Sociologický ústav
Akademie věd ČR

Centrum pro výzkum veřejného mínění

Využití dat webových archivů v sociálních vědách

Paulína Tabery, Matouš Pilnáček
Workshop „Využití dat webových archivů: Možnosti a limity“
21.10.2019

Webové archivy jako výzva pro sociální vědy

- Debata o využití dat archivovaných ve webových archivech pro společenskovědní výzkum se v posledních letech intenzivně rozvíjí (viz Brügger a Schroeder, 2017)
- Webové archivy umožňují srovnání více časových bodů, případně přímo longitudinální analýzu
- Prezentace je zaměřena na kvantitativní metody
- Data lze využít (a skutečně se využívají) i pro kvalitativní výzkum, ale je zapotřebí mít přístup k celým stránkám

Sociologický ústav
Akademie věd ČR

Centrum
pro výzkum
veřejného
mínění

Příklady studií a analýz využívajících stažené (archivované) webové stránky

2005

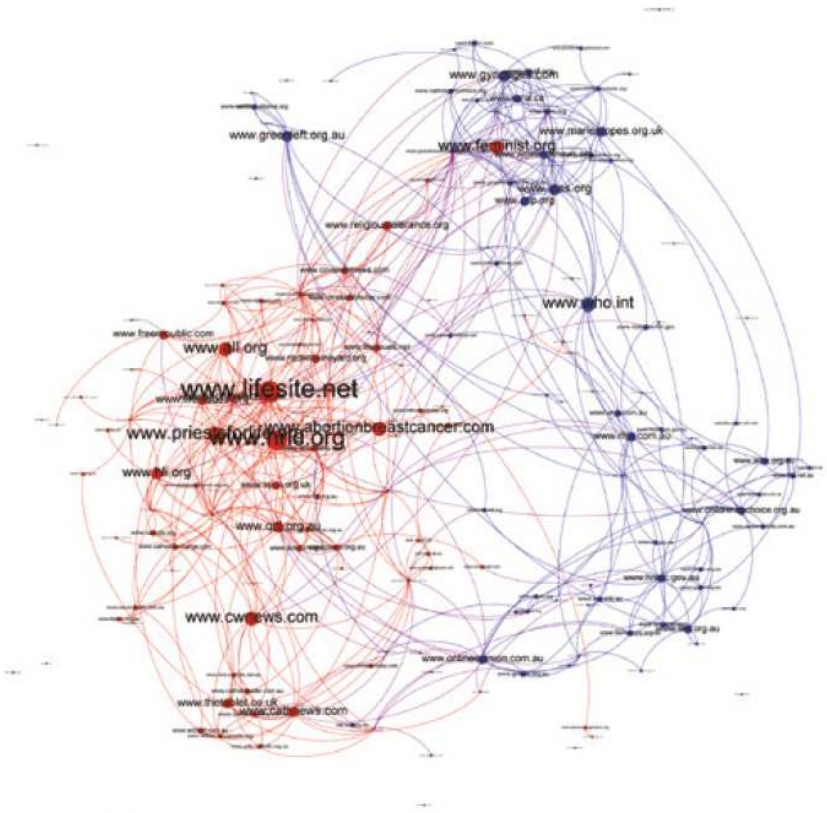


Figure 8.1 Hyperlink network of participants in abortion debate in Australia, 2005. Note: pro-life – red, pro-choice – blue. Node size is proportional to indegree

