

Fenomén #scicomm: využití big data a jejich analýzy při sledování komunikace vědy na internetových stránkách českých výzkumných organizací*

PETRA RAUDENSKÁ^{1**}, RENÁTA TOPINKOVÁ^{1,2}

¹Sociologický ústav AV ČR, v. v. i., Praha

²Ludwig-Maximilians-Universität München

The #scicomm Phenomenon: Using and Analysing Big Data to Track Science Communication on Czech Research Institutional Websites

Abstract: This study focused on science communication on the websites of Czech research institutions. Particularly, we inquired to what extent Czech science is shared with the public on the Internet and what differences can be found between the websites of social and natural science institutions. Textual analysis revealed that on the scientific websites, terms like ‘science’ and ‘popularization’ occurred together with references to scientific institutions, study, and research. In the case of natural sciences, the term ‘popularization’ was more often linked to receiving science awards for science popularization and promotion. Structural web analysis showed that most scientific webs contained hyperlinks to social media such as Facebook, Twitter, YouTube, Instagram, and LinkedIn. Similarly, they often referred to online news outlets such as ceskatelevize.cz, novinky.cz, lidovky.cz, and rozhlas.cz. On the other side, they much less often referred to institutional and government websites. The results suggested that Czech science communication can be characterized as more interactive than canonical.

Keywords: science communication, public research institutions, big data analysis, text analysis, topic models, social network analysis

Sociologický časopis / Czech Sociological Review, 2023, Vol. 59, No. 4: 387–415

<https://doi.org/10.13060/csr.2023.004>

* Tento článek vznikl za podpory Ministerstva kultury ČR v rámci projektu č. DG18P02O-VV016 „Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů“. Autorky děkují recenzentům za podnětné připomínky, kolegovi Matouši Pilnáčkovi za pomoc při přípravě dat k analýze a Tomáši Diviákovi za konzultaci síťových analýz.

** Veškerou korespondenci posílejte na adresu: PhDr. Petra Raudenská, Ph.D., Sociologický ústav AV ČR, v. v. i., Jilská 1, 110 00 Praha 1, e-mail: petra.raudenska@soc.cas.cz.

Úvod

Ještě nedávno věda komunikovala s veřejností primárně prostřednictvím tiskových zpráv či profesionálních médií. Dnes je však pro většinu lidí první zastávkou za (nejen) vědeckými poznatky internet (Purcell et al., 2012). S nástupem internetu, a zejména sociálních médií, mohou jednotliví vědci i celé instituce sdílet výsledky své práce nejrůznější formou, od článků, videí či podcastů po přímou komunikaci prostřednictvím Twitteru či Facebooku. Nová sociální média tak otevírají široké pole možností vědecké spolupráce a zvyšování zájmu veřejnosti o vědu (van Noorden, 2014). Zatímco tiskové zprávy přebírala média od odborníků a zprostředkovávala je veřejnosti, dnes si jejich formu a obsah může upravit kdokoli dalším sdílením či editováním.

Oproti tradičnímu (kanonickému) modelu předávání vědeckých informací, který klade důraz na komunikaci vědeckých zmocněnců s politickými činiteli a mediálními aktéry (Nisbet a Scheufele, 2009), jsou dnes vědci vyzýváni k aktivnímu předávání vědeckých znalostí a angažovanosti ve veřejném dialogu (Sis net n.d.). Důraz je kladen na obousměrnou, interaktivní, otevřenou a méně formální komunikaci vědců a veřejnosti prostřednictvím všech dostupných mediálních platforem (Lee a VanDyke, 2015). Rozdíl nemá být pouze v zapojení vědců, ale i v zapojení samotné veřejnosti; ta nemá být pouhým pasivním konzumentem obsahu, ale naopak má být vybízena k přímé účasti na rodící se vědě (Sis.net). I přes tento apel na větší zapojování veřejnosti se dosud vedou debaty, jakým způsobem tento veřejný dialog smysluplně utvářet, a mnozí vědci své výsledky prezentují spíše jednosměrně, například v komunikaci s vědeckými publicisty (Andrle, 2013; The Royal Society, 2006).

Přestože se již problematice vědecké žurnalistiky a obrazu vědy v českých médiích někteří autoři na poli české vědy věnovali, pozornost byla upřena především na obsahovou analýzu českého tisku (např. Čada et al., 2006; Hrabánková, 2018). Dosud však chybí kvantitativní studie, která by systematicky nebo i jen částečně mapovala komunikaci vědy na českém internetu po nástupu participativního webu. Tato explorační studie se na pozadí rozvoje nových online technologií zaměřuje na komunikaci vědy v prostředí českého internetu, což nám umožní sledovat mnohem širší spektrum možných způsobů její komunikace. Ve studii sledujeme nejen jednostrannou komunikaci vědy, ať už řízenou či neřízenou, ve směru od institucionálních vědeckých webů ke zpravodajským médiím, dostupným českým médiím, vědeckým online přílohám nebo speciálním webům zaměřeným na výklad vědeckých poznatků, ale i směr vědecké komunikace k médiím, která otevřenou diskusi s veřejností umožňují, například ve formě vystoupení vědeckých osobností v médiích nebo odkazů na blogy či sociální média.

Hlavní výzkumná otázka, kterou si v této explorační studii klademe, zní: *Jak je v současnosti komunikována věda na internetových stránkách českých výzkumných organizací?* Studie autorů Ke et al. (2017) naznačuje, že sociální vědy jsou oproti přírodním vědám v online prostředí aktivnější, i když značná část sdíleného obsahu

se netýká jen vědy. Toto zjištění nás dovedlo k sekundární výzkumné otázce, a to *jaký je tedy rozdíl v komunikaci přírodních/technických a společenských/humanitních věd na těchto internetových stránkách?*

K prozkoumání podoby komunikace vědy na *internetových stránkách českých výzkumných organizací* se nabízí využití velkých dat. Analýza velkých dat je značně specifická a vyžaduje značné množství výpočetních mezikroků a časově náročnou přípravu datového souboru. Na druhou stranu lze říci, že přináší zdánlivě jednoduché odpovědi na poměrně originální (a někdy i značně komplexní) výzkumné otázky, které bychom nedokázali analyzovat pomocí standardních dotazníkových šetření. Vedlejším (byť ne nutně druhotným) cílem této stati je proto osvětlit čtenáři, jaké přípravné a výpočetní mezikroky je nutné absolvovat při práci s takovými velkými daty a s jakými peripetiemi se výzkumník může setkat při obsahové a strukturální analýze dat z webových stránek. Zde jsme i značně sebekritické k vlastním výsledkům (sekce Diskuze a shrnutí). Z tohoto úhlu pohledu pokládáme tuto studii nejen za obsahově zajímavou, ale také metodologicky přínosnou a osvětovou.

Teoretické ukotvení: formy komunikace vědy

Věda je v dnešní době nejen důležitým aspektem blahobytu jednotlivců, organizací i národů, ale i klíčovým prvkem demokracie a soudobé kultury znalostní společnosti, neboť tvoří základ mnohých zásadních rozhodnutí a do určité míry ovlivňuje, co si jedinci myslí o aktuálních celospolečenských otázkách (Davies a Horst, 2006; Dijkstra et al., 2015). Samotná věda se neustále vyvíjí a objem nových poznatků kontinuálně narůstá – společně s určitou nejistotou vědě vlastní. Vzniká potřeba se v tomto množství informací zorientovat. Komunikace vědy tak plní zásadní úkol, kterým je učinit vědecké poznatky srozumitelnými a přístupnými pro laickou veřejnost.

Komunikaci vědy (tj. science communication) lze definovat jako organizované jednání za účelem sdělování vědeckých znalostí, metod, procesů a praktik (Borgman a Furner, 2002) nebo jako využití vhodných dovedností, médií, činností a dialogu k vyvolání individuální reakce na vědu (Burns et al., 2003), konkrétně obeznámenosti s novými vědeckými poznatky, emocionální reakce jako potěšení, zájmu, tříbení názorů a porozumění vědě jako takové, jejímu obsahu, procesům a sociálním faktorům. Definice autorů Burns et al. zahrnuje mezi aktéry komunikace vědy nejen samotné vědce, ale i zprostředkovatele, veřejnost nebo členy blízké sociální skupiny (např. vrstevníky).

Jako taková bývá komunikace vědy často synonymně označována nebo diskutována v souvislosti s „obecným povědomím o vědě“ (public awareness of science), „popularizací vědy“ (science popularization), „veřejným porozuměním vědě“ (public understanding of science), „komunikací mezi vědci“ (scholarly communication), „vědeckou kulturou“ (scientific culture), „vědeckou gramot-

ností“ (scientific literacy) či „vědeckou žurnalistikou“ (science journalism) (více k jejím definicím viz Burns et al., 2003); v moderním pojetí lze o ní najít informace pod hashtagy #scicomm, #Scicommunication, #ScienceCommunication.

Komunikace vědy má dva podstatné komunikační cíle: (a) předávání informací za účelem zvyšování obecného povědomí o vědě a vědecké gramotnosti, a (b) zapojení veřejnosti do vědecké debaty ke zvýšení její participace na vědě (Scheufele, 2014; obecná učebnice ke komunikaci vědy v praxi viz Stocklmayer et al., 2002). Probíhá na několika úrovních, od formálního publikování vědeckých studií na akademické půdě přes šíření poznatků na konferencích, vědeckých festivalech a popularizačních akcích až po méně formální komunikaci vědy prostřednictvím tiskových zpráv, popularizačních knih, televize, rádia, tisku a internetu (Bauer et al., 2007; Gu a Widén-Wulff, 2011).

Nielsen et al. (2007) prezentují dva relevantní modely komunikace vědy: kanonický model a model interaktivní, reflexivní. Historicky dříve se pro vysvětlení vztahu mezi vědci a veřejností začal používat tzv. *kanonický model komunikace vědy* (někdy také informačně-diseminační model; viz Hilgartner, 1990), kde vědci sami produkují vědecké poznatky a dále své závěry šíří za účelem vzdělávání nebo i pobavení širší veřejnosti, ale také sociální legitimizace svého vědeckého úsilí. Jde pak především o komunikaci vůči vládním představitelům, institucím a podnikům, vědecké či odborné konference, prezentaci tiskových zpráv a omezenou komunikaci s médii, většinou bez zpětné vazby publika. Navzdory všem konstruktivním námitkám (viz interaktivní model komunikace níže; souhrnně viz Broks, 2006) zůstává kanonický model jednosměrné komunikace vědy pro mnoho vědců příkladem té nejlepší komunikace vědy ve veřejném prostoru.

Dominanci kanonického modelu dokládají i výsledky reprezentativního šetření mezi britskými vědci a inženýry, v němž pouze 12 % považovalo komunikaci se širokou veřejností za skutečně relevantní (The Royal Society, 2006; předchodí výzkumy viz Treise a Weigold, 2002). Naopak komunikaci s veřejnými činiteli hodnotilo jako jednu z nejdůležitějších téměř 90 % vědců. Ačkoliv jako hlavní definici, jak komunikovat s veřejností, uváděli vědci-respondenti „informování, vysvětlování a usilování o správné pochopení faktů“, v otevřené otázce na význam termínu „pochopení vědy veřejností“ neuváděli žádná klíčová slova jako „veřejná diskuse“ nebo „interakce s veřejností“. Podobné výsledky dává i krátká anketa z roku 2013 mezi českými vědci a vědeckými redakcemi na téma popularizace vědy v České republice (Andrle, 2013). Šestnáct respondentů se shodlo, že vědecké poznatky se v dané době komunikovaly především za pomoci tištěných a internetových novin, rozhlasu a televize a že je potřeba si pěstovat kvalitní vztah s novináři a popularizovat dostatečně jednoduchým, ale stále odborně správným způsobem. Pouze tři z dotázaných explicitně zmínili sociální média či skutečný veřejný dialog.

