

DIGILOG: Towards a Monitoring Platform for Digital Transformation of European Communities

Jonathan Gerber^{1,*†}, Jasmin S. Saxer^{1†}, Bruno B. Kreiner^{1†} and Andreas Weiler^{1†}

¹*Institute of Computer Science, Zurich University of Applied Sciences, Technikumstrasse 9, 8401 Winterthur, Switzerland*
<https://www.zhaw.ch/en/engineering/institutes-centres/init/>

Abstract

DIGILOG is an interdisciplinary research project between Computer and Political Science. The goal of the research project is to monitor and evaluate the digital transformation of the local governments of Europe. The project will generate coherent data for a systematic comparison using methodological triangulation, i.e., quantitative and qualitative methods. It will take the form of a regular and automated quantitative survey of all local authorities in 47 European countries (members of the Council of Europe), based on web crawling and machine learning techniques - this is a novel approach in the context of the social sciences - and qualitative research, namely case studies in selected European countries. Renowned scholars from the University of Potsdam, ZHAW, and the Vienna University of Economics and Business, with extensive experience in local government and comparative research, form the consortium of this project.

Key project deliverables will be an openly accessible monitoring platform of digital transformation at the local tier of government, journal articles, an edited volume, and publications for practitioners. The real-time platform “Monitoring Digital Transformation in European Local Governments” will be accessible to researchers and practitioners worldwide and contribute to a better understanding of long-term developments. The duration of the project submitted to the SNSF/DFG is three years; however, by automating the process, the real-time platform will continue to exist and be updated regularly beyond this time frame. The research project will yield policy-relevant knowledge concerning local digitization measures from a European perspective, which can then be utilized to improve policymaking for future public sector modernization.

Keywords

digital transformation, content monitoring, data source evaluation, website embedding

1. Introduction

Digital transformation, a crucial innovation in local government, is anticipated to reshape European public service delivery, administration structures, and overall governance. The recent COVID-19 pandemic underscored the significance of well-prepared digital administration, particularly at the local government level, which plays a pivotal role in digital transformation. However, current comparative research on the digital transformation of state and administration

Joint Proceedings of RCIS 2024 Workshops and Research Projects Track, May 14-17, 2024, Guimarães, Portugal

*Corresponding author.


†These authors contributed equally.

✉ gerj@zhaw.ch (J. Gerber); saxr@zhaw.ch (J. S. Saxer); bapt@zhaw.ch (B. B. Kreiner); wele@zhaw.ch (A. Weiler)

🌐 <https://www.zhaw.ch/de/ueber-uns/person/gerj/> (J. Gerber); <https://www.zhaw.ch/de/ueber-uns/person/saxr/> (J. S. Saxer); <https://www.zhaw.ch/de/ueber-uns/person/bapt/> (B. B. Kreiner);

<https://www.zhaw.ch/de/ueber-uns/person/wele/> (A. Weiler)

<https://www.zhaw.ch/de/ueber-uns/person/wele/> (A. Weiler)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

lacks sufficient investigation into local government levels, creating a knowledge gap on implementation and effects across Europe.

DIGILOG¹ is a research project determined to close this gap. It is an international and interdisciplinary project that consists of political and computer scientists from the University of Potsdam (DE), the Vienna University of Economics and Business (AU), and the Zurich University of Applied Science ZHAW (CH). The Researchers of the project in the field of Computer Science are the contributing authors of this paper. The project is financed by the Swiss National Science Foundation (SNSF / Project Nr. 200839) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The start of the project was in spring 2022 and the end will be in summer 2025. The research project seeks to address this above-mentioned gap by examining two key questions:

- What are the dynamics, scale, and pace of digital transformation in European local governments? Is the change radical, revolutionary, incremental, or evolutionary, and are there identifiable regional differences?
- What effects does digital transformation have on these organizations, specifically in terms of output (service delivery, organization, processes, and resources), outcomes (performance and accountability), and impact (citizen acceptance, governance, and emerging tensions)?

To address these questions comprehensively, data will be collected in different ways from all municipalities in the 46 member states of the Council of Europe. As shown in Figure 1 we collect data for the different communities in three ways.

