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

**PETRA RAUDENSKÁ[1]**, RENÁTA TOPINKOVÁ[1,2]**
[1]Institute of Sociology, Czech Academy of Sicences, Prague
[2]Ludwig-Maximilians-Universität München

**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 characterised 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*

* This work was supported by the Ministry of Culture Czech Republic: 'Development of the centralized interface for the web content and social networks data mining' (Grant No. DG18P02OVV016). Authors would like to thank to reviewers for useful comments and to their colleagues Matouš Pilnáček for the help with data preparation and Tomáš Diviák for the consultation on social network analysis.
** Direct all correspondence to: PhDr. Petra Raudenská, Ph.D., Institute of Sociology, Czech Academy of Sciences, Jilská 1, 110 00 Praha 1, Czech Republic, e-mail: petra.raudenska@soc.cas.cz.

## Introduction

Until recently, science communicated with the public primarily through press releases or professional media. Nowadays, the Internet is the first place most people turn to when searching for scientific (or any other) information (Purcell et al., 2012). With the rise of the Internet, and social media in particular, individual scientists and entire institutions can share the results of their work in a variety of ways, from articles, videos, and podcasts to direct communication via Twitter or Facebook. The new social media thus open up a wide range of possibilities for scientific collaboration and increasing public interest in science (van Noorden, 2014). Whereas in the past, the media used to receive press releases from experts and would communicate their content to the public, today anyone can modify their form and content by sharing or editing them.

In contrast to the traditional (canonical) model of communicating scientific information, in which the focus is on scientific representatives communicating with policy makers and members of the media (Nisbet & Scheufele, 2009), scientists are now called upon to actively communicate scientific knowledge and engage in a dialogue with the public (Sis.net, n.d.). The emphasis is on mutual, interactive, open, and less formal communication between scientists and the public through all media platforms available (Lee & VanDyke, 2015). The difference should be not just that scientists are involved, but that the public is involved as well; as the public should not just be passive consumers of content but should instead be encouraged to participate directly in emerging science (Sis.net, n.d.). Despite this call for the greater engagement of the public, there are still debates on how to lead this public dialogue in a meaningful way and many scientists present their results in more of a one-way manner – for example, by communicating with science journalists (Andrle, 2013; The Royal Society, 2006).

Although the issue of science journalism and the image of science in the Czech media are subjects that have already been addressed by some authors in the field of Czech science, their attention has mainly focused on a content analysis of the Czech press (e.g. Čada et al., 2006; Hrabánková, 2018). However, there is still no quantitative study that has systematically or even partially mapped Czech online communication on science since the rise of the participatory web. Against the backdrop of the development of new online technologies, this exploratory study focuses on Czech online science communication, which will allow us to observe a much wider spectrum of the possible forms this communication can take. In the study, we observe not only one-sided science communication, whether controlled or uncontrolled, that comes from the websites of research institutions and is directed at the news media, available Czech media, scientific online supplements, and special websites focused on the interpretation of scientific knowledge, but also science communication that is directed at media that facilitate an open discussion with the public – for example, in the form of scientific figures who make appearances in the media, or links to blogs and social media.

The main research question we ask in this exploratory study is: *How is science currently communicated on the websites of Czech research institutions?* A study by Ke et al. (2017) suggests that the social sciences are more active in the online

environment than the natural sciences, although much of the content shared is not purely science-related. This finding led us to a secondary research question: *What is the difference between the communication of the natural/technical sciences on these websites and that of the social sciences/humanities?*

To explore the form of science communication that is found on the websites of Czech research institutions, we can use big data. Big data analysis is highly specific and requires a significant number of intermediate computational steps and time-consuming data file preparation. On the other hand, it provides seemingly simple answers to quite original (and sometimes quite complex) research questions that we would not be able to analyse using standard questionnaire surveys. A subsidiary (though not necessarily secondary) aim of this paper is therefore to enlighten the reader about the preparatory and intermediate computational steps that need to be taken when working with big data of this kind and the difficulties that a researcher may encounter when conducting content and structural analysis of data from websites. In this regard, we are also highly self-critical of our own results (see the Discussion and the Summary). From this point of view, we consider this study not only interesting in terms of its content, but also methodologically beneficial.

## Theoretical framework: forms of science communication

Nowadays, science is not only an important aspect of the well-being of individuals, organisations, and nations, it is also a key element of democracy and the contemporary culture of the knowledge society, as it underpins many crucial decisions and to some extent influences what individuals think about current social issues (Davies & Horst, 2006; Dijkstra et al., 2015). Science itself is constantly evolving, and the volume of new findings is continually increasing – along with the element of uncertainty that is inherent to science. This generates a need to be able to navigate this wealth of information. Science communication thus fulfils the essential task of making scientific knowledge understandable and accessible to the general public.

*Science communication* can be defined as organised action for the purpose of communicating scientific knowledge, methods, processes, and practices (Borgman & Furner, 2002) or as the use of appropriate skills, media, activities, and dialogue to elicit an individual response to science (Burns et al., 2003). In more concrete terms, it aims to produce a familiarity with new scientific knowledge, an emotional response, such as enjoyment, interest, and opinion-sharing, and an understanding of science itself and its content, processes, and social factors. According to the definition put forward by Burns et al., the actors in science communication include not only scientists themselves, but also intermediaries, the public, and members of close social groups (e.g. peers).

As such, science communication is often synonymously referred to or discussed in the context of 'public awareness of science', 'the popularisation of science', 'public understanding of science', 'scholarly communication', 'scientific culture', 'scientific literacy', or 'science journalism' (for more on their definitions,

see Burns et al., 2003). In modern terms, information on science communication can be found using the hashtags #scicomm, #Scicommunication, or #ScienceCommunication.

Science communication has two essential communication goals: (a) communicating information in order to increase scientific literacy and a general awareness of science and (b) engaging the public in scientific debate in order to increase public participation in science (Scheufele, 2014; for a general textbook on science communication in practice, see Stocklmayer et al., 2002). It takes place on several levels, from the more formal level of publishing scientific studies in academia and the dissemination of findings at conferences, science fairs, and popularisation events, to the less formal level of communicating science through press releases, books aimed at the general public, and communication activities on television and radio and in print and online (Bauer et al., 2007; Gu a Widén-Wulff 2011).

Nielsen et al. (2007) present two relevant models of science communication: the canonical model and the interactive, reflexive model. Historically, the *canonical model of science communication* (sometimes also referred to as the information-dissemination model; see Hilgartner, 1990) has been used to explain the relationship between scientists and the public, where scientists themselves produce scientific knowledge and further disseminate their findings in order to educate or even entertain the wider public, but also to socially legitimise their scientific endeavours. This mostly relates to communication with government officials, institutions, businesses, and scientific and professional conferences, the issuing of press releases, and limited communication with the media, mostly without any audience response. Despite all the constructive objections (see the interactive communication model below; for a summary, see Broks, 2006), for many scientists the canonical model of one-way science communication is still an example of the best mode of science communication in the public space.

The dominance of the canonical model is also illustrated by the results of a representative survey of British scientists and engineers, in which only 12% considered communication with the general public to be truly relevant (The Royal Society, 2006; for previous research see Treise & Weigold, 2002). On the other hand, communication with public officials was rated as one of the most important activities by almost 90% of scientists. Although respondents among scientists cited 'informing, explaining, and seeking a correct understanding of the facts' as the main definition of how to interact with the public, no key words such as 'public discussion' or 'interaction with the public' were mentioned in an open-ended question about the meaning of the term 'public understanding of science'. Similar results were produced by a short survey of Czech scientists and scientific editors from 2013 on the topic of the popularisation of science in the Czech Republic (Andrle, 2013). Here, 16 respondents agreed that scientific knowledge at the time was communicated mainly through printed and online newspapers, radio, and television, and also noted the need to cultivate good relationships with journalists and to popularise knowledge in a sufficiently simple but still professionally correct way. Only three respondents explicitly mentioned social media or actual public dialogue.

Since 2000, the *interactive, reflexive model* has been presented as an alternative model to the canonical model of science communication. It emphasises the role of two-way communication (Nisbet & Scheufele, 2009), encourages greater public participation in science, and seeks to bring scientists into the arena of public dialogue (Burns et al., 2003; Jünger & Fähnrich, 2020; MacNaghten et al., 2005; Trench & Miller, 2012). It asserts the idea of exchanging the knowledge and competences of scientists with society and calls for the interactive participation of all the components of science communication (Sis.net, n.d.). In contrast to the canonical model, which calls for scientific representatives of universities to communicate with policy makers and the media, the interactive model builds upon a more complex understanding of universities and research institutions as knowledge and cultural institutions and upon the creation of a relationship with the public (Lee & VanDyke, 2015; Scheufele, 2014). Therefore, in today's science communication, in addition to the two dominant actors – *the scientists* who create scientific data and communicate them to the scientific community and *the communicators* who transmit the facts – emphasis is placed foremost on the role of a third group, *the public*; in this regard, the public is seen not as a passive recipient of science communication but as an active agent engaging in dialogue across different media platforms (Sis.net, n.d.)[1].