2015

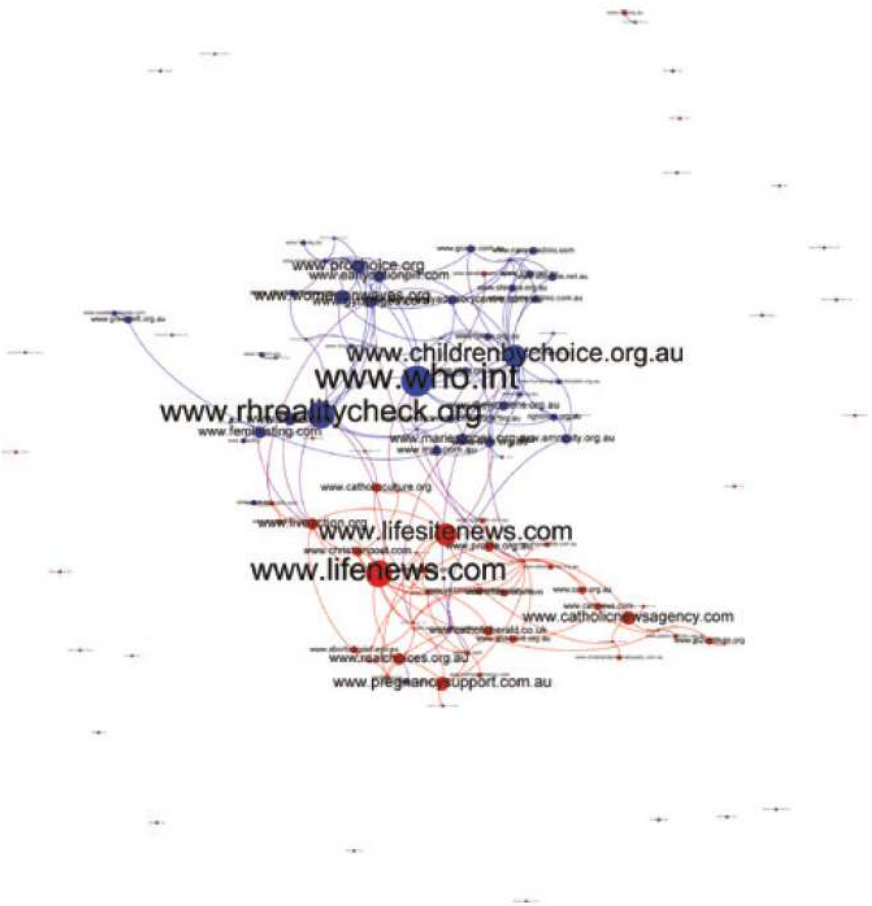


Figure 8.2 Hyperlink network of participants in abortion debate in Australia, 2015. Note: pro-life – red, pro-choice – blue. Node size is proportional to indegree

(Ackland a Evans, 2017)

2005

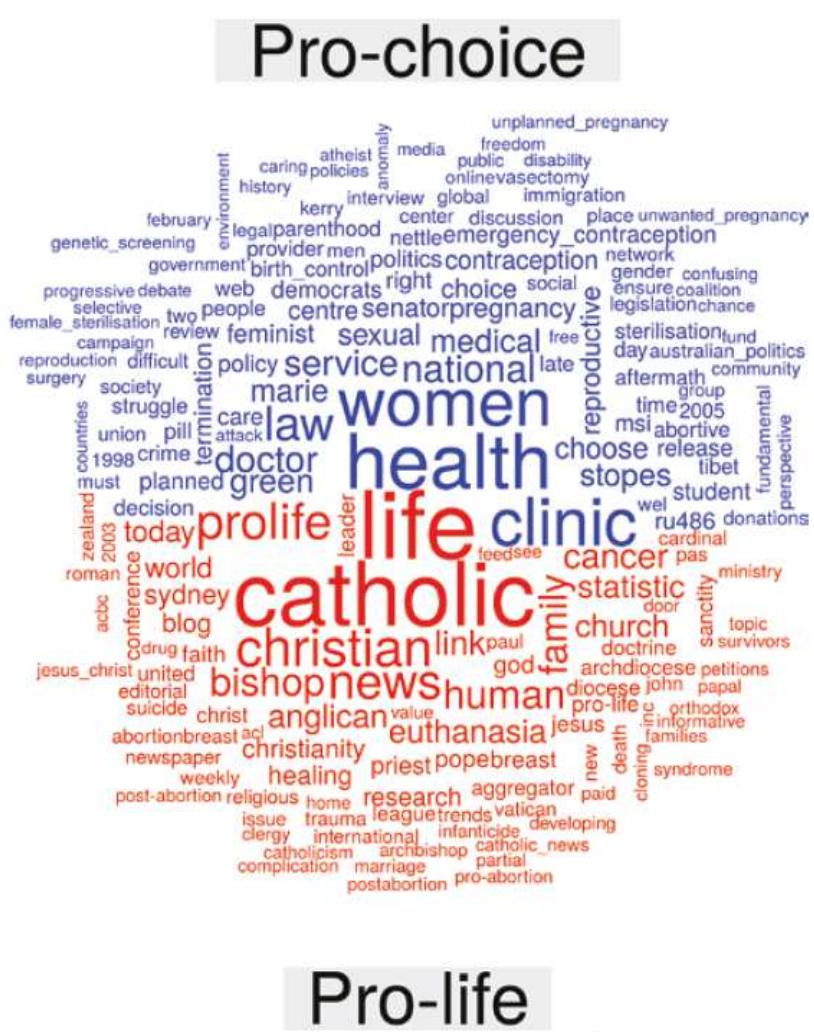


Figure 8.7 Comparison cloud (meta words) – 2005

2015

Figure 8.8 Comparison cloud (meta words) – 2015

(Ackland a Evans, 2017)

