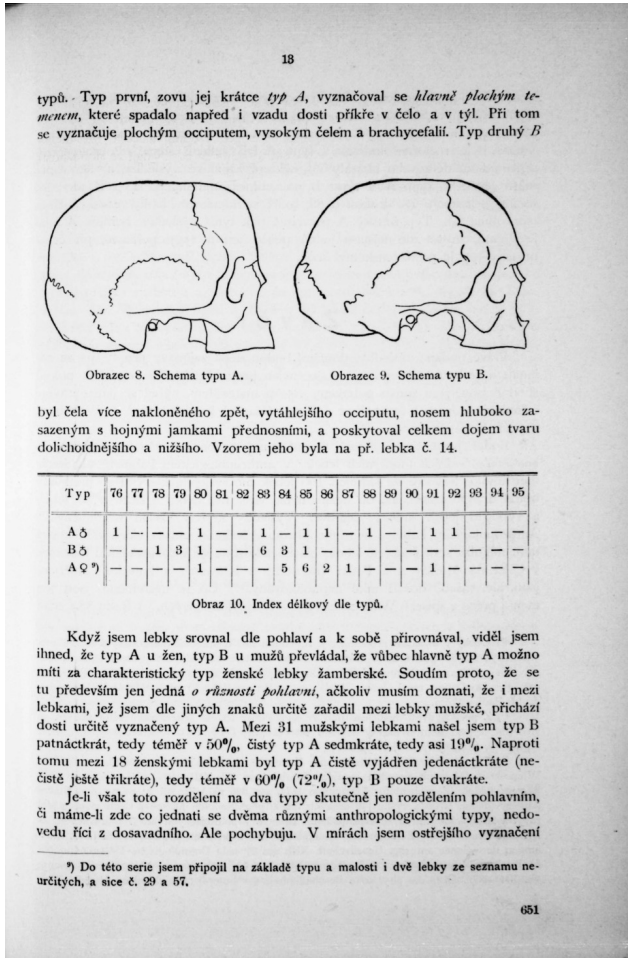
Sklon pánve jest podmíněn differencemi v délce všech 4 končetin a z toho dá se vysvětliti i chůze srostlic, při níž jest nápadno, že Josefa našlapuje levou nohou pouze na hlavičku metatarsu palce a na břišku ostatních prstů, ježto differenci v délce vyrovnává equinosní postavení nohy.

Obr. 3.

Tím, že končetiny dolní nejsou svislými, nýbrž směrem distalním se od sebe rozbíhají, jsou nohy, aby našlapování se dělo celou planta pedis, plantárně flektovány.

Nejméně postižena je pravá noha Rosalie, která jest v postavení skoro pravoúhlém, levá noha Rosalie obnáší ve flexi plantarní 115°, pravá noha Josefy 120°. Táž noha jeví ještě příznaky pes planovalgus s pohyby v talocrurálním kloubu volnými.

Atypickým srůstem v os sacrum vyvolána byla jednak scoliosa, jednak nastala fixace páteře v tom směru, že při předklonu lumbální



typů. Typ první, zovu jej krátce *typ A*, vyznačoval se *hlavně plochým temenem*, které spadalo napřed i vzadu dosti příkře v čelo a v týl. Při tom se vyznačuje plochým occiputem, vysokým čelem a brachycefalií. Typ druhý *B*

Obrazec 8. Schema typu A.

Obrazec 9. Schema typu B.

byl čela více nakloněného zpět, vytáhlejšího occiputu, nosem hluboko zasazeným s hojnými jamkami přednosními, a poskytoval celkem dojem tvaru dolichoidnějšího a nižšího. Vzorem jeho byla na př. lebka č. 14.

Typ	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
A ♂	1	—	—	—	1	—	—	1	—	1	1	—	1	—	—	1	1	—	—	—
B ♂	—	—	1	3	1	—	—	6	3	1	—	—	—	—	—	—	—	—	—	—
A ♀*)	—	—	—	—	1	—	—	—	5	6	2	1	—	—	—	1	—	—	—	—

Obraz 10. Index délkový dle typů.