In this context, the findings of an online survey of Danish natural scientists published in the *Journal of Science Communication* in 2007 are certainly of interest. When the scientists had to evaluate the importance of individual types of media according to their distribution and the size of the target audience, they considered the mass media, interdisciplinary popular-science journals, and public debates to be the most important. The scientists also felt a responsibility to disseminate knowledge, especially new findings and current research outcomes (Nielsen et al., 2007). A subsequent Danish survey of science communicators confirmed that scientists are not only interested in helping the public understand scientific knowledge, they also want to actively contribute to the democratic debate and seek to legitimise science and technology as such (Nielsen, 2010).

The rise of new social media associated with Web 2.0 has transformed the possibilities of informal dialogue among scientists, opened up this space to the lay public and interested parties, and made it possible to increase public engagement in science (Su et al., 2017; Uren & Dadzie, 2015). With the recent development of technology, especially online journalism, e-books, e-conferences, webinars, e-workshops, videoconferences, audio-visual materials, blogs, discussion forums, and new social media, the possibilities for sharing information about science have also expanded (Brossard & Scheufele, 2013; Noruzi, 2008). Science communication in the media is thus increasingly being confronted with the challenged of brining scientists 'face to face' with the public and its interest subgroups. It does so not only through form and content, but also through the growing range of available interactive media (Gu & Widén-Wulff, 2011).

[1] For more on this topic, we recommend reading the special annual issue of *Public Understanding of Science* – Special Issue: Public Engagement in Science (2014).

Despite the unprecedented growth of media coverage of science news in recent decades (Rödder et al., 2012), scientists agree that science communication is still more of a one-way process of communicating information and has not yet fully utilised the potential of new technologies and social media. Trench & Miller (2012) point out that scientists, scientific institutions, and science journalists tend to use the Internet more for professional communication than for mutual communication between scientists or for opening a public dialogue with the support of available multimedia tools (Rybalko & Seltzer, 2010; for a critique of science communication in the online environment, see Davies & Hara, 2017). Collins et al. (2016) confirm that, for example, Twitter is mainly used by scientists to exchange scientific information. On the other hand, Côté & Darling (2018) show that the majority of 'non-scientific' followers of the Twitter microblogs of specific scientists are the media, politicians, and the interested public. Thus, Twitter as a social medium becomes a hybrid tool for science communication that alternates between acting as an information exchange, promoting science, and engaging in a dialogue with individuals or the public (Jünger & Fähnrich, 2020). Although Twitter is used across academic disciplines, the social sciences are more active than the natural sciences, with much of the content shared being not about science but a response to relevant political or social issues (Ke et al., 2017).

This study loosely follows the efforts of several Czech authors to map science communication in the Czech press (Čada et al., 2006) and the image of Czech science in public opinion (Šamanová et al., 2006; cf. an interesting study on the attitudes of the British public towards science by the Office of Science and Technology and the Wellcome Trust, O. O. S. A., 2001). According to the Czech Statistical Office (2010), in 2006, when these studies were published, only 27% of Czech households had access to the Internet; while in 2019, 81% of Czechs over the age of 16 were already using the Internet and 70% were using a smartphone.

Over the past ten years, we have thus seen increasing pressure in the scientific community being put on scientists to communicate with the public and to move this communication into the online environment. For example, the Czech Science Foundation and university grant agencies have allocated financial support to several projects dedicated to the popularisation of science. The Czech Academy of Sciences (CAS) publishes magazines (also online) about science for the public (*Věda a výzkum*, *Živa*), organises the largest popularisation events in the country, such as the Science Fair and the Week of Science and Technology (including the Olomouc-based Academia Film (AFO) festival and Researchers' Night), and organises internships for students and popularisation courses (Open Science). The Centre of Administration and Operations of the CAS runs the popular YouTube channel *Zvěd*. Most institutes and universities have websites and Facebook and Twitter accounts that they actively use. The media itself also engages in extensive popularisation activities, with individual outlets creating various data visualisations (iRozhlas or Novinky), publishing their own extensive surveys (Czech Radio – the survey *Czech Society 30 Years Later*), or having their own scientific editorial offices (e.g. the scientific editorial office of Czech Television that is integrated into its entire news coverage). Czech Television's newscast broadcasts the programme *Hyde Park Civilizace* and the weekly programme *Věda 24* (Science 24), while the

publisher Seznam Zprávy broadcasts the programme *Výzva* (Challenge), where, among other things, scientists answer questions posed by the public; online debates are also frequently organised that invite questions 'from the audience'.

Given the current development of online technologies, this quantitative study rather uniquely focuses mainly on science communication in the Czech online environment. It must be emphasised at this point that the study is fully exploratory and descriptive and seeks to answer the questions of how science is currently communicated on the websites of Czech research institutions and what the difference is between the communication of the natural/technical sciences and the social sciences/humanities on these websites. Massoli (2007) similarly analysed the websites of official European research institutions. However, in addition to the international overlap, her work focused more on the visual and material content of the websites and assessed the possibilities for interaction with the users of the websites, user-friendliness, and the presentation of institutional identity, scientific credibility, available services, etc. In contrast, we focus mainly on the content and structural analysis of the text appearing on the websites of Czech research institutions and in this way observe the nature and form of Czech science communication.

## Data

To answer our research questions, it was necessary to identify the widest possible range of institutional websites – i.e. the websites of research institutions, affiliated scientific and research institutes, universities, grant agencies, specialised offices/institutes, and scientific events. In total, we identified 105 Czech institutional scientific websites on the Internet (see Table 1A in the Online Appendix[2]). Of these original 105 websites, 6 failed to download and 10 could only be partially downloaded. The partially downloaded websites were large websites whose content took more than a day to download and the download failed before the program's completion. Given the amount of data we extracted from these websites, we decided to include them in our analyses anyway, as we believe that most (if not all) of the relevant data were downloaded. In total, we obtained data from 99 websites.

The data were downloaded using Heritrix3,[3] a website archiving program commonly used by web archives. The program crawled hyperlinks from the homepage to a depth of 4 steps on each website. No content was downloaded outside of the domains of the original target websites. We extracted the hyperlinks, information about the presence of videos, and the textual content of the pages. The content was cleared of the boilerplate parts of the websites (e.g., menus, footers, advertisements) using the jusText package (viz Pomikálek, 2011).

In the following analyses, we often speak about the difference between the natural and the social sciences. We classified the sciences into the natural/technical sciences and the social sciences/humanities, in accordance with the common

[2] Online Appendix available at https://doi.org/10.13060/csr.2023.004.
[3] https://github.com/internetarchive/heritrix3

standard (e.g. how they are divided at universities). The natural and technical sciences thus include, for example, physics, mathematics, chemistry, biology, space sciences, and technology and their related subfields, while the social sciences or the humanities usually include history, cultural studies, sociology, philosophy, demography, economics, political science, psychology, linguistics, archaeology, art, law, and ethics. We identified 47 websites in the natural sciences and 27 in the social sciences, while 25 were marked as neutral, as it was not possible to classify them neatly into either category (e.g. the Charles University website, which unites disciplines from both fields).

In this study, we answer the research questions in two ways, that is, with respect to content (what is communicated) and structure (how or to whom it is communicated).

## Description of the methods and results of the analysis

### *Content analysis of the text*

Much of the content on the Internet is recorded in text. This text can take many forms, from digitised documents to news reports or blogs. The traditional method of text analysis in the social sciences is content analysis. Typically, content analysis is done manually by coders: they read the text and then code it. However, as the number of texts to be coded increases, manual content analysis is no longer feasible. For machine-readable texts, automatic processing options are available. Like other types of quantitative analysis, quantitative text analysis consists of two main steps: data pre-processing and then the analysis.

The file that is the basis for our content analysis always contains the domain name, the specific URL link, the keyword that appeared in that URL, and 10 words (context) around that keyword. The selection of keywords was preceded by qualitative pre-research on the websites of Czech research and educational institutions, the websites of the news media on which science is communicated, and specific science-related websites and blogs (a list of which can be provided upon request). The data (i.e. the textual content of the page, hyperlinks, and metadata) for quantitative processing were then downloaded using Heritrix3, which searched the specified institutional websites for the presence of these keywords. We searched the content of the institutional websites from the homepage up to a depth of 4 steps of the hyperlinks appearing on the given website; for the other websites, we focused on articles or content in any way related to science.