1.1. Case Studies

In conjunction with the quantitative surveys, comparative case studies are conducted in selected municipalities, which are also part of the extended survey sample. The case studies are carried out in communities with different administrative cultures to capture the country-specific variance of local administrative systems. The case study approach relies on field research methods, semi-structured expert interviews, and focus groups conducted with local CEOs, Chief Information Officers (CIOs), department heads, employee representatives, and staff. The aim is to gain in-depth insights into the internal processes and actor constellations of the respective digital transformation paths, building on the quantitative part's interim results by capturing the municipalities' organizational realities.

1.2. Survey

In addition to the qualitative case studies, the DIGILOG project is based on two quantitative forms of data collection: a web crawler for analyzing municipal websites and a survey among the leaders of European municipal administrations.

The survey has several objectives. The main goal is to collect information on the status of the surveyed municipal administrations' external and internal digital transformation, from which a

¹<https://www.digilog-project.org/>

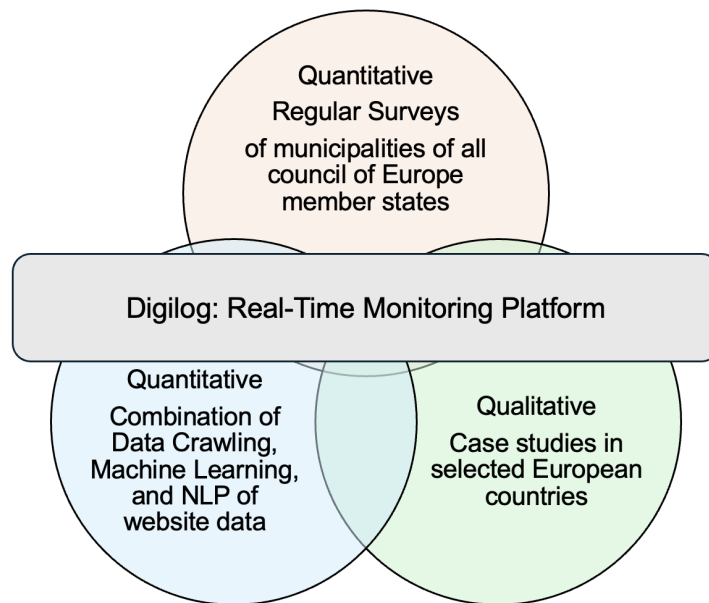


Figure 1: Three different ways of collecting data for the real-time monitoring platform for the digital transformation of municipalities in the 46 member states of the Council of Europe.

Europe-wide index will be created. In the external domain, this primarily includes the digital service offerings of the administrations, classified into various maturity levels according to an established social science model. The categorization spans from basic information provision to options for digital interaction with administrative personnel and completely digital and seamless administrative process handling. The internal domain, on the other hand, covers aspects such as the technical equipment of the administration, forms of internal communication, data management, and the automation of processes and routine decisions.

Furthermore, the survey collects data on various other variables related to digitization. These include factors that can help explain the state of digital transformation in municipalities, such as the size and organizational form of the municipality, as well as those that can reflect the consequences of digitization, such as questions about the efficiency of administration or the satisfaction of citizens with administrative work.

1.3. Web Crawler and Monitoring Platform

Web crawling is a central component of the DIGILOG project. In addition to surveys, automatic crawling and analysis of municipal websites are part of the quantitative analysis. The results of the data analysis described below will be displayed on a dashboard within a monitoring platform. Additionally, the lists of website URLs and email addresses for surveys, if not already provided, are completed through crawling public data sources.

The monitoring platform is based on three main components that interact with each other: web crawling, data storage, and subsequent data analysis. This platform ensures monitoring of the political municipalities' websites during the project duration. To manage the volume

of data, several methods enable targeted data collection with minimal information loss. One project goal is to explore and implement the most efficient method for this task.

Data storage is ensured with two different database systems, a relational and a document-oriented system. The relational system stores database keys and normalized information. Complementarily, the website documents and the analysis results are stored in a document-oriented system. For analysis, clues (e.g., mention of selected services or keywords) indicating digital transformation are extracted and evaluated. Various methods from Natural Language Processing (NLP), a subfield of Machine Learning, are applied.

The analysis, in turn, can provide effective feedback to the intelligent crawler, contributing to its continuous improvement. The quality of the analysis is ensured by domain experts who interpret and contextualize the results for management, political science, and public administration.