Od roku 2000 je jako opoziční model ke kanonickému modelu komunikace vědy předkládán *interaktivní, reflexivní model*. Ten zdůrazňuje roli obousměrné komunikace (Nisbet a Scheufele, 2009), vybízí k větší participaci veřejnosti na

vědě a snaží se dostat vědce do ringu veřejného dialogu (Burns et al., 2003; Jün-ger a Fähnrich, 2020; MacNaghten et al., 2005; Trench a Miller, 2012). Přichází s ideou výměny znalostí a kompetencí vědců se společnostmi a volá po interaktivní účasti všech složek vědecké komunikace (Sis.net, n.d.). Oproti kanonickému modelu, který apeluje na komunikaci vědeckých zástupců univerzit s politickými činiteli a médii, staví interaktivní model na komplexnějším chápání univerzit a vědeckých pracovišť jako znalostních a kulturních institucí a na vytváření vztahu s veřejností (Lee a VanDyke, 2015; Scheufele, 2014). V současnosti je proto v rámci komunikace vědy vedle dvou dominantních aktérů – *vědců*, kteří vědecká fakta vytvářejí a ve vědecké komunitě je komunikují, a *kommunikátorů*, kteří fakta předávají – zdůrazňována především role třetí skupiny, *veřejnosti*, která není v komunikaci vědy považována za pouhého pasivního příjemce, ale za aktivního aktéra zapojujícího se do dialogu napříč různými mediálními platformami (Sis.net, n.d.)¹.

Jistě zajímavé jsou v tomto kontextu závěry internetového dotazníkového šetření mezi dánskými přírodovědci publikované v časopise *Journal of Science Communication* v roce 2007. Pokud měli vědci hodnotit důležitost médií podle jejich rozšíření a velikosti cílové skupiny, jako nejdůležitější se jim jevily právě sdělovací prostředky, populárně vědecké interdisciplinární časopisy a veřejné debaty. Vědci také pociťovali zodpovědnost za šíření znalostí, a to především nových poznatků a aktuálních výzkumných závěrů (Nielsen et al., 2007). Následný dánský průzkum mezi komunikátory vědy potvrdil, že vědci nemají zájem pouze pomáhat veřejnosti porozumět vědeckým poznatkům, ale chtějí také aktivně přispívat do demokratické debaty a usilovat o legitimizaci vědy a technologií jako takových (Nielsen, 2010).

Nástup nových sociálních médií spojených s Webem 2.0 změnil možnosti neformálního dialogu mezi vědci, otevřel tento prostor laikům a zájemcům a umožnil pozvednout veřejnou angažovanost ve vědě (Su et al., 2017; Uren a Dazdie, 2015). S nedávným rozvojem technologií, především online žurnalistiky, elektronických knih, e-konferencí, webinářů, e-workshopů, videokonferencí, audiovizuálních materiálů, blogů, diskusních fór a nových sociálních médií, se rozšířily i možnosti přenosu informací o vědě (Brossard a Scheufele, 2013; Noruzi, 2008). Komunikace vědy v médiích tak v současnosti čím dál častěji stojí před výzvou přímé konfrontace vědců s veřejností a jejími zájmovými podskupinami. Činí tak nejen prostřednictvím formy a obsahu, ale také rostoucí nabídky dostupných interaktivních prostředků (Gu a Widén-Wulff, 2011).

Přes bezprecedentní růst medializovaných vědeckých zpráv v posledních dekádách (Rödder et al., 2012) se vědci shodují, že komunikace vědy je stále spíše jednosměrným procesem předávajícím informace, který ještě plně nevyužívá všech možností nových technologií a sociálních médií. Trench a Miller (2012) po-

¹ K tomuto tématu doporučujeme k četbě speciální výroční číslo časopisu *Public Understanding of Science* – Special Issue: Public Engagement in Science (2014).

ukazují na skutečnost, že vědci, vědecké instituce i vědeckí žurnalisté mají sklon využívat internet spíše k profesionální než osobní komunikaci nebo k otevírání veřejného dialogu za podpory dostupných multimediálních nástrojů (Rybalko a Seltzer, 2010; ke kritice komunikace vědy v online prostředí viz Davies a Hara, 2017). Collins et al. (2016) potvrzují, že kupříkladu Twitter používají vědci převážně k výměně vědeckých informací. Na druhou stranu Côté a Darling (2018) ukazují, že většinu „nevědeckých“ followerů twitterových mikroblogů konkrétních vědců tvoří nejčastěji média, politici a zainteresovaná veřejnost. Tím se Twitter jako sociální médium stává hybridním nástrojem komunikace vědy oscilujícím mezi výměnou informací, propagací vědy a dialogem s jednotlivci či veřejností (Jünger a Fähnrich, 2020). Ačkoliv je Twitter využíván napříč akademickými disciplínami, sociální vědy jsou oproti přírodním vědám aktivnější, přičemž značná část sdíleného obsahu se netýká vědy, ale reaguje na relevantní politická či sociální témata (Ke et al., 2017).

Tato studie volně navazuje na snahy několika českých autorů zmapovat komunikaci vědy v českém tisku (Čada et al., 2006) anebo obraz české vědy ve veřejném mínění (Šamanová et al., 2006; srov. zajímavou studii o postojích britské veřejnosti k vědě Office of Science and Technology and the Wellcome Trust, O. O. S. A., 2001). V roce 2006, kdy byly tyto studie publikovány, mělo dle ČSÚ (2010) internet doma pouhých 27 % českých domácností, v roce 2019 používalo internet již 81 % Čechů starších 16 let, chytrý telefon pak 70 % z nich.

V posledních deseti letech tak ve vědecké komunitě spatřujeme zvyšující se tlak na komunikaci vědců s veřejností a přesun této komunikace do online prostředí. Například Grantová agentura ČR a univerzitní grantové agentury přidělily finanční podporu několika projektům věnovaným popularizaci vědy. AV ČR vydává časopisy (i v online prostředí) o vědě pro veřejnost (*Věda a výzkum, Živa*), pořádá největší popularizační akce v zemi jako *Veletrh vědy* a *Týden vědy a techniky* (mezi ně patří rovněž festival *AFO – Academia Film Olomouc* a *Noc vědců*) a organizuje stáže pro studenty a popularizační kurzy (*Otevřená věda*). Středisko společných činností AV ČR provozuje populární YouTube kanál *Zvěd*. Většina ústavů a univerzit má webové stránky, Facebook a Twitter, které aktivně využívají. Rozsáhlou popularizační činnost vyvíjejí i samotná média, když vytvářejí různé datové vizualizace (iRozhlas nebo Novinky), publikují vlastní rozsáhlé průzkumy (ČRo – průzkum *Česká společnost po 30 letech*) či mají vlastní vědeckou redakci (např. redakce vědy ČT integrovaná do celého zpravodajství). ČT zprávy vysílají pořad *Hyde Park Civilizace* a týdeník *Věda 24* a Seznam zprávy program *Výzva*, kde mj. vědci odpovídají na otázky, které pokládá veřejnost; časté jsou i online debaty s dotazy „z publika“.

Vzhledem k současnému rozvoji online technologií se tato kvantitativní studie zcela ojediněle soustřeďuje především na komunikaci vědy v prostředí českého internetu. Musíme již na tomto místě zdůraznit, že studie je plně explorativní a popisná a snaží se odpovědět na otázky, jak je v současnosti komunikována věda na internetových stránkách českých výzkumných organizací a jaký je

rozdíl v komunikaci přírodních/technických a společenských/humanitních věd na těchto internetových stránkách. Massoli (2007) podobně analyzovala webové stránky oficiálních evropských výzkumných institucí. Její práce se však kromě mezinárodního přesahu více soustředila na vizuální a materiální obsah webových stránek a posuzovala možnosti interakce s uživateli stránek, uživatelskou přívětivost či prezentaci institucionální identity, vědecké kredibility, dostupných služeb atp. Na rozdíl od ní se zaměřujeme především na obsahovou a strukturální analýzu textu vyskytujícího se na internetových stránkách českých výzkumných organizací a sledujeme tak charakter a formu komunikace české vědy.

Data

K odpovědím na položené výzkumné otázky bylo nutné identifikovat co nejširší spektrum tzv. institucionálních webů – tedy webů vědeckých institucí, přidružených vědeckých či výzkumných pracovišť, univerzit, grantových agentur, specifických úřadů/ústavů a vědeckých akcí. Celkem jsme takto identifikovaly 105 institucionálních vědeckých webů na českém internetu (viz Tabulka 1A v online příloze²). Z těchto původních 105 webů se šest nepodařilo stáhnout a 10 se podařilo stáhnout pouze částečně. V případě částečně stažených webů se jednalo o rozsáhlé stránky, kde stahování obsahu trvalo více než den a selhalo před dokončením programu. Vzhledem k množství dat, které jsme z těchto webů vytěžily, jsme se rozhodly je přesto zahrnout do analýz, jelikož se domníváme, že většina podstatného (ne-li vše) byla z webů stažena. Celkem jsme tedy získaly data z 99 webů.

Data byla stažena pomocí programu Heritrix³, sloužícího k archivaci webových stránek a běžně používaného webovými archivy. Program procházel v rámci zadaných webů hypertextové odkazy od domovské stránky do hloubky čtyř kroků. Mimo domény původních cílových webů nebyl stahován žádný obsah. Ze stažených webových stránek byly extrahovány hypertextové odkazy, informace o přítomnosti videa a textový obsah stránky. Od zbývajících textových částí webu (například menu, patička, reklamy) byl obsah očištěn pomocí balíčku jusText (viz Pomikálek, 2011).

V následujících analýzách často hovoříme o rozdílu mezi přírodními a sociálními vědami. Vědy jsme dle běžných standardů (např. rozdělení na univerzitách) klasifikovaly na přírodní/technické a společenské/humanitní. Mezi přírodní a technické vědy tak řadíme např. fyziku, matematiku, chemii, biologii, vědy o vesmíru, technologii a jejich příbuzné podobory, zatímco mezi sociální či humanitní vědy je obvykle zařazena historie, kulturologie, sociologie, filozofie nebo demografie, ekonomie, politologie, psychologie, lingvistika, archeologie,

² Příloha je dostupná na <https://doi.org/10.13060/csr.2023.004>.

³ <https://github.com/internetarchive/heritrix3>

umění, právo a etika. Ve výsledku jsme tak určily 47 přírodovědných a 27 sociálněvědních webů, zatímco 25 jsme označily jako neutrální, jelikož nebylo možné je jednoznačně zařadit ani do jedné kategorie (např. web Univerzity Karlovy, který sdružuje obory z obou oblastí).

V této studii zodpovídáme výzkumné otázky dvěma způsoby, a to s ohledem na obsah (co je komunikováno) a strukturu (jakým způsobem, respektive kam je komunikováno).

Popis metod a výsledky analýzy

Obsahová analýza textu

Velká část obsahu internetu je zaznamenána pomocí textu. Tento text může mít různou formu, od digitalizovaných dokumentů přes zpravodajství až po blogy. Tradiční metodou analýzy textu v sociálních vědách je obsahová analýza. Typicky obsahovou analýzu provádějí kodéři manuálně: daný text čtou a následně jej kódují. S narůstajícím počtem textů ke kódování však přestává být manuální obsahová analýza proveditelná. Pokud je text strojově čitelný, nabízí se možnosti jeho automatického zpracování. Podobně jako u jiných typů kvantitativní analýzy se i kvantitativní analýza textu skládá ze dvou hlavních kroků – předzpracování dat („pre-processing“) a jejich samotné analýzy.

