

# The Internet Never Forgets: A Four-Step Scraping Tutorial, Codebase, and Database for Longitudinal Organizational Website Data

Organizational Research Methods  
1–29

© The Author(s) 2024



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/10944281241284941

[journals.sagepub.com/home/orm](https://journals.sagepub.com/home/orm)**Richard F.J. Haans<sup>1</sup>**  and **Marc J. Mertens<sup>2</sup>** 

## Abstract

Websites represent a crucial avenue for organizations to reach customers, attract talent, and disseminate information to stakeholders. Despite their importance, strikingly little work in the domain of organization and management research has tapped into this source of longitudinal big data. In this paper, we highlight the unique nature and profound potential of longitudinal website data and present novel open-source code- and databases that make these data accessible. Specifically, our codebase offers a general-purpose setup, building on four central steps to scrape historical websites using the Wayback Machine. Our open-access *CompuCrawl* database was built using this four-step approach. It contains websites of North American firms in the Compustat database between 1996 and 2020—covering 11,277 firms with 86,303 firm/year observations and 1,617,675 webpages. We describe the coverage of our database and illustrate its use by applying word-embedding models to reveal the evolving meaning of the concept of “sustainability” over time. Finally, we outline several avenues for future research enabled by our step-by-step longitudinal web scraping approach and our *CompuCrawl* database.

## Keywords

websites, web scraping, Wayback Machine, textual data, Compustat

Websites are an essential tool for organizations to reach customers, attract talent, and disseminate information to stakeholders, such as investors and analysts (Botero et al., 2013; Powell et al., 2016; Santos, 2019). Since information disclosure on websites is voluntary and features more degrees of freedom than, for instance, mandated regulatory filings (Hoberg & Phillips, 2010), websites reveal facets of organizations that the organizations themselves deem salient (Powell et al.,

<sup>1</sup>Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

<sup>2</sup>Chair of Strategic and International Management, University of Mannheim, Mannheim, Germany

## Corresponding Author:

Marc J. Mertens, Chair of Strategic and International Management, University of Mannheim, Mannheim, Germany.

Email: [marc.mertens@uni-mannheim.de](mailto:marc.mertens@uni-mannheim.de)

2016; Trabelsi et al., 2008). This has led to the emergence of organizational websites as “a distinctive genre of collective identity,” serving as an important catalyst of trust and legitimacy in the eyes of stakeholders (Botero et al., 2013; Sillince & Brown, 2009, p. 1835).

Organizations’ websites are multimodal, organically created, and updated frequently, allowing for unique insights that other data sources, such as interviews, keynote speeches, and annual reports, cannot offer. For example, researchers have utilized websites to obtain highly detailed data on firms’ product offerings (Shermon & Moeen, 2022), gain unique insights into their recruitment practices (Stone et al., 2015), and access valuable user-generated data, such as ratings and reviews (Orlikowski & Scott, 2014). Moreover, even when information can also be obtained from other sources, “some data points are actually more valid if taken from a webpage than from other databases” (Powell et al., 2016, p. 117).

The pervasiveness of these “tangible touchpoint[s] between the organization and different audiences” (Santos, 2019, p. 240) has led to a growing interest in accessing, extracting, saving (“scraping”), and analyzing these data at scale. For example, organizational scholars have demonstrated that websites offer reliable insights into diverse organizational phenomena, such as their identity (Botero et al., 2013; Kroezen & Heugens, 2012; Powell et al., 2016; Sillince & Brown, 2009), strategy (e.g., Ebben & Johnson, 2005; Guzman & Li, 2023; Holstein et al., 2018; Jarvis et al., 2019), innovation (Kinne & Lenz, 2021), and networks (Oberg et al., 2009; Powell et al., 2017; Wruk et al., 2020).

Past research relying on website data has, however, faced substantial challenges. First, the sheer volume and variety of websites imply that manual data collection is frequently not an option. Therefore, researchers need to write complex computer code (“scrapers”) to retrieve the desired data in an automated fashion. Second, as websites are highly unstructured, researchers face the time-consuming task of reducing noise in these large and heterogeneous collections of website data. Last, servers only provide the most recent version of a website at the time of data collection. Hence, researchers must either resort to cross-sectional data or wait years to build longitudinal datasets (see Guzman & Li, 2023, for an exception). Given the distinctive characteristics and significant potential of website data, these challenges hinder the exploration of impactful questions in the domain of organizational research.

We aim to address these challenges by, first, describing a general-purpose, four-step approach to collecting longitudinal website data using our open-access codebase written in Python. Our approach makes use of the Wayback Machine developed by the Internet Archive, a nonprofit organization that aims to archive the entire internet. The Wayback Machine contains over 850 billion archived versions of webpages that were collected starting May 12, 1996. Our codebase not only offers streamlined and scalable access to this data source but also provides an end-to-end pipeline for conducting research using website data. Moreover, we provide open access to our *CompuCrawl* database, which was built using our four-step approach and our associated open-access codebase. The *CompuCrawl* database offers researchers immediate access to more than 1.6 million archived webpages between 1996 and 2020 for 11,277 publicly traded firms listed in Compustat North America. Last, we use this database to assess the quality and coverage of the Wayback Machine’s data and demonstrate one research use case of longitudinal website data.

Our work offers several contributions. First, although excellent reviews of web scraping approaches exist (e.g., Boegershausen et al., 2022; Edelman, 2012; Landers et al., 2016), these articles tend to focus on individual-level website data, scraped from a single or a few similarly structured websites. Building on the valuable insights gleaned from these approaches, we discuss the unique challenges that emerge when scraping numerous organizational websites’ unstructured and heterogeneous data. A second contribution emerges from our focus on scraping organizational websites *over time*. Prior work has highlighted the opportunities for scraping longitudinal data and has explicitly suggested using the Wayback Machine (Boegershausen et al., 2022; Landers et al., 2016).

However, we are first to discuss the unique considerations of scraping longitudinal website data and to evaluate the suitability of the Wayback Machine's archive for organizational research. Third, our codebase enables researchers to scrape historical versions of any website. We hope this will enable organizational research to generate and share rich website-based data, as exemplified by our open-access *CompuCrawl* database. As a result, we answer calls for more work that develops and describes datasets for others to use (Ethiraj et al., 2017, 2019). Last, by discussing ways in which future research can utilize longitudinal organizational website data, we augment the research agenda underpinning the "linguistic turn" and the "visual turn" in management research (Boxenbaum et al., 2018; Vaara & Fritsch, 2022).

In the following, we discuss the four key steps underpinning research projects using longitudinal organizational website data, highlighting common considerations and best practices that emerge from our review of prior work. The four steps (sample construction, data collection, data (pre-)processing and cleaning, data description and analysis) are summarized in Table 1. Researchers who want to apply our recommended scraping methods to their own lists of focal websites can download our Python codebase and in-depth documentation of the code at <https://haans-mertens.github.io/code>. Throughout this review, we illustrate the specifics of our approach via an in-depth illustration of the construction of our *CompuCrawl* database. Researchers interested in this ready-made database of longitudinal website data of firms listed in Compustat North America can download our *CompuCrawl* database at <https://haans-mertens.github.io/data>.

## Step 0: Foundations

### Research Question Definition

Before commencing the actual longitudinal web scraping, researchers should first focus on two foundational prerequisites: defining their research question and conducting legal due diligence. For one, a research project's focal research question informs all subsequent web-scraping decisions (Landers et al., 2016). We, therefore, consider clarity on a project's goals and their fit with the focal data source essential to making reasoned methodological choices and ensuring the accuracy and relevance of the collected data. Specifically, researchers are advised to iteratively develop a data source theory, which summarizes their "understanding of what the data source is, provide[s] context for data contained therein, [...] and reveals] whether the information contained in the source can be used to test the research questions" (Landers et al., 2016, p. 483). Moreover, having defined a concrete research question and associated data source theory *a priori* supports researchers in defining suitable scraping parameters, identifying applicable benchmarks against which to compare the coverage of the scraped data, as well as conducting adequate data (pre-)processing, cleaning, and analysis. Therefore, this foundational step enables researchers to identify potential threats to their projects' internal and external validity early on and adjust their scraping decisions accordingly (Boegershausen et al., 2022).

It is worth noting that the specific degree of concreteness of this prerequisite step and the scope of the resulting data collection efforts will depend on the chosen research approach. Hence, projects that employ hypothetico-deductive research akin to "theory-driven web scraping" advocated for by Landers et al. (2016, p. 477) will generally achieve higher initial clarity, allowing for more targeted scraping. Inductive research should similarly start with an *a priori* definition of the focal research question and assess its fit with the data source theory. However, we acknowledge that such endeavors are generally associated with lower initial goal clarity and, thus, may necessitate broader scraping efforts.

In addition to the general aim of illustrating a general approach to working with longitudinal website data, our own scraping project had two goals. First, we sought to create a widely applicable database of longitudinal organizational website data to support future research. Second, our empirical

**Table 1.** Step-By-Step Tutorial for Longitudinal Website Scraping.

General Steps	CompuCrawl Database
<b>Step 0: Foundations</b>	
<ul style="list-style-type: none"> <li>Define research question and associated data source theory.</li> <li>Conduct legal due diligence.</li> </ul>	<ul style="list-style-type: none"> <li>Assessed the suitability of firms' longitudinal website data to test if the meaning of concepts shifts over time.</li> <li>Ensured compliance with local regulations and the Wayback Machine's terms of use.</li> </ul>
<b>Step 1: Sample Construction</b>	
<ul style="list-style-type: none"> <li>Gain access to the website address(es) of the focal organization(s).</li> <li>Clean the website addresses, for example, by enforcing consistent formatting and removing duplicates.</li> </ul>	<ul style="list-style-type: none"> <li>Compustat North America provided data on 17,238 firms with a website.</li> <li>13,575 website addresses remained after automated and manual cleaning that enforced consistent formatting and removed duplicates.</li> </ul>
<b>Step 2: Data Collection</b>	
<ul style="list-style-type: none"> <li>Select the timeframe and collection frequency.</li> <li>Request and scrape archived website versions from the Wayback Machine.</li> <li>Assess coverage over time.</li> <li>Parse the scraped websites for links to subpages.</li> <li>Scrape the subpages.</li> </ul>	<ul style="list-style-type: none"> <li>Time frame: 1996 to 2020, scraping one historic website version for each year.</li> <li>Requested 169,578 archived website versions from the Wayback Machine, successfully scraping 110,379.</li> <li>Confirmed unsuccessful scrapes were the result of websites not yet existing or not related to organization-level factors.</li> <li>Parsed the scraped websites, identifying 2,591,743 subpages.</li> <li>Successfully scraped 2,051,884 subpages.</li> </ul>
<b>Step 3: Data (Pre-)Processing and Cleaning</b>	
<ul style="list-style-type: none"> <li>Convert HTML website files.</li> <li>Exclude invalid or irrelevant webpages.</li> <li>Remove websites not in the desired language.</li> <li>Remove invalid/irrelevant content.</li> <li>Remove excessively short texts.</li> <li>Further process texts based on the needs of the application.</li> </ul>	<ul style="list-style-type: none"> <li>Converted all 2,162,263 HTML website files to plain text.</li> <li>Removed 1,621 frontpages and 468 subpages that were wholly invalid. Additionally used GPT3.5 to classify page structure, removing 255,451 webpages that were duplicates of the frontpage, contained only legal information, technical resources and documentation, or regarded site functionality.</li> <li>Language identification algorithm removed 212,079 non-English webpages.</li> <li>Removed content based on 5,818 (sub-)sentences.</li> <li>Removed 74,969 webpages with 10 or fewer words after cleaning, leaving 1,617,675 webpages and 636,318,656 words.</li> <li>Aggregated frontpage and subpage texts into 86,303 firm-year observations and removed stop words and highly infrequent words for topic modeling.</li> </ul>
<b>Step 4: Data Description and Analysis</b>	
<ul style="list-style-type: none"> <li>Describe and summarize resulting textual data.</li> <li>Use data in project-specific analyses.</li> </ul>	<ul style="list-style-type: none"> <li>Utilized a 125-topic topic model to summarize the contents of the collected website texts.</li> <li>Word-embedding model was used to analyze the change in meaning of the concepts of "sustainability" and "profitability."</li> </ul>

application that utilizes the resulting *CompuCrawl* database explores the research question of whether the meaning of concepts relevant to organizational research evolves over time. We, thus, assessed the suitability of organizations' longitudinal website data to answer the focal research question. This process was informed by Landers et al.'s (2016, p. 484) guiding questions for the development of a data source theory. It, for example, included validating the potential of firm websites to capture relevant shifts in organizational discourse and evaluating the breadth and depth of the available website data to ensure that they sufficiently capture the concepts of interest. Based on these considerations, we concluded that texts from organizations' longitudinal websites archived by the Wayback Machine indeed allow for reliable inferences on whether the meaning of relevant organizational concepts evolves over time.

