Dahlke, Johannes et al.

Working Paper

The WebAI paradigm of innovation research: Extracting insight from organizational web data through AI

ZEW Discussion Papers, No. 25-019

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Dahlke, Johannes et al. (2025) : The WebAI paradigm of innovation research: Extracting insight from organizational web data through AI, ZEW Discussion Papers, No. 25-019, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/319890

// NO.25-019 | 04/2025

DISCUSSION PAPER

// JOHANNES DAHLKE, SEBASTIAN SCHMIDT, DAVID LENZ, JAN KINNE, ROBERT DEHGHAN, MILAD ABBASIHAROFTEH, MORITZ SCHÜTZ, LUKAS KRIESCH, HANNA HOTTENROTT, UMUT NEFTA KANILMAZ, NILS GRASHOF, ARASH HAJIKHANI, LINGBO LIU, MASSIMO RICCABONI, PIERRE-ALEXANDRE BALLAND, MARTIN WÖRTER, AND CHRISTIAN RAMMER

The WebAI Paradigm of Innovation Research: Extracting Insight From Organizational Web Data Through AI

Leibniz Association

ZEW

The WebAI Paradigm of Innovation Research: Extracting Insight from Organizational Web Data Through AI

Johannes Dahlke*[1,2,3], Sebastian Schmidt[1,4,5], David Lenz[1,6], Jan Kinne[1,7,8], Robert Dehghan[1,9], Milad Abbasiharofteh[10], Moritz Schütz[11], Lukas Kriesch[11], Hanna Hottenrott[7,12], Umut Nefta Kanilmaz[4], Nils Grashof[13], Arash Hajikhani[14,15], Lingbo Liu[8], Massimo Riccaboni[16,17], Pierre-Alexandre Balland[18,19], Martin Wörter[3], and Christian Rammer[7]

[1] ISTARI.AI, Mannheim, Germany
[2] University of Twente, Enschede, Netherlands
[3] KOF, Department of Management, Technology, and Economics, ETH Zurich, Zürich, Switzerland
[4] Department of Geoinformatics – Z_GIS, University of Salzburg, Austria
[5] Geosocial Artificial Intelligence, IT:U Interdisciplinary Transformation University Austria, Linz, Austria
[6] Department of Economics, University of Gießen, Germany
[7] Department of Economics of Innovation and Industrial Dynamics, Leibniz Centre for European Economic Research (ZEW), Mannheim, Germany
[8] Center for Geographic Analysis, Harvard University, Cambridge, USA
[9] Department of SME Research and Entrepreneurship, University of Mannheim, Germany
[10] Faculty of Spatial Sciences, University of Groningen, Netherlands
[11] Department of Economic Geography, University of Gießen, Germany
[12] Department of Economics & Policy, Technical University Munich, Munich, Germany
[13] Faculty of Economics and Business Administration, Friedrich Schiller University Jena, Jena, Germany
[14] Quantitative Science and Technology Studies, VTT Technical Research Centre of Finland, Espoo, Finland
[15] School of Business and Management, LUT University, Lappeenranta, Finland
[16] IMT School for Advanced Studies, Lucca, Italy
[17] University School for Advanced Studies IUSS, Pavia, Italy
[18] Centre for European Policy Studies, Brussels, Belgium
[19] Growth Lab, John F. Kennedy School of Government, Harvard University, US
Emails: j.dahlke@utwente.nl, sebastian.schmidt@istari.ai, david.lenz@istari.ai, jan.kinne@istari.ai, robert.dehghan@istari.ai, m.abbasiharofteh@rug.nl, moritz.schuetz@geogr.uni-giessen.de, lukas.kriesch@geogr.uni-giessen.de, hanna.hottenrott@tum.de, umutnefta.kanilmaz@plus.ac.at, nils.grashof@uni-jena.de, arash.hajikhani@vtt.fi, lingboliu@fas.harvard.edu, massimo.riccaboni@imtlucca.it, pierre-alexandre.balland@ceps.eu, mawoerte@ethz.ch, christian.rammer@zew.de

April 16, 2025

Abstract

This paper introduces the WebAI paradigm as a promising approach for innovation studies, business analytics, and informed policymaking. By leveraging artificial intelligence to systematically analyze organizational web data, WebAI techniques can extract insights into organizational behavior, innovation activities, and inter-organizational networks. We identify five key properties of organizational web data (vastness, comprehensiveness, timeliness, liveliness, and relationality) that distinguish it from traditional innovation metrics, yet necessitate careful AI-based processing to extract scientific value. We propose methodological best practices for data collection, AI-driven text

*Corresponding author (j.dahlke@utwente.nl)

analysis, and hyperlink network modeling. Outlining several use cases, we demonstrate how WebAI can be applied in research on innovation at the micro-level, technology diffusion, sustainability transitions, regional development, institutions and innovation systems. By discussing current methodological and conceptual challenges, we offer several propositions to guide future research to better understand i) websites as representations of organizations, ii) the systemic nature of digital relations, and iii) how to integrate WebAI techniques with complementary data sources to capture interactions between technological, economic, societal, and ecological systems.
Keywords: *web data; artificial intelligence; innovation studies; research methods.*
JEL Classification: *C81; C45; B4; O3; R1.*

1 The WebAI Paradigm of Innovation Research

Innovation research faces the challenge of rapidly changing technological, societal, and organizational developments. This affects how we can study drivers of innovation processes as well as innovation-driven economic development at the level of entire economies, industries, regions, or individual organizations (Castellaci et al., 2005; Martin, 2016). For understanding socio-technical change, it is key to analyze how innovations emerge and spread across individual actors (Fagerberg, 2006; Geels, 2020). Organizations—ranging from start-ups and small firms to large corporations, and from universities and research institutes to government agencies and non-profit organizations—are the main entities that develop, implement and use innovations. While organizational change captures these innovation processes, they are stable enough as units of analysis to serve as reference point for this change. Tracking their innovative activities, development, and interactions with other organizations, which manifest as complex economic systems (Foster, 2006), over time and space is the central methodological challenge in innovation research.

Traditional approaches have relied on established indicators for innovation activities such as patents, publications, and survey data (Archibugi & Planta, 1996; Fleming & Sorenson, 2004). However, these sources generally provide only periodic snapshots of organizational development and offer incomplete information about innovation in them. They are typically biased towards larger entities and formal innovation activities, with a particular focus on technological innovations and the front-end of the innovation process (Bogers, Garud, et al., 2022). They struggle to capture forms of 'soft innovations' (Castaldi, 2018; Stoneman, 2010) as well as the diffusion of innovations across organizations. This is also due to the increasingly distributed nature of innovation processes, where organizational boundaries become fluid and span multiple levels between individual organizations and the regimes they are embedded in (Bogers et al., 2017). Traditional approaches of surveying innovation often fail to document the inter-linkages between innovating organizations.

The growing abundance of web data offers a promising resource to address these limitations, enabling researchers to study economic activities and change at unprecedented granularity (Bogers, Garud, et al., 2022; Breithaupt et al., 2024; Rammer & Es-Sadki, 2023). Web data describes the publicly available content of organizational websites (i.e., websites

of companies, research institutions, government authorities, non-profit organizations, regional planning organizations etc.). A website generally consists of a start page and numerous subordinate webpages (Kinne & Axenbeck, 2020). This data commonly represents textual content on organizational activities ranging from the strategic to the operational level. It also comprises technical elements featured in the HyperText Markup Language (HTML) code that can be used to source metadata about the organization, as well as relational data between organizations (e.g., hyperlinks between websites).

The inherent properties of organizational web data, however, can render its processing and interpretation somewhat 'messy' (Vaughan et al., 2007). The different data structures, the breadth of information, as well as sheer scale and dynamic nature of web data pose a central challenge to consistently separate signal from noise. It is, henceforth, crucial to develop robust methodologies that can systematically process and interpret web data. Current artificial intelligence (AI) models based on supervised and unsupervised machine learning (ML) and natural language processing (NLP) techniques are useful instruments to detect patterns within vast datasets. These models help to distill meaningful information on organizational representation grounded in web data by distinguishing individual organizational practices from broader industry tendencies and superficial rhetoric from substantive organizational activities. In this way, AI technology helps to move beyond the surface-level of what organizations present online and detect latent patterns that represent substantial aspects of organizational activity.

The **WebAI paradigm** describes the application of AI models to large-scale organizational web data (ISTARI.AI, 2025). It transcends basic web scraping and lexicometric approaches by leveraging ML to systematically extract, validate, and interpret economic indicators from unstructured web content across multiple dimensions of organizational activity. This paper introduces this WebAI paradigm and provides theoretical and methodological foundations for employing WebAI techniques in innovation studies.

We structure this paper as follows: In Section 2, we discuss the fundamental properties of web data and their value as a data source. In Section 3, we present the technical procedures for deriving robust indicators from web data. In Section 4, we showcase the current state of research applying WebAI techniques in different scientific use cases. In Section 5, we discuss the limitations of WebAI and make propositions for future research. Finally, Section 6 concludes.

2 Properties of Organizational Web Data

Big data is traditionally characterized by the core concepts of volume (sheer quantity), velocity (rate of change), variety (heterogeneity of formats) (De Mauro et al., 2015; Laney, 2001), with additional dimensions bearing relevance (veracity, value, exhaustivity, relationality) (Kitchin & McArdle, 2016). While these concepts are technically accurate, they remain abstract unless interpreted in the context of specific data environments. For developing appropriate processing approaches, it is essential to understand what these di-

mensions entail in relation to organizational web data. The context of innovation, socio-technical change necessitates the development of new terms that better capture the nature of organizational web data. That is, to inform how it can be used and how it must be processed to observe, model, and interpret organizational activity and transformation.

2.1 Vastness

The large volume of organizational web data entails an unprecedented number of observations. Beyond scale, the wide scope of observations captures organizations of diverse sizes and types—from the local bakery to schools, hospitals, or multinational corporations. The resulting *vastness*, which we understand as a large volume exhibiting a wide scope, offers a more exhaustive depiction of entire socio-economic systems rather than sampling a small fraction of the population of organizations (Kitchin & McArdle, 2016). This contrasts with traditional data sources on organizational activities. Annual reports primarily cover large and publicly listed companies; patents are biased towards larger organizations from sectors where patent protection is a viable strategy. Although business surveys can be designed to be representative of the business population, they cover only a small part of the total population, which makes them less suitable for certain research questions (e.g., network analysis). In addition, the results are often available only after a considerable time lag and may suffer from response bias.

This vastness of web data is based on the assumption that most relevant organizations maintain a website. However, the rationale behind describing organizational web data as *vast* versus *exhaustive* is that it is still biased through self-selection. Previous studies of the German company population have shown that around 50% of the organizations listed in commercial registers maintain a website (Kinne & Axenbeck, 2020).[1] The proportion of companies with their own website varies depending on the type of company. Larger (more than 5 employees) and older (older than 2 years) companies are much more likely to operate their own website actively, especially if they belong to technology-related sectors such as mechanical engineering or information and communications technology (ICT). Conversely, non-operational entities without the need for public interaction with external stakeholders, such as certain holding structures, rarely maintain a website. In particular, some small Business-to-Consumer (B2C) companies rely solely on social media profiles. However, over 95% of organizations with characteristics that are particularly relevant for innovation studies (i.e., firms with employees, operating in manufacturing or knowledge-intensive service sectors, economically active) have their own website (Kinne & Axenbeck, 2020).

2.2 Comprehensiveness

The second key property of organizational web data connects to the value dimension of the data retrieved (Marr, 2015), which comes from the *comprehensiveness* that lies in

[1] Note that this population includes one-person businesses.

the informational content of organizational websites. Fundamentally, websites are self-curated representations of organizations. Although inherently self-reported, they are the product of a multi-stakeholder engagement within organizations, involving contributions from different departments and hierarchical levels (Oertel & Thommes, 2018). Furthermore, organizational websites are designed to present the respective entity to diverse stakeholder groups—from investors and customers to potential employees and the general public (Powell et al., 2016). Thus, websites serve as one of the most comprehensive representations of organizational identity (Oertel & Thommes, 2018). They reflect both explicit economic activities (such as product offerings, market presence, partnerships) and implicit signals about firm strategies, technological capabilities, and market positioning. This dual perspective allows researchers to study co-evolving patterns in organizational change, innovation systems, and market dynamics.

In the context of innovation research, organizational web data captures technologies and innovations at various stages—ranging from pre-product concepts to early-stage and mature products that have been successfully commercialized and integrated into business models and organizational identities (Dahlke et al., 2024). For example, a newly founded start-up may showcase a vision of a technology that it has yet to develop, while an established company may report on the latest additions to its fully developed product portfolio.