Soubor, který je základem pro naši analýzu obsahu, obsahuje vždy název domény, konkrétní URL odkaz, klíčové slovo, které se na daném URL vyskytovalo, a 10 slov (kontext) kolem daného klíčového slova. Výběru klíčových slov předcházela kvalitativní předvýzkum internetových stránek českých vědecko-výzkumných a vzdělávacích institucí a ostatních internetových stránek zpravodajských médií, kde se komunikuje věda, specifických webů spojených s vědou a blogů (jejich seznam poskytneme na vyžádání). Data (tj. textový obsah stránky, hypertextové odkazy a metadata) ke kvantitativnímu zpracování byla následně stažena pomocí programu Heritrix3, který prohledával zadané institucionální weby na přítomnost těchto klíčových slov. Obsah institucionálních webů jsme procházely z domovské stránky až do hloubky čtyř kroků hypertextových odkazů objevujících se na daných stránkách; u ostatních webů jsme se zaměřovaly na články či obsah jakkoli spojený s vědou. V kvalitativním předvýzkumu získaného textu jsme se zaměřily na nejčastěji se objevující výrazy spojené s komunikací vědy vůči širšímu publiku, a to jak na jejich specifické znění pro přírodní/technické a společenské/humanitní vědy, tak i na neutrální výrazy. Námi vybraná klíčová slova obsahují slova příbuzná pojmům *vědec*, *věda* a *výzkum* a jejich synonyma jako *odborník*, *expert*, případně oficiální označení pozice vědce v dané instituci (např. *doktor*, *docent*, *profesor* atp.), dále pak označení dané vědecké instituce či disciplíny, aktivity či vědeckého výsledku (kompletní seznam klíčových slov viz Tabulka 2A v online příloze). Nakonec jsme vzhledem k tématu článku zahrnuly i slovo *popularizace*.

V předzpracování textových dat, vygenerovaných již pomocí programu Heritrix3, jsme pak postupovaly následujícím způsobem. Nejdříve jsme identifikovaly duplicitní texty a vyřadily je z analýz. Dále jsme provedly lemmatizaci neboli převedly všechna slova do základního tvaru⁴. Tak například slovo „vědy“ bylo převedeno do základního tvaru „věda“. Pokud bychom tak neučinily, nejčastějšími výrazy v každém textu by byly různé varianty nejčastějších výrazů, například tedy „věda“, „vědy“, „vědě“ apod. Následně jsme odstranily tzv. „stopwords“, což je běžný postup při kvantitativním zpracování přirozeného jazyka⁵ (Silge a Robinson, 2017). Mezi stopwords se typicky řadí nejčastější slova v daném jazyce („všichni“), číslovky („deset“), předložky („bez“) a spojky („a“, „nebo“), které nepřidávají textu mnoho informací navíc. Podobně jako v případě slov před lemmatizací bychom i zde bez vyřazení stopwords získaly jako nejčastější výrazy právě stopwords. Jelikož některé texty byly částečně v angličtině (typicky názvy akademických publikací), rozhodly jsme se vyřadit také anglická stopwords. Nakonec jsme odstranily všechna čísla, interpunkci, slova o méně než třech znacích, slova delší než 25 znaků a slova obsahující speciální znaky⁶ či číslice.

Tento postup čištění dat používáme jak pro popisnou analýzu textových dat, tak i pro následující tematické modely („topic models“).⁷

Výsledky popisné analýzy textových dat

V rámci obsahové analýzy institucionálních webů jsme se nejprve zaměřily na kontext, v jakém se vyskytují vybraná klíčová slova obsahující příbuzná slova pojmu „věda“ a „popularizace“, a na specifika tohoto kontextu v komunikaci přírodních/technických a společenských/humanitních věd. Výraz „věda“ (a slova příbuzná⁸) se vyskytoval celkem ve 258 563 textech, zatímco výraz popularizace byl výrazně méně frekventovaný (n = 2 778). Na prvním místě je třeba poznamenat, že většina výrazů spoluvyskytujících se s klíčovými slovy „věda“ a „popularizace“ se v přírodních vědách a společenských vědách do značné míry překrý-

⁴ Lemmatizaci jsme provedly pomocí UDPipe (Straka a Straková, 2019), resp. balíčku udpipe v R (Wijffels et al., 2021).

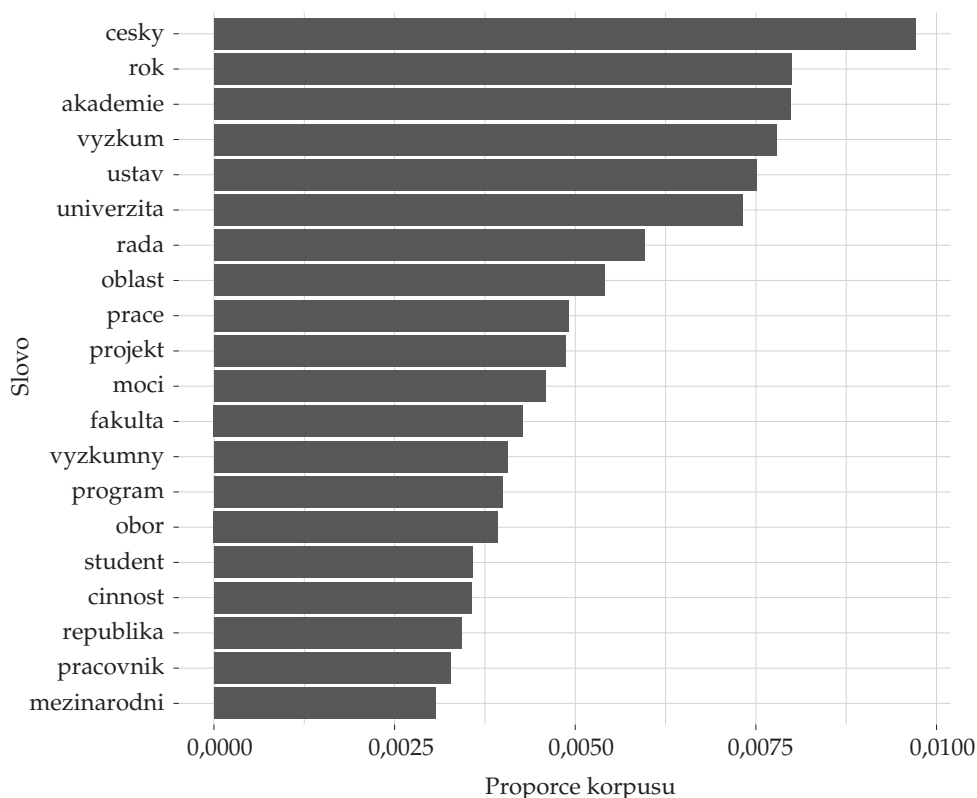
⁵ Pro úplnost uveďme, že se jedná o častý, nikoli však automatický postup. U složitějších výpočetních algoritmů, které zohledňují např. větnou stavbu, může být naopak vyřazení těchto výrazů nevhodné.

⁶ Před tímto krokem bylo nutné z veškerého textu odstranit diakritiku, jelikož písmena s diakritikou nejsou v regulárních výrazech („regular expressions“) považována za písmena, ale za speciální znaky.

⁷ Výpočetní skripty k analýze viz <https://github.com/renatatopinkova/popsci>.

⁸ Soubor příbuzných slov obsahuje všechna vybraná klíčová slova ze seznamu, viz Tabulka A2 v online příloze. Stejně tak obsahuje též všechny jejich slovní tvary (vědy, vědě, vědou, vědci, vědců, apod.), a to díky použití regulárních výrazů a lemmatizace. Pro jednodušší interpretaci a čitelnost textu mluvíme souhrnně o klíčovém slově „věda“, resp. ^věd.*.

Graf 1. Spoluvýskyt slov s klíčovým slovem ^věd.* (věda, vědecký, vědec, vědkyně ve všech tvarech) – všechny vědecké institucionální weby

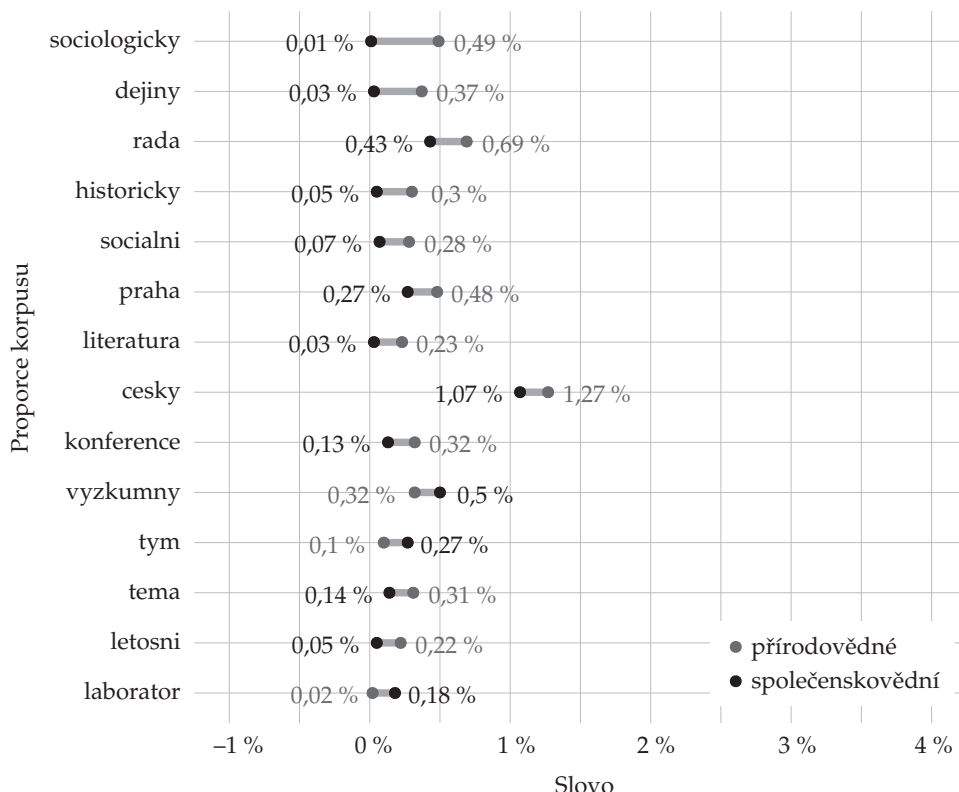


vá. V obou případech se „věda“ a „popularizace“ nejčastěji vyskytují v souvislosti s „českou vědou“, „akademií“, „ústavem“, „univerzitou“ a „výzkumem“ (viz grafy 1 a 3⁹). Není nijak překvapivé, že pojem věda se také nejčastěji objevuje v kontextu dalších institucí, jako je „fakulta“ a „škola“, a s nimi souvisejících činností a nositelů – „program“, „studie“, „student“, „obor“, „program“, „projekt“.

Detailnější pohled potom poskytuje Graf 2, zaměřený specificky na rozdíly mezi společenskovědními/humanitními a přírodovědními/technickými weby. Graf 2 ukazuje, že v online obsahu společenských a humanitních věd se často objevují pojmy „sociologický“ a „věda“ pospolu. Ačkoliv se to může zdát jako za-

⁹ Barevná varianta grafů je k dispozici v online verzi článku.