### ***Legal Due Diligence***

Conducting legal due diligence is the second prerequisite to longitudinal web scraping. Specifically, web scraping legislations differ between jurisdictions and change over time. Therefore, we urge researchers to always assess the legal framework they must abide by and consider ethical norms at the start of any scraping project (see Dykstra et al., 2014, for an example). There are several excellent overviews of common legal and ethical requirements for scraping website data (Braun et al., 2018; Kobayashi et al., 2018; Landers et al., 2016). In general, researchers tend to meet these requirements when they abide by websites' terms of use and general fair use principles (Black, 2016; Dreyer & Stockton, 2013).

As such, we limit our discussion to the specifics pertaining to scraping the Wayback Machine: Whereas researchers typically encounter different terms of use for every website they scrape, one must only abide by the Wayback Machine's terms when scraping its archive. Specifically, the Wayback Machine limits the use of its data to scholarship and research purposes and places few other restrictions on users (Internet Archive, 2022). More fundamentally, the U.S. Copyright Act (17. U.S.C § 108) provides special exemptions for archives, such as the Wayback Machine, from its otherwise stringent regulations. Similarly, the E.U.'s Digital Single Market Directive demands that member states implement exemptions for text and data mining conducted for scientific research (The European Parliament & The Council of the European Union, 2019). Therefore, while websites' terms of use and legal restrictions warrant attention during the inception of any scraping project, researchers have significant opportunities to utilize website data retrieved from the Wayback Machine.

## **Step 1: Sample Construction**

### ***Access Website Addresses***

Following the formulation of the research question, the assessment of its fit with the data source theory, and legal due diligence, researchers must obtain focal organizations' website addresses. For example, commercial databases like Orbis or public repositories provided by Chambers of Commerce provide access to website addresses at scale. The articles in our review utilized a wide range of sources—consistent with the broad potential of website data. For instance, Botero, Thomas, and Graves (2013) obtained lists of family firms and their website addresses from family business organizations in the United States, the United Kingdom, and Australia. Guzman and Li (2023) leveraged the Crunchbase database to access the website addresses of all companies available in that database between 2003 and 2019. Haans (2019) identified organizations' websites through a database of the Dutch Chamber of Commerce. In general, when selecting their source of website addresses, researchers should be aware of “its properties and limits” and the implications that

these will have on the validity of the resulting data (Boyd & Crawford, 2012, p. 668; Landers et al., 2016). Considerations regarding databases' survivorship bias and whether old website addresses are overwritten in the case of website address changes are especially warranted.

For our list of websites to feed into the Wayback Machine, we used Standard and Poor's Compustat North America database. Compustat contains website information for thousands of firms and is a reputable, widely used dataset for firm financials in management research. Indeed, we find that "Compustat" occurs at least once in 29.8% (325) of the 1,090 articles published in the *Strategic Management Journal* and in 25.6% (190) of the 742 articles in the *Academy of Management Journal* between 2011 and 2020. This widespread use highlights that obtaining textual data of the firms listed in Compustat allows for substantial synergies with prior and ongoing research. Moreover, we can expect the websites to contain textual data that cover salient organizational constructs and their change in meaning over time.

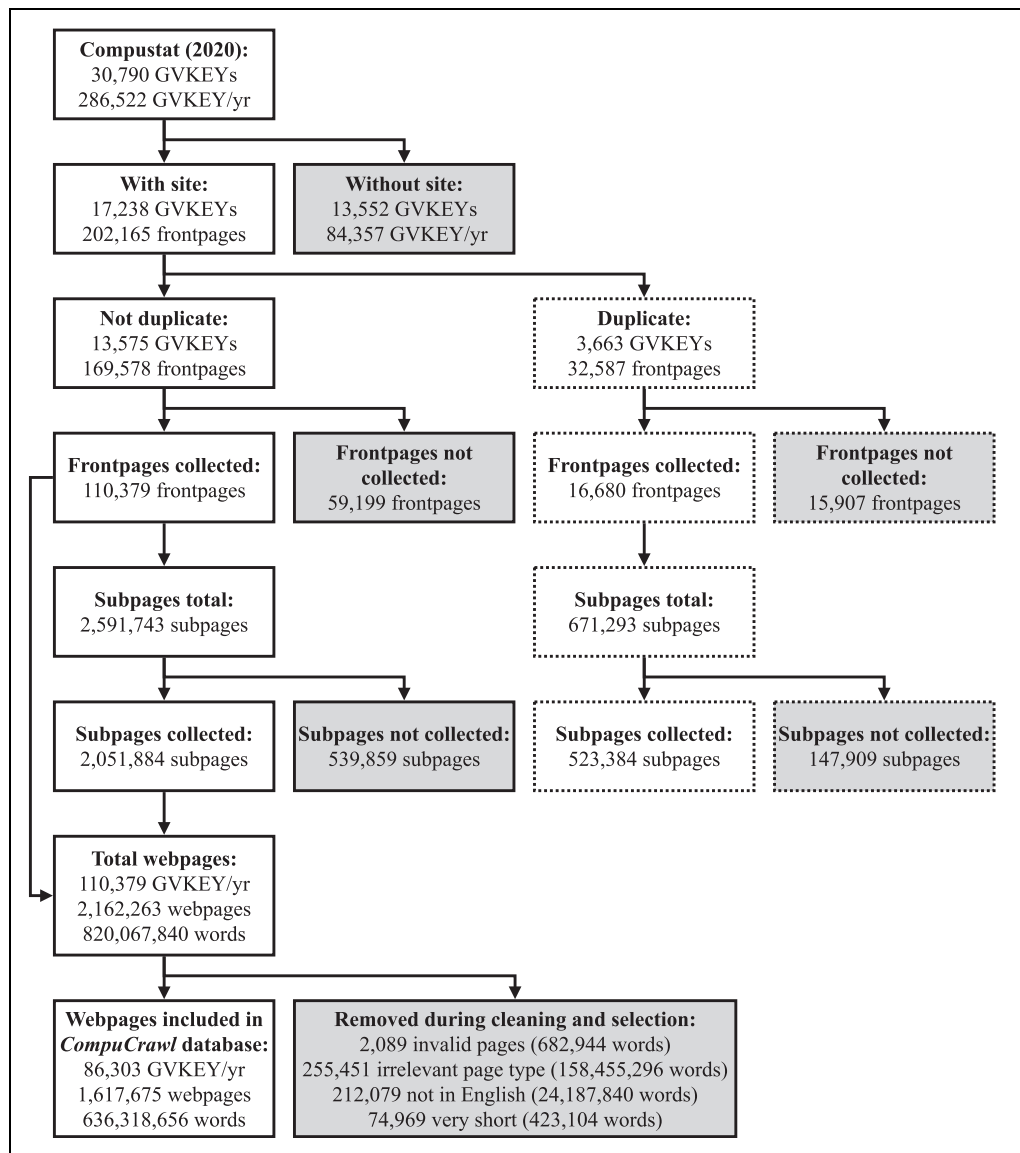
We downloaded the Compustat database in the second half of 2021 to ensure that the information for 2020 was complete and up-to-date. In total, there were 30,790 firms in the Compustat database between 1996 and 2020, each with a unique identifier called "GVKEY." Out of the 286,622 GVKEY/year observations, 84,357 GVKEY/year observations (13,552 GVKEYs) did not contain any website information (see Figure 1).<sup>1</sup> The remaining 202,165 GVKEY/year observations (17,238 GVKEYs) had a website address listed. We reduced the extent of survivorship bias in our sample by also retrieving data from Compustat on firms that have gone out of business. However, Standard and Poor's is known to add and back-fill data of previously successful firms in Compustat (Ball & Watts, 1979; Kothari et al., 1995). This source of survivorship bias is likely still present in our sample of website addresses. Moreover, it is worth noting that the website addresses provided by Compustat were the most recent ones available in the database. Hence, if a firm previously changed its website address, the prior website address was not retained in Compustat, as it was overwritten with the latest address. This implies a potential loss of website addresses, particularly for the initial years of our database. Nevertheless, we assume that such domain switching is relatively rare given the importance of consistency to maintain traffic to one's website.

### Clean Website Addresses

Any data collection effort is only as good as the quality of its inputs. Therefore, researchers should format website addresses consistently and remove duplicate or inactive addresses prior to scraping. In our literature review, examples of this include Kotha et al. (2001) who manually checked each website to determine if it was operational and open for customer interactions as well as Haans (2019) who used an automated approach to validate that websites were active.

We, first, recommend enforcing consistent formatting of the website addresses to ensure that the Wayback Machine correctly identifies archived website versions. Specifically, we advise researchers to convert website addresses to lowercase and strip them of prefixes like "http://" as well as trailing backslashes. Then, website addresses can be uniformly reconstructed according to the following pattern: "https://www. + address." This website address cleaning for best interoperability with the Wayback Machine is conducted automatically as part of our codebase.

An additional benefit of this approach is that duplicate website addresses can be reliably identified. For example, while engaging in this cleaning in preparation for our own scraping for the *CompuCrawl* database, numerous GVKEYs with the same website addresses became apparent. Collecting these websites yields repeating observations for otherwise unique GVKEYs. We advise researchers to not simply drop observations with duplicate website addresses but to explore if patterns in the data explain these duplicates. Clear decision rules should then be formulated on which observations are kept and why. Specifically, we noticed that duplicate entries were highly concentrated; a small number of firms accounted for a disproportionately large percentage of duplicates. Moreover, many of these entries



**Figure 1.** Flowchart of sampling steps.

were associated with financial service firms with separate GVKEY entries for the exchange-traded funds they operate. For example, the website address “www.invesco.com” was linked to 243 unique GVKEYs, but just one GVKEY was associated with the investment management firm Invesco Ltd. The remaining 242 GVKEYs referred to, for example, the Invesco DWA SmallCap Momentum ETF and the Invesco S&P 500 Downside Hedged ETF. Therefore, we manually assessed all duplicate observations to identify the primary GVKEY associated with each website address. If we did not identify a primary GVKEY, we classified all as duplicates. There were 3,663 GVKEYs (32,587 GVKEY/year observations) that became redundant in this way (see Figure 1). We excluded these duplicates from our main database and analyses, leaving us with cleaned, unique website addresses for 13,575 GVKEYs.

## Step 2: Data Collection

### *Determine Scraping Parameters*

Once the addresses of the focal websites have been cleaned, researchers must determine for what period to collect the data, which is a unique consideration for scraping archival organizational websites. As noted earlier, obtaining longitudinal data remains a challenge for researchers interested in leveraging website data. Indeed, nearly all articles we identified relied on cross-sectional data obtained only at the most recent point in time, with only a few exceptions: Powell et al. (2016) manually collected website data for over a decade, while Guzman and Li (2023) leveraged the Wayback Machine to collect a version of organizations' websites in the year after organizations' founding. As discussed previously (see Step 0: Foundations), the focal research question and data source theory inform the entire web scraping project pipeline (Boegershausen et al., 2022; Landers et al., 2016). Hence, we, for example, recommend explicitly considering how long the focal website content existed and how regularly it was generally updated in selecting suitable scraping parameters.