2.3 Timeliness

The *timeliness* of organizational websites based on continuous data streams and updates allows near real-time monitoring of organizations using WebAI techniques, which can capture socio-economic activities and interactions as they unfold. It leverages the velocity of web data in its timely production and availability, which is enabled by the technical infrastructure of the modern web (Marres & Weltevrede, 2013). Providing fresh data flows has become deeply embedded in contemporary web architecture, from instant updates on social media platforms to search engines' increasing prioritization of recent content. In the context of innovation research, it means the capacity to reflect socio-economic activities and discourses in near real-time. Compared to traditional data from surveys or patent offices, it is not hampered by lags caused by procedures of collecting, processing, and publishing relevant information. This is particularly interesting in the context of new and emerging technologies but also during times requiring rapid-responses to crises (Buchanan & Denyer, 2013).

2.4 Liveliness

Beyond this immediate temporal dimension lies a second, more complex property: the *liveliness* of web data. Liveliness describes the capacity of web data to mirror the dynamic and evolving landscape of societal issues and industrial engagement over time. Rather than capturing what is happening currently, liveliness reveals how issues, topics, and engagements evolve, fluctuate, and transform (Marres & Weltevrede, 2013). This property enables researchers to observe meaningful patterns in how socio-economic activities develop, how

networks of associations shift, and how engagement with particular topics varies over time (Oberg et al., 2022). The technical architecture of the web, with its Uniform Resource Locators (URLs), hyperlinks, and timestamps paired with the richness of unstructured (textual) data makes it possible to trace these evolutionary patterns and transformations.

Although both properties benefit from the strong incentive for organizations to keep their websites constantly updated, the distinction between timeliness and liveliness is not merely theoretical but has profound implications for how we apply WebAI techniques. While timeliness demands technical capabilities for continuous data capture and processing, liveliness requires methods to detect changes over time, track variations, and understand long-term dynamics. In this respect, the development and maintenance of web archives that store older versions of websites have become a critical tool for the analysis of liveliness (Schafer & Winters, 2021). Together, these two properties make web data a valuable resource for studying innovation and societal transitions, offering both immediate insights into current developments, and deeper understanding of how issues and engagements evolve over time.

2.5 Relationality

The capacity of web data to capture meaningful relationships at the individual level (e.g. through social media data) has long been discussed (Boyd & Crawford, 2012). At the level of organizations, *relationality* of web data can be understood through both direct and indirect dimensions. Direct relations between organizations can be captured by retrieving hyperlink connections between websites as well as by detecting textual references to other organizations. Hyperlinks have been shown to represent direct relations between organizations (Vaughan et al., 2007) and to be economically relevant (Vaughan & Wu, 2004). In the context of innovation studies, hyperlink networks have been shown to exhibit emergent properties and scaling characteristics of innovation systems (Katz & Cothey, 2006), suggesting their potential in capturing information flows between organizations.

Indirectly, organizational websites relate to one another through the comparability of their content (i.e., their similarity). The inherent technical structures and socio-economic functions of websites as communication channels offer the opportunity to source data on various topics from various types of organizations (such as public and private ones) in a consistent way (Oberg et al., 2022), which would otherwise be fragmented across different data bases and formats. While the veracity of big data also applies to the raw information scraped from (differently formatted and curated) websites, websites can offer a rather consistent unit of observation across a vast population of organizations, especially when processed adequately.

3 Methodologies for Extracting and Processing Organizational Web Data

Transforming the vast, unstructured, and dynamic nature of organizational web data into reliable research insights necessitates a robust and carefully considered methodological pipeline (Figure 1). This chapter presents an overview of the most important analysis steps—from the identification of relevant organizations to data retrieval, processing through AI and ML techniques, and validation.

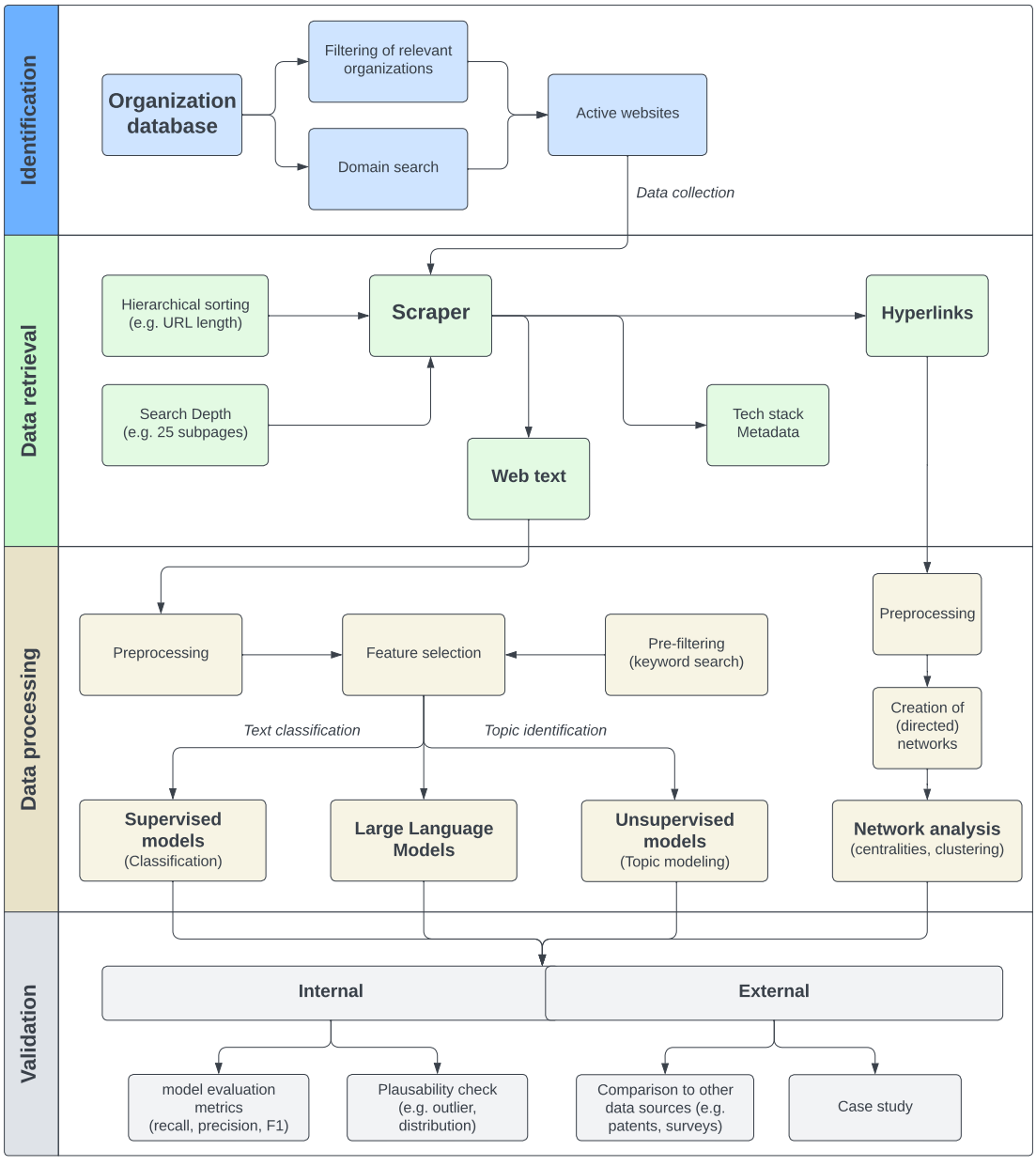

Figure 1: The WebAI pipeline: From organizational websites to validated insights.

3.1 Identification of organizational websites

The initial step in carrying out analyses according to the WebAI paradigm is to create a comprehensive base dataset of organizations and their associated websites. The quality of this foundation directly determines the scope and validity of subsequent analyses. One approach is to use company databases (e.g. ORBIS, S&P, Infogroup), which often contain not only classic information such as revenue, number of employees, and the address of a company but also its URL (Kinne & Axenbeck, 2020). Other sources may include commercial registries, member lists of industrial associations, databases with grant recipients, or open source point of interest (POI) databases. Depending on the research question and database content, the initial number of websites can be reduced by pre-filtering (e.g., based on sector, size, or region), minimizing scraping effort and enhancing subsequent AI-based information retrieval. However, researchers must be aware of potential limitations: coverage gaps (especially for small and medium-sized enterprisess (SMEs) or newer firms), outdated URLs, and potential biases inherent in how the database was compiled (Nathan & Rosso, 2015; Rammer & Es-Sadki, 2023).

In the absence of available URLs for organizations of interest, automated searches using search engine Application Programming Interfaces (APIs) (e.g., Google Search API, Bing API) can identify potential candidates based on organizational names and addresses. This often requires sophisticated fuzzy matching or approximate string matching algorithms (Navarro, 2001) to handle variations in naming conventions and address formats, a process related to the broader field of record linkage (Fellegi & Sunter, 1969). The returned candidates can be evaluated based on their contents, i.e. comparison against the known information about the company at hand. For studies requiring high precision, particularly those involving niche organizations or specific sectors, manual verification or even curation of a URL list may be necessary. Often, a hybrid approach combining automated URL discovery with targeted manual validation yields the best balance between scale and accuracy, reflecting principles of human-in-the-loop system design (Mosqueira-Rey et al., 2023).

3.2 Web scraping: Efficient data retrieval from organizational websites

Once the target organizational websites are identified, the next crucial step is to retrieve their content through web scraping. This process involves retrieving the raw HTML source code from a potentially large number of websites. A broad web scraping algorithm can process diverse types of websites and website contents, unlike 'focus' web scraping algorithms which target specific information (e.g. prices and product names from a single e-commerce website). An exemplary open-source solution is Automated Robot for Generic Universal Scraping (ARGUS), a web scraping framework that offers customizable scraping depth (i.e. the number of webpages that will be collected), automatic redirect detection, and parallel processing capabilities for efficiency. Several established libraries for creating custom web scraping solutions exist, e.g. *Scrapy* or *Beautiful Soup*.

While previous research suggests a scrape limit of 250 pages per website captures

around 90% of corporate website content (Kinne & Axenbeck, 2020), researchers can significantly reduce this by implementing more resource-efficient, strategic scraping approaches with targeted URL pattern matching (e.g., for subpages containing '/products/', '/technology/', '/research/'), if properly informed by their empirical focus. Excluding subpages with a legal focus, duplicates in other languages, or subpages dedicated to site functionality can be considered, since they generally provide little additional information (Haans & Mertens, 2024). For websites heavily reliant on dynamic content loaded via JavaScript, tools controlling headless browsers (e.g., Selenium, Puppeteer) may be necessary, though often slower and more resource-intensive. A targeted scraping approach reduces computational overhead while maintaining data quality, as innovation-relevant information tends to be concentrated in specific website sections. To balance coverage and cost, we thus recommend implementing an adaptive depth strategy, starting with a 25-50 page limit augmented by targeted URL matching (Kinne & Axenbeck, 2020). Unlike traditional approaches, language preferences are less critical when using modern Large Language Models (LLMs) and embedding techniques, which can handle multilingual content effectively, at least for high-resource languages (Li et al., 2024). However, considering language tags in the HTML head of websites can still be useful as a potential filtering mechanism. Furthermore, implementing comprehensive logging and proper error handling for timeouts, connection issues, or redirects is essential, as some domains may be temporarily unreachable or inactive, which could disrupt data collection.

Another approach to acquire textual data from organizational websites is to extract information from pre-crawled webcorpora, such as Common Crawl (Kriesch, 2023). Since CommonCrawl has archived a large number of historical websites, retrospective time series analyses of the same website are possible, which can be used e.g. to map trends in relation to certain technologies or to identify domains associated with organizations. However, researchers must be prepared to handle potential data inconsistencies, archival gaps, and varying coverage quality across time and sources (Thelwall & Vaughan, 2004).

3.3 Processing web text data

The results of web scraping are large volumes of raw HTML, i.e. relatively noisy and unstructured text data, that requires automated processing. ML and AI methods lend themselves very well to the extraction of meaningful, standardized information from such data (Gentzkow et al., 2019). This involves several stages, from initial cleaning to applying sophisticated analytical models. The most important methods for a WebAI analysis are presented in the following subchapters.