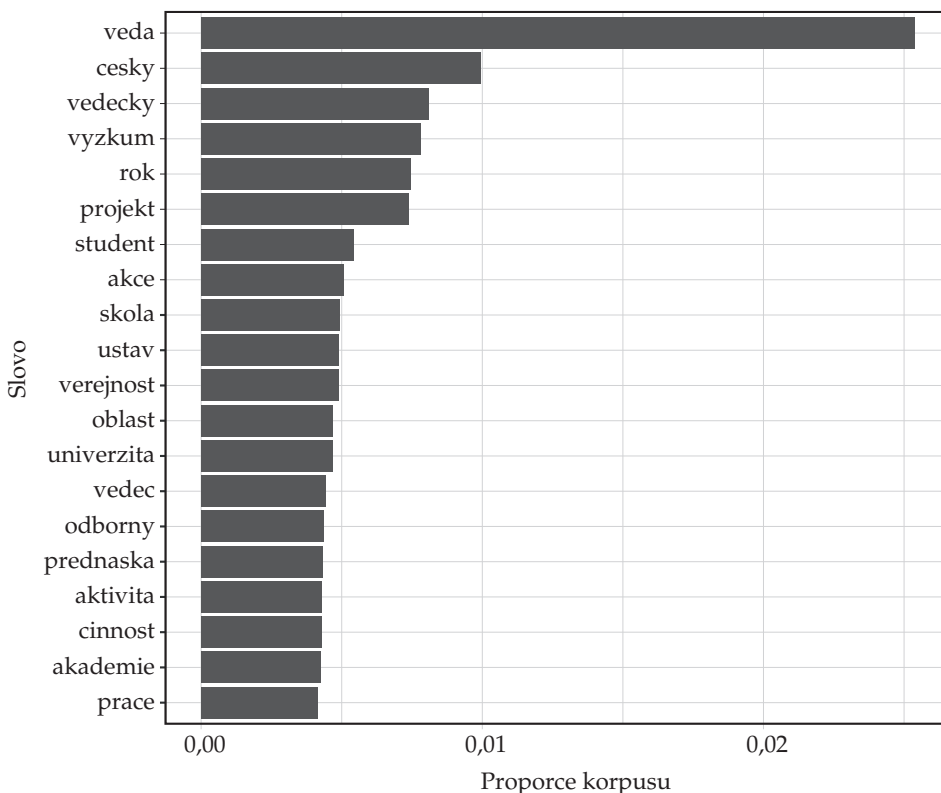
Graf 2. Spoluvýskyt slov s klíčovým slovem ^věd.* (věda, vědecký, vědec, vědkyně ve všech tvarech) – rozdíly mezi společenskovědními a přírodovědními weby



jímavý poznatek, jde o artefakt dat. Web Sociologického ústavu AV ČR je totiž jedním z nejrozsáhlejších webů, které jsme stahovaly, a 78 % textů obsahujících slovo „sociologický“ pochází právě odtud. Pokud bychom web Sociologického ústavu z analýz vyřadili, výraz „sociologický“ by se nedostal ani do první stovky nejčastějších slov vyskytujících se se slovem „věda“. Naopak „dějiny“, „společnost“, „sociální“ nebo „historický“ jsou slova vyskytující se v souvislosti s vědou napříč sociálněvědními weby. Další časté výrazy zahrnují odkazy na pořádání či účastnění se akcí („konference“, „přednáška“), ale také „časopis“, „literatura“ nebo „umění“.

U přírodovědných/technických webů potom věda často souvisí se slovy „tým“, „získat“, „cena“ nebo „mladý“, což patrně reflektuje prezentaci vědeckých úspěchů. Dále také vidíme slova související s prováděním přírodovědného

Graf 3. Spoluvýskyt slov s výrazem „^populari.*“ (popularizace, popularizovat, popularizační etc.) – všechny vědecké institucionální weby

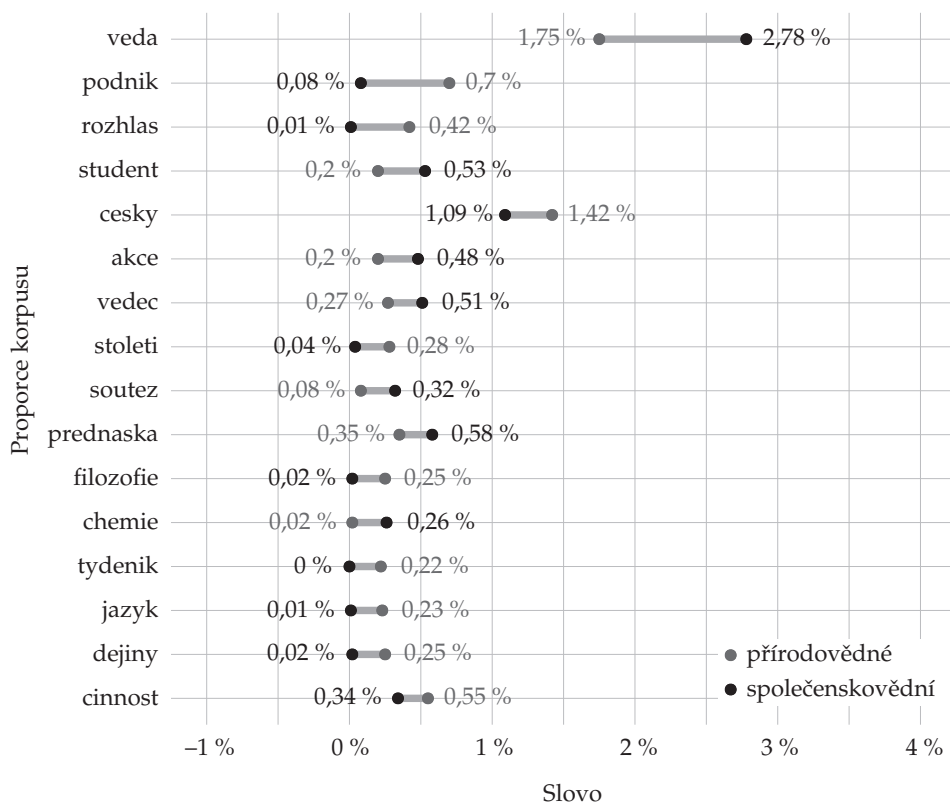


či technického výzkumu („metoda“, „laboratoř“, „technika“). Na rozdíl od sociálních věd, které „vědu“ častěji spojují se „společností“, hovoří přírodní vědy častěji o „člověku“.

U obou vědních odvětví byl zaznamenán silný spoluvýskyt slov s institucionálním obsahem, jako je akademie, ústav, konference, výzkum, projekt, univerzita a studium. Stejně tak se jak u sociálněvědních, tak přírodovědných webů objevují slova „spolupráce“, „mezinárodní“ a „evropský“.

Spoluvýskyt slov s výrazem „popularizace“ v online obsahu institucionálních webů se do značné míry kryje se slovy souvisejícími s výrazem „věda“ (student, věda, přednáška, výzkum, obor aj., více Graf 3). Pokud se ovšem podíváme na rozdíly mezi společenskými a přírodovědnými weby, je situace mnohem

Graf 4. Spoluvýskyt slov s výrazem „popularizace“ (popularizace, popularizovat, popularizační etc.) – rozdíly mezi společenskovědními a přírodovědními weby



rozmanitější (viz Graf 4). V případě společenskovědních/humanitních webů se s pojmem „popularizace“ objevují často výrazy odkazující k určitým výstupům (časopis, dokument, kniha, materiál), komunikaci s médii (rozhlas, týdeník) nebo obecnějším konceptům jako století, dějiny, filozofie, jazyk, kultura či literatura. Naopak v případě přírodovědných/technických webů se s pojmem „popularizace“ často objevují výrazy jako „věnovat“, „cena“, „ocenění“ a „propagace“, které většinou odkazují k ocenění konkrétního vědce za popularizaci a propagaci vědy. Ačkoliv tato slova nacházíme i na webech společenských věd, děje se tomu výrazně méně často. Výrazně častěji se také objevují slova „soutěž“, „podpořit“, „video“ nebo „film“. Dále se častěji objevují slova spojená s pořádáním veřejných akcí, jako například „akce“, „přednáška“, „aktivita“, „návštěvník“ nebo „vele-

trh“. Dalším častým a pro přírodní vědy specifickým slovem je „Petr“. Nejedná se však o chybu v datech, pod tímto jménem se totiž skrývají různí popularizátoři vědy, nejčastěji vulkanolog Petr Brož nebo biolog Jaroslav Petr.

Tematické modely

V klasické kvantitativní analýze je k dispozici celá řada testů a modelů, jichž lze v závislosti na kladené výzkumné otázce využít. Nejinak je tomu i v analýze textu – od popisných analýz využívajících počítání slov přes slovníkové metody (např. analýza sentimentu) a hledání skrytých témat v textu (např. tematické modely) až po složitější metody využívající strojové učení (např. deep learning). Vzhledem k tomu, že nás zde zajímají případné odlišnosti v obsahu komunikace přírodních a sociálních věd, jsme zvolily tematické modely (konkrétně pomocí výpočtu LDA – Latent Dirichlet Allocation). Tematické modely využívají tzv. „bag of words“ přístup. Ten nezohledňuje pořadí slov, ve kterém se v rámci daného textu vyskytují (Blei, 2012) – nejčastěji se ilustruje tak, že dáme všem slova vyskytující se v určitém textu, například v tomto odstavci, do krabice a touto krabicí zatřese. Z klasických sociologických metod mají tematické modely nejblíže ke klastrové analýze. Jak však uvádí Bail (2014, 2015)¹⁰, oproti klastrové analýze nemusí být každé pozorování (dokument, text) přiřazeno jen k jednomu tématu (klastru), ale naopak je mu přiřazena pravděpodobnost, se kterou přísluší ke každému z témat. Dále se tematické modely liší ve způsobu výpočtu, kdy má na začátku každé pozorování přiřazenu náhodnou pravděpodobnost příslušnosti k jednotlivým tématům, přičemž tyto pravděpodobnosti se zpřesňují s přibývajícím množstvím zpracovaných dat. Jedná se tedy o iterativní bayesiánskou techniku. Výsledky tematických modelů potom ukazují dvě věci: slova, která jsou nejčastěji spojována s daným tématem, a pravděpodobnost, s jakou každý dokument obsahuje jednotlivá témata (např. Blei, 2012; Silge a Robinson, 2017).

V našem případě jsou jednotkou zkoumání kontextové úryvky kolem klíčových slov. Každý text (dokument) obsahuje klíčové slovo a kontext maximálně 10 slov před a za klíčovým slovem. Zároveň jsme stanovily vrchní a spodní limity výskytu slov v korpusu, kdy se každé slovo smělo vyskytovat maximálně v 60 % dokumentů a zároveň muselo být nejméně v 500 různých dokumentech. Výsledný korpus obsahuje celkem 1 676 629 dokumentů a 4 813 slov.

Stejně jako v případě některých algoritmů klastrové analýzy (např. k-means) je nutné předem stanovit počet témat, do kterých má algoritmus daná slova a texty rozřadit, což je rozhodnutí se zásadním vlivem na výsledky analýz. V ideálním případě by měl mít výzkumník konkrétní teoretická očekávání ohledně počtu témat (a priori), jež se v dokumentech vyskytují (Bail, 2014, 2015). Naše

¹⁰ https://cbail.github.io/textasdata/topic-modeling/rmarkdown/Topic_Modeling.html

práce je však explorativní, a tak žádné vhodné, teoreticky podložené očekávání o počtu témat nemáme k dispozici. V analýzách jsme tak vyzkoušely varianty s různými počty témat, konkrétně 10, 20, 30, 40 a 50. Model k interpretaci jsme zvolily na základě kvalitativní validizace, tj. pročtení výsledků modelů z hlediska 10–15 slov s nejvyšší pravděpodobností souvislosti s každým tématem a pročtení vzorku textů, které je reprezentují, a následně jsme porovnaly modely z hlediska sémantické koherence a exkluzivity, což jsou metriky zaměřující se na kvalitu témat. Sémantická koherence je maximalizována, když se nejpravděpodobnější slova v daném tématu často vyskytují společně. Jedná se o metriku dobře korepondující s lidským úsudkem o kvalitě tématu (Mimno et al., 2011). Exkluzivita pak dosahuje vyšších hodnot, když je více slov výlučných pro odpovídající tématu (Roberts et al., 2014). Tyto dvě metriky spolu negativně korelují, volba modelu na jejich základě je tak otázkou kompromisu.