To illustrate our approach's full potential and observe concepts' shift in meaning over an extended period, we chose the timeframe of 1996 up to and including 2020. The Wayback Machine started archiving websites in 1996, and the most recent, complete Compustat dataset was from 2020. This yielded 339,375 potential GVKEY/year observations (25 years multiplied by 13,575 GVKEYs with a unique, cleaned website address). However, we only scraped the Wayback Machine's archive starting with the year when the firm was first present in Compustat up to and including the last year it was listed. For example, Shutterstock Inc. ([www.shutterstock.com](http://www.shutterstock.com)) was present in Compustat from 2010 onwards, and our scraper requested an archived page only for its active years. We, thereby, reduced the total number of potential GVKEY/year observations to 169,578. Such parsimonious scraping is a best practice, as it conserves researchers' resources and limits the strain on the servers that are scraped (Landers et al., 2016).

Moreover, researchers must determine for which point in time the archived website version is requested from the Wayback Machine. The requested timestamp can be flexibly adjusted in our codebase, for instance, setting it to the founding dates of firms (see Guzman & Li, 2023) or major exogenous events. Since the scraper collects the closest archived website version to the specified timestamp, we advise researchers to set the timestamp to the midway point of the period of interest. For example, researchers interested in daily website versions should set the requested timestamp to noon of the respective days. We opted for yearly intervals as the resulting data for our *CompuCrawl* database promise to accommodate many researchers' needs. Moreover, we expected the temporal shifts in organizational concepts' meaning to occur relatively slowly over time. For such yearly scraping, July 2—the midway point of the year—should be selected.

### *Scrape Frontpages*

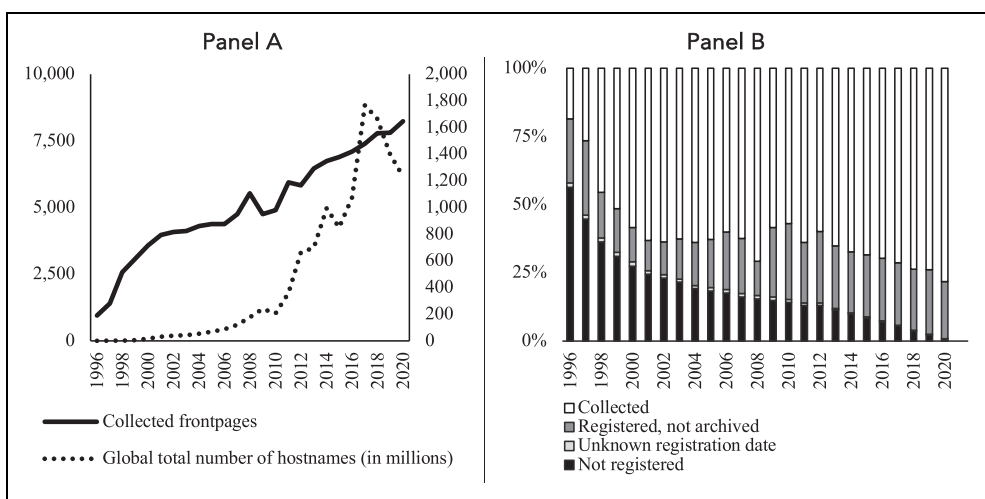
Once these parameters are decided upon, the next step is to download ("scrape") the websites associated with the focal website addresses. Although some authors did not use automated scraping (commonly working with small sets of websites, e.g., Bertels et al., 2014; Borah et al., 2021; Botero et al., 2013; Ebben & Johnson, 2005; Hales et al., 2021), we find that most large-scale applications leverage code-based approaches to automatically collect the focal websites. In our application, we accessed the Wayback Machine to download the available archived websites during the specified period. The main/landing pages of these websites will be referred to as "frontpages." This initial phase yielded a total of 110,379 frontpages, implying that 59,199 frontpages were not collected (see Figure 1).



## Describing Coverage

It is essential to assess the coverage of scraped website data to draw justified conclusions from these data (Landers et al., 2016). However, whereas assessing if a page is available at the time of scraping is relatively straightforward for cross-sectional scraping projects, scraping organizational website data over time introduces specific challenges. We, therefore, encourage researchers using archival website data to diligently assess and transparently describe the coverage of their data. Specifically, we recommend benchmarking the data coverage against the total number of unique domains on the entire World Wide Web and the data coverage of relevant comparable databases. Moreover, a rigorous assessment of data coverage should include an explicit distinction between missingness due to websites not existing yet compared to websites not having been archived by the Wayback Machine. Finally, researchers should probe for systematic reasons that specific websites went unarchived.

We start the comparative assessment of the coverage of our longitudinal web scraping using Figure 2. Panel A of Figure 2 shows the number of frontpages that were collected for our database over time together with the total number of unique domains on the entire World Wide Web (Netcraft, 2021). Panel B of the same figure shows rates of (un)successful scraping by year. Cases of missing websites are broken down into two primary reasons for missingness: On the one hand, the website may not have existed yet in the focal year even if the organization already existed. On the other hand, the website may not have been archived in the focal year even though it existed. We advise researchers to distinguish between these two reasons for missingness by visiting each domain's "who.is" page. "who.is" provides domain registration information, including the timestamp of when a domain was registered, which we retrieved for each domain in our sample. In Panel B of Figure 2, "Not registered" comprises missing observations for which the requested website versions predate their respective registration dates. Observations for which "who.is" shows that the websites were registered yet the Wayback Machine did not archive them are categorized as "Registered, not archived." The few websites without registration information on "who.is" are listed as "Unknown registration date" in Panel B.



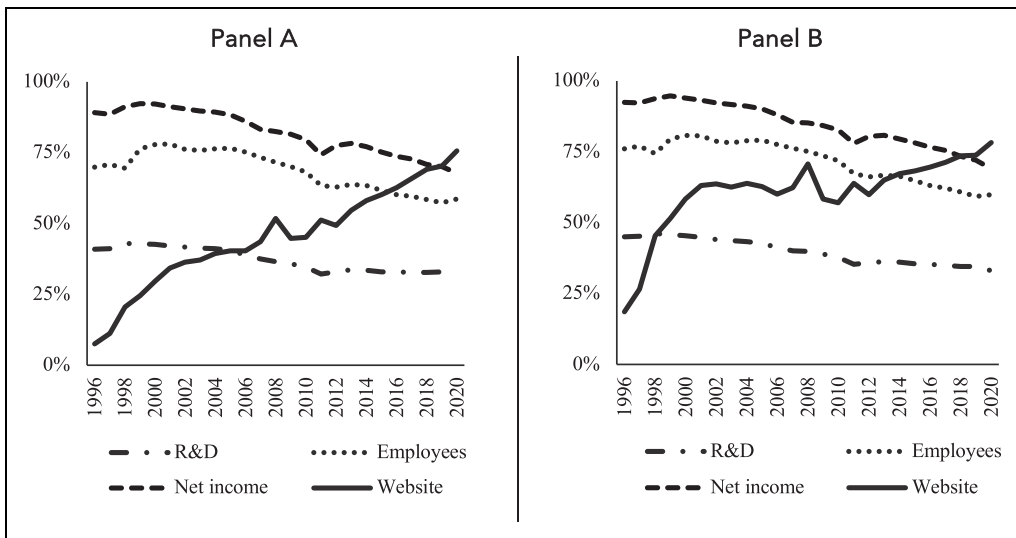
**Figure 2.** Data coverage over time.

Note: Panel A shows the total number of collected frontpages and unique domains on the entire World Wide Web in July of each year. Panel B shows yearly rates of (un)successful data collection.

Jointly, the two panels in Figure 2 show that the raw number of collected webpages has grown rapidly and consistently in more recent years. Thus, there are more firm websites available to be successfully scraped as time progresses—consistent with the tremendous growth of the World Wide Web over time. Accordingly, the rate of missing observations due to websites not existing yet represents a large share of all scraping attempts in initial years and rapidly diminishes over time. It is also evident that the ratio of registered websites that were not archived by the Wayback Machine remains fairly consistent over time, hovering at around twenty percent. Hence, there appears to be a random chance of about twenty percent that the Wayback Machine does not archive a website in a given year.

We advise researchers to statistically validate such inference by probing for systematic reasons that specific websites went unarchived, which would be problematic. We, therefore, estimated conditional logistic regressions predicting that a registered website went unarchived. The number of employees, total revenues, net income, and the return on equity of the associated firm served as explanatory variables.<sup>2</sup> Each of these variables is representative of firm visibility and/or success, and less visible/successful firms' websites might have a lower chance of being archived. None of the four explanatory variables (separately or jointly in a single model, with  $p$ -values ranging from .135 to .790) explain the probability that a registered website went unarchived. In all, this suggests that the pattern of registered websites that were not archived by the Wayback Machine is indeed seemingly random.

A logical next question that researchers should also explore is how their archival website data compare to alternative longitudinal data sources in terms of coverage. Therefore, Figure 3 compares the percentage of successfully scraped websites in each year to the percentage of available observations for several commonly used variables from the Compustat database. Panel A covers the entire database—including observations without a website address—while Panel B focuses on the subsample of observations with a website available in Compustat. Both panels show that website data coverage was initially limited, which was primarily due to firms' websites not existing yet, as described above. However, absolute as well as relative coverage compared to commonly used variables from



**Figure 3.** Comparison of coverage over time.

Note: The graphs show the percentage of observations in the Compustat database that contained an observation for the respective variables. "Website" indicates a successful scrape of the Wayback Machine. Panel A shows coverage for all observations in the database. Panel B shows coverage for those with a website in Compustat.

Compustat significantly improved as time progressed. In recent years, the coverage of website data even surpassed that of net income, the most widely available comparative variable under consideration.

### *Scrape Subpages*

Having collected the archived frontpages of websites, a common way to enrich the data is to go further down the page hierarchy by collecting so-called “subpages” (for instance: <https://www.website.com/about>). For example, Kinne and Lenz (2021) and Guzman and Li (2023) scraped frontpages and up to 25 and 10 additional subpages, respectively. Similarly, Haans (2019) collected frontpages and all subpages that were accessible from the frontpage. Following such practice, our code continues by walking through all website addresses that were listed on the successfully collected archived frontpages and scrapes those that point to the same domain.<sup>3</sup> In other words, for each archived frontpage, the scraper attempts to collect all archived subpages that the frontpage linked to in the year under consideration. In our application, this yielded 2,591,743 additional valid subpage addresses, of which 2,051,884 were also archived by and successfully scraped from the Wayback Machine. On average, each frontpage, therefore, contained 18.58 additional archived subpages, resulting in a joint database of 2,162,263 webpages including the collected frontpages (see Figure 1). Such large-scale scraping projects require significant time to collect the focal data (Landers et al., 2016). In our case, data collection ran at an average pace of about 2,100 webpages per hour, taking about 1,000 hours or 43 days to complete.

### **Step 3: Data (Pre-)Processing and Cleaning**

After scraping the available longitudinal frontpages and subpages from the Wayback Machine, the resulting files are converted “to ensure that the structure and form of the data collected match the structure and form intended” (Landers et al., 2016, p. 485). Specifically, the scraped websites are initially stored as HyperText Markup Language (HTML), the standard web language for creating and defining websites’ structure and content. Given that the vast majority of research applications using website data are centered on textual content (e.g., Guzman & Li, 2023; Jarvis et al., 2019; Sillince & Brown, 2009), our codebase converts the scraped HTML files to one of the most flexible data formats: plain text (.txt).<sup>4</sup> It is important to note here that the websites under consideration were created using multiple versions of the HTML standard, which evolved over time (see Landers et al., 2016, p. 476, for a discussion of these developments). As a result, we took special care in our codebase to accurately process the broad range of possible page structures to yield valid textual data. This is yet another unique feature of longitudinal web scraping compared to cross-sectional scraping projects, where researchers will generally encounter a much more limited variety of HTML standards.