3.3.1 Web text pre-processing

While state-of-the-art LLMs have demonstrated a remarkable ability to process raw or minimally processed HTML directly (Tan et al., 2025), careful pre-processing remains highly valuable, particularly for reducing the amount of text to be process and thus optimizing computational efficiency and lowering computational costs. Encoder-based models still

benefit from structured pre-processing steps, including noise reduction, removal of boilerplate content and text chunking (Penedo et al., 2025; Wenzek et al., 2019). This is because these models lack the extensive contextual reasoning capabilities of autoregressive LLMs and rely more on clean, structured input for optimal performance (Devlin et al., 2019; Raffel et al., 2020).

Effective pre-processing aims to reduce noise and structure the data without losing vital information, i.e. reduce the amount of data that contain no relevant information (e.g. cookie or data privacy texts). Common steps include parsing the HTML to extract the main textual content while potentially preserving structural tags (`<p>`, headings, lists) as semantic cues. Identifying and removing boilerplate elements such as navigation menus, headers, footers, and consent banners can also improve subsequent model performance. For models with fixed input lengths, text chunking is necessary; techniques like semantic chunking, which splits texts based on topical coherence using embeddings, are often superior to fixed-size methods as they better preserve context (Qu et al., 2024). Further noise reduction may include filtering out overly short texts through length thresholds. Considering the symbol-to-word ratio, specific stop words, filtering out documents in which a predefined proportion of lines begin with a bullet point or end with an ellipsis, as well as imposing a threshold requiring a minimum proportion of words to contain at least one alphabetic character, can also enhance the quality of the text database (Rae et al., 2021). However, there is always a trade-off between the level of quality of the pre-processing results of website data, i.e. formatting consistency, and the ability to generalize to a variety of websites across industries, countries, and cultures.

Effective feature selection is a crucial step in the pre-processing of web texts, ensuring that only the most relevant and informative content is retained for further analysis (Thirumoorthy & Muneeswaran, 2022). Feature selection often involves keyword-based methods to identify and filter content based on specific terms, such as domain-specific terminology relevant to the research. Additionally, frequency-based keyword filtering can ensure that only texts with a sufficient number of relevant terms are included, while distinct keyword-based filtering assesses the diversity of unique terms within a text to avoid overly repetitive content (Z. Wang et al., 2021). A significant, emerging challenge is the increasing prevalence of AI-generated content online. Future pre-processing pipelines must incorporate methods to detect and handle such potentially low-substance text to maintain data integrity.

3.3.2 Supervised approaches

When the research objective involves identifying specific, predefined concepts within the web text (e.g., classifying firms by technology adoption), supervised learning offers a powerful approach. Modern NLP models predominantly leverage fine-tuning models based on the Transformer architecture, which was introduced by Vaswani et al. (2017). This architecture revolutionized the field by utilizing self-attention mechanisms, enabling models to process sequences more efficiently and capture long-range dependencies in text (Cui

et al., 2019). The Transformer architecture serves as the foundation for advanced models such as Bidirectional encoder representations from Transformers (BERT)—including the widely used modifications *Sentence-BERT* (Reimers & Gurevych, 2019), *RoBERTa* (Y. Liu et al., 2019) and *DistilBERT* (Sanh et al., 2020) -, Generative Pre-trained Transformer (GPT) (Radford et al., 2018), and T5 (Raffel et al., 2020), which are capable of various natural language understanding and text generation tasks. To apply these models to organizational web data, the established procedure is to fine-tune pre-trained architectures. For this purpose, training data must be created to provide the models with domain-specific knowledge.

Selecting the appropriate models and training methodologies is critical for effectively processing organizational website texts. Multilingual models are crucial for cross-language applications, whereas language-specific models may yield more accurate results for targeted analyses due to their tailored training data (Zhu et al., 2024). However, selecting the appropriate model involves trade-offs, such as balancing computational efficiency with accuracy and customizing models for domain-specific language. As the success of fine-tuning is highly dependent on the quality of training data, it is important to follow general best practices, ensuring that datasets are consistent, representative and relevant. To this end, clear labeling guidelines should be developed and inter-annotator agreement considered. Implementing active learning strategies can further enhance efficiency by focusing on the most uncertain or ambiguous samples, thereby reducing the overall labeling effort (Mosqueira-Rey et al., 2023). Moreover, iterative model refinement ensures continuous performance improvement through repeated evaluation and adjustment. If resources are limited, adopting sample-efficient fine-tuning frameworks such as SetFit can significantly reduce the need for large labeled datasets, making model adaptation more efficient and scalable. This is achieved by fine-tuning a pre-trained Sentence Transformer on a small set of text pairs in a contrastive manner, generating rich embeddings for a lightweight classification head, and enabling fast, prompt-free adaptation with far fewer parameters than existing methods (Tunstall et al., 2022).

3.3.3 Unsupervised approaches

To explore underlying, coherent topics or themes from a web data corpus without relying on predefined categories, unsupervised ML methods are highly valuable (Churchill & Singh, 2022). In the case of unstructured text data, this is often referred to as topic modeling. Traditional probabilistic methods like Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Maier et al., 2018) remain useful, but more recent approaches such as BERTopic, which leverage contextualized text embeddings for clustering, often capture semantic nuances more effectively (Grootendorst, 2022). For more fine-grained analysis, hierarchical topic modeling can be employed, i.e. the implementation of various subsequent topic modeling steps, allowing for the exploration of broader topics and related subtopics (Xu et al., 2018). Additionally, dynamic topic models can capture how topics evolve over time, which makes them especially useful for tracking shifts in themes or trends across different

time periods (Blei & Lafferty, 2006). Hierarchical topic modeling approaches applied to web-scraped textual data have successfully linked firm-level content with established scientific knowledge bases, showcasing the benefits of integrating NLP methodologies into organizational research for detailed sectoral analysis (Hajikhani et al., 2022).

However, topic modeling faces some challenges such as selecting the optimal number of topics, ensuring the interpretability of results, maintaining topic stability, and managing memory efficiency (Kherwa & Bansal, 2018). Another issue lies in the hard assignment of document to a single dominant topic, which can introduce biases when constructing quantitative indicators. Soft assignments or mixed-memberships (e.g., the topic proportions from LDA (Blei et al., 2003)), which represent documents as distributions over topics, might offer a more nuanced basis for such analyses. To address some of these challenges, the integration of supervised and unsupervised learning offers a complementary approach that enhances both model interpretability and performance. While unsupervised topic models autonomously identify latent structures in text, they often require manual tuning to determine the optimal number of topics. Supervised methods, on the other hand, can guide topic modeling toward more stable and semantically meaningful outcomes but rely on labeled data. Combining these approaches improves topic stability and memory efficiency, leading to more robust and scalable solutions (Talukdar & Biswas, 2024).

3.3.4 Large Language Models (LLMs)

Recent advances in AI and ML research have enabled more personalized and interactive analysis methods for text data. A key development in this field is the rise of LLMs, which have gained significant attention, especially since the release of ChatGPT in November 2022. The term LLM encompasses a diverse range of large neural network models, but current developments, particularly those gaining widespread attention (e.g., GPT-4, Claude 3, Llama 3), are predominantly autoregressive, decoder-heavy models (Radford et al., 2018). These models excel at generative tasks due to their architecture, making them highly capable of complex question-answering requiring synthesis, summarization, translation, creative text generation, and synthetic data generation (Singhal et al., 2025). Their strength lies in few-shot or zero-shot learning via sophisticated prompt engineering (P. Liu et al., 2023), allowing for flexible application without task-specific fine-tuning. In this regard, careful prompt design (e.g., using techniques like chain-of-thought) and selecting appropriate context window parameters are crucial, as they determine how much of the surrounding text the model considers, which can significantly affect the coherence and relevance of the generated response (Chen et al., 2023). Additionally, incorporating Retrieval-augmented Generation (RAG) frameworks allows LLMs to access and integrate external knowledge sources, which can improve the reliability of generated content (Wu et al., 2025).

Despite their capabilities, LLMs are susceptible to so-called hallucinations, i.e. the fabrication of facts (H. Liu et al., 2024), and have been shown to exhibit various biases from training data (Gallegos et al., 2024). Consequently, choosing between general-purpose and

specialized language models requires careful consideration, as general-purpose models such as GPT-4 or LLaMA may introduce biases or sensitivity issues that specialized models mitigate more effectively, particularly in tasks requiring precise and accurate content interpretation (Hajikhani & Cole, 2024). It should also be noted that generative AI solutions for text classification are (still) quite expensive for use in research contexts. In contrast, training a customized BERT model is much cheaper in terms of computation, time, and emissions (Laurer, 2024). Another challenge is ensuring reproducibility of analyses based on closed-source LLMs due to limited access to model parameters. This makes it difficult to verify results, replicate experiments, or assess the impact of different training configurations. These limitations mandate critical evaluation, extensive validation protocols, and considerable caution when relying on LLMs for research.

3.4 Hyperlink extraction and network construction

Studying collaboration between organizations has long been a core interest in innovation research (Belderbos et al., 2004; Hottenrott & Lopes-Bento, 2016). Studies using co-inventions, co-publications, joint ventures, or R&D agreements consistently show a positive link to innovation performance (Caloghirou et al., 2003; Xie et al., 2023). A key limitation of these approaches is their reliance on outcomes and largely formal collaborations. Organizational web data, however, also allows capturing informal connections through hyperlinks, which can serve as proxies for inter-organizational relationships.

To extract hyperlinks from organizational websites, we recommend a robust pipeline to extract all outgoing links on each webpage, based on the HTML $< a >$ tag and its $href$ attribute, and retain only those pointing to other relevant organizational domains, e.g. excluding social media platforms, news sites, file downloads, or other non-company destinations. Additionally, researchers should implement domain validation to ensure extracted hyperlinks lead to legitimate organizational websites, avoiding parked domains or spam sites. Redirect chains require special attention, as outdated URLs can create misleading connections and should therefore be excluded from the analysis. To determine genuine inter-firm connections, Abbasiharofteh, Krüger, et al. (2023) suggest a weighted approach where reciprocal hyperlinks receive a higher weight than unidirectional hyperlinks, acknowledging the stronger connection implied by mutual recognition. Link strength can also be assessed by frequency, i.e. multiple links between the same organizations. Furthermore, analyzing the ambient text surrounding a hyperlink using NLP techniques can help classify the nature of the connection. They generally fall into three categories: (1) business relations (e.g., partnerships, suppliers, customers), (2) non-business relations (e.g., industry associations, regulatory bodies), and (3) purely informational links (e.g., news references, general information).

The resulting hyperlink network—where organizations are nodes and hyperlinks are edges—can be analyzed using established network science metrics. Important node-level metrics include network centralities, such as the betweenness centrality, which measures a firm's influence on information flow, and the degree centrality, which reflects its poten-

tial reach (Wasserman & Faust, 1994). Clustering coefficients reveal local network density (Latapy et al., 2008), while path lengths, i.e. the distances between network nodes, help assess overall structure and information flow potential. Community detection algorithms can identify clusters of closely connected firms, potentially revealing industry structures, regional business networks, or innovation ecosystems (Abbasiharofteh, Kogler, & Lengyel, 2023). Since firms are localized in geographic space, these analyses can also be interpreted spatially, e.g. highlighting areas with a high concentration of key nodes. In addition to the spatial dimension, a temporal resolution of network data enables tracking link stability over time, where persistent connections might signal stronger relationships. Integrating this spatio-temporal dimension generates large datasets, requiring advanced methods such as Graph ML. At the node level, Graph ML enables the identification of complex spatio-temporal clusters (Y. Zhang & Cheng, 2020), offering insights into how firm connections evolve over geographic space and time. A key application of Graph ML is link prediction (M. Zhang & Chen, 2018), which estimates the likelihood of future or missing connections between nodes based on network structure and attributes.

3.5 Validation of WebAI insights

Establishing the credibility and robustness of insights derived from the WebAI paradigm is essential. This involves both internal assessment of model performance and external comparison against independent benchmarks.

3.5.1 Internal validation

Internal validation focuses on the technical performance of the ML models used, particularly supervised classifiers. Established evaluation metrics such as precision, recall and F1-score can also be applied for the analysis of organizational websites. To ensure the reliability and robustness of ML models, cross-validation techniques are essential for assessing their performance across different data splits. This helps mitigate overfitting and dataset biases. Key methods include K-fold validation, which evaluates model stability by dividing the dataset into multiple folds for training and testing, and out-of-sample validation to verify the model's performance on unseen data, providing an extra layer of robustness (Montesinos López et al., 2022). Beyond quantitative metrics, qualitative error analysis, i.e. systematically examining instances where the model makes mistakes, provides invaluable insights into model weaknesses, potential biases in the data or labeling, and specific areas needing refinement.