Když jsem lebky srovnal dle pohlaví a k sobě přirovnával, viděl jsem ihned, že typ A u žen, typ B u mužů převládal, že vůbec hlavně typ A možno míti za charakteristický typ ženské lebky žamberské. Soudím proto, že se tu především jen jedná *o různosti pohlavní*, ačkoliv musím doznati, že i mezi lebkami, jež jsem dle jiných znaků určitě zařadil mezi lebky mužské, přichází dosti určitě vyznačený typ A. Mezi 31 mužskými lebkami našel jsem typ B patnáctkrát, tedy téměř v 50%, čistý typ A sedmkráte, tedy asi 19%. Naproti tomu mezi 18 ženskými lebkami byl typ A čistě vyjádřen jedenáctkráte (nečistě ještě třikráte), tedy téměř v 60% (72%), typ B pouze dvakráte.

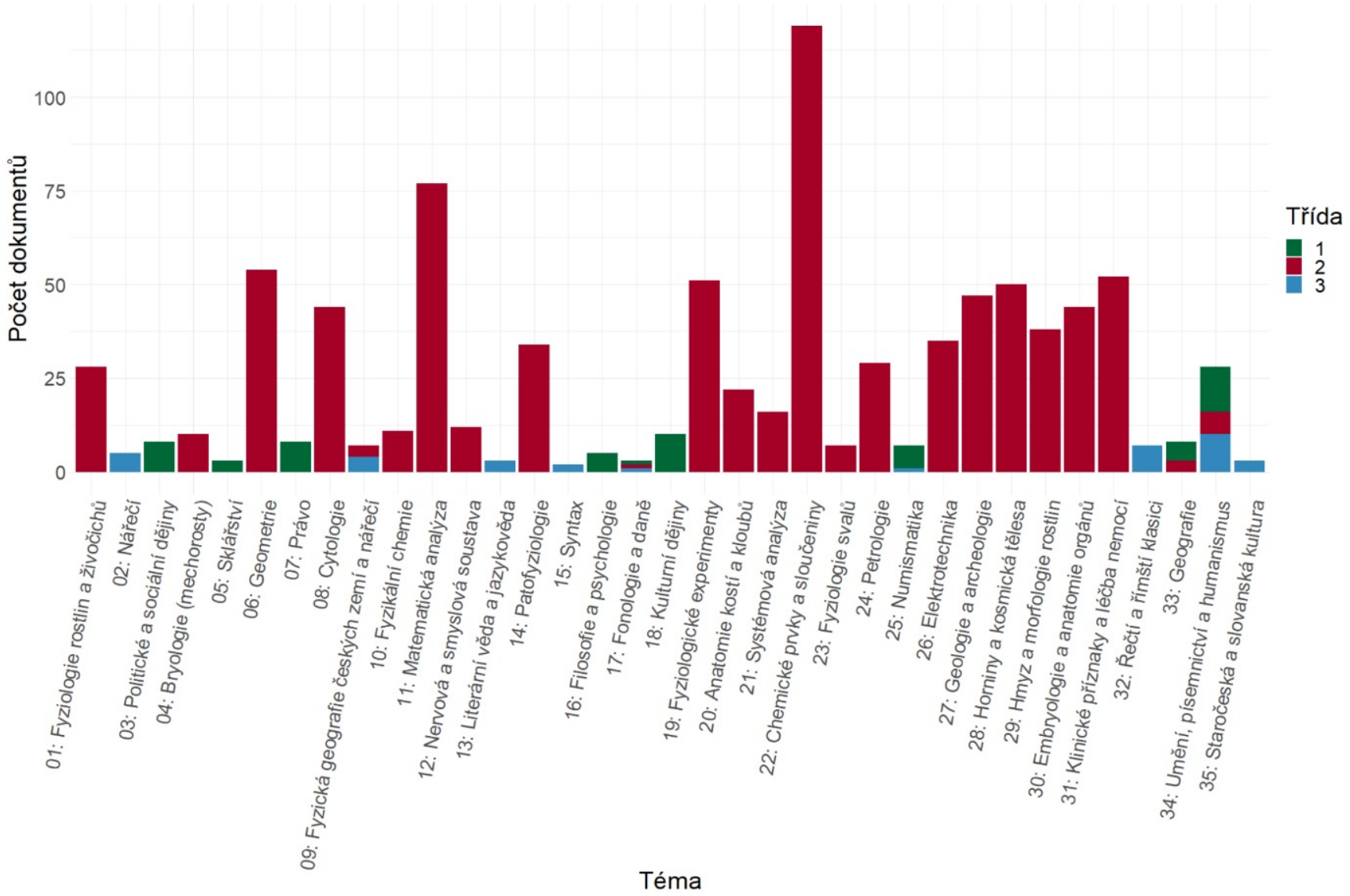
Je-li však toto rozdělení na dva typy skutečně jen rozdělením pohlavním, či máme-li zde co jednati se dvěma různými anthropologickými typy, nedovedu říci z dosavadního. Ale pochybuju. V mírách jsem ostřejšího vyznačení