Počet odkazů ze zpravodajského webu BBC na pět nejčastějších domén zemí

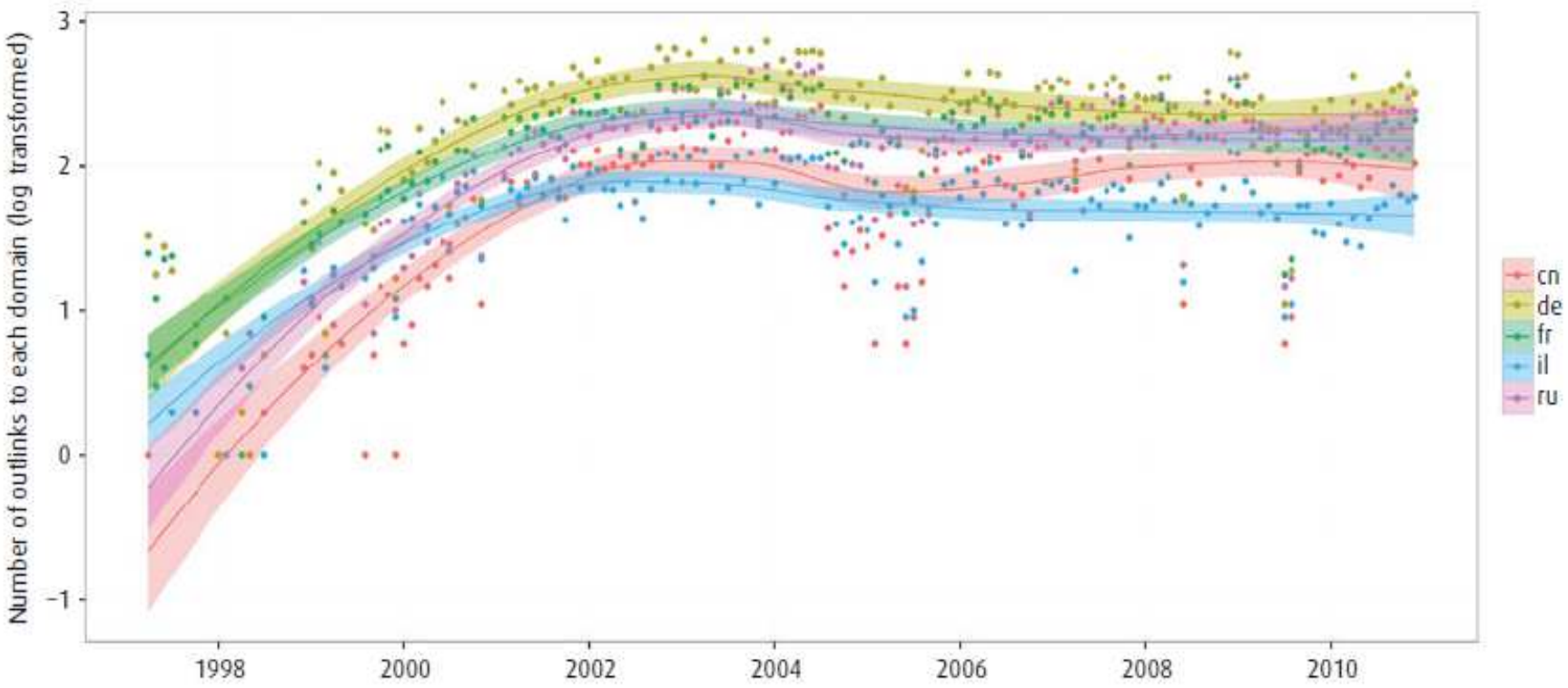


Figure 5.1 Evolution of outlinks to top five country domains over time

(Cowls a Bright, 2017)