Then, a crucial step is to (pre-)process, select, and clean the textual data. Although there are many viable approaches, we find that most empirical applications engage in at least one of three general steps: excluding irrelevant webpages (Borah et al., 2021; Botero et al., 2013; Guzman & Li, 2023; Haans, 2019; Kotha et al., 2001), excluding webpages in languages not of interest to the researchers (Guzman & Li, 2023; Jancsary et al., 2017), and cleaning as well as processing the remaining texts (see, e.g., Braun et al., 2018; Hickman et al., 2022, for reviews on practices related to this step).<sup>5</sup>

For one, to identify irrelevant (types of) webpages, we recommend two main considerations: removing entirely invalid webpages and removing subpages that are of irrelevant types. For the former, we manually checked all websites where there was only a frontpage available, yet no further subpages were available or successfully collected. This may occur, for instance, if website addresses were registered but did not yet contain any content except for placeholder text. In

earlier years, it was also sometimes the case that the website required the use of specific browsers that the Wayback Machine's archiving crawlers did not use, resulting in error messages. Manually checking all websites without subpages yielded 1,621 wholly invalid websites (shown in the sheet "pages" in the "exclusion\_list.xlsx" file, available at <https://haans-mertens.github.io/data>).<sup>6</sup>

Next, we removed subpages of irrelevant type (e.g., site functionality subpages). We, therefore, utilized recent developments in large language models to assign all subpages to one of sixteen categories: products and services; news and events; home; contact and locations; investor relations; about us; legal; resources, support, and documentation; sustainability and social responsibility; site functionality; donate and support; jobs and opportunities; partners and affiliates; team and leadership; testimonials and reviews; and an "other" category for subpages not fitting the above categories. We created this set of sixteen categories by first engaging in a broad web search to identify the most common subpage types. We also inductively assessed a random set of subpages in our database—adding any common categories that we had not covered yet. We then iteratively clustered the resultant categories into 16 higher-level categories.

To assign all subpages to one of these 16 categories, we first manually classified 600 randomly selected subpages based on their HTML title and their full website address.<sup>7</sup> We also used OpenAI's GPT3.5 model to classify these 600 subpages using their HTML title and subpage path (excluding the domain name). The performance of GPT 3.5 was satisfactory: Whereas the two authors had an 88% agreement rate in the random sample of 600 subpages, the pair-wise agreement rates with GPT were 80.3% and 81.7%. This implies a Krippendorff's alpha of 0.787 or a moderate to strong level of agreement. Therefore, we used GPT3.5 to classify all subpages into the 16 categories.<sup>8</sup> Definitions of these categories and details on their prevalence in the database over time are shown in Table A in the Online Appendix.

We consider the following four page categories irrelevant for most applications in organizational research, including the research question of our application below: home (52,773 subpages), as these subpages are duplicates of the already collected frontpages and thus contain redundant information; legal (81,900 subpages), as these subpages tend to provide boilerplate legal information rather than more substantive content; resources, support, and documentation (75,755 subpages), as qualitative assessments suggested that these subpages provide highly technical and overly specific information and a high degree of nontextual content; and site functionality (45,024 subpages), as these subpages provide little to no relevant content, centering on technical details of the functioning of the sites. Removing all subpages that are associated with these four categories from the subsequent analyses further removes 255,451 subpages. However, our codebase provides researchers with full flexibility to make their own project-specific decisions regarding this and all other (pre-)processing and cleaning decisions.

Second, we follow common practice by removing non-English texts. We do so by using the "detect\_langs" command from the "langdetect" package in Python, which is a port of Google's language detection library (Danilk, 2021). To ensure the replicability of the language estimation process, we set the seed to 123456789.<sup>9</sup> We only retained webpages identified as English with at least 85% certainty, as there was a distinct knee in the certainty score distribution at this value: 212,079 additional webpages were excluded via this selection decision.

Third, the texts that remained after the above selection and cleaning sometimes still contained faulty content, such as error messages (e.g., "Ouch! Your JavaScript is disabled.") or content that we deemed otherwise irrelevant for organizational research, such as cookie notifications (e.g., "read more about our use of cookies"). Sometimes, such faulty content was all that was shown on the webpage, whereas other times it was only a fragment of the text. As such, we recommend taking a two-stage approach to cleaning the remaining texts. For one, invalid content should be removed. Therefore, we manually went through scraped webpages that seemed to contain faulty or irrelevant content to identify specific (sub-)sentences that could be removed to improve data

quality. We identified these webpages by iteratively checking very short webpages (<100 words long) as well as those that contained specific keywords, such as “frames,” “error,” “sorry,” “Netscape,” “under construction,” “under development,” “javascript,” “cookies,” and “does not support.” This enabled us to identify a set of 5,818 (sub-)sentences that we recommend be removed from website texts, as they contain invalid content. Applying this cleaning step to our data removes 5,357,823 out of 636,741,758 words.

Moreover, some texts contained no or little content after removing these (sub-)sentences, but we also observed webpages that contained either no or only minimal content from the start. Given the limited added value of excessively short texts, we recommend removing them outright. To decide which short texts to remove, we checked a random selection of webpage texts of differing lengths (sampling 200 webpages with 1–10 words, 200 webpages with 11–20 words, etc. up through 91–100 words) and manually coded if these texts were invalid. Whereas about 64% of texts in the first length bracket were wholly invalid, this rate quickly dropped in subsequent brackets. Therefore, we decided to exclude all webpage texts with 10 or fewer words after cleaning, removing 74,969 out of the remaining 1,692,644 webpages. Our codebase includes this as well as the preceding selection and cleaning steps. Finally, to move these data to the GVKEY/year level for further analysis, we merged frontpage texts with their respective lower-level subpage texts for a given GVKEY/year into one aggregated file, yielding the consolidated *CompuCrawl* database representing 86,303 GVKEY/year observations, 1,617,675 webpages, and a total of 636,318,656 words (see Figure 1).<sup>10</sup> Our full database, documentation of the codebase scripts, frequently asked questions, and all intermediate data steps can be accessed at <https://haans-mertens.github.io>.

In general, it is important to emphasize that the aforementioned selection and cleaning decisions are, to some degree, subjective, specific to the collected dataset, and fundamentally informed by one’s data source theory (Landers et al., 2016). To this end, we recommend researchers to carefully consider which types of webpages and content should be removed for their own purposes, how their data are affected by this removal (for instance, by reporting the associated change in the number of observations), and assess the implications of this cleaning for their subsequent analyses (we do so at the end of our own application below). Anticipating the need for custom selection and cleaning decisions, while also recognizing the importance of rigorously documenting these choices (Boegershausen et al., 2022; Landers et al., 2016), we designed our codebase to allow researchers to flexibly make and keep track of alternative choices to move from the uncleaned texts to their own cleaned texts.

## Step 4: Data Description and Analysis

### *Describing and Summarizing Textual Content*

Once the cleaned and processed dataset has been constructed, an important next step is to describe the collected data before turning to (statistical) analyses. For research leveraging website texts, we view topic modeling as a particularly useful tool for this task (see Haans, 2019; Powell et al., 2016, for examples). Topic models are probabilistic models designed to discover and analyze latent themes in large-scale textual data. Utilizing documents and their words, which are observed, topic models reveal unobserved topic structures—returning individual topics, their prevalence per document, and the prevalence of words per topic as outputs (see Hannigan et al., 2019; Schmiedel et al. 2019, for overviews).<sup>11</sup>

To estimate our topic model, we utilized the “stm” package in “R” (Roberts et al., 2019), which offers an efficient implementation of various topic modeling algorithms. Specifically, we applied a Correlated Topic Model, which has been shown to outperform alternative algorithms by allowing for topics to exhibit correlation (Blei & Lafferty, 2007).<sup>12</sup> It is common practice in topic modeling

to remove both extremely frequent (i.e., stop words) and infrequent terms for both substantive and computational reasons (Blei & Lafferty, 2007, p. 28). Hence, we removed all stop words as defined by Python's *nlTK* package and terms that occur fewer than 100 times in all website texts, yielding a dataset of 86,299 GVKEY/year observations.

One crucial choice in topic modeling is the number of topics to be estimated, as there is no single best number for a given collection of texts (Roberts et al., 2019). To this end, we assessed two common fit measures: semantic coherence and exclusivity (see Roberts et al., 2019). Semantic coherence focuses on whether topics are internally consistent, containing words that are similar in meaning. Yet by itself, semantic coherence is easily achieved by having just a few topics dominated by the most common terms. In contrast, exclusivity captures if salient terms within a topic are unique to that topic, having low probabilities of appearing in other topics (see Roberts et al., 2019). High-quality topic models feature topics with words that are both semantically coherent and high in exclusivity. Following Hannigan et al. (2019), we estimated models ranging from 25 to 300 topics, which suggested that the 125-topic model offers the best trade-off between coherence and exclusivity (shown in Table B in the Online Appendix).

Table 2 shows the 25 most prevalent topics across all scraped website texts. For each topic, the table also lists the five words with the highest likelihood of occurrence. We also provide labels for all topics based on these top words. Given that this labeling is fundamentally interpretive (see also Hannigan et al., 2019; Schmiedel et al., 2019), each author independently labeled the full set of 125 topics, and we queried GPT 4-Turbo to generate labels for each topic based on its words with the highest likelihood of occurrence. Subsequently, we went through the three sets of labels to decide on the most appropriate label for each topic. Table C in the Online Appendix provides an overview of all topics, together with additional clarifications of specific words.

In organizational research contexts, topic models, generally, capture the symbolic and material attributes of products, organizations, and industries that are both shared among actors and that distinguish these entities from others (Durand & Thornton, 2018). Indeed, dominant topics pertain to, for instance, executive leadership (topic 53: "president," "officer," and "chief"), organizational culture and employment (topic 105: "work," "people," and "employees"), and product offerings (topic 42: "solutions," "technology," and "products"). Some topics are also anchored in specific industries (topic 41: "bank," "banking," "deposit," or topic 83: "clinical," "drug," and "development"). Therefore, our data contain a wide variety of rich textual information about how firms present themselves to the outside world. This further supports the suitability of our database for addressing the research question posed in our empirical application below.

## Application

To briefly demonstrate a use-case of our approach and highlight the unique benefits of leveraging organizational website data over time, we extend recent work that employs word-embedding models to study the meaning assigned to salient organizational concepts (Poschmann et al., 2023). The use of word-embedding models in organizational research is a relatively recent methodological innovation (see Aceves & Evans, 2024, for a review). The key notion underpinning these models is that the meaning of a word can be inferred by considering the words that frequently appear in its immediate context (Firth, 1957; Harris, 1954). Therefore, based on shared words in their immediate contexts, words can have a similar meaning even if they do not occur in the same texts. For instance, even though "doctors" and "lawyers" rarely co-occur in texts, they both frequently have words such as "work," "cases," and "appointment" in their immediate context, implying a similar meaning. In contrast, topic models identify meaning solely based on co-occurrences of words within the same texts.