3.5.2 External validation

External validation extends beyond model metrics to assess whether the WebAI-derived indicators correspond to or correlate with external, real-world measures or established knowledge. In this context, traditional indicators, such as survey results, patent statistics, R&D expenditures, or official statistics, are valuable baselines, though they may often

cover only a subset of analyzed organizations (Ashouri et al., 2022; Kinne & Lenz, 2021). Having domain experts review samples of model outputs for validity, coherence, and semantic accuracy can help identify subtle errors or misinterpretations. Such plausibility checks should at least include a review of the results and methods by experts, benchmarking against established standards, a baseline, or a historical trend analysis to check consistency over time. For some use cases, validation with complimentary geodata, e.g. remote sensing data for sustainability-related analyses (Schmidt et al., 2022), or news articles might also be sensible.

3.6 Methodological challenges

Beyond limitations specific to methods employed at each processing stage, researchers using the WebAI paradigm must navigate several overarching methodological challenges that can influence the validity and reliability of their findings:

- **Text identification:** Reliably identifying relevant textual content in the noise of diverse website structures remains difficult. Evaluating semantic content and information density, removing non-informative elements (navigation, ads, boilerplate), and standardizing formats across heterogeneous sites are crucial but non-trivial steps (Penedo et al., 2025; Wenzek et al., 2019).

- **Technical limitations:** While vastness is one of the greatest strengths of organizational web data, it also presents practical constraints. Achieving sufficient scraping depth without excessive computational cost requires careful strategies (Kinne & Axenbeck, 2020; Sayles et al., 2022). Handling dynamic content (rendered via JavaScript) often necessitates more complex acquisition methods, while some content formats (e.g., embedded applications, complex PDFs) may remain inaccessible (Valova et al., 2023).

- **Bias and representativeness:** There are several potential biases that need to be considered in evaluating WebAI results. Large companies generally maintain more comprehensive websites, in terms of the sheer quantity and liveliness of textual content and its semantic diversity. Similarly, companies with technology-related business activities generally have a stronger web presence and are therefore more suitable for WebAI-based analyses than companies in traditional industries (e.g. agriculture) (Kinne & Axenbeck, 2020). As urban areas usually have better website coverage due to higher digital adoption rates (Vicente & López, 2011), geography might also be an influencing factor. Given that organizations generally curate their online presence, selective emphasis or omission of information should be expected. Disentangling substantive activity from more strategic communication remains a key challenge.

- **Temporal Consistency:** For longitudinal studies using web archives or repeated scrapes, ensuring methodological consistency over time is paramount. Changes in website design trends, scraping technologies, data processing steps, or AI models

may introduce artificial variations that obscure genuine temporal trends (Haans & Mertens, 2024).

- **Legal and ethical considerations:** Although data acquisition from publicly accessible websites is generally less problematic than e.g. the scraping social media data (Brown et al., 2024; Gilga et al., 2025), some ethical and legal considerations remain essential. This includes compliance with data privacy regulations like General Data Protection Regulation (GDPR) (for data collection and storage in the European Union) and developing robust data anonymization protocols to safeguard sensitive information (Weitzenboeck et al., 2022). Ethical scraping practices aim to minimize the impact on servers and adhere to established guidelines. This involves the implementation of rate limiting to prevent server overload, strict compliance with *robots.txt* directives and site-specific policies, as well as the responsible management of server loads through the use of distributed scraping techniques (Brown et al., 2024).

4 Use Cases for WebAI Methods in Innovation Studies

Innovation lies at the heart of societal progress. In today's world, the process of generating and diffusing innovation is increasingly complex, distributed, and embedded in evolving socio-technical systems (Geels, 2020; Geels et al., 2023). Addressing grand challenges such as climate change, digital transformation, and inequality requires us to understand innovation not as a linear outcome, but as an emergent process shaped by interactions across diverse actors, institutions, and sectors (Foster, 2006; Mainzer, 2009; Rittel & Webber, 1973). Traditional data sources struggle to capture this complexity. The WebAI paradigm—leveraging large-scale web data and advanced AI techniques—provides a framework for studying innovation in a comprehensive way by capturing both innovation-related activities of organizations and links between innovative organizations, while allowing for the characterization of innovation by technologies used, sustainability impacts, and the regional dimension.

The five key properties outlined above make organizational web data uniquely suited to this task. First, its vastness provides broad coverage across organizational types and sizes, enabling system-level insights that go beyond narrow samples (Kinne & Axenbeck, 2020). Second, its comprehensiveness captures rich and self-curated information about organizational identity, strategy, and technological positioning—crucial for tracing innovation trajectories across stages of maturity (Dahlke et al., 2024; Oertel & Thommes, 2018). Third, its timeliness allows for near real-time observation of emerging trends and innovation signals, especially during periods of rapid change (Buchanan & Denyer, 2013; Marres & Weltevrede, 2013). Fourth, liveliness makes it possible to track how discourses, technologies, and organizational roles evolve over time, thereby supporting dynamic analyses of innovation ecosystems (Oberg et al., 2022). Finally, relationality captures both explicit and latent connections between organizations—through hyperlinks or content similarity—revealing the underlying structure and flow of knowledge and influence within

innovation networks (Abbasiharofteh, Krüger, et al., 2023; Katz & Cothey, 2006; Vaughan et al., 2007).

Together, these properties position the WebAI paradigm as a powerful approach to study innovation in the context of socio-economic complexity. The following sections highlight how this approach enables novel insights into the distributed and evolving nature of innovation across different research fields in innovation studies.

4.1 Measuring firm-level innovation in times of rapid change

The innovativeness of a company is traditionally measured on the basis of publications, patents, or through surveys (Archibugi & Planta, 1996). As a precursor to the WebAI paradigm, Gök et al. (2015) find that studying R&D activities in innovative firms by accessing websites from the Wayback Machine and performing a keyword-based analysis can yield valid results. Several studies based on the WebAI paradigm, i.e., the extension of such simple NLP approaches to more complex semantic analysis methods, have since shown that organizational web data is a suitable alternative or complementary data source for innovation research (Kinne & Axenbeck, 2020).

Building on a framework by Kinne and Axenbeck (2020), Kinne and Lenz (2021) develop a methodology to predict the level of innovativeness of a company based on its website text for the first time. They develop a method that uses artificial neural networks to classify firms as product innovators, leveraging data from the German Community Innovation Survey to train the model. Comparing the predictions to traditional benchmarks like patent statistics, survey extrapolation, and regional innovation indicators, they find the web-based approach to be reliable, cost-efficient, and capable of providing greater coverage and regional detail.

Organizational web data has also been shown to be well suited to investigating very specific topics, which can be of particular interest for innovation-related analyses. Focusing on the diffusion of a specific standard within the German firm population, Mirtsch et al. (2021) examine the adoption of the ISO/IEC 27001 information security management standard in Germany, using web mining to analyze the websites of 2,664 firms from the Mannheim Enterprise Panel dataset, finding that larger and more innovative firms are more likely to obtain the certification.

Against the backdrop of the COVID-19 pandemic, a study by Dörr et al. (2022) shows one of the major advantages of the WebAI paradigm, its timeliness. By analyzing corporate websites, they are able to provide early insights into the pandemic's diverse impacts and predict firm-level outcomes, such as credit rating changes. Similarly, Koenig et al. (2025) analyze the content of five million corporate websites worldwide to quantify the extent and nature of the impact of the COVID-19 pandemic on companies. They show that this web-based impact indicator at the firm level is highly correlated with pandemic containment measures and reliably predicts the performance of firms during the pandemic. In fields of innovation such as artificial intelligence, which change quickly and frequently, web data can provide timely information, thereby overcoming for instance the typical de-

lay in patent data. This enables the large-scale monitoring of product innovation activities. Using a similar concept, Borchert et al. (2023) use NLP methods, such as BERT, to integrate website contents into business failure prediction models, finding that this improves model performance.

4.1.1 Tracking emerging technologies and their diffusion

A particular focus in innovation studies is on analyzing new technologies, especially with regard to their emergence and diffusion (Comin & Mestieri, 2014). However, traditional methods, such as patents and surveys, generally capture technological change with significant time lags, making it difficult to track emerging technologies in near real time. Additionally, static patent classifications and predefined survey categories can struggle to keep pace with the rapid advances of new technologies (Martin, 2016). Moreover, product digitization derived from web-scraped organizational data has been empirically associated with productivity improvements among European high-tech companies, highlighting the economic significance of digital transformations captured via website data (Deschryvere et al., 2023). The lively nature of web data offers a unique perspective on how technologies emerge and become (re)defined by the economic actors using them. More than that, the comprehensive information in organizational web data helps to distinguish between different groups of actors engaging with the technology, such as technology producers or technology users, and accounting for the interactions between them. This is particularly important because technology usage often provides deeper insights into economic development than production alone, for example, regarding technology diffusion in networks of adopters or regarding the resulting economic dependencies between producers and users (Franco et al., 2023).

Dahlke et al. (2024) derive a web-based indicator for AI adoption among firms from Germany, Austria, and Switzerland, using textual data from over 1.1 million websites. They also leverage the relationality of web data by constructing a hyperlink network of more than 380,000 firms, which allows them to incorporate social capital and network embeddedness into epidemic models of inter-firm diffusion of AI technology. Based on the developed indicator and constructed networks, they identify three key mechanisms influencing AI adoption: co-location in AI-related industrial and regional hubs, direct exposure to deep AI knowledge; and relational embeddedness in the AI knowledge network. In their study, (Schwierzy et al., 2022) analyze the spread of additive manufacturing in Germany. For this, they distinguish different types of key actors in the diffusion process, including manufacturers, service providers, retailers, and information providers. Additionally, they find indications that the proximity to pioneer firms and technical universities could influence the diffusion of 3D printing. Similarly, Gschnaidtner et al. (2024) analyze companies with blockchain-related products and services in Germany, Austria, and Switzerland. Their findings identify regional blockchain clusters close to important financial centers, despite blockchain being an inherently decentralized and distributed technology.

4.1.2 Discerning sustainability engagement across concepts and space

One of the key challenges in contemporary innovation research is studying the green transition, i.e. the transformation of business models and processes towards more sustainable and environmentally friendly ones (Martin, 2016). Here, two key issues persist: One of scope and one of legitimacy. Regarding the former, the wickedness of today's problems (Ritala, 2024; Rittel & Webber, 1973) requires consideration of a broad range of sustainability aspects and their interrelations. This range is particularly reflected by the UN's Sustainable Development Goalss (SDGs), which address the ecological, economic, and social dimensions of the concept of sustainability (Bennich et al., 2020). The comprehensive content sourced from organizational websites can capture this range across the diverse areas of sustainability, as well as their co-evolution, since companies generally use their websites to present their sustainable products and practices.

Following this WebAI paradigm, Wildnerova et al. (2024) analyze content from over one million firm websites across 15 OECD countries and develop a metric to identify the sustainability engagement of SMEs. They find that approximately one-third of SMEs are sustainability-engaged, i.e. offer sustainability-related products or services, though this varies significantly by country, with solar energy, recycling, and energy efficiency being the most frequently cited actions. Building on the metric of Wildnerova et al. (2024), Kinne et al. (2024) exploit the vastness, comprehensiveness, and relationality of organizational web data to study the twin transition for all US firms by analyzing the co-occurrence of blockchain knowledge and sustainability engagement in the context of firm ecosystems. For this, they create measures based on the hyperlink networks and physical co-location of firms. To quantify location factors (e.g., proximity to transport infrastructure), they use data from OpenStreetMap (OSM). While they find that blockchain remains a niche technology (Gschnaidtner et al., 2024), they show a much higher level of sustainability engagement for Blockchain firms than is exhibited across the general firm population. They also show the high importance of deep embedding in entrepreneurial ecosystems for the promotion of the twin transition. Kriesch and Losacker (2024) use website texts from CommonCrawl to analyze firm activities related to the bioeconomy. They find that such firms are mostly located in rural areas and close to their biomass feedstocks, whereas innovations in the bioeconomy are found in urban areas. In a more technical study, Auzepy et al. (2023) classify climate-related financial disclosures on corporate websites, introducing a zero-shot approach. They find that climate-related reporting has risen, while the extent of the reporting varies considerably. Focusing on the social dimension of sustainability, W. Wang et al. (2023) use structural topic modeling to identify key elements in corporate diversity statements, in particular with respect to racism.