In the qualitative pre-research of the retrieved text, we focused on the most frequently occurring terms associated with science communication to a wider audience, including both neutral terms and c wordings of them specific to the natural/technical sciences or the social sciences/humanities. Our selected keywords included words related to the terms *vědec*, *věda* and *výzkum* (scientist, science, and research) and their synonyms, such as *odborník* or *expert* (specialist or expert), or the official designation of a scientist's position in a given institution (e.g., *doktor*, *docent*, *profesor* [doctor, associate professor, professor], etc.), as well as the desig-

nation of a given research institution or discipline, activity, or scientific result (see Table 2A in the Online Appendix for a complete list of keywords). Finally, given the topic of the article, we also included the word *popularizace* (popularisation).

In the pre-processing of the text data obtained with Heritrix3, we proceeded as follows. First, we identified duplicate texts and excluded them from the analyses. Next, we performed lemmatisation, which means we converted all the words into their base form.[4] For example, the word *vědy* (the declined form of the word 'science') was converted into its base form *věda* (science). If we had not done this, the most frequent expressions in each text would have been the different declined variants of the most frequent expressions – for example, *věda*, *vědy*, *vědě* (different declensions of the word 'science'), etc. We then removed the 'stopwords', a common practice in quantitative natural language processing[5] (Silge & Robinson, 2017). Stopwords typically include the most common words in the language ('all'), numerals ('ten'), prepositions ('without'), and conjunctions ('and', 'or'), which do not add much information to the text. As in the case of words before lemmatisation, stopwords would have again come up as the most frequent expressions had they not been removed. Since some texts were partly in English (typically titles of academic publications), we also decided to exclude stopwords in English. Finally, we removed all numbers, punctuation, words of fewer than 3 characters, words longer than 25 characters, and words containing special characters[6] or digits. This data cleaning procedure was used for the descriptive analysis of text data and for the topic models outlined below.[7]

*Results of the descriptive analysis of text data*

As part of the content analysis of the institutional websites, we first focused on the context in which the keywords *věda* (science) and *popularizace* (popularisation) occur, and the specifics of this context in the communication of the natural/technical sciences and the social sciences/humanities. The term věda[8] appeared in a total of 258,563 texts, while the term popularizace was significantly less frequent (n = 2778). In the first place, it should be noted that many of the terms co-

[4] We performed lemmatisation using UDPipe (Straka & Straková, 2019), specifically the udpipe package in R (Wijffels et al., 2021).

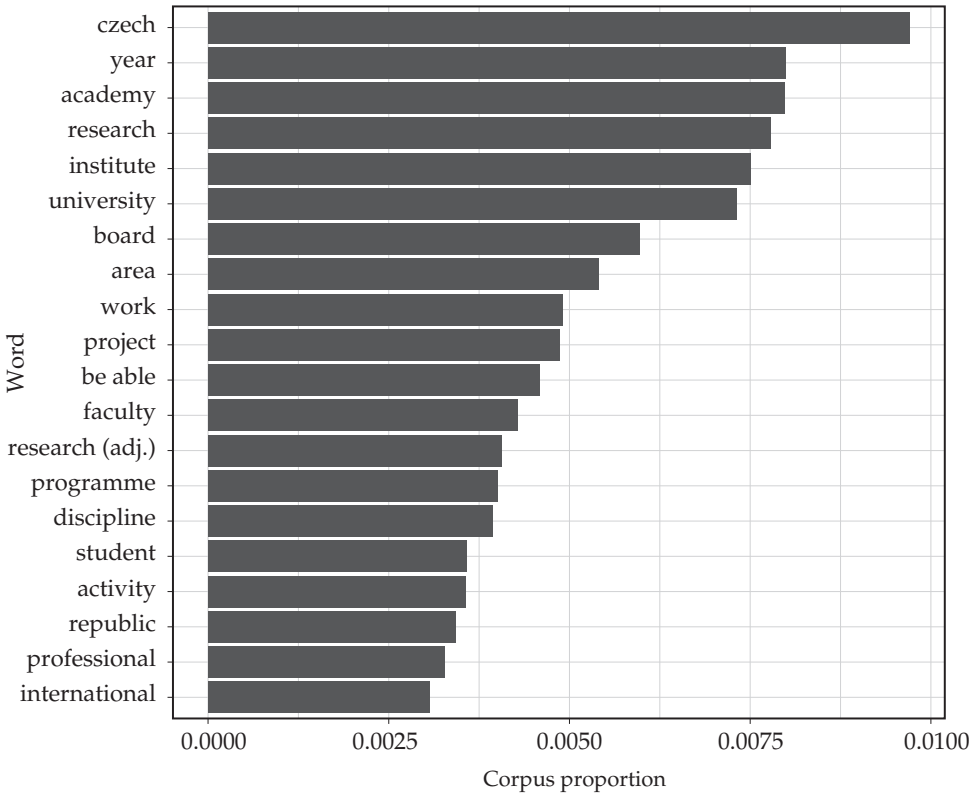
[5] For the sake of completeness, we should note that this is a common but not automatic practice. For more complex computational algorithms that take into account, for example, sentence structure, discarding these expressions may actually be inappropriate.

[6] Before this step, it was necessary to remove the diacritics from all the text, because letters with diacritics are not considered letters in regular expressions, but special characters.

[7] For the calculation scripts for the analysis see https://github.com/renatatopinkova/popsci.

[8] The set of related words contains all selected keywords from the list, see Table 2 in the Online Appendix. Similarly, it also contains all their word forms (the related words of the Czech word věda: *vědy*, *vědě*, *vědou*, *vědci*, *vědců*, apod.), since we use regular expressions and lemmatisation. For easier interpretation and readability of the text, we refer collectively to the keyword *věda* or ^*věd*.

**Figure 1. Co-occurrence of words with the keyword ^věd.* (^sci.*– science, scientific, scientist, etc. in all forms based on the root) – all websites of research institutions**

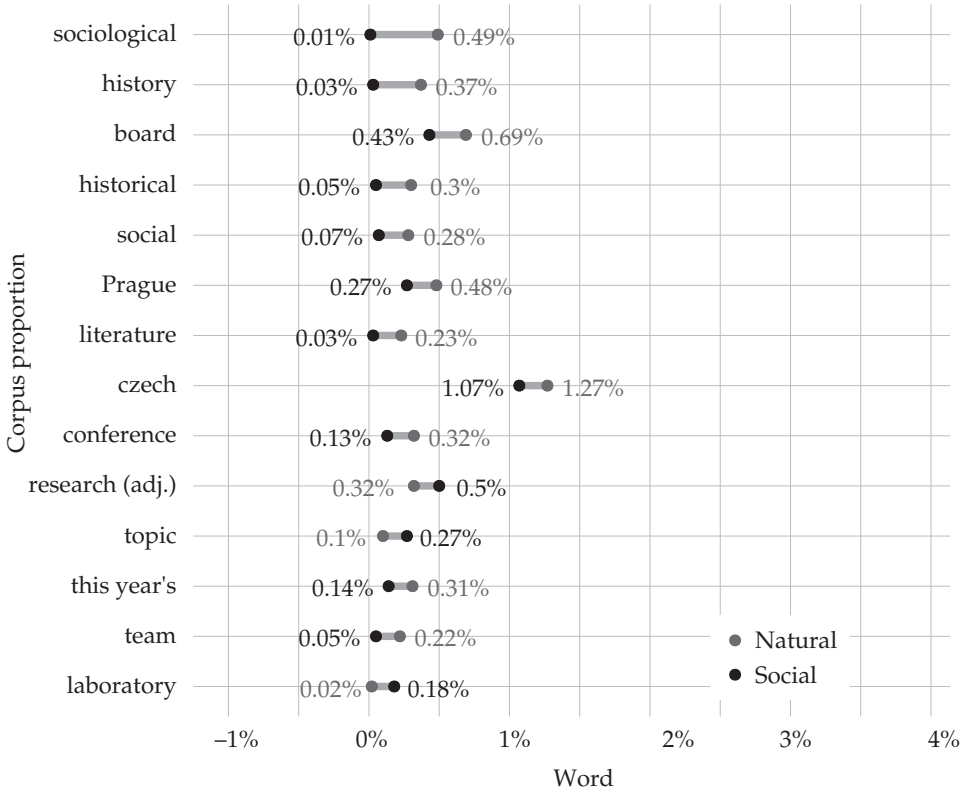



occurring with the keywords *věda* and *popularizace* overlap in the natural sciences and social sciences. In both cases, *věda* and *popularizace* are most often found in the context of česká věda (Czech science), akademie (academy), ústav (institute), univerzita (university), and výzkum (research) (see Figures 1 and 3).[9] It comes as no surprise that the term věda also appears most often in the context of other institutions such as fakulta (faculty), škola (school), and related activities and personnel – program (programme), studie (study), student (student), obor (discipline), program (programme), and projekt (project).