**Table 2.** Most Prevalent Topics in the Final Database.

Topic	Prevalence	Label	Word 1	Word 2	Word 3	Word 4	Word 5
53	3.70%	Executive leadership	President	Vice	Officer	Chief	Director
105	3.64%	Firm culture and employment	Employees	People	Work	Company	Business
42	3.18%	Technology solutions	Solutions	Technology	Technologies	Systems	Products
94	3.02%	Customer support	Get	Help	Need	One	Make
73	2.86%	General firm information	Products	Sales	Product	Company	Business
109	2.60%	Contact details	Email	Please	Contact	Phone	Name
41	2.24%	Banking and financial services	Account	Bank	Banking	Business	Credit
22	2.13%	Global operations	Inc	Company	Group	International	Companies
84	2.04%	Website technologies	Web	Site	Internet	Online	Service
83	1.93%	Pharmaceutical research	Clinical	Drug	Patients	Development	Disease
55	1.91%	Oil and gas: Exploration	Oil	Gas	Energy	Production	Natural
70	1.88%	Oncology research	Cancer	Clinical	Cell	Patients	Cells
115	1.87%	Investor relations	Investor	Relations	Stock	Report	Press
104	1.87%	Months	June	May	March	April	July
76	1.78%	Mining: Exploration	Gold	Exploration	Mining	Project	Property
40	1.62%	Data use agreement	Information	May	Data	Use	Must
30	1.61%	Staff qualifications	Experience	Team	Work	Job	Service
114	1.60%	Contact details and customer service hours	Drive	Monday	Friday	Services	Street
44	1.58%	Corporate governance and disclosure	Statements	Forward	Looking	Company	Results
81	1.51%	Financial performance	Million	Net	Quarter	Income	Cash
66	1.46%	Software management systems	Management	Software	Data	Business	Solutions
26	1.42%	Real estate: Commercial	Real	Estate	Property	Properties	Office
79	1.35%	Business consulting	Services	Clients	Management	Solutions	Business
13	1.33%	Modes of contact	Fax	Tel	United	Phone	Europe
58	1.30%	Mining: Exploitation	Mining	Gold	Mine	Project	Mineral

Note: Topic numbers are determined by the algorithm and do not correlate with the overall importance of the topics. Prevalence is based on the average topic loading across the entire collection of texts.

To estimate the word embedding vectors, we utilized the widely used continuous bag-of-words (CBOW) Word2Vec architecture (Mikolov et al., 2013), which has a straightforward implementation in Python and has been shown to perform well in a range of settings (Aceves & Evans, 2024). The intuition behind this approach is as follows: Each unique word in a collection of texts is randomly positioned in high-dimensional, shared vector space<sup>13</sup> (Hovy, 2020). The CBOW algorithm goes through all texts word by word. The words that precede and succeed a target word are used to predict that word.<sup>14</sup> A word that repeatedly predicts a focal word is nudged closer to it in the vector space. As a result, the algorithm generates so-called embeddings, high-dimensional vectors that represent each word based on its association with other words. In turn, these vectors can be used to understand the meaning of specific words by comparing them to the vectors of other words—for instance by considering which words are closest in vector space based on their cosine similarity.<sup>15</sup>

There are two key decisions when applying the Word2Vec algorithm: the number of dimensions that should be used to represent words (i.e., the size of the embeddings to be learned) and the size of the context window that the algorithm uses (i.e., how many words to the left and right of each target word to consider). The developers of the Word2Vec algorithm identified 300 as the optimal number of dimensions to represent words (Mikolov et al., 2013), which later work confirmed (Rodriguez & Spiraling, 2022), such that we followed these recommendations. For the context window size, prior work identified five or six words as optimal (Le & Mikolov, 2014; Rodriguez & Spiraling, 2022). We set the window size to six, as larger windows encode finer-grained differences. Unlike for our topic model, we did not remove stop words or infrequent words in our pre-processing as this would alter the context windows used by the word-embedding model. In addition, given that words on separate webpages are conceptually not part of the same context, we used individual webpage texts as input documents.<sup>16</sup>

The goal of our application is to assess the assumption made by Poschmann et al. (2023, p. 10) “that the semantics of individual words [are...] relatively stable within similar textual contexts”. Accordingly, they task future research to explore if the meaning of concepts evolves over time (endnote 4, p. 23). To do so, we focus on two important concepts for organizations: “sustainability” and “profitability.” We selected these concepts because we expected the discourse around sustainability to have changed substantially over the years, whereas the meaning of profitability is unlikely to have changed. Specifically, the concept of sustainability first meaningfully entered the organizational domain in the mid-1990s following the 1987 Brundtland Report—which primarily emphasized an environmental lens to sustainability. The United Nations Millennium Development Goals set in 2000 then widened the scope of the concept to include the three interconnected pillars of environmental, economic, and social sustainability. This, in turn, led to the emergence of different standards to capture progress in these three dimensions. Finally, the Sustainable Development Goals adopted by the UN in 2015 led to further crystallization of relevant standards and the substantive meaning of the concept of sustainability. Thus, we expect a relatively large shift in the meaning of sustainability from the starting point of our sample period relative to later years. In contrast, we do not expect such a shift in meaning of the concept of profitability due to its more objective nature and established operationalization in accounting.

To test these conjectures, we split our dataset into five time periods of 5 years each (e.g., 1996–2000) and separately applied the word embedding algorithm to each of these subperiods (Hamilton et al., 2016). Respectively, this yielded 7,125 firm/year observations, 37,068 unique webpages, and 8,693,766 total words for the 1996–2000 period, 11,617/85,435 / 21,593,183 for the 2001–2005 period, 15,994/195,797/59,239,480 for the 2006–2010 period, 23,631/570,869/218,366,018 for the 2011–2015 period, and 27,936/728,506/328,426,220 for the 2016–2020 period.<sup>17</sup> One complication is that embedding models trained on different sets of texts will not be aligned to the same coordinate axes. While this does not impact the word representations within a period, it obscures the change in the meaning of concepts between periods. Therefore, following Hamilton et al. (2016),



we enabled the comparison of vectors over time by using orthogonal Procrustes to align periods' learned dimensions to those of the prior period.<sup>18</sup>

Table 3 shows the five most similar word vectors to “sustainability” and “profitability” over time. For “sustainability,” the high fluctuation in the list of most similar word vectors over the five time periods supports our suspicion that this concept's meaning evolved. Specifically, in the 1996–2000 period, “sustainability” was most similar to the terms “habitat,” “fostering,” “vitality,” “welfare,” and “political.” In the 2001–2005 period, “sustainability” shifted in meaning to become aligned with “Environment, Health, and Safety” standards (EHS and HSE in the table). In the last three periods, “sustainability” aligned closely with the concept of corporate social responsibility (CSR) and organizational citizenship. Confirming this pattern, the vector of “sustainability” is initially quite dissimilar to its own vector in the preceding period (e.g., 0.301 cosine similarity, comparing the 2001–2005 vector to the 1996–2000 vector) and stabilizes in later periods (e.g., 0.916 cosine similarity, comparing the 2016–2020 vector to the 2011–2015 vector). In contrast, the meaning of “profitability” has remained mostly constant over time (cosine similarity > 0.74 between all periods). This strongly suggests that the shifts in meaning observed for sustainability are not merely an artifact of the empirical approach taken, as we would then expect to see the meaning of all concepts change over time. Importantly, as shown in Table E in the Online Appendix, these patterns are unchanged when not incorporating the various aforementioned selection and cleaning decisions (other than the removal of duplicate frontpages)—suggesting that these steps did not have substantive implications for our overall conclusion.

In all, this brief application highlights the unique insights that can emerge when leveraging longitudinal organizational website data. It reveals that the meaning of some—but not all—concepts in organizational discourse, as captured by how organizations discuss them on their websites, changes over time. Thus, work that analyzes textual data generated by organizations should critically consider the extent to which the fundamental meaning of focal concepts differs depending on the time period under consideration.

## Discussion and Conclusion

Organizational websites offer researchers access to unique perspectives compared to other data sources, such as firm disclosures (e.g., Hoberg & Phillips, 2010). Nevertheless, collecting and analyzing the unstructured big data necessary for such research remains a major bottleneck—especially concerning website data over time. We provide a framework and toolkit that enables researchers to leverage the latent potential of these data—offering a four-step tutorial, an open-access codebase, and a novel database. Following suggestions in recent work (Boegershausen et al., 2022; Edelman, 2012; Landers et al., 2016; Powell et al., 2016), we have designed our approach around Archive.org's Wayback Machine and critically evaluated the data it provides. As a result, we offer, to the best of our knowledge, the first systematic empirical validation of this rich database's data quality and coverage.

Our approach to collecting longitudinal website data at scale that works with any website sparks novel research avenues that leverage websites' texts, their multimodal nature, and/or their HTML code. Table 4 summarizes the nonexhaustive list of research avenues that we discuss below. First, future research may explicitly consider how the effects of variables constructed using organizations' website texts compare to those generated using other data sources. For example, while organizational distinctiveness can be operationalized using many different variables (Deephouse, 1999; Zhao et al., 2017), work to date has largely remained agnostic about the implications of doing so. Yet, positioning oneself as different from competitors online should have substantially different consequences than distinctively arranging strategic resources: The former is highly visible yet may be merely rhetorical, while the latter is less visible yet entails actual resource allocation decisions. Therefore,

**Table 3.** Most Similar Word Vectors to “Sustainability” and “Profitability” Over Time.

Period	Prior	Word 1	Word 2	Word 3	Word 4	Word 5
<i>Sustainability</i> 1996–2000	n.a.	Habitat (0.406)	Fostering (0.387)	Vitality (0.378)	Welfare (0.375)	Political (0.370)
2001–2005	0.301	Stewardship (0.496)	Sustainable (0.495)	HSE (0.495)	EHS (0.473)	Environmental (0.463)
2006–2010	0.705	Sustainable (0.579)	Stewardship (0.560)	Environmental (0.518)	Citizenship (0.498)	CSR (0.480)
2011–2015	0.847	Sustainable (0.634)	Stewardship (0.587)	Environmental (0.553)	CSR (0.546)	Citizenship (0.544)
2016–2020	0.916	CSR (0.672)	Sustainable (0.630)	Stewardship (0.620)	ESG (0.582)	Environmental (0.560)
<i>Profitability</i> 1996–2000	n.a.	Profits (0.640)	Efficiencies (0.602)	Margins (0.586)	Competitiveness (0.575)	Productivity (0.536)
2001–2005	0.744	Profits (0.569)	Efficiencies (0.557)	Productivity (0.522)	Competitiveness (0.514)	Margins (0.514)
2006–2010	0.742	Competitiveness (0.569)	Profits (0.561)	Productivity (0.560)	Margins (0.532)	Efficiencies (0.531)
2011–2015	0.806	Productivity (0.604)	Efficiencies (0.600)	Competitiveness (0.565)	Profits (0.560)	Margins (0.559)
2016–2020	0.874	Efficiencies (0.651)	Competitiveness (0.634)	Margins (0.633)	Productivity (0.630)	Revenue (0.593)

Note: Column “Prior” indicates the cosine similarity of the vector of the focal word to its own vector one period prior. Cosine similarities of the listed word and the focal word are provided in parentheses. HSE = health, safety, and environment; EHS = Environment, health, and safety; CSR = corporate social responsibility; ESG = environmental, social, and governance.

**Table 4.** Examples of Approaches to Using Website-Based Data for Future Organizational Research.

Domain	Construct	Operationalization
Optimal distinctiveness	Distinctiveness	Deviation of organizations' website-based topic weights from industries' mean topic weights
Strategic groups	Industry association	Clusters of organizations with similar values for their website-based topic weights in an n-dimensional space
Organizational ecology	Industry concentration	Density of organizations in an n-dimensional space where an organization's position is determined by its website-based topic weights
	Industry turbulence	Rate of change of organizations' association to clusters in an n-dimensional space where organizations' position is determined by their website-based topic weights
Competitive dynamics	Closest competitor	Highest pairwise similarity between organizations' website-based topic weights and the topic weights of all other organizations in the industry
Organizational behavior and change	Responsiveness	Time between an exogenous shock and the first mention of the event on an organization's website
Organizational reputation	For example, reputation for customer service	Prevalence of keywords on organizations' websites that indicate customer-centric operations
Domain-spanning	Within-concept change in meaning over time	Word embedding vectors' cosine similarity between periods to proxy the degree of change in terms' meaning over time
	Between-concept similarity	Word embedding vectors' cosine similarity between terms to proxy their perceived similarity
	Salience	Operationalize the salience of constructs for organizations based on the percentage of their subpages devoted to, e.g., website category 9: <i>Sustainability &amp; Social Responsibility</i>
Human resource management	Recruitment practices	Qualitative content analysis (assisted by generative artificial intelligence) of organizations' talent acquisition, as described on subpages in website category 12: <i>Jobs &amp; opportunities</i>
Organizational identity	For example, differentiation-centric organizational identity	Prevalence of FREX words (FREquent in the website and EXclusive to it) on organizations' websites
Networks	Strength of network link	Number of outgoing hyperlinks from organization A's website to organization B's website
	Network status	Total number of outgoing hyperlinks across all other websites that direct to a focal organization's website
Innovation and information systems	Organizational readiness for digital innovation	Duration between announcement of a new World Wide Web Consortium HTML standard and its implementation on the focal website
Stakeholder strategy	Stakeholder communication effort	General website complexity (e.g., number of lines of code, number of dependencies); the use of boilerplate websites compared to custom-written sites

operationalizations of distinctiveness based on differing data sources promise to have strong theoretical implications (Durand & Haans, 2022).