A second issue in measuring sustainability orientation is one of legitimacy. A critical issue that needs to be discussed in the context of sustainability analyses is so-called greenwashing. This refers to the misrepresentation of environmental practices for advertising purposes. Against the backdrop of the ever-increasing importance of sustainability and a 'green image', greenwashing is becoming an increasingly common practice (De Freitas

Netto et al., 2020; Y. Liu et al., 2023). In order to study the application and diffusion of truly sustainable economic practice, the processing of web data by means of AI models takes center stage. Training such models can help to discern superficial engagement from a true commitment to different dimensions of sustainability as displayed in an organization's web data.

Schmidt et al. (2022) address this topic by studying the sustainability engagement of the US metal industry, which they evaluate based on the WebAI paradigm. They find a correlation between the location of the metal industry and sulfur dioxide concentrations in the atmosphere, which they derive from Sentinel-5P satellite data. Using a spatial regression analysis, they place the presence of sustainability-engaged companies in the context of these air pollutants, but find no statistically significant correlation and therefore no evidence of widespread greenwashing. This study serves as a first contribution to evaluating the legitimacy of sustainability reporting on websites through earth observation data, i.e., an external data source.

4.2 Mapping innovation systems through hyperlink networks

The systemic approach to innovation has been conceptualized in the literature with different focal points in national (Lundvall et al., 2002), regional (Asheim et al., 2011) and technological innovation systems (Carlsson & Stankiewicz, 1991), as well as business ecosystems (Gawer & Cusumano, 2014). A common methodological challenge unifying these streams is the difficulty in delineating the boundaries of such systems, as well as of determining measures for their aggregate performance (Carlsson et al., 2002). Due to their complex nature, studies on innovation systems often rely on qualitative approaches confined to narrow contexts, where all elements and outputs of such systems can be mapped (Bergek et al., 2015). Pairing the relational nature of organizational web data with the vast and comprehensive information on its individual elements has the potential to translate the systemic approach to innovation to a larger scale.

First, hyperlinks between organizational websites can be seen as a proxy for collaboration and cooperation between entities. Reminiscent of the basic idea of business clusters (Delgado et al., 2014; Grashof & Fornahl, 2021; Porter, 1998) and innovation networks (Chesbrough & Rosenbloom, 2002; Katz & Cothey, 2006), classical methods of network analysis can be applied to this data. This allows to uncover system boundaries in the geographic and relational space of companies endogenously. Second, the economic developments unfolding within these networks can be measured through aggregating the thematic properties of its individual elements based on the comprehensiveness of organizational websites. For example, capturing network dynamics such as social capital, co-location, and relational embeddedness, web data can provide insights into the clustered nature of technology adoption (Dahlke et al., 2024).

Developing the concept of a *digital layer*, Abbasiharofteh, Krüger, et al. (2023) reveal that the intensity and quality of firms' hyperlinks are strongly correlated with their innovation capabilities, particularly in relation to geographically distant and cognitively close

firms. The *digital layer* framework is designed to address limitations of traditional datasets, which often overlook dynamic, real-time, and informal interactions, particularly among smaller or less-publicized firms.

Technically, the *digital layer* is constructed by scraping hyperlinks and textual content from organizational websites. These organizations are then represented as nodes in a network, with hyperlinks between organizations forming edges. Edge weights reflect link reciprocity, assigning higher importance to bidirectional connections. Several attributes are derived from this network, including the total number of hyperlinks (link count), geographic distances (calculated as Euclidean distances between firms), and cognitive distances (measured via cosine similarity of website text representations using TF-IDF). Hyperlinks are further categorized into "business" or "non-business" types using machine learning models trained on contextual data. In a follow-up study, Abbasiharofteh, Kinne, and Krüger (2023) show that strong inter-firm connections are formed involving common third-party partners and inter-regional relationships and are positively linked to higher innovation levels.

The *digital layer* concept and adjacent approaches have been applied in several contexts. Using hyperlink and textual data from corporate websites, C. Liu et al. (2024) investigate how multidimensional proximities —geographical, cognitive, organizational, institutional, social, and technological — interact to influence innovation for over three million technology firms in China. Arifi et al. (2023) explore how connection strategies of innovative firms differ between Twitter and hyperlink networks. Analyzing a sample of 11,892 IT firms, the study compares the two networks across four dimensions: network structure, information flow patterns, geographic and cognitive proximity, and the influence of company characteristics. The results suggest that innovative companies tend to align their connection strategies across online networks, while highlighting the importance of geographic and cognitive proximity for online connections. Sayles et al. (2022) investigate how different search depths (i.e. number of analyzed subpages) impact the creation of hyperlink networks, focusing on environmental organizations. In a similar context, Bogers, Biermann, et al. (2022) examine the impact of the creation of the SDGs on the hyperlink networks of 276 international organizations, revealing that fragmentation has increased, which counteracts the intended objective. Humalisto et al. (2021) study the policy on the circular economy in Finland through an analysis of hyperlink clusters. Hyperlink network analyses also form integral parts of previously described studies on technology diffusion (Dahlke et al., 2024; Kinne et al., 2024).

4.3 Identifying spatial patterns and geographies of innovation

Mapping and analyzing the spatial dynamics of production, consumption, distribution, and exchange in economies have been fundamental topics in regional studies and economic geography (Clark et al., 2018). Scholars in these fields have mostly relied on standardized classifications of economic activities (e.g., NACE codes) and geo-coded firm data to approximate the spatial distribution of economic activities. While this approach is

widely used for convenience, it has limitations that can introduce biases in empirical analyses. One key limitation is that these classifications assign firms to a single sector, failing to account for the fact that companies often engage in multiple economic activities. This approach contradicts the evolutionary perspective on economic activities, which suggests that firms build on their routines and relationships, diversifying into new activities (Nelson & Winter, 1982). It becomes particularly problematic when researchers seek to identify firms operating in market niches or commercializing newly developed technologies. For instance, a study found that the ICT sector in the UK is 40% larger than estimates based on standardized classifications of firm activities (Nathan & Rosso, 2015). Here, advanced WebAI techniques have significantly enhanced the granularity and linkage to traditional classification systems like NACE, capturing additional complexities and innovation activities that traditional industry classifications may overlook (Hajikhani et al., 2023).

Advancing new techniques to specify the economic activities of firms and regions is a crucial focus in today's economic geography and regional studies. While most studies and policies have focused on the technological capabilities of regions, less attention has been given to market applications, overlooking the capabilities of a significant share of regions—particularly peripheral ones outside major technology hubs (Breznitz, 2021; Tödtling et al., 2020). Addressing this gap is a key component of the normative turn in the geography of innovation research (Binz & Castaldi, 2024). This shift offers valuable insights into place-based policies that leverage the capabilities of local firms as a window of opportunity for a just and green transition of local economies (Lema et al., 2020). The application of WebAI is a timely enabling method for tackling the challenges described above. Websites and the data they provide can be linked to organizational characteristics through the uniqueness of URLs, including organizational location. The location can be determined through external data sources (e.g., company databases) or by geo-coding the address provided in the website's legal notice. This geo-referencing provides higher spatial accuracy than other web-based data (e.g., from social media) (Honzák et al., 2024), enabling entity-level analyses, such as individual companies (Kinne et al., 2024). WebAI-based analyses offer higher spatial resolution than traditional economic aggregates, which often suffer from the modifiable areal unit problem (Fotheringham & Wong, 1991). With precise geo-location of company data, regional analyses can be performed through spatial aggregation at district, nuts, or federal state levels, complemented by additional data sets (e.g., socio-demographic data) (Dahlke et al., 2024; C. Liu et al., 2024; Schmidt et al., 2022; Schwierzy et al., 2022). WebAI powered by recently developed LLMs can provide insights into the economic activities of firms (and regions at the aggregate level), as web text includes information about firms' products. This approach can be combined with goods and services descriptors provided by trademark applicants, serving as a benchmark for LLM-driven classification of real-world market applications (Abbasiharofteh et al., 2022; Castaldi, 2019). For example, a recent study analyzes firms' web text and used market application benchmarks to identify firms offering eco-friendly and AI-related products (Abbasiharofteh & Kriesch, 2024).

Corporate websites provide a valuable, vast data source for assessing the digital divide between regions (Mazzoni et al., 2024), a phenomenon that reflects differences in digital capabilities and online engagement. Different website features can be used as proxies for underlying digital capabilities and engagement. Individual features are often combined into a composite index, providing a more holistic and nuanced measure of a company's digital capabilities than any single indicator in isolation. This allows researchers to make comparisons across various company characteristics and geographic regions. The resulting analysis reveals patterns and trends in digital readiness, highlighting disparities between different groups of companies and territories.

Given the near-real-time and highly granular nature of WebAI insights, this methodology can significantly improve the targeting and assessment of innovation and regional development policies. Policymakers can leverage WebAI-generated indicators to monitor the diffusion of new technologies, identify emergent innovation clusters, and evaluate the effectiveness of place-based or industry-specific policies with significant timeliness and detail. Moreover, by combining WebAI results with external validation sources (e.g., survey and patent data), researchers and policymakers can develop more accurate assessments of policy impacts, adjusting and refining interventions dynamically rather than retrospectively.

WebAI provides particularly valuable input for place-based policymaking, such as smart specialization strategies for sustainable and inclusive growth. By offering insights into the market applications of firms within local economies and their diversification, policymakers can prioritize strategic investments (Balland et al., 2019) identify gaps in complementary capabilities that may be lacking in the region. These gaps can be addressed by funding research locally or by fostering interregional collaborations (Balland & Boschma, 2021).

4.3.1 Analyzing institutions at granular scale

Another pivotal challenge in economic geography and regional studies is understanding the function of institutions in influencing economic development (Boschma & Frenken, 2009; Gertler, 2010; Rodríguez-Pose, 2013). While institutions are widely acknowledged as key enablers of innovation, mutual learning and productivity growth, the empirical evaluation of these institutions and their impact at the sub-national level remains difficult. Significant strides have been made in this regard through research on institutional quality and, more specifically, 'quality of government' (Rodríguez-Pose, 2020). Governments, in general, and local and regional governments, in particular, are the primary organizations that establish the rules of the game at the local level. Consequently, the quality of local governments will influence economic development and assist in shaping the efficiency and returns of public investments (Rodríguez-Pose & Garcilazo, 2015).

Existing measurement approaches based on expert surveys and composite indices, such as the European Quality of Government Index (EQI) (Charron et al., 2014), provide valuable insights but face significant limitations. Survey-based indicators rely on the will-

ingness to participate by respondents, are costly to update regularly, and are often based on small samples which limits the extent to which findings can be aggregated at relatively large spatial scales (such as NUTS2 regions) and hinders the capture of local institutional variation (Charron et al., 2019). Furthermore, these measures primarily concentrate on formal institutions, such as laws and governance structures, while largely overlooking informal institutions—norms, networks and localized social practices—that are equally relevant to economic development (Rodríguez-Pose, 2013).

This complexity poses significant challenges in effectively measuring and operationalizing these institutions at local and regional levels. In this context, analyzing text data from local government websites provides a novel opportunity to gain deeper insights into the functioning of regional institutions. As demonstrated by (Youngblood & Mackiewicz, 2012), government websites have been shown to be utilized for the purpose of enhancing the perception held by external stakeholders. Furthermore, these websites reflect the government's priorities and strategies (Feeney & Brown, 2017; Sandoval-Almazan & Gil-Garcia, 2012). Making use of the WebAI paradigm, particularly its timeliness and thematic comprehensiveness (Gentzkow et al., 2019), it is now possible to establish a novel data source comprising county-level web text data, demonstrating that this data source can be utilized to study local institutions and to monitor government priorities and policies across a wide range of issues.

5 Limitations and Propositions for Future Research

While organizational web data represent a wealth of information, the shift from traditional, structured indicators (e.g., patents, surveys) to web-based proxies marks a fundamental methodological and epistemological transformation. The properties of web data—vastness, comprehensiveness, timeliness, liveliness, and relationality—afford new observational capacities, but also introduce new and echo known challenges to innovation measurement validity and theoretical grounding. We would propose to focus on several areas of development, addressing issues of representation of i) individual organizations, ii) innovation systems, and iii) multi-system interactions through WebAI techniques.

5.1 Representation of organizational identity and firm-level innovation

The WebAI paradigm builds on the assumption that organizational websites serve as comprehensive representations of organizational identity, offering data to extract insights into their strategy, capabilities, and innovations (Oertel & Thommes, 2018; Powell et al., 2016). However, this demands careful and systematic analysis of the mechanisms, completeness, and frequency with which these identities are (re)constructed and maintained online, and how they reflect innovation activities.