Lineární regrese vysvětlující množství zmínek zemí na webu BBC

Variable	Coefficient
Population (log transformed)	0.63***
Trade with UK (log transformed)	−0.09
GDP per capita (log transformed)	0.32
Distance from UK (log transformed)	0.08
Homicide rate	−0.02*
Peace Index	0.53*
Disaster risk	−3.13
Commonwealth ember	−0.34
Internet penetration	0
English as an official or primary language	0.89*
adj. R-squared	0.43
N	148

Lineární regrese vysvětlující množství odkazů na stránky s doménou zemí z webu BBC

Variable	Coefficient
Population (log transformed)	0.52***
Trade with UK (log transformed)	0.02
GDP per capita (log transformed)	0.11
Distance from UK (log transformed)	0.04
News mentions (log transformed)	0.16*
Homicide rate	0.01
Peace Index	−0.44*
Disaster risk	−3.34
Commonwealth member	−0.41
Internet penetration	0.02***
English as an official or primary language	0.66*
adj. R-squared	0.69
N	148

(Cowls a Bright, 2017)

Sociologický ústav
Akademie věd ČR

Centrum
pro výzkum
veřejného
mínění

Jaké jsou problémy a jaká jejich řešení při využití webových archivů pro sociální vědy?

1. Etická omezení

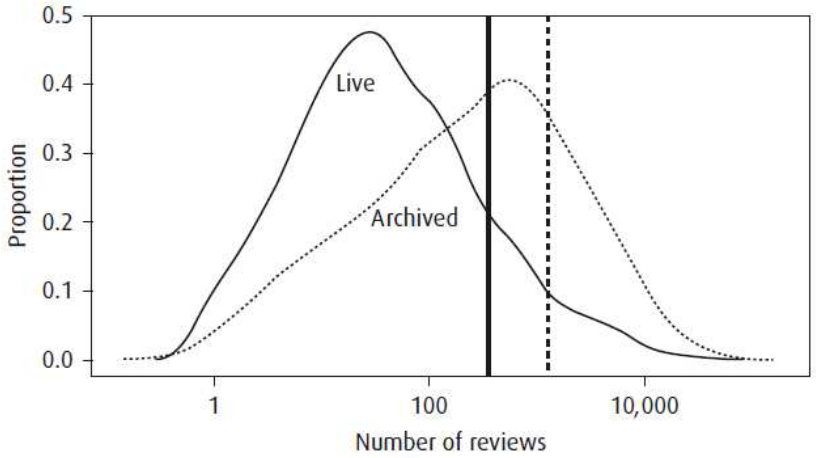
- Ochrana osobních údajů
- Sběr a analýza dat může vytvořit novou kvalitu
- Centralizované rozhraní tento problém eliminuje

2. Nereprezentativita dat (celoplošné sklizně)

- Kapacita archivu je omezena a není proto možné archivovat veškerý obsah českého internetu
- Program stahující stránky prochází obsah na základě odkazů z jiných stránek
- Stránky na které je častěji odkazováno, jsou tak častěji archivovány

- V tuto chvíli nemá jasné řešení
- Nabízí se dělat z archivu reprezentativní výběr
  - Je ovšem zapotřebí kvalitní opora výběru
- Pro několik témat lze využít tematické sklizně
  - Volby, Uprchlická krize, Povodně 2013, apod.

Distribuce počtu archivovaných stránek webu TripAdvisor z hlediska počtu uživatelských recenzí na živém webu a v britském webovém archivu (Hale, Blank a Alexander, 2017))