Výsledky analýzy tematických modelů

Graf 5 ukazuje, že vhodný počet témat by se pohyboval kolem 20 až 30. Pro další interpretaci jsme se rozhodly využít model s 20 tématy, který byl lépe interpretovatelný než model se 30 tématy. Graf 6 znázorňuje prevalenci 20 témat, která algoritmus v textech našel, a jaká slova mají nejvyšší pravděpodobnost výskytu v daných tématech.

Výsledky ukazují, že nejčastěji zastoupeným tématem v našem korpusu je téma č. 6, které reprezentují výrazy „oblast“, „základní“, „znalost“, „obor“, „teoretický“, „absolvent“, „odborný“. V tématu č. 13 se objevují slova jako „materiál“, „proces“, „technologie“, „chemický“, „vývoj“, „metoda“ nebo „struktura“. Toto téma nalezneme nejčastěji na webech technicky či technologicky zaměřených webů – Středoevropského technologického institutu, Fyzikálního ústavu, Ústavu analytické chemie, Matematického ústavu, Ústavu struktury a mechaniky hornin, ale také popularizační akce Týden vědy a techniky nebo online magazínu iForum.

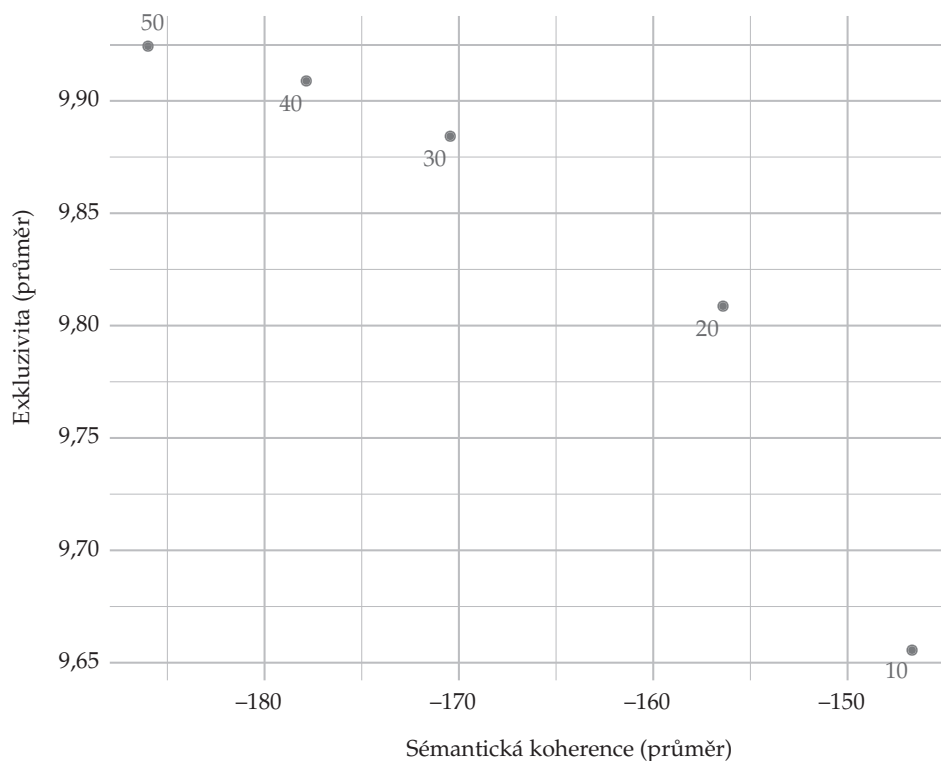
Dále jsou hojně zastoupena témata č. 16, 3 a 17, která se týkají studia na vysoké škole a která nalezneme na webech univerzit (Univerzita Karlova, Masarykova univerzita, Technická univerzita v Liberci). V tématu č. 5 se dohromady vyskytují pojmy pro oblast kultury: „literatura“, „dějiny“, „jazyk“, „umění“, „český“ a „historický“. Jeho zástupci jsou jak humanitní weby zaměřené na jazyky a literaturu, zejména Ústav pro jazyk český a literaturu AV ČR, Janáčkova akademie múzických umění, Filozofický ústav AV ČR nebo Historický ústav AV ČR. Podobně téma 14 obsahuje pojmy „věda“, „český“, „akademie“, „republika“, „Praha“, „společnost“ – a i je reprezentují spíše humanitní vědy. V tématu č. 19 se společně objevují pojmy spojené s veřejným míněním: „veřejný“, „otázka“, „prostředí“, „šetření“ a nalézáme je nepřekvapivě zejména na webu Centra pro výzkum veřejného mínění Sociologického ústavu AV ČR nebo Centra pro otázky životního prostředí.

Naopak specifické pro přírodovědné, respektive spíše technicky zaměřené instituce je téma č. 15 („systém“, „analýza“, „model“, „metoda“, „data“), spojující

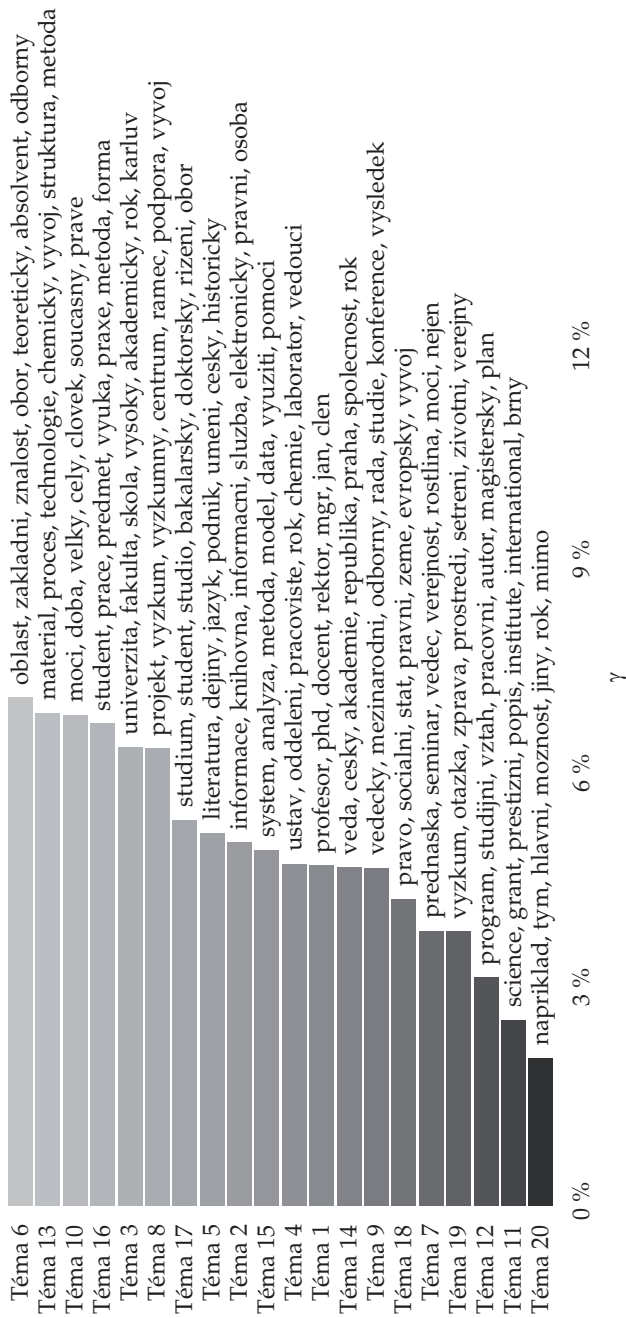
mj. instituce Ústav fyziky atmosféry AV ČR, Ústav teorie informace a automatizace nebo Matematický ústav AV ČR.

Téma č. 8 obsahuje výrazy „projekt“, „výzkum“, „centrum“, „rámec“, „podpora“, „vývoj“. Objevuje se napříč různými disciplínami, včetně přírodovědně zaměřených institucí (např. Fyziologický ústav AV ČR), i společenskovedních institucí (např. CERGE-EI, Centrum pro otázky životního prostředí), ale také na webech grantových agentur (GA ČR a TA ČR). Zvláště zajímavé je téma č. 7, které odkazuje na pořádání akcí pro veřejnost („přednáška“, „seminář“, „vědec“ a „veřejnost“) a objevuje se zejména na webech přírodovědných institucí (Institut (experimentální) botaniky AV ČR, Vysoká škola polytechnická Jihlava) nebo popularizačních akcí (Týden mozku, Otevřená věda), ale nikoliv na webech společenskovedních institucí.

Graf 5. Porovnání sémantické koherence a exkluzivity modelů



Graf 6. Tematické modely: Prevalence témat v korpusu



Analýza multimediálního obsahu

Dále nás zajímalo sdílení audio či video obsahu. Přítomnost takového obsahu na webech určujeme podle výskytu html tagů `iframe`, `amp-iframe`, `video` nebo `audio`, popřípadě již zastaralých tagů `object` a `embed` u webů se starší architekturou¹¹.

Jednoznačně nejvíce audio a video obsahu sdílely weby explicitně zaměřené na popularizaci vědy, jakými jsou stránky Týdne vědy a techniky nebo Akademie věd ČR. Porovnání průměrů proporcí sdílených videí či audií dále ukazuje, že přírodovědné či technické institucionální weby sdílejí mírně větší množství multimediálního obsahu (4 %) než vědy společenské či humanitní (2 %). Zde ale musíme uvést, že použitý způsob detekce audio či video obsahu je velice hrubý a nespolehlivý. Ačkoliv se nám postup založený na html tazích, v nichž by měla tato média být, jevil jako intuitivní, v praxi příliš nefungoval, byl náchylný na chyby a zároveň prakticky neumožňoval tyto chyby detekovat. Pro ilustraci: pokud má webová stránka video v postranním panelu, který se zobrazuje při každém prokliku uvnitř webu, náš program toto jedno video započítá tolikrát, kolik prokliků na webu provede. Podobně některé weby mohou být postavené na `iframe` či `object` tagu nezávisle na tom, zda se na nich nějaký multimediální obsah nachází.

Analýza odkazů

Klíčovou vlastností internetu jako média je jeho propojenost. Ta se projevuje mnoha způsoby, včetně toho, že jednotlivé stránky často obsahují odkazy na jiné. Tato propojenost je tak často zachycována jako síť, v níž uzly reprezentují určité weby (či uživatele, diskusní fóra apod.) a hrany reprezentují vazby mezi nimi (např. hyperlinky, interakce, lajky) (Ackland, 2013; Robins, 2015). Analýza sociálních sítí pak představuje možný způsob, jak popsat a modelovat strukturu a topologii takových sítí.

Prostřednictvím strukturní analýzy hyperlinků institucionálních webů jsme se snažily zodpovědět otázku, zda se současná česká věda stále popularizuje v uzavřeném módu (tj. převažuje kanonický model online komunikace), nebo již v otevřeném módu, vstříc veřejnosti a přístupně k diskusi, a zda existuje rozdíl v otevřenosti či interaktivnosti přírodních a sociálních věd. V uzavřeném módu komunikace bychom očekávaly, že budou institucionální weby odkazovat primárně na další weby vědeckých institucí, případně na weby státních úřadů

¹¹ Zde je nutné poznamenat, že se nejedná o stoprocentní přístup k identifikaci video a audio obsahu. Například tag `object` se ukázal být nepoužitelný u jednoho z webů, postaveného výhradně na javascriptových „objects“, které nemají s videí či audii nic společného. Rovněž `iframe` a `amp-iframe` jsou často využívány k jiným účelům, např. vkládání interaktivního odkazu na mapy.