A more detailed overview is provided in Figure 2, which focuses specifically on the differences between the social sciences/humanities and the natural/technical websites. Figure 2 shows that the terms *sociologický* (sociological) and

[9] A colour version of the figures is available in the Online Appendix on the website of the *Czech Sociological Review*.

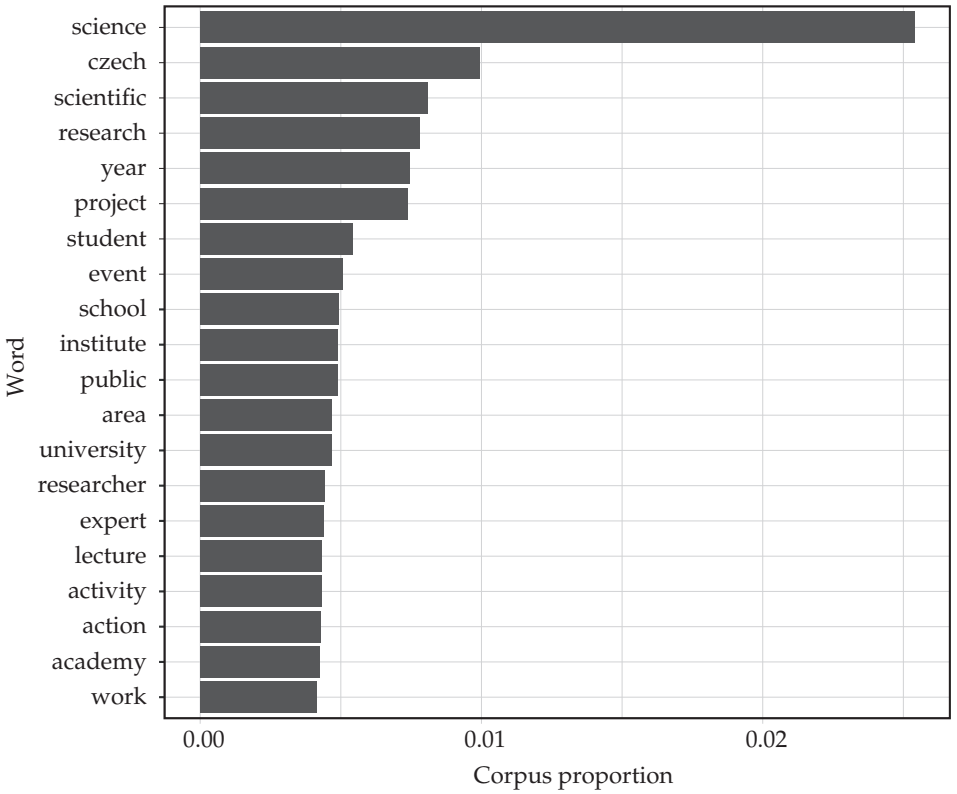
**Figure 2. Co-occurrence of words with the keyword ^věd.* (^sci.*– science, scientific, scientist, etc., in all forms based on the root) – differences between social and natural science websites**




*věda* (science) often appear together in the online content of the social sciences and humanities. While this may seem like an interesting finding, it is a data artefact. The website of the Institute of Sociology CAS is actually one of the largest websites we downloaded, and 78% of the texts containing the word *sociologický* come from this very website. If we were to exclude the website of the Institute of Sociology from the analysis, the term *sociologický* would not even be in the top 100 most frequent words that occur in combination with the word *věda*. Conversely, *dějiny* (history), *společnost* (society), *sociální* (social), and *historický* (historical) are words found in the context of science across social science and humanities websites. Other common terms include references to organising or attending events (*konference* [conference], *přednáška* [lecture]), as well as the words *časopis* [magazine], *literatura* [literature], and *umění* [art]).

In the case of the natural/technical science websites, the word 'science' is often used in connection with the words *tým* (team), *získat* (win), *cena* (prize), and

**Figure 3. Co-occurrence of words with the keyword ^populari.* (popularisation, popularise, popularising, etc.) – all websites of research institutions**

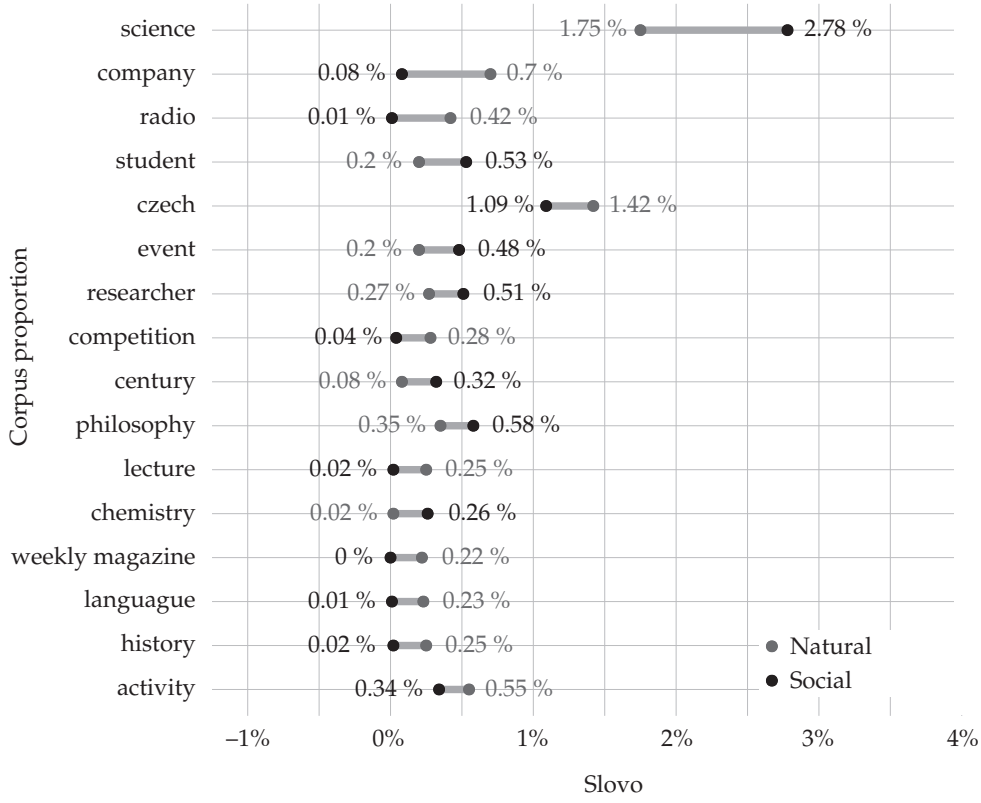



*mladý* (young), which seems to reflect the showcasing of scientific achievements. We also find words relating to the work of scientific and technical research (*metoda* [method], *laboratoř* [laboratory], *technika* [technique]). Unlike the social sciences, which more often associate 'science' with *společnost* (society), the natural sciences more often speak of *člověk* (human being).

For both scientific disciplines, we detected a strong co-occurrence of words with institutional content, such as academy, institute, conference, research, project, university, and study. Similarly, the words *spolupráce* (collaboration), *mezinárodní* (international), and *evropský* (European) appear on both social and natural science websites.

The co-occurrence of words with the term *popularizace* (popularisation) largely coincides with words related to the term *věda* (student, science, lecture, research, discipline, etc., see Figure 3). However, if we look at the differences between social and natural science websites, the situation is much more diverse (see Figure 4). In

**Figure 4. Co-occurrence of words with the keyword ^populari.* (popularisation, popularise, popularising, etc.) – differences between social and natural science websites**




the case of social science/humanities websites, the term *popularizace* is often associated with terms referring to specific outputs (journal, document, book, material), communication with the media (radio, weekly), and more general concepts such as century, history, philosophy, language, culture, and literature. Conversely, in the case of natural and technical science websites, the term *popularizace* is often associated with terms such as *věnovat* (donate), *cena* (prize), *ocenění* (award), and *propagace* (promotion), which usually refer to an award given to a specific scientist for popularising and promoting science. Although these words are also found on social science and humanities websites, this happens significantly less often. The words *soutěž* (competition), *podpořit* (support), *video* (video), and *film* (film) are significantly more common. Words associated with organising public events, such as *akce* (event), *přednáška* (lecture), *aktivita* (activity), *návštěvník* (visitor), and *veletrh* (fair), also appear more frequently. Another common word specific to the natural

sciences is 'Petr', a common Czech male name. However, this is not an error in the data. 'Petr' is the first name of a number of science popularisers, most notably the volcanologist Petr Brož and the biologist Jaroslav Petr.