A second research opportunity regarding website texts lies in how they enable the categorization of organizations into groups based on more fine-grained and multidimensional data than traditional approaches. Indeed, much work in organization research hinges on comparisons between firms, requiring researchers to define which firms are considered and compared. Here, work commonly relies on established industry classifications, such as SIC Codes, to determine the set of relevant firms. More recently, however, scholars have attempted to develop inductive classifications using dictionary-based approaches (Hoberg & Phillips, 2010), topic modeling (Guo et al., 2017; Shi et al., 2016), and embeddings (Guzman & Li, 2023). Organizational website texts enable researchers to devise such bottom-up, positioning-based classifications for all organizations with a website. For example, the topic model that we used to summarize the textual content of firms' websites also enables the granular identification of industries, their concentration, and changes therein over time. Moreover, website texts can be used to operationalize firms' strategic positioning, including their closest competitors (Guzman & Li, 2023). In contrast to approaches using mandatory disclosures as their source of organizational textual data (e.g., Hoberg & Phillips, 2010), website-based insights can also be generated for third-sector industries and private entities with minimal disclosure requirements. Website data can, thus, offer novel insights on topics such as strategic groups, organizational ecology, and competitive dynamics for types of organizations not captured in other data sources (Carroll, 1984; Hannan & Freeman, 1977; McNamara et al., 2003; Porac et al., 1989).

We also see valuable contributions emerging from the longitudinal website data in our *CompuCrawl* database, given the widespread use of Compustat in organizational research. Compustat is frequently augmented with other databases, such as those containing data on patent citations (Arora et al., 2018; Hall et al., 2001; Yu et al., 2019), to study innovativeness and technological exploration. However, even those data extensions only capture a fraction of organizations' portfolios of actions. Accordingly, Morandi Stagni et al. (2021, p. 25) highlighted how incorporating "other domains of corporate activity" would nuance our understanding of organizational behavior, for example, firms' reactions to competitive shocks. We view the website texts offered in our database as uniquely valuable to address these shortcomings and broaden our understanding of, amongst others, organizational behavior and change.

Similarly, scholars studying the antecedents and implications of organizational reputation frequently rely on financial metrics, such as organizations' market share and asset quality (Deephhouse & Carter, 2005; Shamsie, 2003). However, organizational reputations are inherently multifaceted, and organizations can, for example, be known for "being diversified or focused, environmentally friendly, or high-technology oriented" (Blagoeva et al., 2020, p. 1737). Yet, scholars have been limited in operationalizing such reputational multidimensionality when solely utilizing financial data (Lange et al., 2011)—despite the significant practical and scholarly implications of studying conflicting reputations (Parker et al., 2019). Hence, employing website texts could allow for a more holistic assessment of organizational reputations.

Moreover, longitudinal website texts hold great potential to reveal elusive yet highly relevant organizational phenomena when paired with state-of-the-art machine learning approaches. For one, we agree with Aceves and Evans (2024) that word-embedding models enable substantial advances in the measurement and theory of organizational research. As demonstrated in our brief application, word embeddings of longitudinal website texts can be used to quantify the temporal change in meaning of otherwise difficult-to-measure constructs. Similarly, word-embedding models can be used to assess the similarity between different concepts and the temporal development thereof (Charlesworth et al., 2022; Garg et al., 2018; Hamilton et al., 2016). These powerful techniques can be paired with impactful research questions in the domain of organizational research that could otherwise not be answered. Second, the large-scale website data generated by our approach

pair well with generative artificial intelligence applications. In our project, Open AI's GPT approached human-level performance in assigning websites to predefined categories. Moreover, it completed the task at a fraction of the time and cost that human coders would have required. The resulting categorization of webpages into sixteen categories can, for example, enable researchers to operationalize the salience of sustainability in organizations' public self-representations over time. More generally, generative AI models can be readily integrated into research workflows that utilize large-scale website data. The potential applications range from simple categorization tasks to, for example, human resource scholars cooperatively using generative AI to support their qualitative content analysis of firms' jobs and opportunities subpages.

While our application and the preceding future research avenues have put website texts front and center, websites provide additional, highly valuable research data. For one, websites are inherently multimodal, integrating visual elements, such as photos and videos, with text. Accessing these visuals is straightforward with our codebase and database, which provide access to historical websites' HTML files and the contained links to the visual elements archived by the Wayback Machine. Due to the rise of the visual turn in organizational research (Boxenbaum et al., 2018), which leverages organizations' visual artifacts, we see great potential in researchers studying visuals on organizations' websites—especially in combination with website texts. Such multimodal research has been repeatedly called for in the literature (Höllerer et al., 2018; Meyer et al., 2013; Quattrone et al., 2021). For example, organizational identity researchers' access to longitudinal website texts and visual artifacts at scale can support their use of these metaphorical windows into organizations. Specifically, websites' textual and/or visual artifacts have been utilized to research firms' (multiple) identities, their use of narratives, and the formation of legitimacy as well as stakeholder trust (Bell & Davison, 2013; Bertels et al., 2014; Botero et al., 2013; Jancsary et al., 2017; Meyer et al., 2013; Santos, 2019; Sillince & Brown, 2009). However, research in this domain has mostly resorted to cross-sectional analyses or short observation periods and small samples due to the need for repeatedly collecting live website versions. Hence, the exploration of this “distinctive genre of collective identity” can substantially benefit from longitudinal, multimodal website data at scale and the resulting ability to analyze identity shifts over time (Sillince & Brown, 2009, p. 1835).

In addition to websites' visitor-centric texts and visuals, websites' HTML files can be leveraged to operationalize otherwise elusive constructs. For example, a stream of research in network theory operationalizes network ties through outgoing links on organizational websites (e.g., Powell et al., 2016, 2017; Wruk et al., 2020). These studies are based on the premise that outgoing links represent relational resources and, thus, indicate embeddedness. However, researchers in this domain have been confined to cross-sectional analyses, and were, thus, “technically not able to confirm an ongoing process” (Oberg et al., 2009, p. 6; Wruk et al., 2020). Therefore, longitudinal website data enable more rigorous assessments of network associations, for example, via stochastic actor-oriented models (Snijders, 2017) to help illuminate processes.

Moreover, what is considered modern website code changes over time (Landers et al., 2016). Hence, researchers can exploit events such as the transition from HTML4 to HTML5 as the officially recommended website markup language by the World Wide Web Consortium on October 28, 2014. The lag between new HTML versions becoming state-of-the-art and organizations adopting them on their websites may reflect organizations' readiness to embrace technological advances (Lokuge et al., 2019). Similarly, HTML files can inform researchers of the effort that organizations exert in their stakeholder communication through their websites. For example, HTMLs reveal websites' complexity and whether organizations utilize customized code or resort to boilerplate websites provided by website-building companies like Squarespace.

To conclude, numerous high-potential organizational research streams have emerged from the increased importance of language and visual artifacts (Bencherki et al., 2021; Boxenbaum et al., 2018; Cornut et al., 2012; Krautzberger et al., 2021; Kwon et al., 2014; Quattrone et al., 2021;

Spee & Jarzabkowski, 2011). However, collecting the big data necessary for exploring many associated research questions has frequently remained out of researchers' grasp (Simsek et al., 2019). We hope this paper as well as the associated codebase and database spark future organizational research by offering an open-source approach to tapping into the rich, multimodal, and ever-changing data that websites represent.

## Acknowledgments

We are grateful to the Editor Lisa Schurer Lambert, Associate Editor Michael Withers, and the anonymous reviewers for their valuable feedback and guidance, which significantly improved this paper.


## Declaration of Conflicting Interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

## ORCID iDs

Richard F.J. Haans  <https://orcid.org/0000-0002-4868-1488>

Marc J. Mertens  <https://orcid.org/0000-0002-0199-8939>

## Supplemental Material

Supplemental material for this article is available online. Our full database, documentation of the codebase scripts, frequently asked questions, and all intermediate data steps can be accessed at <https://haans-mertens.github.io>.

## Notes

1. It might be possible to manually collect websites for some of these cases. To assess the extent to which such an effort would yield additional data, we engaged in a Google search for the firm name, the term "website," and the firm's city for a random sample of 100 firms with a missing website. This effort took around three hours due to the need to thoroughly verify potential matches yet yielded only 33 valid websites—suggesting, in the case of Compustat, limited added value (an expected 4,400 websites) relative to the manual effort required to collect these websites (around 400 hours).
2. We transformed these variables since they are all highly skewed. Because some of them can be negative or zero, precluding a log transformation, we utilized the cube-root transformation (Cox, 2011).
3. It is, of course, possible to continue further down the page hierarchy to collect lower-level subpages from links listed on first-level subpages. While this would add many subpages (an additional 20,869,923 unique subpages in the case of our database), the added value appears limited. First, there are a handful of websites that yield an extreme number of additional subpages at this level: The three largest have 400,751, 453,052, and 489,402 additional subpages (compared to 2,304, 2,361, and 6,433 subpages at the current page hierarchy depth). Second, this level of analysis seemingly goes further into the page hierarchy than is desirable: It yields a median number of "/" in the website addresses of three (two in the current level of depth, e.g., [www.shutterstock.com/blog/design](http://www.shutterstock.com/blog/design)), with a 95th percentile of six (four at the current depth), and a maximum of 90 (23). Thus, the current approach to scrape subpages based on frontpage links already

captures website data in great detail. Nevertheless, we offer code to go further down the page hierarchy in our codebase.

4. While we focus on websites' textual content in the subsequent discussion and our application, websites are inherently multimodal, frequently featuring visual elements, such as photos and videos together with text. Therefore, we intentionally provide access to websites' HTML files from which links to these archived visuals can be extracted. The resulting links can be used to download these visuals from the Wayback Machine.
5. It is also common practice in natural language processing to remove stop words (such as "the" and "a") as well as highly infrequent words (e.g., Pina & Tether, 2016). However, because such terms do not represent inherently invalid content, we recommend that these decisions be informed by the subsequent use case of these data. Word embedding approaches, for instance, assess terms' latent meaning based on their context, and removing terms, such as stop words, affects these contexts. Hence, most embedding-based projects do minimal cleaning (Rodriguez & Spirling, 2022). In contrast, "bag of words" approaches, such as most topic models, do not rely on the sequence of words but, for example, on the co-occurrence of words within documents. Accordingly, researchers typically remove both stop words and infrequent terms for such applications due to substantive and computational reasons (e.g., Blei & Lafferty, 2007, p. 28). We, thus, leave this particular cleaning step to be implemented according to the specific research application.
6. This sheet can also be used to remove additional webpages in a straightforward manner. We used this opportunity to remove 468 subpages that manual assessments in subsequent cleaning steps, discussed below, determined to contain invalid data.
7. An HTML title is the text enclosed within <title> tags in most HTML documents. This text is displayed to website users in their browsers' tab bar and usually provides a concise description of the webpage's content.
8. We used GPT 3.5 due to its satisfactory performance and because, at the time of writing, there was no API implementation of more recent models. The total classification task cost \$ 1,627.79 and took over two weeks to complete, which would have been thirtyfold with, for instance, GPT4 due to pricing and rate limit differences.
9. In probabilistic machine learning applications, such as `detect_langs`, outcomes are not deterministic. While these applications do not require a seed, setting one during initialization ensures consistent results, facilitating collaboration and reproducibility. The chosen seed value is arbitrary and interchangeable with any other integer without impacting the application's functionality.
10. Although these numbers suggest that these are relatively big data, we have tested all scripts on a (at the time of writing) seven-year-old workstation with an AMD Ryzen 7 1700 3.0 GHz processor and 16 GB of 3200 MHz memory, which handled all reported steps with ease. As such, we anticipate that most machines are capable of working with the general approach and the database described in this paper.
11. In brief, the process begins with a collection of documents (e.g., website texts). It is assumed that each document contains a mixture of various topics. For example, a firm's investor relations page may feature topics related to financial performance, financial targets, and stakeholder communication. Moreover, it is assumed that each topic has a distinct distribution of words. For instance, the topic related to financial performance might have a high probability for words like "profits," "revenue," and "margin." To arrive at these associations between topics and words, words are iteratively associated with topics. Topic-word associations are retained if they reflect the observed patterns in the underlying documents, as measured by fit statistics. For example, associating the word "income" with the topic related to financial performance may fit the data better than associating the word "travel" with it. Through this iterative process, the algorithm refines the associations between topics and words. Finally, it returns a prespecified number of topics that are characterized by their associations with words from the input documents. The algorithm also outputs the prevalence of the topics in each of the documents (e.g., the salience of the topic related to financial performance in every website text).
12. As the name suggests, Correlated Topic Models explicitly incorporate the correlations that exist between topics to better reflect the underlying documents. Returning to the example from the previous note, the