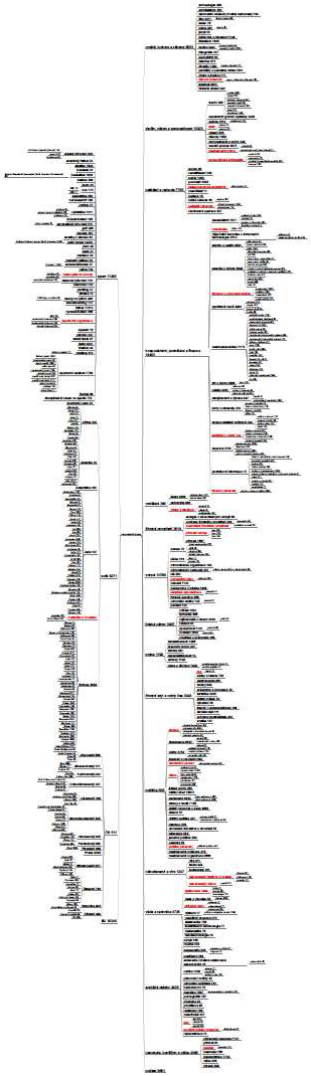
3. Výběr dat z archivu

- Reprezentativní výběr by zároveň snížil nároky na výpočetní výkon
- Na základě jakých proměnných dělat reprezentativní výběr?
  - potřebujeme mít proměnnou v datech a informaci o distribuci v realitě; např.:
  - velikost webu/stránky
  - žánr webu
  - návštěvnost
  - …
  - bylo by zapotřebí budovat přehled základních charakteristik českého webu
    - ideálně tříděného podle témat

- Jak zúžit zkoumaný obsah webu?
  - na základě času
  - na základě klíčových slov
  - výběr konkrétních webů
  - na základě tématu webu určeného tvůrci archivu
  - (na základě žánru webu – fórum, prezentační stránky, e-shop apod.)

4. Kategorizace stránek podle tématu

- Automatická klasifikace témat (učení s učitelem)
- Vytvoření stromu témat pro celý český internet
  - základ tvoří strom zpravodajských témat poskytnutý KKY ZČU
    - Východisko IPTC
  - došlo k rozšíření o nezpravodajská témata často se vyskytující se na internetu
    - Zdroje: https://odkazy.seznam.cz/, http://odp.org/
- Zdroj učících dat
  - vlastní set vytvořený pro Webarchiv
  - zpravodajství
  - Wikipedie

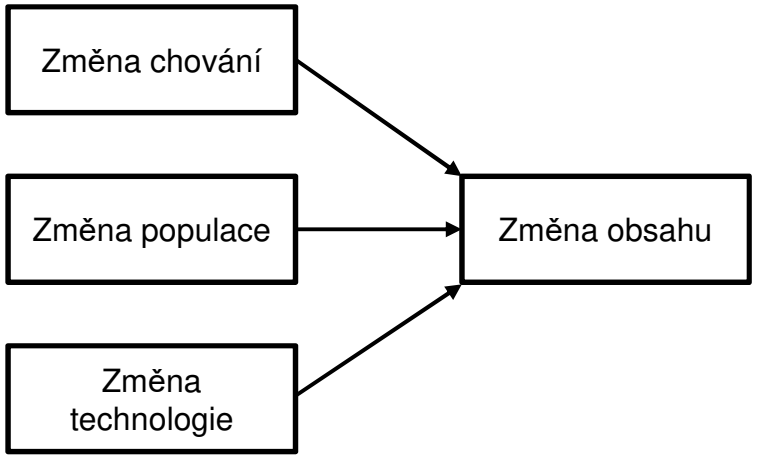


5. Interpretace výsledků

- Při používaní rozhraní nemá výzkumník kontrolu nad celým postupem analýzy
- Výstupem jsou analýzy obsahu bez znalosti motivací autorů, kteří obsah vytvářeli

- Pečlivá a přístupná dokumentace sběru a zpracování dat
- Export syntaxe umožňující replikaci analýzy v budoucnu
- Kombinace s tradičními metodami sociálněvědního výzkumu (dotazníkové šetření, hloubkové rozhovory s tvůrci)

Přehled důvodů pozorované změny obsahu (podle Salganik 2017: 33)



Zdroje

Ackland, Robert, a Ann Evans. „Using the web to examine the evolution of the abortion debate in Australia, 2005– 2015". In *The Web as History*, editoval Niels Brügger a Ralph Schroeder, 159–89. London: UCL Press, 2017.

Brügger, Niels, a Ralph Schroeder, ed. *The Web as History*. London: UCL Press, 2017.

Cowls, Josh, a Jonathan Bright. „International hyperlinks in online news media". In *The Web as History*, editoval Niels Brügger a Ralph Schroeder, 101–16. London: UCL Press, 2017.

Hale, Scott A., Grant Blank, a Victoria D. Alexander. „Live versus archive: Comparing a web archive to a population of web pages". In *The Web as History*, editoval Niels Brügger a Ralph Schroeder, 45–61. London: UCL Press, 2017.

Salganik, Matthew. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press, 2017.

Děkujeme za pozornost

paulina.tabery@soc.cas.cz
matous.pilnacek@soc.cas.cz

Sociologický ústav
Akademie věd ČR

Centrum
pro výzkum
veřejného
mínění