*) Do této serie jsem připojil na základě typu a malosti i dvě lebky ze seznamu neurčitých, a sice č. 29 a 57.

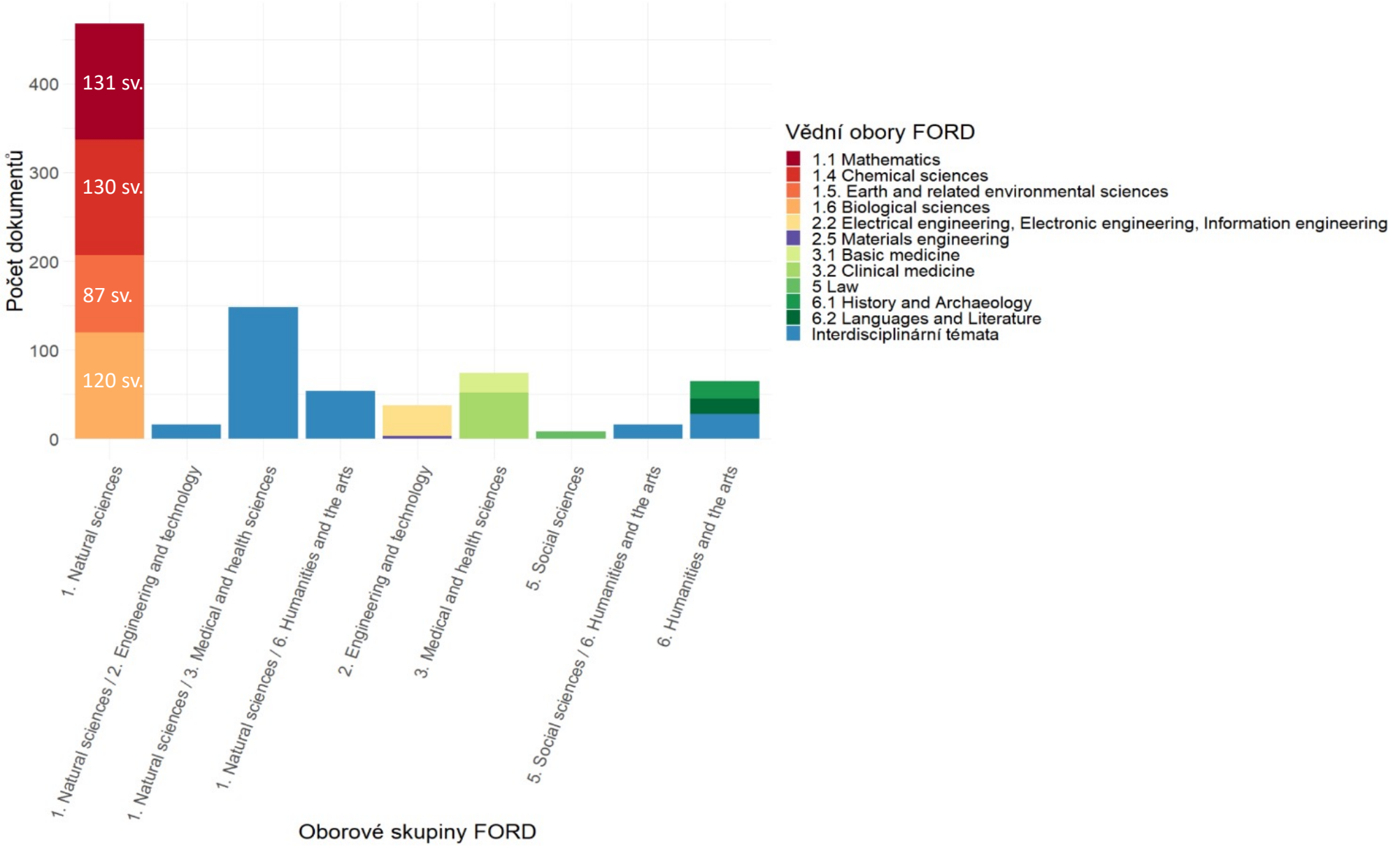
Rozpravy České akademie císaře Františka Josefa pro vědy, slovesnost a umění. Třída II, Mathematicko-přírodnická. Praha: Česká akademie císaře Františka Josefa pro vědy slovesnost a umění, 1910, 19(42). ISSN 1210-874X. Dostupné také z: https://kramerius.lib.cas.cz/uuid/uuid:c221fe93-435d-11dd-b505-00145e5790ea