topic related to financial performance likely correlates positively with the topic related to stakeholder communication. When one is salient in a website text, the other frequently is as well. By incorporating such inter-topic correlations, the Correlated Topic Model presents an improvement over classic latent Dirichlet allocation (LDA) topic models. It, thus, “gives a more realistic model of the latent topic structure” in real-world texts (Blei & Lafferty, 2007, p. 19).

13. A high-dimensional vector space describes a coordinate system with numerous dimensions in which the words are represented. The vector space is “shared” as all words in the focal collection of texts are positioned within the same vector space, allowing for comparisons and analyses across the entire collection of texts.
14. For example, consider a context window size of three and the sentence “Our recent profitability has rewarded shareholders for their trust.” When the sliding context window of the CBOW algorithm reaches the word “shareholders,” it attempts to predict this word using the three preceding and succeeding words (“profitability,” “has,” “rewarded,” “for,” “their,” and “trust”). Across all texts, “shareholders” may be predicted more reliably by the word “profitability” than by the word “trust.”
15. Consider a simplified scenario with just three words: “sustainability,” “profitability,” and “shareholders.” The words are represented by vectors, so-called embeddings, in a two-dimensional shared vector space. For example, “shareholders” could be represented as vector  $a = [0.3, 0.6]$ , “sustainability” as vector  $b = [0.9, 0.4]$ , and “profitability” as vector  $c = [0.2, 0.7]$ . The respective values result from the words’ relationships with other words as learned by the embedding model. The similarity between the three words can then be calculated based on the vectors’ cosine similarity. Specifically, “shareholders” is closer to “profitability” (cosine similarity =  $S_{\cosine}(a, c) = \frac{a \cdot c}{\|a\| \times \|c\|} = \frac{0.3 \times 0.2 + 0.6 \times 0.7}{\sqrt{0.3^2 + 0.6^2} \times \sqrt{0.2^2 + 0.7^2}} = 0.983$ ) than “shareholders” is to “sustainability” ( $S_{\cosine}(a, b) = 0.772$ ).
16. To further ensure that words are in the same context, rather than, for instance, separated by images or other types of content within a given page, we additionally ran our Word2Vec application taking individual sentences as input documents. The results of this analysis are shown in the Online Appendix Table D; we thank an anonymous reviewer for this suggestion.
17. Altszyler et al. (2017) showed that the threshold above which Word2Vec outperforms simpler approaches is around one million words. All periods are well above this threshold. Indeed, Altszyler et al. (2017, p. 179) noted that a dataset of 8 million words, which corresponds to the size of the dataset for the smallest period in our application, is a “medium size corpus.”
18. Orthogonal Procrustes is a mathematical technique that aligns two sets of vectors via scaling, shifting, and rotation while preserving their original similarities and angles. In our application, this technique is used to map a period’s word embedding vectors to that of the prior period without distorting the associations between vectors. As a result, we can make inter-period comparisons without altering the meaning of words.

## References

- Aceves, P., & Evans, J. A. (2024). Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, 35(3), 788–814. <https://doi.org/10.1287/orsc.2023.1686>
- Altszyler, E., Ribeiro, S., Sigman, M., & Fernández Slezak, D. (2017). The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition*, 56, 178–187. <https://doi.org/10.1016/j.concog.2017.09.004>
- Arora, A., Belenzon, S., & Pataconi, A. (2018). The decline of science in corporate R&D. *Strategic Management Journal*, 39(1), 3–32. <https://doi.org/10.1002/smj.2693>
- Ball, R., & Watts, R. (1979). Some additional evidence on survival biases. *The Journal of Finance*, 34(1), 197–206. <https://doi.org/10.1111/j.1540-6261.1979.tb02080.x>
- Bell, E., & Davison, J. (2013). Visual management studies: Empirical and theoretical approaches\*. *International Journal of Management Reviews*, 15(2), 167–184. <https://doi.org/10.1111/j.1468-2370.2012.00342.x>



- Bencherki, N., Sergi, V., Cooren, F., & Vásquez, C. (2021). How strategy comes to matter: Strategizing as the communicative materialization of matters of concern. *Strategic Organization*, 19(4), 608–635. <https://doi.org/10.1177/1476127019890380>
- Bertels, S., Hoffman, A. J., & DeJordy, R. (2014). The varied work of challenger movements: Identifying challenger roles in the US environmental movement. *Organization Studies*, 35(8), 1171–1210. <https://doi.org/10.1177/0170840613517601>
- Black, M. L. (2016). The world wide web as complex data set: Expanding the digital humanities into the twentieth century and beyond through internet research. *International Journal of Humanities and Arts Computing*, 10(1), 95–109. <https://doi.org/10.3366/ijhac.2016.0162>
- Blagoeva, R. R., Kavusan, K., & Jansen, J. J. P. (2020). Who violates expectations when? How firms' growth and dividend reputations affect investors' reactions to acquisitions. *Strategic Management Journal*, 41(9), 1712–1742. <https://doi.org/10.1002/smj.3155>
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-AOAS114>
- Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of gold: Scraping web data for marketing insights. *Journal of Marketing*, 86(5), 1–20. <https://doi.org/10.1177/00222429221100750>
- Borah, D., Malik, K., & Massini, S. (2021). Teaching-focused university–industry collaborations: Determinants and impact on graduates' employability competencies. *Research Policy*, 50(3), 104172. <https://doi.org/10.1016/j.respol.2020.104172>
- Botero, I. C., Thomas, J., Graves, C., & Fediuk, T. A. (2013). Understanding multiple family firm identities: An exploration of the communicated identity in official websites. *Journal of Family Business Strategy*, 4(1), 12–21. <https://doi.org/10.1016/j.jfbs.2012.11.004>
- Boxenbaum, E., Jones, C., Meyer, R. E., & Svejenova, S. (2018). Towards an articulation of the material and visual turn in organization studies. *Organization Studies*, 39(5–6), 597–616. <https://doi.org/10.1177/0170840618772611>
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>
- Braun, M. T., Kuljanin, G., & DeShon, R. P. (2018). Special considerations for the acquisition and wrangling of big data. *Organizational Research Methods*, 21(3), 633–659. <https://doi.org/10.1177/1094428117690235>
- Carroll, G. R. (1984). Organizational ecology. *Annual Review of Sociology*, 10(1), 71–93. <https://doi.org/10.1146/annurev.so.10.080184.000443>
- Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, 119(28), e2121798119. <https://doi.org/10.1073/pnas.2121798119>
- Cornut, F., Giroux, H., & Langley, A. (2012). The strategic plan as a genre. *Discourse & Communication*, 6(1), 21–54. <https://doi.org/10.1177/1750481311432521>
- Cox, N. J. (2011). Stata tip 96: Cube roots. *The Stata Journal: Promoting Communications on Statistics and Stata*, 11(1), 149–154. <https://doi.org/10.1177/1536867X1101100112>
- Danilk, M. M. (2021). *Langdetect*. <https://pypi.org/project/langdetect/>
- Deephouse, D. L. (1999). To be different, or to be the same? It's a question (and theory) of strategic balance. *Strategic Management Journal*, 20(2), 147–166. [https://doi.org/10.1002/\(SICI\)1097-0266\(199902\)20:2<147::AID-SMJ11>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-0266(199902)20:2<147::AID-SMJ11>3.0.CO;2-Q)
- Deephouse, D. L., & Carter, S. M. (2005). An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, 42(2), 329–360. <https://doi.org/10.1111/j.1467-6486.2005.00499.x>
- Dreyer, A. J., & Stockton, J. (2013). Internet “data scraping”: A primer for counseling clients. *New York Law Journal*. <https://www.law.com/newyorklawjournal/almID/1202610687621/>

- Durand, R., & Haans, R. F. J. (2022). Optimally distinct? Understanding the motivation and ability of organizations to pursue optimal distinctiveness (or not). *Organization Theory*, 3(1), 263178772210793. <https://doi.org/10.1177/26317877221079341>
- Durand, R., & Thornton, P. (2018). Categorizing institutional logics, institutionalizing categories: A review of two literatures. *Academy of Management Annals*, 12(2), 631–658. <https://doi.org/10.5465/annals.2016.0089>
- Dykstra, S., Dykstra, B., & Sandefur, J. (2014). *We just ran twenty-three million queries of the World Bank's website* (Working Paper 362, pp. 1–20). Center for Global Development. <https://doi.org/10.2139/ssrn.2458086>
- Ebben, J. J., & Johnson, A. C. (2005). Efficiency, flexibility, or both? Evidence linking strategy to performance in small firms. *Strategic Management Journal*, 26(13), 1249–1259. <https://doi.org/10.1002/smj.503>
- Edelman, B. (2012). Using internet data for economic research. *Journal of Economic Perspectives*, 26(2), 189–206. <https://doi.org/10.1257/jep.26.2.189>
- Ethiraj, S. K., Gambardella, A., & Helfat, C. E. (2017). Improving data availability: A new SMJ initiative. *Strategic Management Journal*, 38(11), 2145–2146. <https://doi.org/10.1002/smj.2690>
- Ethiraj, S. K., Gambardella, A., & Helfat, C. E. (2019). Articles on datasets. *Strategic Management Journal*, 40(5), 713–714. <https://doi.org/10.1002/smj.3000>
- Firth, J. R. (1957). *Papers in linguistics*. Oxford University Press.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), 3635–3644. <https://doi.org/10.1073/pnas.1720347115>
- Guo, L., Sharma, R., Yin, L., Lu, R., & Rong, K. (2017). Automated competitor analysis using big data analytics: Evidence from the fitness mobile app business. *Business Process Management Journal*, 23(3), 735–762. <https://doi.org/10.1108/BPMJ-05-2015-0065>
- Guzman, J., & Li, A. (2023). Measuring founding strategy. *Management Science*, 69(1), 101–118. <https://doi.org/10.1287/mnsc.2022.4369>
- Haans, R. F. J. (2019). What's the value of being different when everyone is? The effects of distinctiveness on performance in homogeneous versus heterogeneous categories. *Strategic Management Journal*, 40(1), 3–27. <https://doi.org/10.1002/smj.2978>
- Hales, S., Riach, K., & Tyler, M. (2021). Close encounters: Intimate service interactions in lap dancing work as a nexus of 'self-others-things.' *Organization Studies*, 42(4), 555–574. <https://doi.org/10.1177/0170840619830127>
- Hall, B., Jaffe, A., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. National Bureau of Economic Research. p. 8498.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change). In *Proceedings of the 54th annual meeting of the association for computational linguistics*, Berlin, August 2016 (Vol. 1: Long Papers, pp. 1489–1501). <https://doi.org/10.18653/v1/P16-1141>
- Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American Journal of Sociology*, 82(5), 929–964. <https://doi.org/10.1086/226424>
- Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchaljian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13(2), 586–632. <https://doi.org/10.5465/annals.2017.0099>
- Harris, Z. S. (1954). Distributional structure. *Word & World*, 10(2–3), 146–162. <https://doi.org/10.1080/00437956.1954.11659520>
- Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1), 114–146. <https://doi.org/10.1177/1094428120971683>
- Hoberg, G., & Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, 23(10), 3773–3811. <https://doi.org/10.1093/rfs/hhq053>