5.1.1 Measuring intra-firm diffusion of innovation

While WebAI indicators have proven to be plausible and reliable to measure inter-firm diffusion of innovation at larger scale (see chapter 4), a key limitation in current approaches lies in their inability to adequately capture the intra-organizational diffusion of technologies and practices. As Battisti and Stoneman (2003) emphasize, innovation does not conclude at the point of adoption; rather, it unfolds through the uneven and incremental spread of new technologies within organizations. Current web indicators primarily reflect externally directed signals—mentions of technologies, initiatives, or strategic goals—without providing a clear picture of how deeply or widely these practices are embedded within organizational routines. While current measurements can be used to calculate intensities at the website level (Dahlke et al., 2024; Kinne et al., 2024), future work could extend the analysis by analyzing specific web pages (e.g., subdomains for business units) to measure the diffusion of a technology within the organization.

5.1.2 Accounting for strategic signaling

Website's function as inherently strategic communication channels further complicates efforts to approximate the ground truth of intra-firm diffusion of innovations. Organizations increasingly align digital messaging with policy and market trends—such as "sustainability," "digital transformation," or "AI leadership" irrespective of actual internal capability. This raises the issue of strategic signaling and reputational inflation. Much like patents filed for strategic signaling (Griliches, 1990), web content may overstate innovation engagement, particularly when incentives are high to appear aligned with fashionable trends. To reduce the risk of producing falsely positive measurements, the applied research using WebAI techniques has introduced AI models able to distinguish between a superficial and sincere engagement of organization with certain innovations, which goes beyond a detection of keywords and builds on curated training data to make inferences based on more complex patterns.[2] Besides the progressing capabilities of generative AI models allowing for flexible and adaptive usage, this highlights value of training stable supervised models to detect latent patterns of specific concepts.

5.1.3 Distinguishing innovation inputs and outputs

When aggregating WebAI metrics at the level of entire organizations, they run the risk of conflating innovation inputs (e.g., expressions of intent, R&D emphasis, human capital, or capability signaling) with outputs (e.g., realized products, new services, or operationalized practices). While this conflation has been recognized in innovation measurement more broadly (Rogers, 1998), web-based approaches have yet to establish reliable methods for distinguishing between the two. Here, the comprehensiveness of organizational web content presents untapped potential for further differentiation: job postings may serve as

[2]The reverse issue of accounting for false negatives is more salient as some organizations may avoid reporting activities in certain technology fields for strategic purposes.

signals of investment and capacity building (inputs) also in the field of R&D, while product announcements or project cases may reflect realized innovation outcomes (outputs). However, current indicators lack the conceptual clarity and computational precision to operationalize these distinctions at scale. In particular, the curation of more targeted sets of training data and the use of LLMs to contextualize calculated indicators constitute promising approaches to meet this end.

5.1.4 Distinguishing product, process, and business model innovation

Drawing from the Oslo Manual's (Eurostat, 2018) distinction between product and process innovations, it is important to recognize that websites may differentially reflect these categories. Product innovations, often more visible and customer-facing, are likely overrepresented in web narratives, while process innovations—embedded in backend systems or organizational routines—may be underreported. This warrants more thorough investigations into what types of innovation WebAI metrics can capture. As a starting point, Dahlke et al. (2024) applied topic modeling to website paragraphs exhibiting high scores for their AI indicator and found that both product- and process-related topics were associated with it. Further research should explore ways to infer process innovations from indirect signals such as supply chain topics, operational case studies, or recruitment for technical roles. The comprehensiveness of organizational websites could also offer a valuable resource to measure innovation to business models.

5.1.5 Accounting for methodological drift

The liveliness of web data, with its continuous updating and responsiveness to sociopolitical signals, enables the early detection of emerging issues and discursive shifts (Marres & Weltevrede, 2013). However, this same volatility poses challenges for temporal consistency. As websites evolve in structure and content and/or as AI models are updated or retrained, indicators can drift. This represents a fundamental methodological shift from the use of stable, archival data to fluid, conceptually evolving corpora. Without robust versioning of both data and models, longitudinal studies lose comparability (Schafer & Winters, 2021). Techniques such as anomaly detection, temporal consistency checks, and contextual semantic analysis are a way forward to mitigate these biases.

5.1.6 Propositions

Future research working within the WebAI paradigm must address the following representational issues at the organizational level:

- Develop ontologies and taxonomies aligned with innovation theory to improve indicator interpretability.
- Investigate mechanisms of digital self-representation to distinguish rhetorical innovation from "de facto" innovation (particularly along the dimension of process innovation).

- Use domain-adapted LLMs to extract and classify innovation claims with higher semantic specificity.
- Explore techniques to address representation biases in web data, including imputation techniques, as well as robustness tests based on survey data and administrative data.
- Develop infrastructures for longitudinal tracking of WebAI metrics, including timestamped indicators, panel datasets, version-controlled models, and techniques to check for temporal consistency of web data.
- Map innovation input-output trajectories across different types of data by correlating WebAI indicators with patent and publication records or job postings and exploring their temporal sequence.

5.2 Representation of innovation systems

5.2.1 Exhaustiveness and bias in capturing economic systems

We have argued that the WebAI paradigm exceeds traditional innovation measurement practices through removing limitations in data collection, allowing to move from the analysis of small sample sizes to capturing larger parts of economic systems. However, this coverage is likely to still carry systematic bias. Despite the vastness of organizational web data, many firms—especially micro-enterprises, young ventures, or entities in traditional service sectors—are less likely to (actively) maintain a web presence. Even among registered firms, web presence may be limited to inactive landing pages or social media profiles (Kinne & Axenbeck, 2020). As such, WebAI indicators skew toward digitally mature, resource-rich firms, introducing a representation bias that echoes known limitations in traditional data sources, even if this bias may arguably be less severe.

5.2.2 Structural and functional characteristics of hyperlink networks

Hyperlinks, co-mentions, or semantic similarity of websites encode a latent network of innovation-related relationships. This relationality can enable WebAI to model information flows, inter-firm proximity, and embeddedness (Dahlke et al., 2024). However, the digital architecture of these networks is shaped by rules that are only partially understood. Hyperlink connections may indicate formal partnerships, supply chain relations, shared ecosystem membership, or simply content management strategies. While digital relational data has the potential to reflect sectoral or national innovation systems (Katz & Cothey, 2006; Vaughan et al., 2007), this requires linking patterns of connection to underlying system functions (e.g., knowledge development, resource mobilization, market formation, etc.). Without such functional mapping, network metrics risk becoming descriptive rather than explanatory. Regarding indirect relations, co-occurrence of topics across websites could indicate diffusion patterns (epidemic effects) but could also reflect parallel

signaling to external audiences. This ambiguity complicates the interpretation of WebAI-generated network indicators. Further exploration should combine longitudinal data with more advanced network analyses to develop a sequential picture of relationships in order to determine (investigate) their causality.

5.2.3 Propositions

Future research working within the WebAI paradigm must, thus, address the following representational issues at the level of socio-economic systems:

- Analyze the logic behind digital link formation, including the representation of different forms of organizational relationships and strategic linking behavior.
- Use advanced AI techniques to infer latent relations and knowledge flows from textual and structural web data.
- Understand temporal patterns in hyperlink network including link formation between individual organization and structural properties of networks.
- Situate digital networks within established systems-of-innovation frameworks to uncover their structural-functional properties.

5.3 Multi-system and multi-data integration

Given the potentials and limitations mentioned above, we argue that the WebAI paradigm is best understood as an empirical perspective on innovation that complements traditional measurements. In particular, we argue that the vast, comprehensive, and lively properties of web-based measures could offer richer insights on broader industrial regimes. That is, because WebAI techniques are particularly well suited to identify organizations as *users* of innovation alongside organizations as *producers* of innovation (which are also well covered by traditional metrics). This advantage may be particularly salient for soft and intangible forms of innovations (i.e., as often found in the service sectors).

We are, nevertheless, conscious of the fact that innovation is embedded in multi-level, interdependent systems that span economic, social, scientific, and ecological domains (Geels, 2020). Organizational innovation and industrial practice, as captured through WebAI techniques, is deeply entangled with other regimes that form within institutional, technological, and ecological systems (Geels, 2005; Geels et al., 2023). Understanding this complexity of modern economic systems necessitates an integrative approach that adopts a system of systems perspective. Using the relational properties of web data and combining it with other data sources such as patents, publications, as well as institutional and environmental data can be a promising way for future research to capture the interaction between such different (industrial, technological, socio-political, and ecological) systems.

5.3.1 Propositions

To integrate WebAI techniques into more holistic concepts of innovation and transition research, future work should:

- Design multi-level models capable of capturing micro-level organizational activity, meso-level system dynamics, and macro-level trends by integrating multiple data sources.
- Explore how innovation signals propagate across systems (e.g., from science to technology to organizational practice and public perception, or regulation).

6 Conclusions

This paper aims to spark a discussion on a new WebAI paradigm in innovation studies, business analytics, and informed policymaking. We presented suggestions on how to extract insight from organizational web data by using state-of-the-art AI techniques. We introduced five central properties of organizational web data (vastness, comprehensiveness, timeliness, liveliness, and relationality) that set it apart from traditional data sources such as surveys or patents, and make it relevant for studying innovation. These properties also necessitate rigorous processing techniques to separate signal from noise. To this end, we provided best practices and outlined current challenges of the methods that can be employed to extract and process organizational web data. The potential of the WebAI paradigm is highlighted through several scientific use cases that demonstrate how WebAI enables the analysis of innovation in a comprehensive way, including micro-level innovation activities, technology diffusion, sustainability engagement, innovation systems, organizational networks, institutions, and the geography of innovation.

While organizational web data represent a wealth of information, it may also be susceptible to selective self-presentation, potentially introducing biases into analyses derived from WebAI methods. Applying the WebAI paradigm in future research should focus on sharpening processing techniques to further distinguish rhetorical signaling from actual innovation activities. We propose that it should also focus on developing methods to precisely measure different types of innovation. While the WebAI paradigm unlocks a more exhaustive coverage of economic systems than traditional metrics can offer, future research should focus on accounting for remaining representation biases in hosting or maintaining organizational websites (e.g., for micro-enterprises, traditional service providers). Furthermore, the relational nature of web data (e.g., hyperlinks) needs to be explored and linked to specific functions (e.g., knowledge sharing) to move beyond merely descriptive network measures.

Finally, innovation is embedded in multi-level systems that include organizational, industrial, societal, and ecological contexts. We argue that WebAI should not replace traditional data (e.g., patents, publications), but rather complement them to trace innovation pathways and understand how innovation signals spread from scientific discovery to

technological application and economic outcomes, as well as how they relate to the socio-political discourse. This requires models that leverage the relationality of web data and explore its integration with other data sources to link innovation signals at the micro level with higher-level developments and regimes in socio-technical systems.

Competing interests: David Lenz, Robert Dehghan and Jan Kinne founded and, together with Sebastian Schmidt, are employed by istari.ai. The company did not fund this research but is involved in the commercial provision of services that are related to the methods described in this paper. The authors affirm that this did not influence the paper's contents, which are presented in full adherence to scientific principles and best practices.

References

Abbasiharofteh, M., Castaldi, C., & Petralia, S. G. (2022). From patents to trademarks: Towards a concordance map.

Abbasiharofteh, M., Kinne, J., & Krüger, M. (2023). Leveraging the digital layer: The strength of weak and strong ties in bridging geographic and cognitive distances. *Journal of Economic Geography*. https://doi.org/10.1093/jeg/lbad037

Abbasiharofteh, M., Kogler, D. F., & Lengyel, B. (2023). Atypical combinations of technologies in regional co-inventor networks. *Research Policy*, *52*, 104886. https://doi.org/10.1016/j.respol.2023.104886

Abbasiharofteh, M., & Kriesch, L. (2024). Not all twins are identical: The digital layer of "twin" transition market applications.

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, *18*(4), 507–529. https://doi.org/10.1080/17421772.2023.2193222

Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–519. https://doi.org/10.1016/0166-4972(96)00031-4

Arifi, D., Resch, B., Kinne, J., & Lenz, D. (2023). Innovation in hyperlink and social media networks: Comparing connection strategies of innovative companies in hyperlink and social media networks (C. (J. Gomez, Ed.). *PLOS ONE*, *18*(3), e0283372. https://doi.org/10.1371/journal.pone.0283372

Asheim, B. T., Smith, H. L., & Oughton, C. (2011). Regional innovation systems: Theory, empirics and policy. *Regional Studies*, *45*(7), 875–891.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., & Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, *42*, 108246.