či oficiálních databází vědeckých časopisů, knihoven a nakladatelství. Naopak v otevřeném, interaktivním módu by měly převažovat odkazy na sociální sítě, zpravodajská média, popularizační weby s vědeckou tematikou a blogy ve zpravodajských médiích. Jelikož náš primárně zajímal rozdíl mezi přírodními a sociálními vědami jako dvěma skupinami, nesledujeme konkrétní institucionální weby, ale spíše to, na jaké weby či typy webů odkazuje nejvíce institucionálních webů uvnitř těchto skupin. Z tohoto důvodu také pracujeme pouze s weby sociálněvědních/humanitních institucí ($n = 27$) a přírodovědných/technických institucí ($n = 47$).

Provedly jsme analýzu sociálních sítí, v níž uzly reprezentovaly weby a hrany odkazy (hyperlinky) mezi nimi. Přítomnost odkazu na jiný web jsme určily pomocí anchor tagu, tj. zachytily jsme odkazy z webu na web. Pro účely analýzy jsme všechna URL zkrátily na jejich doménu. To znamená, že pokud náš program navštívil URL „www.soc.cas.cz/projekty“, zkrátily jsme URL na „soc.cas.cz“, aby byly všechny prokliky v rámci jednoho webu reprezentovány jedním uzlem. Analyzovat celou síť by však bylo velice náročné, a to jak z hlediska interpretace, tak objemu dat. Proto jsme se zaměřily na weby jednotlivých institucí a weby, na které tyto přímo odkazují, čehož jsme docílily prostřednictvím analýzy egocentrických sítí. Ve středu každé sítě tak leží výchozí institucionální web (ego), dále jsou potom v síti odkazy na jiné weby ve vzdálenosti = 1 (ego–alter vazby), tj. weby, na které institucionální web přímo odkazuje. Každá egocentrická síť také obsahuje vazby alter–alter, tedy pokud institucionální web odkazuje na web A a na web B a zároveň web A odkazuje na web B, je do dané sítě zahrnuta i vazba mezi A a B.

Sítě sociálněvědních a přírodovědných institucí se od sebe příliš strukturálně neliší. Nejmenší sociálněvědní síť má pouze dva odkazy na jiné weby než sebe sama, nejmenší přírodovědná síť má 38 odkazů na celkem osm externích webů. Naopak největší sociálněvědní síť odkazuje na 452 unikátních domén (počet vazeb 91 681) a největší přírodovědná síť na 451 unikátních domén (počet vazeb 91 981). Sítě sociálněvědních webů měly v průměru menší počet uzlů a hran (171 oproti 198, resp. 18 028 oproti 20 374) a vyšší průměrný stupeň (276 oproti 220) než přírodovědné sítě.

Tabulka 1. Průměrné statistiky egocentrických sítí dle typu instituce

Typ webů	Počet	Průměrný počet uzlů	Průměrný počet hran	Průměrný stupeň
Přírodovědné	47	198	20 374	220
Sociálněvědní	27	171	18 028	276

Tabulka 2. Prvních 30 nejčastějších hyperlinků – výňatek z celkové analýzy (přírodovědné/technické weby n = 6 087, společenskovední/humanitní weby n = 3 381)

Společenskovední/humanitní weby				Přírodovědné/technické weby			
	Název	n	prop		Název	n	prop
1	facebook.com	22	0,81		facebook.com	43	0,91
2	youtube.com	21	0,78		youtube.com	43	0,91
3	twitter.com	16	0,59		ceskatelevize.cz	39	0,83
4	avcr.cz	15	0,56		avcr.cz	37	0,79
5	ceskatelevize.cz	15	0,56		doi.org	35	0,74
6	instagram.com	13	0,48		sciencedirect.com	35	0,74
7	youtu.be	13	0,48		dx.doi.org	32	0,68
8	doi.org	12	0,44		nature.com	31	0,66
9	msmt.cz	12	0,44		link.springer.com	29	0,62
10	cs.wikipedia.org	11	0,41		onlinelibrary.wiley.com	29	0,62
11	docs.google.com	11	0,41		twitter.com	29	0,62
12	lib.cas.cz	11	0,41		youtu.be	26	0,55
13	novinky.cz	11	0,41		msmt.cz	25	0,53
14	rozhlas.cz	11	0,41		tydenvedy.cz	24	0,51
15	academia.cz	10	0,37		instagram.com	23	0,49
16	ec.europa.eu	10	0,37		scopus.com	23	0,49
17	vyzkum.cz	10	0,37		ncbi.nlm.nih.gov	22	0,47
18	gacr.cz	9	0,33		mdpi.com	21	0,45
19	mapy.cz	9	0,33		gacr.cz	20	0,43
20	mzv.cz	9	0,33		linkedin.com	20	0,43
21	scopus.com	9	0,33		pubs.acs.org	20	0,43
22	archiv.ihned.cz	8	0,3		google.com	19	0,4
23	bit.ly	8	0,3		rvvi.cz	19	0,4
24	cas.cz	8	0,3		vesmir.cz	19	0,4
25	cesnet.zoom.us	8	0,3		ec.europa.eu	18	0,38
26	jstor.org	8	0,3		lidovky.cz	18	0,38
27	rvvi.cz	8	0,3		natur.cuni.cz	18	0,38
28	search.ebscohost.com	8	0,3		novinky.cz	18	0,38
29	vltava.rozhlas.cz	8	0,3		ct24.ceskatelevize.cz	17	0,36
30	aleph22.lib.cas.cz	7	0,26		lib.cas.cz	17	0,36

Poznámky: Prop = % webů z kategorie, které odkazují na danou adresu. První řádek říká, že 81 % společenskovedních/humanitních webů a 91 % přírodovědných/technických webů odkazuje na stránky facebook.com. Celá tabulka k dispozici na vyžádání.

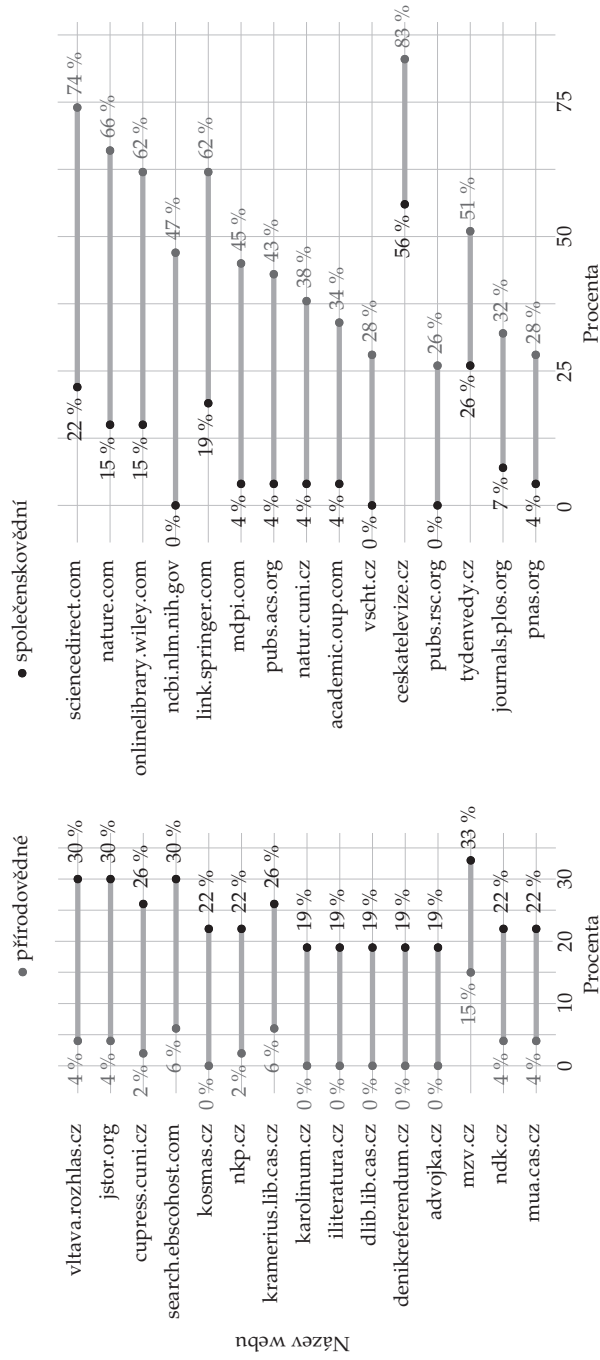
Výsledky analýzy odkazů

Celkově jsme identifikovaly 3 381 unikátních domén, na které odkazují sociálněvědní/humanitní weby a 6 087 unikátních domén odkazovaných přírodovědnými/technickými weby. V rámci našich výzkumných otázek nás zajímalo, na které weby odkazuje nejvíce institucionálních webů, což zachycuje Tabulka 2. První řádek tak podává informaci o tom, že 81 % společenskovědních/humanitních webů (v absolutním počtu 22) a 91 % přírodovědných/technických webů odkazuje na sociální médium facebook.com. Dále 78 % společenskovědních/humanitních webů a 91 % přírodovědných/technických webů odkazuje na link youtube.com. Třetí nejodkazovanější link je potom v rámci společenskovědních/humanitních webů twitter.com (59 %) a v rámci přírodovědných/technických webů ceskatelevize.cz (83 %). Nejodkazovanějším zdrojem mezi institucionálními vědeckými weby se tedy stávají sociální média, resp. komunitní fóra (např. Facebook, LinkedIn), podcastové/video platformy (např. YouTube, Spotify) a blogy/mikroblogy (např. Twitter). Téměř polovina společenskovědních/humanitních a přírodovědných/technických webů rovněž odkazuje na instagram.com, na kterém se sdílejí obrázky, fotografie a videa, a okolo 40 % na wikipedia.org. Nechybí ani četné hyperlinky na zpravodajská média typu ceskatelevize.cz, novinky.cz, rozhlas.cz, lidovky.cz, ihned.cz atp., na které odkazuje okolo 40 až 80 % vědeckých webů. Fakt, že na tyto weby odkazuje většina institucionálních webů, signalizuje využívání nových technologií a médií v rámci prezentace a popularizace české vědy, což by mohlo ukazovat na formy komunikace umožňující interaktivní diskusi, vstup veřejnosti do diskuse a přímou konfrontaci vědců či výzkumných institucí s veřejností.

Dále jsme se zaměřily na weby, které jsou nejvíce nadreprezentované v odkazech společenskovědních institucí oproti přírodovědným a naopak (viz Graf 7). Z Grafu 7 je patrné, že odkazy na některé zpravodajské weby jsou unikátní pro společenskovědní instituce. Zatímco na Deník Referendum, A2 a Český rozhlas Vltava odkazovalo mezi 19 a 30 % všech společenskovědních institucí, z přírodovědně zaměřených institucí na ně neodkazovaly žádné nebo pouhé jednotky. Specifická jsou také některá nakladatelství, konkrétně Kosmas a Nakladatelství Karolinum, a vědecké databáze jako EBSCO nebo JSTOR. Naopak většina přírodovědných institucí (62–74 %) sdílela odkazy na weby ScienceDirect, Nature, Wiley, Springer (oproti 15–22 % společenskovědních institucí). Takřka unikátní pro přírodovědné instituce byly potom odkazy na web VŠCHT, Přírodovědecké fakulty Univerzity Karlovy, nakladatelství MDPI a Royal Society of Chemistry nebo časopisy PNAS a PLoS. Zajímavostí jsou také odkazy na popularizační akci Týden vědy a techniky, nalezené u poloviny (51 %) přírodovědných webů, ale pouze u čtvrtiny (26 %) webů společenskovědních.