*Topic models*

In classical quantitative analysis, there are a number of tests and models that can be used depending on the research question. This is also the case for text analysis – from descriptive analyses using word counting through dictionary methods (e.g., sentiment analysis) and searching for hidden themes in the text (e.g., topic models) to more complex methods using machine learning (e.g. deep learning). Given that we are interested in possible differences in the content of the communication of the natural and social sciences, we chose topic models (specifically by computing LDA – Latent Dirichlet Allocation). Topic models use the 'bag of words' approach. This approach does not take into account the order in which words occur within a given text (Blei, 2012). The bag of words approach is most often illustrated by imagining putting all the words from a given text – for example, this paragraph – into a box and then shaking the box. Among the classical sociological methods, topic models are closest to cluster analysis. However, as Bail (2014, 2015)[10] points out, in contrast to cluster analysis, each observation (document, text) does not have to be assigned to just one topic (cluster); instead, it is assigned the probability with which it belongs to each topic. Topic models also differ in the way they are calculated, where each observation is at the beginning assigned a random probability of belonging to each topic, with these probabilities being refined as the amount of processed data increases. It is therefore an iterative Bayesian technique. The results of the topic models then indicate two things: the words that are most often associated with a given topic, and the likelihood with which each document contains each topic (e.g., Blei, 2012; Silge & Robinson, 2017).

In our case, the contextual snippets around the keywords are the unit of observation. Each text (document) contains a keyword and a context of up to ten words before and after the keyword. At the same time, we set upper and lower limits on the occurrence of the words in the corpus, where each word was allowed to appear in no more than 60% of the documents and at the same time had to appear in at least 500 different documents. The resulting corpus contains a total of 1,676,629 documents and 4,813 words.

As in the case of some cluster analysis algorithms (e.g. k-means), the number of topics into which the algorithm should classify words and texts must be determined in advance, which is a decision that has a major impact on the results of the analyses. Ideally, the researcher should have specific theoretical expectations about the number of topics (a priori) that occur in the documents (Bail, 2014, 2015). However, since our work is exploratory, we do not have any suitable theoretically based expectations about the number of topics. Therefore, in the analyses we calculated models with different numbers of topics, namely 10, 20,

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

30, 40 and 50. We selected a final model based on qualitative validation, i.e. looking at the 10–15 words with the highest probability of being associated with each topic and reading a sample of texts from each topic. We then compared the models in terms of semantic coherence and exclusivity, which are metrics that focus on topic quality. Semantic coherence is maximised when the most likely words in a given topic often occur together. This is a metric that correlates well with human judgement about the quality of a topic (Mimno et al., 2011). Exclusivity then reaches higher values when more words are exclusive to the corresponding topics (Roberts et al., 2014). These two metrics are negatively correlated, so choosing a model based on them is a matter of compromise.

**Results of the analysis of topic models**

Figure 5 shows that a suitable number of topics would be around 20 to 30. For further interpretation, we decided to use the model with 20 topics, which was more interpretable than the model with 30 topics. Figure 6 shows the prevalence of the twenty topics that the algorithm found in the texts, and which words are most likely to be found in those topics.

The results show that the most frequent topic in our corpus is topic 6, represented by the terms *oblast* (area), *základní* (basic), *znalost* (knowledge), *obor* (field), *teoretický* (theoretical), *absolvent* (graduate), and *odborný* (expert). Words such as *materiál* (material), *proces* (process), *technologie* (technology), *chemický* (chemical), *vývoj* (development), *metoda* (method), and *struktura* (structure) appear in topic 13. This topic is most often found on technically or technologically oriented websites – such as those of the Central European Institute of Technology, the Institute of Physics, the Institute of Analytical Chemistry, the Institute of Mathematics, and the Institute of Rock Structure and Mechanics, and also on the website devoted to the Week of Science and Technology popularisation event and the website of the online magazine *iForum*.

Topics 16, 3, and 17, which are related to studying at a university and can be found on the websites of universities (Charles University, Masaryk University, Technical University of Liberec), are also well represented. In topic 5, the cultural terms *literatura* (literature), *dějiny* (history), *jazyk* (language), *umění* (art), *český* (Czech), and *historický* (historical) occur together. This topic can be found on humanities websites devoted to the study of languages and literature, such as the Institute of the Czech Language and the Institute of Czech Literature CAS, the Janáček Academy of Music and Performing Arts, the Institute of Philosophy CAS, and the Institute of History CAS. Similarly, topic 14 contains the terms *věda* (science), *český* (Czech), *akademie* (academy), *republika* (republic), *Praha* (Prague), and *společnost* (society), which are also more commonly found on humanities websites. Terms associated with public opinion appear together in topic 19: *veřejný* (public), *otázka* (question), *prostředí* (environment), and *šetření* (survey) and are found, unsurprisingly, on the websites of the Public Opinion Research Centre, the Institute of Sociology, and the Charles University Environment Centre.

Conversely, topic 15 (*systém* [system], *analýza* [analysis], *model* [model], *me-*

*toda* [method], *data* [data]) is specific to institutions in the natural and technical sciences, such as the Institute of Atmospheric Physics CAS, the Institute of Information Theory and Automation, and the Institute of Mathematics CAS.

Topic 8 contains the terms *projekt* (project), *výzkum* (research), *centrum* (centre), *rámec* (framework), *podpora* (support), and *vývoj* (development). It appears across various disciplines, including institutions in the natural sciences (e.g. the Institute of Physiology CAS) and institutions in the social sciences (e.g. CERGE-EI, the Charles University Environment Centre), but it is also found on the websites of grant agencies (the Czech Science Foundation and the Technology Agency of the Czech Republic). Particularly interesting is topic 7, which refers to the organising of events for the public (*přednáška* [lecture], *seminář* [seminar], *vědec* [scientist], and *veřejnost* [public]) and appears primarily on the websites of natural science institutions (the Institute of Experimental Botany CAS, the College of Polytechnics Jihlava) and the websites of popularisation events (Week of the Brain, Open Science), but is not found on the websites of any social sciences institutions.