Auzepy, A., Tönjes, E., Lenz, D., & Funk, C. (2023). Evaluating TCFD reporting—A new application of zero-shot analysis to climate-related financial disclosures (S. Kaddoura, Ed.). *PLOS ONE*, *18*(11), e0288052. https://doi.org/10.1371/journal.pone.0288052

Balland, P.-A., & Boschma, R. (2021). Complementary interregional linkages and smart specialisation: An empirical study on european regions. *Regional Studies*, *55*(6), 1059–1070. https://doi.org/10.1080/00343404.2020.1861240

Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the european union: Relatedness, knowledge complexity and regional diversification. *Regional studies*, *53*(9), 1252–1268.

Battisti, G., & Stoneman, P. (2003). Inter- and intra-firm effects in the diffusion of new process technology. *Research Policy*, *32*(9), 1641–1655. https://doi.org/10.1016/S0048-7333(03)00055-6

Belderbos, R., Carree, M., & Lokshin, B. (2004). Cooperative r&d and firm performance. *Research Policy*, *33*(10), 1477–1492. https://doi.org/https://doi.org/10.1016/j.respol.2004.07.003

Bennich, T., Weitz, N., & Carlsen, H. (2020). Deciphering the scientific literature on SDG interactions: A review and reading guide. *Science of The Total Environment*, *728*, 138405. https://doi.org/10.1016/j.scitotenv.2020.138405

Bergek, A., Hekkert, M., Jacobsson, S., Markard, J., Sandén, B., & Truffer, B. (2015). Technological innovation systems in contexts: Conceptualizing contextual structures and interaction dynamics. *Environmental innovation and societal transitions*, *16*, 51–64.

Binz, C., & Castaldi, C. (2024). Toward a normative turn in research on the geography of innovation? evolving perspectives on innovation, institutions, and human well-being. *Progress in Economic Geography*, *2*, 100018. https://doi.org/10.1016/j.peg.2024.100018

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning research*, *3*(Jan), 993–1022.

Bogers, M., Garud, R., Thomas, L., Tuertscher, P., & Yoo, Y. (2022). Digital innovation: Transforming research and practice. *Innovation*, *24*(1), 4–12.

Bogers, M., Zobel, A.-K., Afuah, A., Almirall, E., Brunswicker, S., Dahlander, L., Frederiksen, L., Gawer, A., Gruber, M., Haefliger, S., et al. (2017). The open innovation research landscape: Established perspectives and emerging themes across different levels of analysis. *Industry and Innovation*, *24*(1), 8–40.

Bogers, M., Biermann, F., Kalfagianni, A., Kim, R. E., Treep, J., & De Vos, M. G. (2022). The impact of the Sustainable Development Goals on a network of 276 international organizations. *Global Environmental Change*, *76*, 102567. https://doi.org/10.1016/j.gloenvcha.2022.102567

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. *European Journal of Operational Research*, *306*(1), 348–357. https://doi.org/10.1016/j.ejor.2022.06.060

Boschma, R., & Frenken, K. (2009). Some notes on institutions in evolutionary economic geography. *Economic Geography*, *85*(2), 151–158. Retrieved March 7, 2025, from http://www.jstor.org/stable/40377293

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

Breithaupt, P., Hottenrott, H., Rammer, C., & Römer, K. (2024). Linked employer–employee data from xing and the mannheim enterprise panel. *Jahrbücher für Nationalökonomie und Statistik*. https://doi.org/doi:10.1515/jbnst-2024-0070

Breznitz, D. (2021). *Innovation in real places: Strategies for prosperity in an unforgiving world*. Oxford University Press.

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). Web scraping for research: Legal, ethical, institutional, and scientific considerations. https://doi.org/10.48550/arXiv.2410.23432

Buchanan, D. A., & Denyer, D. (2013). Researching tomorrow's crisis: Methodological innovations and wider implications. *International Journal of Management Reviews*, *15*(2), 205–224. https://doi.org/https://doi.org/10.1111/ijmr.12002

Caloghirou, Y., Ioannides, S., & Vonortas, N. S. (2003). Research joint ventures. *Journal of Economic Surveys*, *17*(4), 541–570. https://doi.org/https://doi.org/10.1111/1467-6419.00204

Carlsson, B., & Stankiewicz, R. (1991). On the nature, function and composition of technological systems. *Journal of evolutionary economics*, *1*, 93–118.

Carlsson, B., Jacobsson, S., Holmén, M., & Rickne, A. (2002). Innovation systems: Analytical and methodological issues. *Research Policy*, *31*(2), 233–245.

Castaldi, C. (2018). To trademark or not to trademark: The case of the creative and cultural industries. *Research Policy*, *47*(3), 606–616.

Castaldi, C. (2019). All the great things you can do with trademark data. taking stock and looking ahead. *Strategic Organization*, *18*, 472–484. https://doi.org/10.1177/1476127019847835

Castellaci, F., Grodal, S., Mendonca, S., & Wibe, M. (2005). Advances and challenges in innovation studies. *Journal of Economic Issues*, *39*(1), 91–121. https://doi.org/10.1080/00213624.2005.11506782

Charron, N., Dijkstra, L., & Lapuente, V. (2014). Regional governance matters: Quality of government within european union member states. *Regional Studies*, *48*(1), 68–90. https://doi.org/10.1080/00343404.2013.770141

Charron, N., Lapuente, V., & Annoni, P. (2019). Measuring quality of government in eu regions across space and time. *Papers in Regional Science*, *98*(5), 1925–1954. https://doi.org/10.1111/pirs.12437

Chen, S., Wong, S., Chen, L., & Tian, Y. (2023). Extending context window of large language models via positional interpolation. https://doi.org/10.48550/arXiv.2306.15595

Chesbrough, H., & Rosenbloom, R. S. (2002). The role of the business model in capturing value from innovation: Evidence from xerox corporation's technology spin-off companies. *Industrial and corporate change*, *11*(3), 529–555.

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1–35. https://doi.org/10.1145/3507900

Clark, G. L., Feldman, M. P., Gertler, M. S., & Wójcik, D. (Eds.). (2018). *The new oxford handbook of economic geography*. Oxford University Press.

Comin, D., & Mestieri, M. (2014). Technology Diffusion: Measurement, Causes, and Consequences. In *Handbook of Economic Growth* (pp. 565–622, Vol. 2). Elsevier. https://doi.org/10.1016/B978-0-444-53540-5.00002-1

Cui, B., Li, Y., Chen, M., & Zhang, Z. (2019). Fine-tune bert with sparse self-attention mechanism. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3546–3551. https://doi.org/10.18653/v1/D19-1361

Dahlke, J., Beck, M., Kinne, J., Lenz, D., Dehghan, R., Wörter, M., & Ebersberger, B. (2024). Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. *Research Policy*, *53*(2), 104917. https://doi.org/10.1016/j.respol.2023.104917

De Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. D. L. (2020). Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, *32*(1), 19. https://doi.org/10.1186/s12302-020-0300-3

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? a consensual definition and a review of key research topics. *AIP Conference Proceedings*, *1644*(1), 97–104. https://doi.org/10.1063/1.4907823

Delgado, M., Porter, M. E., & Stern, S. (2014). Clusters, convergence, and economic performance. *Research Policy*, *43*(10), 1785–1799.

Deschryvere, M., Schubert, T., Ashouri, S., Jäger, A., Visentin, F., Cunningham, S. W., Hajikhani, A., Pukelis, L., & Suominen, A. (2023). The role of product digitization for productivity: Evidence from web-scraping european high-tech company websites. *Poster session presented at DRUID 2023, Lisbon, Portugal*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *PLOS ONE*, *17*(2). https://doi.org/10.1371/journal.pone.0263898

Eurostat, O. (2018). *Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation, 4th edition: The measurement of scientific, technological and innovation activities*. https://doi.org/10.1787/9789264304604-en

Fagerberg, J. (2006, January). Innovation: A guide to the literature. In *The oxford handbook of innovation*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286805.003.0001

Feeney, M. K., & Brown, A. (2017). Are small cities online? Content, ranking, and variation of U.S. municipal websites. *Government Information Quarterly*, *34*(1), 62–74. https://doi.org/10.1016/j.giq.2016.10.005

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic management journal*, *25*(8-9), 909–928.

Foster, J. (2006). Why is economics not a complex systems science? *Journal of Economic Issues*, *40*(4), 1069–1091.

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space*, *23*(7), 1025–1044. https://doi.org/10.1068/a231025

Franco, S. F., Graña, J. M., Flacher, D., & Rikap, C. (2023). Producing and using artificial intelligence: What can europe learn from siemens's experience? *Competition & Change*, *27*(2), 302–331.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gawer, A., & Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. *Journal of product innovation management*, *31*(3), 417–433.

Geels, F. (2005). Processes and patterns in transitions and system innovations: Refining the co-evolutionary multi-level perspective [Transitions towards Sustainability through System Innovation]. *Technological Forecasting and Social Change*, *72*(6), 681–696. https://doi.org/https://doi.org/10.1016/j.techfore.2004.08.014

Geels, F. (2020). Micro-foundations of the multi-level perspective on socio-technical transitions: Developing a multi-dimensional model of agency through crossovers between social constructivism, evolutionary economics and neo-institutional theory. *Technological Forecasting and Social Change*, *152*, 119894.

Geels, F., Kern, F., & Clark, W. C. (2023). System transitions research and sustainable development: Challenges, progress, and prospects. *Proceedings of the National Academy of Sciences*, *120*(47), e2206230120.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Gertler, M. S. (2010). Rules of the game: The place of institutions in regional economic change. *Regional Studies*, *44*(1), 1–15. https://doi.org/10.1080/00343400903389979

Gilga, C., Hochwarter, C., Knoche, L., Schmidt, S., Ringler, G., Wieland, M., Resch, B., & Wagner, B. (2025). Legal and ethical considerations for demand-driven data collection and ai-based analysis in flood response. *International Journal of Disaster Risk Reduction*, 105441. https://doi.org/10.1016/j.ijdrr.2025.105441

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*, 653–671. https://doi.org/10.1007/s11192-014-1434-0

Grashof, N., & Fornahl, D. (2021). "to be or not to be" located in a cluster?—a descriptive meta-analysis of the firm-specific cluster effect. *The Annals of Regional Science*, *67*(3), 541–591.

Griliches, Z. (1990). *Patent statistics as economic indicators: A survey part i* (Working Paper No. 3301). National Bureau of Economic Research. Cambridge, MA. https://www.nber.org/papers/w3301

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/ARXIV.2203.05794

Gschnaidtner, C., Dehghan, R., Hottenrott, H., & Schwierzy, J. (2024). Adoption and Diffusion of Blockchain Technology. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4775993

Haans, R. F., & Mertens, M. J. (2024). The internet never forgets: A four-step scraping tutorial, codebase, and database for longitudinal organizational website data. *Organizational Research Methods*, 10944281241284941. https://doi.org/10.1177/10944281241284941

Hajikhani, A., & Cole, C. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized ai. *Quantitative Science Studies*, *5*(3), 736–756.

Hajikhani, A., Cunningham, S. W., Pukelis, L., & Suominen, A. (2023). Measuring innovation and complexity in industry. *Academy of Management Proceedings*, *2023*(1), 15893.

Hajikhani, A., Pukelis, L., Suominen, A., Ashouri, S., Schubert, T., Notten, A., & Cunningham, S. W. (2022). Connecting firm's web scraped textual content to body of science: Utilizing microsoft academic graph hierarchical topic modeling. *MethodsX*, *9*, 101650.

Honzák, K., Schmidt, S., Resch, B., & Ruthensteiner, P. (2024). Contextual Enrichment of Crowds from Mobile Phone Data through Multimodal Geo-Social Media Analysis. *ISPRS International Journal of Geo-Information*, *13*(10), 350. https://doi.org/10.3390/ijgi13100350

Hottenrott, H., & Lopes-Bento, C. (2016). R&D partnerships and innovation performance: Can there be too much of a good thing? *Journal of Product Innovation Management*, *33*(6), 773–794. https://doi.org/https://doi.org/10.1111/jpim.12311

Humalisto, N., Valve, H., & Åkerman, M. (2021). Making the circular economy online: A hyperlink analysis of the articulation of nutrient recycling in Finland. *Environmental Politics*, *30*(5), 833–853. https://doi.org/10.1080/09644016.2020.1817291

ISTARI.AI. (2025). ISTARI.AI webAI. https://istari.ai

Katz, J. S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2), 85–95.