- Höllerer, M. A., Jancsary, D., & Grafström, M. (2018). A picture is worth a thousand words': Multimodal sensemaking of the global financial crisis. *Organization Studies*, 39(5–6), 617–644. <https://doi.org/10.1177/0170840618765019>
- Holstein, J., Starkey, K., & Wright, M. (2018). Strategy and narrative in higher education. *Strategic Organization*, 16(1), 61–91. <https://doi.org/10.1177/1476127016674877>
- Hovy, D. (2020). *Text analysis in Python for social scientists: Discovery and exploration*. Cambridge University Press.
- Internet Archive. (2022). *Terms of use*. <https://archive.org/about/terms.php>
- Jancsary, D., Meyer, R. E., Höllerer, M. A., & Barberio, V. (2017). Toward a structural model of organizational-level institutional pluralism and logic interconnectedness. *Organization Science*, 28(6), 1150–1167. <https://doi.org/10.1287/orsc.2017.1160>
- Jarvis, L. C., Goodrick, E., & Hudson, B. A. (2019). Where the heart functions best: Reactive–affective conflict and the disruptive work of animal rights organizations. *Academy of Management Journal*, 62(5), 1358–1387. <https://doi.org/10.5465/amj.2017.0342>
- Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, 16(4), e0249071. <https://doi.org/10.1371/journal.pone.0249071>
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods*, 21(3), 733–765. <https://doi.org/10.1177/1094428117722619>
- Kotha, S., Rindova, V. P., & Rothaermel, F. T. (2001). Assets and actions: Firm-specific factors in the internationalization of U.S. Internet Firms. *Journal of International Business Studies*, 32(4), 769–791. <https://doi.org/10.1057/palgrave.jibs.8490994>
- Kothari, S. P., Shanken, J., & Sloan, R. G. (1995). Another look at the cross-section of expected stock returns. *The Journal of Finance*, 50(1), 185–224. <https://doi.org/10.2307/2329243>
- Krautzbberger, M., Fohim, E., Cooren, F., & Schumacher, T. (2021). The communicative constitution of institutional change in expression games. *Strategic Organization*, 19(4), 667–692. <https://doi.org/10.1177/1476127020959253>
- Kroezen, J. J., & Heugens, P. P. M. A. R. (2012). Organizational identity formation: Processes of identity imprinting and enactment in the Dutch microbrewing landscape. In M. Schultz, S. Maguire, A. Langley, & H. Tsoukas (Eds.), *Constructing identity in and around organizations* (pp. 89–127). Oxford University Press.
- Kwon, W., Clarke, I., & Wodak, R. (2014). Micro-level discursive strategies for constructing shared views around strategic issues in team meetings. *Journal of Management Studies*, 51(2), 265–290. <https://doi.org/10.1111/joms.12036>
- Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, 21(4), 475–492. <https://doi.org/10.1037/met0000081>
- Lange, D., Lee, P. M., & Dai, Y. (2011). Organizational reputation: A review. *Journal of Management*, 37(1), 153–184. <https://doi.org/10.1177/0149206310390963>
- Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of Machine Learning Research*, 32(2), 1188–1196. <https://doi.org/10.48550/arXiv.1405.4053>
- Lokuge, S., Sadera, D., Grover, V., & Dongming, X. (2019). Organizational readiness for digital innovation: Development and empirical calibration of a construct. *Information & Management*, 56(3), 445–461. <https://doi.org/10.1016/j.im.2018.09.001>
- McNamara, G., Deephouse, D. L., & Luce, R. A. (2003). Competitive positioning within and across a strategic group structure: The performance of core, secondary, and solitary firms. *Strategic Management Journal*, 24(2), 161–181. <https://doi.org/10.1002/smj.289>
- Meyer, R. E., Höllerer, M. A., Jancsary, D., & Van Leeuwen, T. (2013). The visual dimension in organizing, organization, and organization research: Core ideas, current developments, and promising avenues. *Academy of Management Annals*, 7(1), 489–555. <https://doi.org/10.5465/19416520.2013.781867>

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv Preprint arXiv:1301.3781.
- Morandi Stagni, R., Fosfuri, A., & Santaló, J. (2021). A bird in the hand is worth two in the bush: Technology search strategies and competition due to import penetration. *Strategic Management Journal*, 42(8), 1516–1544. <https://doi.org/10.1002/smj.3277>
- Netcraft. (2021). *December 2021 Web Server Survey*. <https://news.netcraft.com/archives/category/web-server-survey/>
- Oberg, A., Schoellhorn, T., & Woywode, M. (2009). Isomorphism in organizational self-representation in the world wide web? Institutionalization process regarding internet presentation of organizations. *Academy of Management Proceedings*, 2009(1), 1–6. <https://doi.org/10.5465/ambpp.2009.44246572>
- Orlikowski, W. J., & Scott, S. V. (2014). What happens when evaluation goes online? Exploring apparatuses of valuation in the travel sector. *Organization Science*, 25(3), 868–891. <https://doi.org/10.1287/orsc.2013.0877>
- Parker, O., Krause, R., & Devers, C. E. (2019). How firm reputation shapes managerial discretion. *Academy of Management Review*, 44(2), 254–278. <https://doi.org/10.5465/amr.2016.0542>
- Pina, K., & Tether, B. S. (2016). Towards understanding variety in knowledge intensive business services by distinguishing their knowledge bases. *Research Policy*, 45(2), 401–413. <https://doi.org/10.1016/j.respol.2015.10.005>
- Porac, J. F., Thomas, H., & Baden-Fuller, C. (1989). Competitive groups as cognitive communities: The case of Scottish knitwear manufacturers. *Journal of Management Studies*, 26(4), 397–416. <https://doi.org/10.1111/j.1467-6486.1989.tb00736.x>
- Poschmann, P., Goldenstein, J., Büchel, S., & Hahn, U. (2023). A vector space approach for measuring relationality and multidimensionality of meaning in large text collections. *Organizational Research Methods*, 1–31. <https://doi.org/10.1177/10944281231213068>
- Powell, W. W., Horvath, A., & Brandtner, C. (2016). Click and mortar: Organizations on the web. *Research in Organizational Behavior*, 36, 101–120. <https://doi.org/10.1016/j.riob.2016.07.001>
- Powell, W. W., Oberg, A., Korff, V., Oelberger, C., & Kloos, K. (2017). Institutional analysis in a digital era: Mechanisms and methods to understand emerging fields. In G. Krücken, C. Mazza, R. Meyer, & P. Walgenbach (Eds.), *New themes in institutional analysis* (pp. 305–344). Edward Elgar Publishing.
- Quattrone, P., Ronzani, M., Jancsary, D., & Höllerer, M. A. (2021). Beyond the visible, the material and the performative: Shifting perspectives on the visual in organization studies. *Organization Studies*, 42(8), 1197–1218. <https://doi.org/10.1177/01708406211033678>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2), 1–40. <https://doi.org/10.18637/jss.v091.i02>
- Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1), 101–115. <https://doi.org/10.1086/715162>
- Santos, F. P. (2019). Websites and the discursive legitimization of new ventures: Embracing conformity and distinctiveness. In F.-X. de Vaujany, A. Adrot, E. Boxenbaum, & B. Leca (Eds.), *Materiality in institutions* (pp. 223–253). Springer International Publishing.
- Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, 22(4), 941–968. <https://doi.org/10.1177/1094428118773858>
- Shamsie, J. (2003). The context of dominance: An industry-driven framework for exploiting reputation. *Strategic Management Journal*, 24(3), 199–215. <https://doi.org/10.1002/smj.291>
- Shermon, A., & Moeen, M. (2022). Zooming in or zooming out: Entrants' product portfolios in the nascent drone industry. *Strategic Management Journal*, 43(11), 2217–2252. <https://doi.org/10.1002/smj.3407>
- Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, 40(4), 1035–1056. <https://doi.org/10.25300/MISQ/2016/40.4.11>
- Sillince, J. A. A., & Brown, A. D. (2009). Multiple organizational identities and legitimacy: The rhetoric of police websites. *Human Relations*, 62(12), 1829–1856. <https://doi.org/10.1177/0018726709336626>

- Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., & Shaw, J. D. (2019). New ways of seeing big data. *Academy of Management Journal*, 62(4), 971–978. <https://doi.org/10.5465/amj.2019.4004>
- Snijders, T. A. B. (2017). Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, 4(1), 343–363. <https://doi.org/10.1146/annurev-statistics-060116-054035>
- Spee, A. P., & Jarzabkowski, P. (2011). Strategic planning as communicative process. *Organization Studies*, 32(9), 1217–1245. <https://doi.org/10.1177/0170840611411387>
- Stone, R. W., Baker-Eveleth, L., & Eveleth, D. (2015). The influence of the firm's career-website on job-seekers' intentions to the firm. *International Journal of Human Resource Studies*, 5(3), 111. <https://doi.org/10.5296/ijhrs.v5i3.8172>
- The European Parliament & The Council of the European Union. (2019). DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL—of 17 April 2019—On copyright and related rights in the Digital Single Market and amending Directives 96/ 9/ EC and 2001/ 29/ EC. *Official Journal of the European Union*, 34, 1–34. <http://data.europa.eu/eli/dir/2019/790/oj>
- Trabelsi, S., Labelle, R., & Dumontier, P. (2008). Incremental voluntary disclosure on corporate websites, determinants and consequences. *Journal of Contemporary Accounting & Economics*, 4(2), 120–155. [https://doi.org/10.1016/S1815-5669\(10\)70032-1](https://doi.org/10.1016/S1815-5669(10)70032-1)
- Vaara, E., & Fritsch, L. (2022). Strategy as language and communication: Theoretical and methodological advances and avenues for the future in strategy process and practice research. *Strategic Management Journal*, 43(6), 1170–1181. <https://doi.org/10.1002/smj.3360>
- Wruk, D., Schöllhorn, T., & Oberg, A. (2020). Is the sharing economy a field? How a disruptive field nurtures sharing economy organizations. In I. Maurer, J. Mair, & A. Oberg (Eds.), *Research in the sociology of organizations* (pp. 131–162). Emerald Publishing Limited.
- Yu, W., Minniti, M., & Nason, R. (2019). Underperformance duration and innovative search: Evidence from the high-tech manufacturing industry. *Strategic Management Journal*, 40(5), 836–861. <https://doi.org/10.1002/smj.2988>
- Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness: Broadening the interface between institutional theory and strategic management. *Strategic Management Journal*, 38(1), 93–113. <https://doi.org/10.1002/smj.2589>

## Authors' Biographies

**Richard F.J. Haans** is an Associate Professor of Strategic Management and Entrepreneurship at the Rotterdam School of Management, Erasmus University Rotterdam. His primary research interest is competitive dynamics, specifically, the question of how different organizations (should) strive to be different from competitors to attain optimal performance. He is also interested in methodological advances, such as those pertaining to curvilinear relationships and text analysis using machine learning.

**Marc J. Mertens** is a Research Associate and Ph.D. Candidate at the University of Mannheim, where he works at the Chair of Strategic and International Management. His research centers on stakeholder strategy, with a particular focus on strategic interactions between firms and their key stakeholders, such as activist hedge funds. He holds an M.Sc. degree from the Rotterdam School of Management, Erasmus University, and a B.Sc. from Maastricht University.