**Figure 5. Comparison of the semantic coherence and semantic exclusivity models**

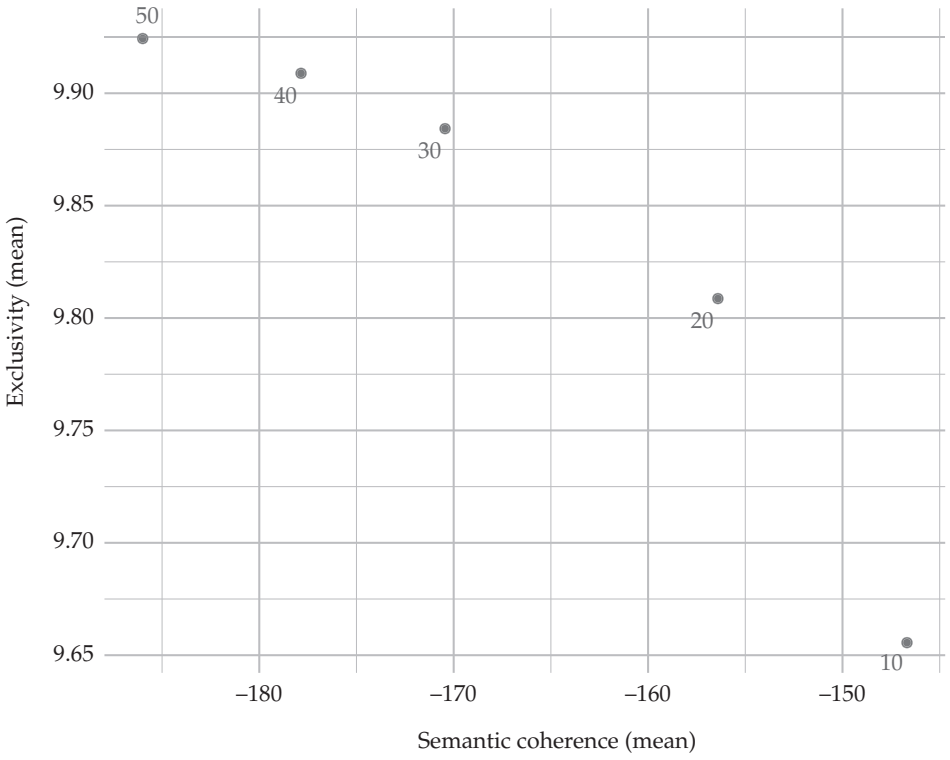
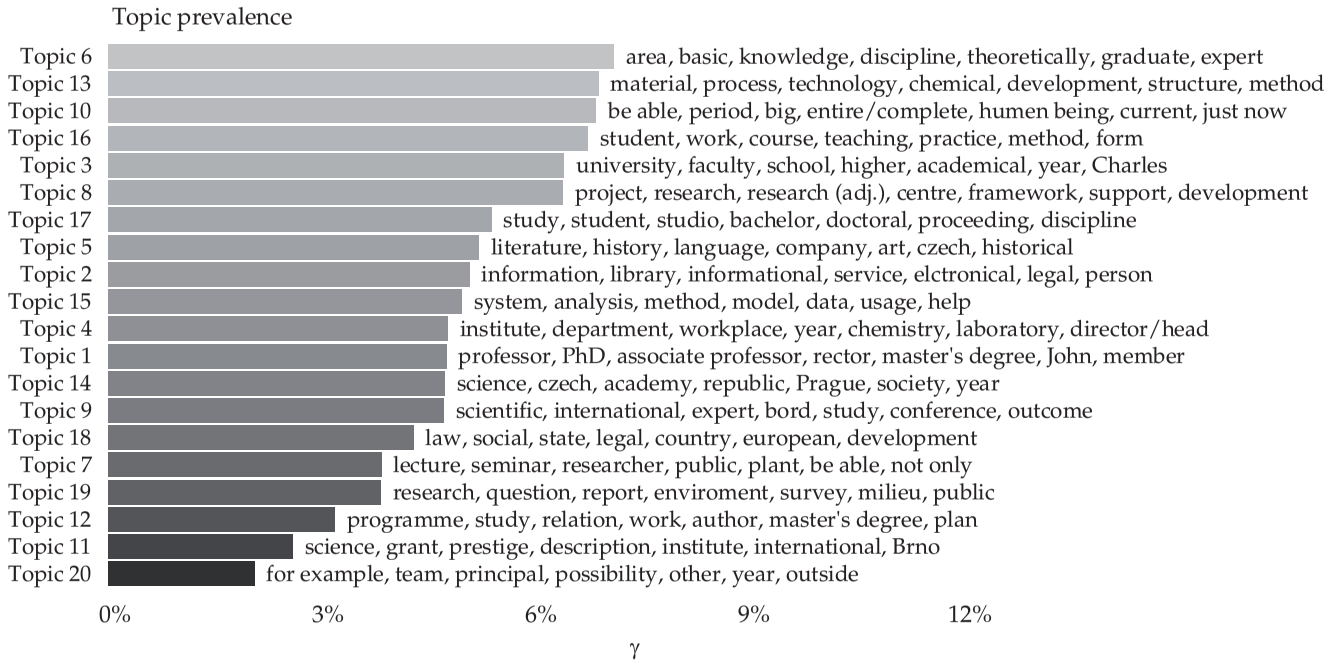


**Figure 6. Topic models: Prevalence of topics in the corpus**



*Analysis of multimedia content*

We were also interested in the sharing of audio or video content. The presence of audio or video content on websites is determined by the presence of the HTML tags iframe, amp-iframe, video or audio, or the obsolete tags <object> and <embed> on websites with older architecture.[11]

By far the most audio and video content was shared by websites explicitly aimed at the popularisation of science, such as the website of the Week of Science and Technology and that of the Czech Academy of Sciences. A comparison of the average proportion of shared video or audio content also shows that the websites of institutions in the natural and technical sciences share slightly more multimedia content (4%) than websites in the social sciences and humanities (2%). However, we must state here that the method we used to detect audio or video content is very crude and unreliable. Although an approach based on the HTML tags that embed these media seemed intuitive to us, in practice it did not work very well. It was prone to errors and made the error detection itself very difficult. To illustrate: if a website has a video in the sidebar that is displayed every time a person clicks on the website, our program will count this same video every time we access any URL from the website. Similarly, older websites may be built using an iframe or <object> tag regardless of whether they include any multimedia content.

*Analysis of hyperlinks*

A key characteristic of the Internet as a medium is its interconnectedness. This manifests itself in many ways, including the fact that individual pages often contain links to other pages. This interconnectedness is thus often depicted as a network in which nodes represent specific websites (or users, discussion forums, etc.) and edges represent the links between them (e.g. hyperlinks, interactions, likes) (Ackland, 2013; Robins, 2015). Social network analysis is then a possible way of describing and modelling the structure and topology of such networks.

Through a structural analysis of the hyperlinks of institutional websites, we attempted to answer the question of whether contemporary Czech science is still being popularised in a closed mode (i.e. whether the canonical model of online communication prevails) or is now being popularised in an open mode, in a way that is open to the public and open to discussion, and whether there is any difference between the natural sciences and the social sciences in terms of the openness and interactivity of their communication. In a closed mode of communication,

[11] It should be noted here that this is not a perfect approach for identifying video and audio content. For example, the 'object' tag turned out to be unusable on one of the websites, which was built entirely on javascript 'objects' that have nothing to do with video or audio content. The tags iframe and amp-iframe are also often used for other purposes for example, inserting interactive links to maps.

we would expect institutional websites to link primarily to the websites of other research institutions, or to the websites of state authorities or of the official databases of scientific journals, libraries, and publishers. On the other hand, in an open, interactive mode, links to social networks, news media, popular science websites, and blogs in the news media should predominate. Since our primary interest was in the differences in the science communication of the natural sciences and the social sciences, we are not focusing on specific institutional websites, but rather on which websites or which types of websites are linked to most by the majority of institutional websites within these groups. For this reason, we are only working with the websites of social science/humanities institutions (n = 27) and those of natural/technical science institutions (n = 47).

We performed a social network analysis in which nodes represented the websites and edges represented the links (hyperlinks) between them. We detected the presence of a link to another website using an anchor tag, i.e. we captured links from one website to another. For analysis purposes, we shortened all the URLs to their domain. This means that if our program visited the URL 'www.soc.cas.cz/projekty', we shortened the URL to 'soc.cas.cz', so that all clicks within a single website are represented by a single node. However, analysing the entire network would be very challenging, both in terms of interpretation and data volume. Therefore, we focused on the websites of individual institutions and the websites they directly link to by mapping their egocentric networks. At the centre of each network lies the original institutional website (ego), then there are links to other websites in the network at a distance = 1 (ego-alter ties), which are websites to which the institutional website links directly. Each egocentric network also contains alter-alter ties, i.e. if an institutional website links to website A and to website B, and website A also links to website B, the tie between A and B is included in the network.

The networks of social science and natural science institutions do not differ much structurally. The smallest social science network only has two links to websites other than itself, while the smallest natural science network has 38 links to a total of 8 external websites. In contrast, the largest social science network links to 452 unique domains (91,681 ties) and the largest natural science network links to 451 unique domains (91,981 ties). On average, the social science networks had fewer nodes and edges (171 versus 198 and 18,028 versus 20,374, respectively) and a higher average degree (276 versus 220) than the natural science networks.

**Table 1. Average statistics of egocentric networks by type of institution**

| Website type | Number | Avg. number of nodes | Avg. number of edges | Avg. degree |
|---|---|---|---|---|
| Natural sciences | 47 | 198 | 20 374 | 220 |
| Social sciences | 27 | 171 | 18 028 | 276 |

**Table 2. The 30 most frequent hyperlinks – excerpt from the full analysis (natural/technical science websites n = 6087, social science/humanities websites n = 3381)**

| | Social science/humanities webs | | | Natural/technical science webs | | |
|---|---|---|---|---|---|---|
| | Name | n | prop | Name | 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 |

Note: Prop = % of websites in the category that link to the given website. The first line shows that 81% of social science/humanities websites and 91% of natural/technical science websites link to facebook.com. The full table is available upon request.