Celkově institucionální weby odkazovaly na více než 10 000 unikátních webů (resp. domén). Jelikož navštívit každý z webů a zakódovat jeho typ nebylo v rámci projektu možné, zaměřily jsme se na weby, na které bylo odkazováno nejčastěji (tj. nejvíce weby sociálněvědních a přírodovědných institucí). Konkrét-

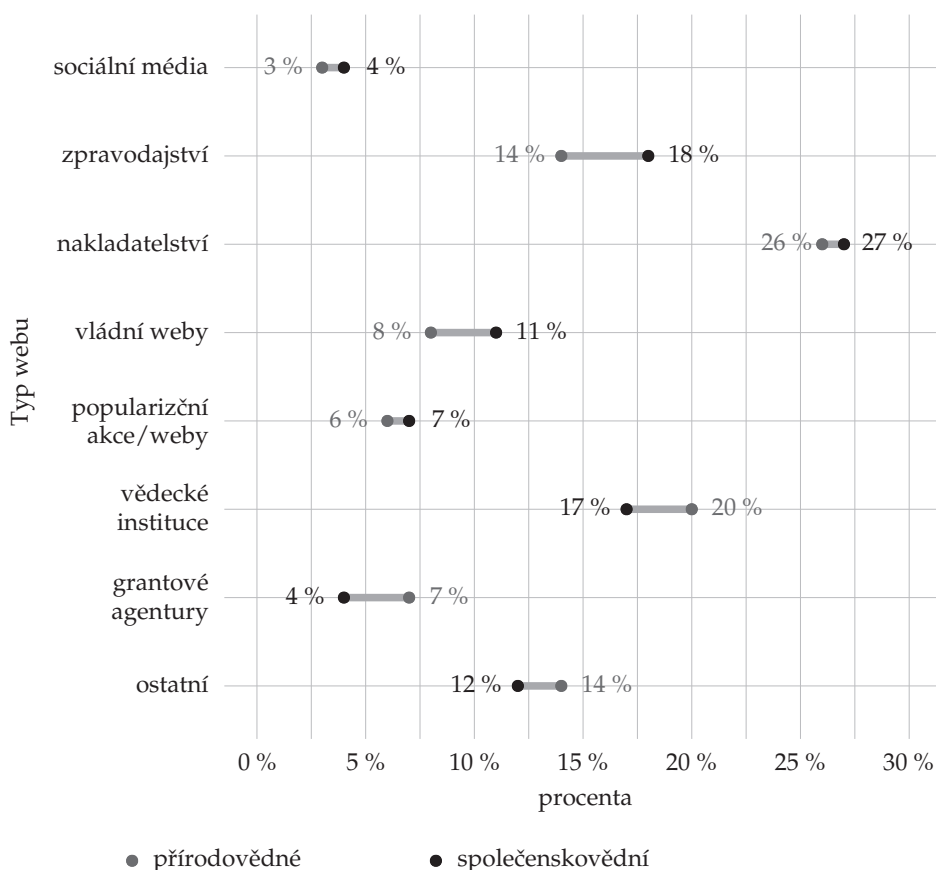
Graf 7. Nejvíce nadreprezentované weby dle typu instituce



Poznámka: Graf udává, kolik % institucionálních webů z dané kategorie na konkrétní web odkazuje. Vlevo odkazy nadreprezentované u společenskovedních institucí, vpravo odkazy nadreprezentované u přírodovědných institucí.

ně jsme okódovaly ty weby, na které odkazovalo alespoň 10 % institucionálních webů společenských věd (n = 242) nebo alespoň 10 % institucionálních webů přírodních věd (n = 239). Rozlišujeme následujících 8 kategorií webů: (1) sociální média, (2) zpravodajské weby, (3) knihkupectví, nakladatelství a knihovny, (4) vládní/ministerské weby, (5) popularizační weby/časopisy/akce, (6) jiné vědecké instituce/univerzity, (7) grantové agentury a (8) ostatní. Každá z autorek nezávisle navštívila každý z webů a okódovala jej. Při prvním kódování byla dosažena 77% shoda u webů, na které odkazovaly společenskovědní instituce a 71,3% shoda u webů, na které odkazovaly přírodovědné instituce. Následně byly probrány weby, u nichž se autorky při prvním kódování neshodly, a i u nich bylo dosaženo shody.

Graf 8. Struktura nejčastějších 250 hyperlinků na webech pro společenské/humanitní a přírodní/technické vědy



Ačkoliv se může na první pohled zdát, že jak sociálněvědní, tak přírodovědné institucionální weby málo odkazují na sociální média (3 %), je nutné si uvědomit, že velkých hráčů na poli sociálních médií je jen velmi omezený počet, a tudíž nemohou zastávat větší proporce. Druhou nejčastěji zastoupenou kategorií jsou odkazy na knihkupectví, nakladatelství a knihovny (25 % a 26 %), kam se řadí mj. odkazy na akademické publikace. Sociálněvědní weby potom odkazují na více zpravodajských (19 % oproti 14 %) a vládních webů (13 % oproti 9 %) než ty přírodovědné, zatímco přírodovědné weby odkazují na více webů jiných vědeckých institucí (20 % oproti 17 %), grantových agentur (6 % oproti 3 %) a v neposlední řadě také popularizačních webů (8 % oproti 6 %). Výsledky zachycuje Graf 8.

Diskuse a shrnutí

Tato studie se primárně zaměřila na téma komunikace vědy na internetových stránkách českých výzkumných organizací. Zajímalo nás, jaký charakter a formu tato komunikace české vědy v online prostředí má a do jaké míry je česká věda v online prostředí sdílena s veřejností – a jaký je v tomto ohledu rozdíl mezi společenskovědními/humanitními a přírodovědnými/technickými weby. Výsledky ukázaly, že online obsah vědeckých organizací se nijak nevymyká běžným standardům, kdy jsou pojmy „vědy“ a „popularizace“ komunikovány především v souvislosti s dalšími vzdělávacími institucemi, studiem a výzkumem. Oproti sociálním vědám je v případě přírodních věd častěji popularizace spojována s oceňováním vědců právě za propagaci a popularizaci vědy, což může na jedné straně ukazovat na větší incentive k popularizaci v těchto oborech. Může to ale být i naopak – pokud popularizace do značné míry chybí, tak ji nelze ani oceňovat.

Strukturální analýza webů nám pak dala možnost nahlédnout komunikaci české vědy spojenou s novými sociálními médii. Rozbor hyperlinků naznačuje, že naprostá většina vědeckých webů odkazuje na sociální média typu Facebook, Twitter, YouTube, Instagram a LinkedIn. Podobně je v online obsahu vědeckých institucí odkazováno na zpravodajská média typu *ceskatelevize.cz*, *novinky.cz*, *lidovky.cz* nebo *rozhlas.cz* – tedy na online noviny, televizi i rozhlas. To potvrzuje i pohled na strukturu hyperlinků sdílených největším počtem sociálněvědních nebo přírodovědných institucionálních webů – zpravodajská média se objevují nejčastěji po hyperlincích týkajících se akademických publikací. Výsledky tedy naznačují, že komunikace české vědy v online prostředí směřuje k otevřenému, interaktivnímu sdílení vědy s veřejností.

Přestože mají výše shrnuté výsledky svou výpovědní hodnotu, je nutné také upozornit na jejich úskalí. Již k prvnímu zkreslení může dojít při hledání institucionálních vědeckých webů, respektive k jejich nedostatečnému pokrytí při výběru výzkumníkem. Totéž platí pro (subjektivní) volbu klíčových slov, na základě

nichž se potom stahuje množství online obsahu k další analýze. Nutno říci, že naším původním záměrem bylo provést obsahovou a strukturální analýzu všech internetových zdrojů, kde se komunikuje věda (těch jsme identifikovaly přibližně 166). V nich jsme potom měly v úmyslu analyzovat na základě klíčových slov počty příspěvků věnovaných společenským či přírodním vědám (tj. tiskové zprávy, aktuality, krátké zprávy, popularizační články, vědecké zpravodajství v médiích, blogy o vědeckých tématech, rozhovory s vědci a vystoupení vědců v televizním či rozhlasovém pořadu). To ale nebylo z důvodu velikosti dat, technické náročnosti (dostupné techniky či zpracování dat jako takových) ani časových možností v rámci našeho projektu možné.

Naše studie má řadu limitů. Prvním z nich je, že k webovým stránkám přistupujeme, jako by byly statické, což nemusí nutně platit u každého z webů. Následkem tohoto přístupu tak potenciálně přicházíme o část informací z dynamických stránek, které generují obsah na základě chování uživatele, např. tedy načítají další informace po kliknutí na určité tlačítko, aniž by uživatel opustil dané URL.

Druhé omezení vyvstává ze zpracování textových dat. Ačkoliv jsme věnovaly čištění dat značnou pozornost, nachází se v datech nadále částečné duplicity. Některé texty obsahovaly více než jedno klíčové slovo, například ve spojení „popularizace vědy v prostředí biologie“ jsou klíčová slova tři („popularizace“, „věda“ a „biologie“), přičemž ke každému z nich stahujeme kontext 10 slov okolo. Výsledkem jsou tři texty, které se do značné míry překrývají, ale nejsou zcela totožné. Vyřadit ovšem tyto částečné duplicity by bylo výpočetně (a tím pádem i časově) náročné – obnáší to porovnat všechny texty mezi sebou, spočítat pravděpodobnost, se kterou se jedná o částečné duplicity, a následně tyto duplicity sloučit. Vzhledem k tomu, že náš datový soubor obsahuje více než milion takových textů, zabral by tento krok sám o sobě několik dnů či týdnů. Dalším vylepšením by také bylo zahrnutí sousloví (např. „Česká republika“), která nám v modelu pracujícím pouze s unigramy, tj. jednoslovnými výrazy, unikají. Dále se ve zpracování textových dat objevily specifické komplikace – kupříkladu problémy s termínem „sociologický“, uvedené v popisné analýze textových dat, poukazují na to, že pro srovnávací účely by bylo vhodné počty vážit velikostí webových stránek.¹²

Třetím omezením je, že nedokážeme určit stáří stahovaných URL, protože většina webů tuto informaci buď neobsahuje, nebo je generována dynamicky při otevření webu. Pokud by bylo naším cílem kompletně zmapovat proměny komunikace české vědy na institucionálních webech nebo na českém internetu obecně v delším časovém horizontu, vyžadovalo by to longitudinální studii internetového obsahu. K tomu bychom do budoucna mohly využít nově vznikající softwarový nástroj webového archivu Národní knihovny, umožňující procházet obsah licencovaného českého internetu zpětně až do roku 2000.

¹² Za tyto přínosné poznámky děkujeme recenzentovi.

Další bariérou byla nemožnost sledovat vědeckou komunikaci přímo na sociálních médiích, neboť kupříkladu Twitter umožňuje sbírat text mikroblogů, ale Facebook či Instagram nikoliv. V tomto ohledu by bylo zajímavé analyzovat sledovanost jednotlivých vědeckých příspěvků, počet lajků, velikost diskuse a rozsah dalšího sdílení. Někteří autoři tak uvádějí, že i když vědecké instituce již akceptovaly sociální média jako primární nástroj komunikace s veřejností, stále je ještě využívají spíše jako jednosměrný komunikační kanál, informují o vědeckých poznatcích a pokroku, avšak podceňují jejich potenciál pro vedení dialogu a navazování diskuse s veřejností (Dudo a Besley, 2016; Lovejoy a Saxton, 2012).