Rozpravy České akademie císaře Františka Josefa pro vědy, slovesnost a umění. Třída II, Mathematicko-přírodnická.. Praha: Česká akademie císaře Františka Josefa pro vědy slovesnost a umění, 1892, 1(31). ISSN 1210-874X. Dostupné také z: https://kramerius.lib.cas.cz/uuid/uuid:c256a329-435d-11dd-b505-00145e5790ea

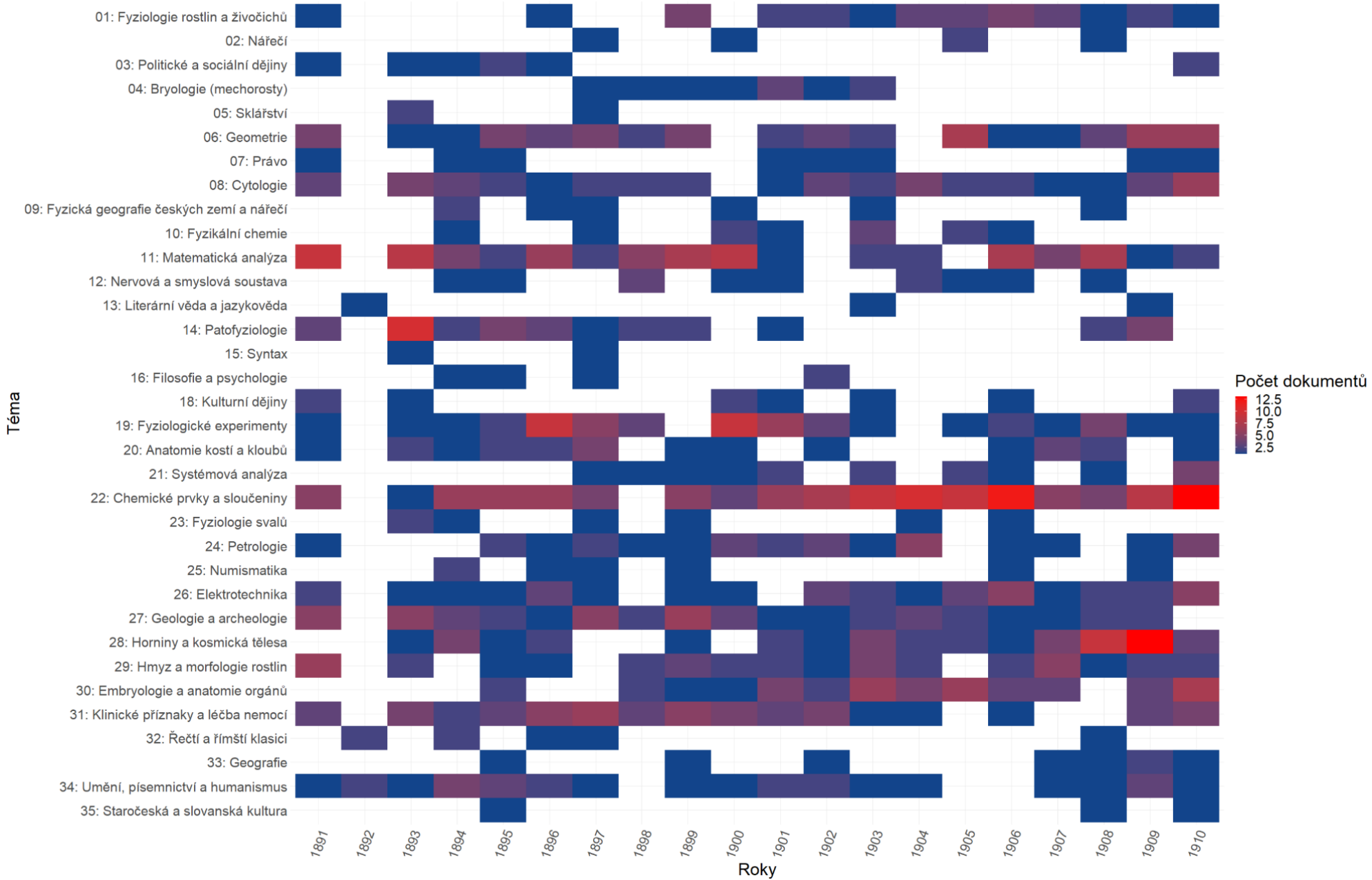
Porovnání počtu dokumentů v tématech



Počet svazků podle klasifikace FORD



Přehled počtu svazků Rozprav podle témat v průběhu let 1890–1910



Výsledky

- Provedená analýza podává přehled **konkrétních témat, kterým se Česká akademie ve sledovaném období věnovala**,
  - jejich sémantické podobnosti,
  - počtu svazků zařazených do každého z témat,
  - a proměn počtu vydaných svazků v jednotlivých letech.

Výsledky

- Provedená analýza podává přehled **konkrétních témat, kterým se Česká akademie ve sledovaném období věnovala**,
  - jejich sémantické podobnosti,
  - počtu svazků zařazených do každého z témat,
  - a proměn počtu vydaných svazků v jednotlivých letech.
- Práce **představuje podrobný pohled na publikační činnost České akademie jako celek** a umožňuje rozšířit dosavadní stav poznání oblastí zájmů této instituce.

Výsledky

- Provedená analýza podává přehled **konkrétních témat, kterým se Česká akademie ve sledovaném období věnovala**,
  - jejich sémantické podobnosti,
  - počtu svazků zařazených do každého z témat,
  - a proměn počtu vydaných svazků v jednotlivých letech.
- Práce **představuje podrobný pohled na publikační činnost České akademie jako celek** a umožňuje rozšířit dosavadní stav poznání oblastí zájmů této instituce.