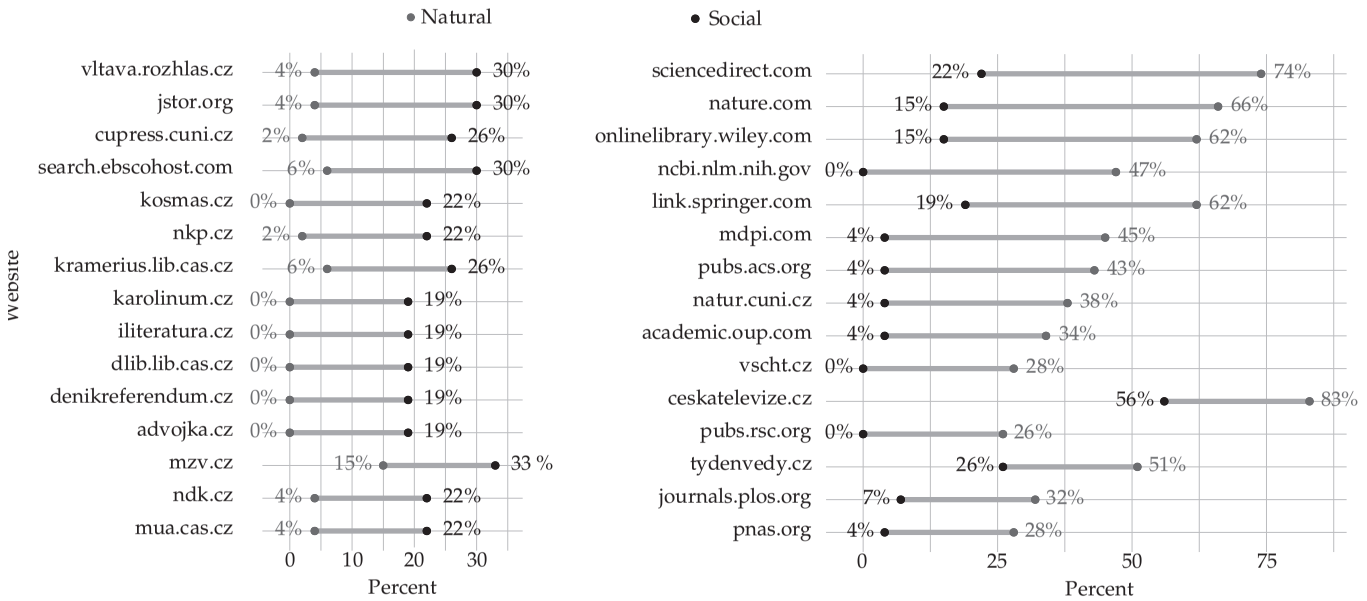
Results of the analysis of links

In total, we identified 3381 unique domains that were linked to by social science/humanities websites and 6087 unique domains that were linked to by natural/technical science websites. As part of our research questions, we were interested in which websites were linked to by the most institutional websites, as shown in Table 2. The first row in the table shows that 81% of the social science/humanities websites (22 in absolute numbers) and 91% of the natural/technical science websites link to the social media site facebook.com. Furthermore, 78% of the social science/humanities websites and 91% of the natural/technical science websites link to youtube.com. The third most linked website among the social science/humanities websites is twitter.com (59%) and among the natural/technical science websites it is ceskatelevize.cz (83%). The websites linked most by both the natural and the social sciences are social media or community forums (e.g. Facebook, LinkedIn), podcast/video platforms (e.g. YouTube, Spotify), and blogs/microblogs (e.g. Twitter). Almost half of the social science/humanities websites and the natural/technical science websites also contain links to instagram.com, a social media site for sharing images, photos, and videos, and around 40% link to wikipedia.org. There are also numerous hyperlinks to news media, such as ceskatelevize.cz, novinky.cz, rozhlas.cz, lidovky.cz, ihned.cz, etc., which are linked to by around 40% to 80% of scientific websites. The fact that the majority of institutional websites link to these websites signals the use of new technologies and media in the presentation and popularisation of Czech science, which could indicate forms of communication enabling interactive discussion, public input into the debate, and more direct contact between scientists or research institutions and the public.

Next, we focused on the websites that are the most overrepresented in the links of social science institutions compared to natural science institutions (see Figure 7). Figure 7 shows that links to some of the news sites are unique to social science institutions. While between 19% and 30% of all social science institutions linked to *Deník Referendum*, *A2,* and *Český rozhlas Vltava*, none or only a few of the natural science institutions linked to them. Some publishers, namely Kosmas and Karolinum Press, and databases such as EBSCO and JSTOR are specific to the social sciences. In contrast, the majority of natural science institutions (62–74%) shared links to the websites ScienceDirect, Nature, Wiley, and Springer (compared to 15–22% of social science institutions). Almost unique to natural science institutions were links to the websites of the University of Chemistry and Technology in Prague, the Faculty of Science of Charles University, the publishers MDPI and the Royal Society of Chemistry, and the journals *PNAS* and *PLoS*. It is also interesting to note that the website of, the Week of Science and Technology popularisation event was linked to by half (51%) of the natural science websites but by only a quarter (26%) of the social science websites.

In total, institutional websites linked to more than 10,000 unique websites (or domains). Since it was not possible to visit each of the websites and encode their type within the project, we focused on the websites that were linked to most frequently. Specifically, we encoded websites that were linked to by at least 10%

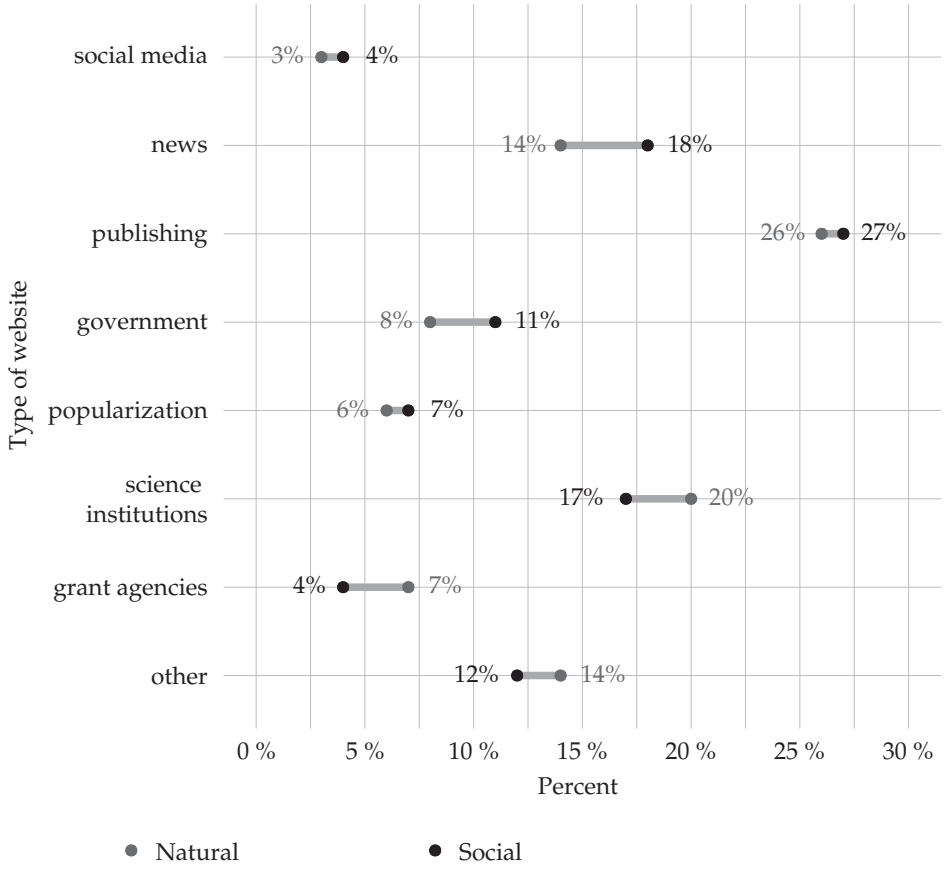
**Graf 7. Most overrepresented websites by type of institution**




Note: The figure shows the percentage of institutional websites in a given category that link to a particular website. On the left, links overrepresented for social science institutions, on the right, links overrepresented for natural research institutions.

of institutional social science websites (n = 242) or at least 10% of institutional natural science websites (n = 239). We distinguish the following 8 categories of websites: (1) social media, (2) news websites, (3) bookstores, publishers and libraries, (4) government/ministerial websites, (5) popularisation websites/journals/events, (6) other scientific institutions/universities, (7) grant agencies, and (8) others. Each of the authors independently visited each of the websites and encoded their type. During the initial coding, 77% agreement was reached for the websites linked to by social science institutions and a 71.3% agreement for the websites linked to by natural science institutions. Subsequently, websites on which the authors disagreed during the first coding were discussed and an agreement was reached for them as well.

**Figure 8. Structure of the 250 most frequent hyperlinks on social science/humanities and natural/technical science websites**



Although it may seem at first glance that institutional websites in both the social and natural sciences do not link much to social media (3% of all links shared), it is important to remember that there is a very limited number of major players in the field of social media and they cannot thus represent a larger proportion. The second most represented category consists of links to bookshops, publishers, and libraries (25% and 26%, respectively), a category that also includes links to academic publications. Social science websites link to more news (19% vs 14%) and government websites (13% vs 9%) than do natural science websites, while natural science websites link to more websites of other scientific institutions (20% vs 17%), grant agencies (6% vs 3%), and popularisation websites (8% vs 6%). The results are captured by Figure 8.