Závěrem lze podotknout, že i přes zmíněné nedostatky prezentovaného textu se jedná o první kvantitativní explorační studii, která ke zmapování charakteru online komunikace české vědy využila analýzu velkých dat. Doufáme, že detailně pojatý metodologický oddíl a kritické zhodnocení limitů a nastínění dalších možností analýzy online obsahu webových stránek vědeckých organizací poslouží dalším výzkumníkům i jako úvodní představení práce s velkými daty v sociologii, včetně jejich kladů i záporů.

PETRA RAUDENSKÁ *pracuje jako vědecká pracovnice Sociologického ústavu AV ČR. K jejím výzkumným zájmům patří metodologie sociálních výzkumů, vzdělanostní nerovnosti, lidský kapitál a problematika hodnotových orientací. Je autorkou monografie Srovnatelnost postojevých škál v komparativním výzkumu (UP v Olomouci, 2015). Publikovala v časopisech Social Science Research, Poetics, International Sociology a Innovation: The European Journal of Social Science Research.*

ORCID: 0000-0002-0330-139X

Research ID: Q-2330-2016

RENÁTA TOPINKOVÁ *je doktorandkou na Katedře sociologie Filozofické fakulty UK, pracovala v oddělení Sociální stratifikace Sociologického ústavu AV ČR a v současnosti působí na LMU v Mnichově v oddělení Computational Social Sciences. Ve své dizertaci se zabývá homofilií v online seznamování. Výzkumně se zajímá o sociologii rodiny a computational social science. Publikovala v časopisech Journal of Family Issues, Journal of Family Research a Sociological Research Online.*

ORCID: 0000-0003-0362-7290

Literatura

- Ackland, R. (2013). *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. Sage.
- Andrle, M. (2013). Současné přístupy k popularizaci vědy v České republice. *Teorie vědy/Theory of Science*, 35(1), 113–125.
- Bail, C. A. (2014). The Cultural Environment: Measuring Culture with Big Data. *Theory and Society*, 43(3), 465–482. <https://doi.org/10.1007/s11186-014-9216-5>
- Bail, C. A. (2015). Lost in a Random Forest: Using Big Data to Study Rare Events. *Big Data & Society*, 2(2), 2053951715604333. <https://doi.org/10.1177/2053951715604333>
- Bauer, M. W., Allum, N. a Miller, S. (2007). What Can We Learn from 25 Years of PUS Survey Research? Liberating and Expanding the Agenda. *Public Understanding of Science*, 16(1), 79–95. <https://doi.org/10.1177/0963662506071287>
- Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4), 77–84.
- Borgman, C. L. a Furner, J. (2002). Scholarly Communication and Bibliometrics. *Annual Review of Information Science and Technology*, 36(1), 2–72. <https://doi.org/10.1002/aris.1440360102>
- Broks, P. (2006). *Understanding Popular Science*. Open University Press.
- Brossard, D., Dietram, E. a Scheufele, A. (2013). Science, New Media, and the Public. *Science*, 339(40), 40–41. <https://doi.org/10.1126/science.1232329>
- Brossard, D. a Scheufele, D. A. (2013). Science, New Media, and the Public. *Science*, 339(6115), 40–41. <https://doi.org/10.1126/science.1232329>
- Burns, T. W., O'Connor, D. J. a Stocklmayer, S. M. (2003). Science Communication: a Contemporary Definition. *Public Understanding of Science*, 12(2), 183–202.
- Collins, K., Shiffman, D. a Rock, J. (2016). How Are Scientists Using Social Media in the Workplace? *PLoS ONE*, 11(10), e0162680. <https://doi.org/10.1371/journal.pone.0162680>
- Côté, I. M. a Darling, E. S. (2018). Scientists on Twitter: Preaching to the Choir or Singing from the Rooftops? *FACETS*, 3(1), 682–694. <https://doi.org/10.1139/facets-2018-0002>
- Čada, K., Červinková, A., Linková, M., Řeháčková, D. a Stöckelová, T. (2006). *Věda jako věc veřejná: vědní politiky a média*. Sociologický ústav AV ČR.
- ČSÚ. (2010). *Kolik domácností v ČR má počítač a internet?* Český statistický úřad. https://www.czso.cz/csu/czso/kolik_domacnosti_v_cr_ma_pocitac_a_internet
- Davies, S. R. a Hara, N. (2017). Public Science in a Wired World: How Online Media are Shaping Science Communication. *Science Communication*, 39(5), 563–568. <https://doi.org/10.1177/1075547017736892>
- Davies, S. R. a Horst, M. (2016). *Science Communication: Culture, Identity and Citizenship*. Springer.
- Dijkstra, A. M., Roefs, M. M. a Drossaert, C. H. (2015). The Science-media Interaction in Biomedical Research in the Netherlands. Opinions of Scientists and Journalists on the Science-media Relationship. *Journal of Science Communication*, 14(2), 1–21. <https://doi.org/10.22323/2.14020203>
- Dudo, A. a Besley, J. C. (2016). Scientists' Prioritization of Communication Objectives for Public Engagement. *PLoS ONE*, 11(2), e0148867. <https://doi.org/10.1371/journal.pone.0148867>
- Gu, F. a Widén-Wulff, G. (2011). Scholarly Communication and Possible Changes in the Context of Social Media: A Finnish Case Study. *The Electronic Library*, 29(6), 762–776. <https://doi.org/10.1108/02640471111187999>
- Hilgartner, S. (1990). The Dominant View of Popularization: Conceptual Problems, Political Uses. *Social Studies of Science*, 20(3), 519–539. <https://doi.org/10.1177/030631290020003006>

- Hrabánková, M. (2018). Způsob prezentace přírodních věd ve vybraných českých médiích v roce 2013. *Mediační studia*, 12(01), 115–132.
- Jünger, J. a Fähnrich, B. (2020). Does Really No One Care? Analyzing the Public Engagement of Communication Scientists on Twitter. *New Media & Society*, 22(3), 387–408. <https://doi.org/10.1177/1461444819863413>
- Kee, Q., Ahn, Y. Y. a Sugimoto, C. R. (2017). A Systematic Identification and Analysis of Scientists on Twitter. *PLoS one*, 12(4), e0175368. <https://doi.org/10.1371/journal.pone.0175368>
- Lee, N. M. a VanDyke, M. S. (2015). Set It and Forget It: The One-way Use of Social Media by Government Agencies Communicating Science. *Science Communication*, 37(4), 533–541. <https://doi.org/10.1177/1075547015588600>
- Lovejoy, K. a Saxton, G. D. (2012). Information, Community, and Action: How Nonprofit Organizations Use Social Media. *Journal of Computer-Mediated Communication*, 17(3), 337–353. <https://doi.org/10.1111/j.1083-6101.2012.01576.x>
- MacNaghten, P., Kearnes, M. B. a Wynne, B. (2005). Nanotechnology, Governance, and Public Deliberation: What Role for the Social Sciences? *Science Communication*, 27(2), 268–291. <https://doi.org/10.1177/1075547005281531>
- Massoli, L. (2007). Science on the Net: An Analysis of the Websites of the European Public Research Institutions. *Journal of Science Communication*, 6(3), A03.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M. a McCallum, A. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (s. 262–272). <https://aclanthology.org/D11-1024.pdf>
- Nielsen, H. K. 2010. More than ‘Mountain Guides’ of Science: A Questionnaire Survey of Professional Science Communicators in Denmark. *Journal of Science Communication*, 9(2), p.A02. <https://doi.org/10.22323/2.09020202>
- Nielsen, K. H., Kjaer, C. R. a Dahlggaard, J. 2007. Scientist and Science Communication: A Danish Survey. *Journal of Science Communication*, 6(1), 1–12. <https://doi.org/10.22323/2.06010201>
- Nisbet, M. C. a Huge, M. (2006). Attention Cycles and Frames in the Plant Biotechnology Debate: Managing Power and Participation through the Press/Policy Connection. *Harvard International Journal of Press-Politics*, 11(2), 3–40. <https://doi.org/10.1177/1081180x06286701>
- Nisbet, M. C. a Scheufele, D. A. (2009). What’s Next for Science Communication? Promising Directions and Lingering Distractions. *American Journal of Botany*, 96(10), 1767–1778. <https://doi.org/10.3732/ajb.0900041>
- Noruzi, A. (2008). Editorial: Science Popularization through Open Access. *Webology*, 5(1), editorial 15.
- Office of Science and Technology and the Wellcome Trust, O. O. S. A. (2001). Science and the Public: A Review of Science Communication and Public Attitudes toward Science in Britain. *Public Understanding of Science*, 10(3), 315–330.
- Pomikálek, J. (2011). *JusText* [software]. Masaryk University, NLP Centre. <https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-000D-F696-9>.
- Public Understanding of Science. (2014). Special Issue: Public Engagement in Science. <https://journals.sagepub.com/toc/pus/23/1>
- Purcell, K., Brennen, J. a Rainie, L. (2012). *Search Engine Use 2012*. Pew Research Center. <https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012/>
- The Royal Society. (2006). *Science Communication. Survey of Factors Affecting Science Communication by Scientists and Engineers*. The Royal Society, Research Councils UK, and Wellcome Trust.

- Robins, G. (2015). *Doing Social Network Research: Network-based Research Design for Social Scientists*. Sage.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>
- Rödger, S., Franzen, M. a Weingart, P. (eds.). 2012. *The Sciences' Media Connection—public Communication and Its Repercussions* (Vol. 28). London: Springer Science & Business Media.
- Rybalko, S. a Seltzer, T. (2010). Dialogic Communication in 140 Characters or Less: How Fortune 500 Companies Engage Stakeholders Using Twitter. *Public Relations Review*, 36(4), 336–341. <https://doi.org/10.1016/j.pubrev.2010.08.004>
- Sis.net. (n.d.). *Science Communication Policy Brief*. <https://www.sisnetwork.eu/media/althjodasvid/Policy-Brief-SCIENCE-COMMUNICATION-FINAL.pdf>
- Scheufele, D. A. (2014). Science Communication as Political Communication. *Proceedings of the National Academy of Sciences*, 111(Supplement 4), 13585–13592. <https://doi.org/10.1073/pnas.1317516111>
- Silge, J. a Robinson, D. (2017). *Text Mining with R: A Tidy Approach* (First edition). O'Reilly.
- Stockmayer, S., Bryant, Ch. a Gore, M. M. (2002). *Science Communication in Theory and Practice*. Kluwer Academic Publishers.
- Straka, M. a Straková, J. (2019). Universal Dependencies 2.5 Models for UDPipe (2019-12-06), LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-3131>
- Su, L. Y. F., Scheufele, D. A., Bell, L., Brossard, D. a Xenos, M. A. (2017). Information-sharing and Community-building: Exploring the Use of Twitter in Science Public Relations. *Science Communication*, 39(5), 569–597. <https://doi.org/10.1177/1075547017734226>
- Šamanová, G., Škodová, M. a Vinopal, J. (2006). *Obráz vědy v českém veřejném mínění*. Sociologické studie 2006:8. Sociologický ústav AV ČR.
- Treise, D. a Weigold, M. (2002). Advancing Science Communication: A Survey of Science Communicators. *Science Communication*, 23(3), 310–322.
- Trench, B. a Miller, S. (2012). Policies and Practices in Supporting Scientists' Public Communication through Training. *Science and Public Policy*, 39(6), 722–731. <https://doi.org/10.1093/scipol/scs090>
- Uren, V. a Dadzie, A. S. (2015). Public Science Communication on Twitter: A Visual Analytic Approach. *Aslib Journal of Information Management*, 67(3), 337–355. <https://doi.org/10.1108/ajim-10-2014-0137>
- Van Noorden, R. (2014). Scientists and the Social Networks. *Nature news*, 512(7513), 126–130. <https://doi.org/10.1038/512126a>
- Wijffels, J., Straka, M. a Straková, J. (2021). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit* (R package version 0.8.6) [software]. <https://cran.r-project.org/web/packages/udpipe/index.html>