- Diplomová práce ukázala, že **dokumenty dostupné v digitálních knihovnách lze použít k výzkumu v digitálních humanitních vědách**, přestože hlavními důvody jejich digitalizace je zajištění jejich dlouhodobé ochrany a možnosti online zpřístupnění běžným čtenářům.

Závěr

- Součástí diplomové práce jsou 3 přílohy:
  - **Příloha č. 1:** Seznam publikací vydaných Českou akademií věd císaře Františka Josefa pro vědy, slovesnost a umění z období let 1890–1918.
  - **Příloha č. 2:** Postup stažení a přípravy dat a modelování témat
    - Data a scripty byly zveřejněny také na platformě GitHub: https://github.com/kerschfilip/tematicke_modelovani_cavu
    - Veškerá data jsou dostupná v repozitáři Zenodo: https://doi.org/10.5281/zenodo.10395970
  - **Příloha č. 3**: Seznam odkazů na digitalizované svazky časopisu Rozpravy rozřazený podle rozpoznaných témat

Diskuse

- Identifikovaná témata považovat za **úplný** přehled oblastí zájmu akademie:
  - Analyzovány pouze *Rozpravy* – ČAVU vydávala i monografie a odborné časopisy.
  - Není pokryta publikační činnost jiných učených společností – model lze ale rozšířit.

- Zapojení většího množství digitalizovaných dokumentů do analýzy znesnadňuje jejich rozptýlení po relativně velkém množství digitálních knihoven.
  - Kromě dostupnosti představuje určitý limit pro automatickou analýzu dat i kvalita zejména starších digitalizátů.

- **Evaluace témat** (Možnosti matematického ověření koherence / posuzování lidmi)

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Diplomová práce

Odborné fórum, ÚISK FF UK, 4. 3. 2024

Filip Kersch, FilipKersch@gmail.com, kersch@knav.cz