## Discussion and summary

The main focus of this study was the topic of science communication on the websites of Czech research institutions. We were interested in the nature and form of Czech science communication online and in finding out to what extent Czech science is shared with the public in the online environment and what differences exist between social science/humanities and natural/technical science websites in this regard. The results showed that the online content of research institutions does not deviate from the usual standards, where the terms 'science' and 'popularisation' are communicated mainly in the context of other educational institutions, studies, and research. In contrast to the social sciences, in the natural sciences popularisation is more often associated with rewarding scientists for the promotion and popularisation of science, which on the one hand may indicate greater incentives for popularisation in these fields. However, it could also be the other way around: if popularisation is largely absent, it can hardly be rewarded.

A structural analysis of the websites then gave us an insight into Czech science communication that uses the new social media. An analysis of the hyperlinks suggests that the vast majority of scientific websites share social media content, as they contain hyperlinks to Facebook, Twitter, YouTube, Instagram, and LinkedIn. Similarly, most websites of research institutions share links to news media such as ceskatelevize.cz, novinky.cz, lidovky.cz, and rozhlas.cz – i.e. online newspapers, television, and radio. This is also confirmed by the structure of the hyperlinks that were shared by the largest number of institutional websites in both the social sciences and the natural sciences – hyperlinks to news media are the second most common after hyperlinks to academic publications. The results thus suggest that Czech science communication online is moving towards an open and interactive mode of sharing science with the public.

Although the results summarised above have a certain informative value, it is also necessary to point out some weaknesses. The first distortion can already occur in the search for research institutional websites or in the insufficient cover-

age of websites during the researcher's selection. The same applies to the (subjective) choice of keywords, based on which a large amount of online content is then downloaded for further analysis. It must be said that our original intention was to conduct a content and structural analysis of all Internet resources where science is communicated (we identified approximately 166 of them). We then intended to analyse, on the basis of keywords, the number of posts devoted to the social or natural sciences in these sources (i.e. press releases, news items, short reports, popularising articles, coverage of science in the media, blogs on scientific topics, interviews with scientists, and appearances of scientists on TV and radio). However, this was not possible within our project due to the size of the data, technical complexity (the available technology or data processing), and time limitations.

Our study has a number of limitations. The first one is that we treat websites as if they were static, which is not necessarily true for every website. A result of this approach is that we may lose some of the information from dynamic websites that generate content based on user behaviour – for example, they retrieve additional information after a certain button is clicked, without the user leaving the current URL.

The second limitation arises from the processing of text data. Although we paid considerable attention to data cleaning, there are still partial duplications in the data. Some texts contained more than one keyword; for example, there are three keywords in the phrase 'popularisation of science in the field of biology' ('popularisation', 'science', and 'biology'), and we downloaded a context of 10 words around each keyword. This resulted in three texts that overlap to a large extent but are not completely identical. However, eliminating these partial duplications would be complicated in terms of computation (as well as in terms of time), as it would involve comparing all the texts with each other, calculating the probability with which they are partial duplications, and then merging these duplications. Given that our data file contains more than a million such texts, this step alone would take several days or weeks. Another improvement would be the inclusion of phrases (e.g. 'Czech Republic'), which we miss out on by using a model that only works with unigrams, i.e. one-word expressions. Furthermore, specific complications have arisen in the processing of text data – for example, the problems with the term *sociologický* (sociological) described in the descriptive analysis of the text data indicate that, for comparative purposes, it would be appropriate to weight the numbers by the size of the individual websites.[12]

The third limitation lies in the fact that we cannot determine the age of the downloaded URLs because most websites either do not include this information or it is generated dynamically when the website is opened. If our goal were to completely map the changes in the communication of Czech science on institutional websites or on the Czech Internet in general over a longer time horizon, it would require a longitudinal study of Internet content. For this purpose, in the

[12] We would like to thank the reviewer for these helpful notes.

future we could use a software tool of the web archive of the Czech National Library that is currently being developed, as it can be used to browse the content of the licensed Czech Internet back to the year 2000.

Another barrier was the impossibility of following scientific communication directly on social media, as, for example, Twitter allows the collection of texts on microblogs, but Facebook or Instagram do not. In this regard, it would be interesting to analyse the size of the audience of individual scientific posts, the number of likes, the size of the discussion, and the extent of further sharing. Some authors thus argue that even though research institutions have already accepted social media as a primary tool for communication with the public, they still use it more as a one-way communication channel, reporting on scientific knowledge and progress, but underestimating its potential for engaging in a dialogue and discussion with the public (Dudo & Besley, 2016; Lovejoy & Saxton, 2012).

In conclusion, we would like to point out that in spite of the mentioned shortcomings of the presented text, this is the first quantitative exploratory study that uses big data analysis to map the nature of Czech online science communication. We hope that the detailed methodological section, the critical evaluation of the study's limitations, and the outline of further possibilities for analysing the online content of s the websites of research institutions can serve other researchers as an introduction to working with big data in sociology and the pros and cons of this task.

Petra Raudenská *works as a senior researcher at the Institute of Sociology of the Czech Academy Sciences. Her research interests are survey methodology, social and educational inequality, human capital, life satisfaction, and human values. She is the author of the monograph* Comparability of Attitudinal Scales in Comparative Research *(in Czech, 2015) and published in* Social Science Research, Poetics, International Sociology, Innovation: The European Journal of Social Science Research.
ORCID: 0000-0002-0330-139X
Research ID: Q-2330-2016

Renáta Topinková *is a Ph.D. student in sociology at Charles University, and a researcher in Computational Social Science at LMU Munich. Prior to that, she worked at the Institute of Sociology of the Czech Academy of Sciences. Her dissertation examines the role of homophily in online dating. Her research interests include computational social science and sociology of the family. She has published in* Journal of Family Issues, Journal of Family Research, *and* Sociological Research Online.
ORCID: 0000-0003-0362-7290

## Reference


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 [Current approaches to the popularization of science in the Czech Republic]. *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., & 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., & 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., & Scheufele, A. (2013). Science, New Media, and the Public. *Science*, *339*(40), 40–41. https://doi.org/10.1126/science.1232329

Brossard, D., & 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., & Stocklmayer, S. M. (2003). Science Communication: a Contemporary Definition. *Public Understanding of Science*, *12*(2), 183–202.

Collins, K., Shiffman, D., & 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., & 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

Czech Statistical Office. (2010). *Kolik domácností v ČR má počítač a internet?* [How Many Households in the Czech Republic Have a Computer and the Internet?]. Czech Statistical Office. https://www.czso.cz/csu/czso/kolik_domacnosti_v_cr_ma_pocitac_a_internet

Čada, K., Červinková, A., Linková, M., Řeháčková, D., & Stöckelová, T. (2006). *Věda jako věc veřejná: vědní politiky a média* [Science as a Public Matter: Science Policies and the Media]. Institute of Sociology of the Czech Academy of Sciences.

Davies, S. R., & 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., & Horst, M. (2016). *Science Communication: Culture, Identity and Citizenship*. Springer.

Dijkstra, A. M., Roefs, M. M., & 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., & 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., & 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 [The way natural sciences are presented in selected Czech media in 2013]. *Mediální studia*, *12*(01), 115–132.
Jünger, J., & Fähnrich, B. (2020). Does Really No One Care? Analyzing the Public Engagement of Communication Acientists on Twitter. *New Media & Society*, *22*(3), 387–408. https://doi.org/10.1177/1461444819863413
Ke, Q., Ahn, Y. Y., & 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., & 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., & 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., & 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., & McCallum, A. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 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., & Dahlgaard, 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., & 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., & 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., & 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 Enginners*. 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

Rödder, S., Franzen, M., & Weingart, P. (eds.). 2012. *The Sciences' Media Connection–public Communication and Its Repercussions* (Vol. 28). London: Springer Science & Business Media.

Rybalko, S., & 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., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach* (First edition). O'Reilly.

Stocklmayer, S., Bryant, C., & Gore, M. M. (2002). *Science Communication in Theory and Practice*. Kluwer Academic Publishers.

Straka, M., & 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., & 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., & Vinopal, J. (2006). *Obraz vědy* v *českém veřejném mínění* [The Image of Science in Czech Public Opinion]. Sociologické studie 2006:8. Institute of Sociology of the Czech Academy of Sciences.

Treise, D., & Weigold, M. (2002). Advancing Science Communication: A Survey of Science Communicators. *Science Communication*, *23*(3), 310–322.

Trench, B., & 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., & 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., & 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