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, *0*(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*, 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Kinne, J., Dehghan, R., Schmidt, S., Lenz, D., & Hottenrott, H. (2024). Location factors and ecosystem embedding of sustainability-engaged blockchain companies in the US. A web-based analysis. *International Journal of Information Management Data Insights*, *4*(2), 100287. https://doi.org/10.1016/j.jjimei.2024.100287

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, *16*(4). https://doi.org/10.1371/journal.pone.0249071

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951716631130

Koenig, M., Rauch, J., & Woerter, M. (2025). Real-time monitoring of economic shocks using company websites. https://arxiv.org/abs/2502.17161

Kriesch, L. (2023). *Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie* [Doctoral dissertation, Universitätsbibliothek Gießen]. https://doi.org/10.22029/JLUPUB-15686

Kriesch, L., & Losacker, S. (2024). Bioeconomy firms and where to find them. *REGION*, *11*(1), 55–78. https://doi.org/10.18335/region.v11i1.523

Laney, D. (2001). *3d data management: Controlling data volume, velocity and variety* (Research Note No. 6). META Group.

Latapy, M., Magnien, C., & Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, *30*(1), 31–48. https://doi.org/10.1016/j.socnet.2007.04.006

Laurer, M. (2024). Synthetic data: Save money, time and carbon with open source. https://huggingface.co/blog/synthetic-data-save-costs

Lema, R., Fu, X., & Rabellotti, R. (2020). Green windows of opportunity: Latecomer development in the age of transformation toward sustainability. *Industrial and Corporate Change*, *29*(5), 1193–1209. https://doi.org/10.1093/icc/dtaa045

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., & Du, M. (2024). Language ranker: A metric for quantifying llm performance across high and low-resource languages. https://doi.org/10.48550/arXiv.2404.11553

Liu, C., Peng, Z., Liu, L., Wu, H., Kinne, J., Cai, M., & Li, S. (2024). XAI in geographic analysis of innovation: Evaluating proximity factors in the innovation networks of Chinese technology companies through web-based data. *Applied Geography*, *171*, 103373. https://doi.org/10.1016/j.apgeog.2024.103373

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). A Survey on Hallucination in Large Vision-Language Models.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. Retrieved December 13, 2022, from http://arxiv.org/abs/1907.11692

Liu, Y., Li, W., Wang, L., & Meng, Q. (2023). Why greenwashing occurs and what happens afterwards? a systematic literature review and future research agenda. *Environmental Science and Pollution Research*, *30*(56), 118102–118116.

Lundvall, B.-Å., Johnson, B., Andersen, E. S., & Dalum, B. (2002). National systems of production, innovation and competence building. *Research Policy*, *31*(2), 213–231.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Mainzer, K. (2009). Challenges of Complexity in the 21st Century. An Interdisciplinary Introduction. *European Review*, *17*(2), 219–236. https://doi.org/10.1017/S1062798709000714

Marr, B. (2015). *Big data: Using smart big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons.

Marres, N., & Weltevrede, E. (2013). Scraping the social? issues in live social research. *Journal of cultural economy*, *6*(3), 313–335.

Martin, B. R. (2016). Twenty challenges for innovation studies. *Science and Public Policy*, *43*(3), 432–450. https://doi.org/10.1093/scipol/scv077

Mazzoni, L., Pinelli, F., & Riccaboni, M. (2024). Measuring corporate digital divide through websites: Insights from italian firms. *EPJ data science*, *13*(1), 51.

Mirtsch, M., Kinne, J., & Blind, K. (2021). Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100. https://doi.org/10.1109/TEM.2020.2977815

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 109–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_4

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Nathan, M., & Rosso, A. (2015). Mapping digital businesses with big data: Some early findings from the uk. *Research Policy*, *44*(9), 1714–1733. https://doi.org/10.1016/j.respol.2015.01.008

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Nelson, R. R., & Winter, S. G. (1982). The schumpeterian tradeoff revisited. *The American Economic Review*, *72*(1), 114–132.

Oberg, A., Powell, W. W., & Schöllhorn, T. (2022). Representations of self in the digital public sphere: The field of social impact analyzed through relational and discursive moves. In *Digital transformation and institutional theory* (pp. 167–196, Vol. 83). Emerald Publishing Limited.

Oertel, S., & Thommes, K. (2018). History as a source of organizational identity creation. *Organization Studies*, *39*(12), 1709–1731.

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C. A., Von Werra, L., Wolf, T., et al. (2025). The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, *37*, 30811–30849.

Porter, M. E. (1998). Clusters and competition. *On Competition*, *7*, 91.

Powell, W. W., Horvath, A., & Brandtner, C. (2016). Click and mortar: Organizations on the web. *Research in Organizational Behavior*, *36*, 101–120.

Qu, R., Tu, R., & Bao, F. (2024). *Is Semantic Chunking Worth the Computational Cost?* arXiv: 2410.13070 [cs]. https://doi.org/10.48550/arXiv.2410.13070

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators - a literature review. *Technological Forecasting and Social Change*, *197*, 122874. https://doi.org/https://doi.org/10.1016/j.techfore.2023.122874

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. Retrieved June 7, 2023, from http://arxiv.org/abs/1908.10084

Ritala, P. (2024). Grand challenges and platform ecosystems: Scaling solutions for wicked ecological and societal problems. *Journal of Product Innovation Management*, *41*(2), 168–183. https://doi.org/https://doi.org/10.1111/jpim.12682

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, *4*(2), 155–169.

Rodríguez-Pose, A. (2013). Do institutions matter for regional development? *Regional Studies*, *47*(7), 1034–1047. https://doi.org/10.1080/00343404.2012.748978

Rodríguez-Pose, A. (2020). Institutions and the fortunes of territories. *Regional Science Policy & Practice*, *12*(3), 371–386. https://doi.org/10.1111/rsp3.12277

Rodríguez-Pose, A., & Garcilazo, E. (2015). Quality of government and the returns of investment: Examining the impact of cohesion expenditure in european regions. *Regional Studies*, *49*(8), 1274–1290. https://doi.org/10.1080/00343404.2015.1007933

Rogers, M. (1998). *The definition and measurement of innovation* (Working Paper No. 10/98). Melbourne Institute of Applied Economic and Social Research. The University of Melbourne. https://melbourneinstitute.unimelb.edu.au

Sandoval-Almazan, R., & Gil-Garcia, J. R. (2012). Are government internet portals evolving towards more interaction, participation, and collaboration? revisiting the rhetoric of e-government among municipalities. *Government Information Quarterly*, *29*, S72–S81. https://doi.org/10.1016/j.giq.2011.09.004

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

Sayles, J. S., Furey, R. P., & Ten Brink, M. R. (2022). How deep to dig: Effects of web-scraping search depth on hyperlink network analysis of environmental stewardship organizations. *Applied Network Science*, *7*(1), 36. https://doi.org/10.1007/s41109-022-00472-0

Schafer, V., & Winters, J. A. (2021). The values of web archives. *International Journal of Digital Humanities*, *2*, 129–144. https://doi.org/10.1007/s42803-021-00037-0

Schmidt, S., Kinne, J., Lautenbach, S., Blaschke, T., Lenz, D., & Resch, B. (2022). Greenwashing in the US metal industry? A novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining. *Science of the Total Environment*, *835*, 155512. https://doi.org/10.1016/j.scitotenv.2022.155512

Schwierzy, J., Dehghan, R., Schmidt, S., Rodepeter, E., Stömmer, A., Uctum, K., Kinne, J., Lenz, D., & Hottenrott, H. (2022). Technology mapping using WebAI: The case of 3D printing. https://arxiv.org/pdf/2201.01125

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., … Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, *31*(3), 943–950. https://doi.org/10.1038/s41591-024-03423-7

Stoneman, P. (2010). *Soft innovation: Economics, product aesthetics, and the creative industries*. Oxford University Press.

Talukdar, W., & Biswas, A. (2024). Synergizing unsupervised and supervised learning: A hybrid approach for accurate natural language task modeling. *International Journal of Innovative Science and Research Technology (IJISRT)*, 1499–1508. https://doi.org/10.38124/ijisrt/IJISRT24MAY2087

Tan, J., Dou, Z., Wang, W., Wang, M., Chen, W., & Wen, J.-R. (2025). Htmlrag: Html is better than plain text for modeling retrieved knowledge in rag systems. https://doi.org/10.1145/3696410.3714546

Thelwall, M., & Vaughan, L. (2004). A fair history of the web? examining country balance in the internet archive. *Library & information science research*, *26*(2), 162–176.

Thirumoorthy, K., & Muneeswaran, K. (2022). Feature selection for text classification using machine learning approaches. *National Academy Science Letters*, *45*(1), 51–56. https://doi.org/10.1007/s40009-021-01043-0

Tödtling, F., Trippl, M., & Frangenheim, A. (2020). Policy options for green regional development: Adopting a production and application perspective. *Science and Public Policy*, *47*(6), 865–875. https://doi.org/10.1093/scipol/scaa063

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. https://doi.org/10.48550/arXiv.2209.11055

Valova, I., Mladenova, T., Kanev, G., & Halacheva, T. (2023). Web Scraping - State of Art, Techniques and Approaches. *2023 31st National Conference with International Participation (TELECOM)*, 1–4. https://doi.org/10.1109/TELECOM59629.2023.10409723

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Retrieved August 9, 2023, from http://arxiv.org/abs/1706.03762

Vaughan, L., Kipp, M., & Gao, Y. (2007). Why are websites co-linked? the case of canadian universities. *Scientometrics*, *72*(1), 81–92.

Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information. *Scientometrics*, *60*(3), 487–496.

Vicente, M. R., & López, A. J. (2011). Assessing the regional digital divide across the European Union-27. *Telecommunications Policy*, *35*(3), 220–237. https://doi.org/10.1016/j.telpol.2010.12.013

Wang, W., Dinh, J. V., Jones, K. S., Upadhyay, S., & Yang, J. (2023). Corporate Diversity Statements and Employees' Online DEI Ratings: An Unsupervised Machine-Learning Text-Mining Analysis. *Journal of Business and Psychology*, *38*(1), 45–61. https://doi.org/10.1007/s10869-022-09819-x

Wang, Z., Wang, D., & Li, Q. (2021). Keyword extraction from scientific research projects based on srp-tf-idf. *Chinese Journal of Electronics*, *30*(4), 652–657. https://doi.org/10.1049/cje.2021.05.007

Wasserman, S., & Faust, K. (1994, November). *Social Network Analysis: Methods and Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The gdpr and unstructured data: Is anonymization possible? *International Data Privacy Law*, *12*(3), 184–206. https://doi.org/10.1093/idpl/ipac008

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Wildnerova, L., Menon, C., Dehghan, R., Kinne, J., & Lenz, D. (2024, July). *Which SMEs are greening?: Cross-country evidence from one million websites* (OECD SME and En-

trepreneurship Papers No. 60) (Series: OECD SME and Entrepreneurship Papers Volume: 60). https://doi.org/10.1787/ddd00999-en

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N., & Xue, C. J. (2025). *Retrieval-Augmented Generation for Natural Language Processing: A Survey*. arXiv: 2407.13193 [cs]. https://doi.org/10.48550/arXiv.2407.13193

Xie, X., Liu, X., & Chen, J. (2023). A meta-analysis of the relationship between collaborative innovation and innovation performance: The role of formal and informal institutions. *Technovation*, *124*, 102740. https://doi.org/https://doi.org/10.1016/j.technovation.2023.102740

Xu, Y., Yin, J., Huang, J., & Yin, Y. (2018). Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications*, *103*, 106–117. https://doi.org/10.1016/j.eswa.2018.03.008

Youngblood, N. E., & Mackiewicz, J. (2012). A usability analysis of municipal government website home pages in alabama. *Government Information Quarterly*, *29*(4), 582–588. https://doi.org/10.1016/j.giq.2011.12.010

Zhang, M., & Chen, Y. (2018). Link Prediction Based on Graph Neural Networks. *Advances in Neural Information Processing Systems*, *31*. Retrieved March 17, 2025, from https://proceedings.neurips.cc/paper/2018/hash/53f0d7c537d99b3824f0f99d62ea2428-Abstract.html

Zhang, Y., & Cheng, T. (2020). Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. *Computers, Environment and Urban Systems*, *79*, 101403. https://doi.org/10.1016/j.compenvurbsys.2019.101403

Zhu, S., Supryadi, Xu, S., Sun, H., Pan, L., Cui, M., Du, J., Jin, R., Branco, A., & Xiong, D. (2024). Multilingual large language models: A systematic survey. https://doi.org/10.48550/arXiv.2411.11072

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de