1.4. Measurement of Digital Transformation

Several relevant indices on digital service provision exist, offering country rankings and potentially serving as a valuable foundation for an index on local digital service provision within the scope of this project. The European Commission publishes the Digital Economy and Society Index Report on digital public services; however, it lacks specificity for the local government tier [1]. The Digital Adoption Index by the World Bank, a composite index gauging the adoption of digital technologies globally, focuses on the government sector, with sub-indices covering core administrative systems, online public services, and digital identification. The United Nations' E-Government Development Index assesses the effectiveness of public service delivery, identifying patterns in e-government development and regional challenges. Despite its Local Online Service Index focusing on the local level, evaluating the scope and quality of online services, telecommunication infrastructure development, and human capital, it only assesses portals in a selection of 100 cities worldwide, overlooking smaller local governments [2]. The E-Government Monitor, conducted through a representative survey of populations in Germany, Austria, and Switzerland, explores the usage and satisfaction related to e-government services. Results indicate a pronounced use of e-government services in Austria, followed by Switzerland and Germany [3]. Nonetheless, once again, this index lacks specificity for the local government tier. The German Index of Digitalization (Deutschland-Index Digitalisierung) scrutinizes digital infrastructure, the use of digital services, the digital economy, and e-government in individual German states but is confined to Germany [4].

2. Related Work

The project described is interdisciplinary. It intersects with the research area of political Science and Information Retrieval in Computer Science. However, we only focus on the related work of Information Retrieval related to this project.

There is already work claiming to measure the level of digital transformation within local governments. Garcia-Sanchez *et al.* [5] presents an analysis of the development of e-governments of 102 Spanish municipalities where they select features from various papers and frameworks. Pina *et al.* [6] conducted an empirical study about the effect of e-government on transparency,

openness, and hence accountability in 15 countries of the EU and a total of 318 government websites. This task of assessing websites even finds its application in other domains such as health [7].

Since we focus on website content to measure digital transformation, we note the importance of existing work on website processing, classification, and embedding, which is the encoding of data into a lower-dimensional representation in such a way that preserves some relationship in the data. We might focus on a website's visual or textual aspects, or even both, and leverage machine learning for our digitalization measurements. It's not surprising that recent work often uses Large Language Models (LLMs) and Convolutional Neural Networks (CNNs). Other classical machine-learning approaches rely more on feature engineering. However, they do not generalize as well as the state-of-the-art models due to their lack of flexibility regarding structural changes of an HTML page. A large amount of related work exists in the field of text-based embedding and classification of websites, which might help us categorize certain website elements. Kowsari *et al.* [8] and Minaee *et al.* [9] provide reviews on past work on text classification in general, while Hashemi [10] gives us a survey on web page classification. While "classification" refers to categorizing websites, before making the final prediction, we need to transform website data into a more manageable form which can involve creating embeddings for the websites. These website embeddings can be compared based on numerical similarity for various use cases. The classification models can be used to detect important digitalization elements on the website while also giving us insight into how to process websites effectively.

2.1. Visual, text, and mixed Website Classification

Visual-only classifications are, in many cases, applied to the detection of harmful content such as propaganda of terrorism [11], alcohol, adult content, weapons [12, 13] or just food, fashion or landscapes [14]. These classes all have distinct visual features. However, in many cases, these approaches can't distinguish between visually similar pages (e.g., municipality homepage vs. tourism page of the same municipality).

In text-based website classification, some approaches rely on classical machine learning [15, 16]. However, the majority are based on neural networks [17, 18, 19, 20] and the more recent approaches are transformers architecture [21, 22, 23, 24]. Most notably, [23] proposes MarkupLM for document understanding tasks based on the raw text and markup language, which is also used to code websites.

A mixed approach using both textual and visual features can be seen in [25] and [26]. The ladder encodes multiple parts of a website, such as a screenshot and metadata, and combines them to feed it into a neural network as input. The model is trained to categorize websites into 14 different classes. While previous work gives insight into how websites are processed and represented numerically, we must apply this knowledge to our specific data. How exactly website data is handled is not a solved problem. Kiesel *et al.*[27], for example, compares different web page segmentation algorithms. Dividing the page into individual segments might provide more concentrated information sources for our future algorithms. Finally, recent AI chatbots such as ChatGPT or open-source variants are capable of understanding a wide range of instructions. Recent developments have made it possible for the models to even react to image input while understanding user instruction, making them large multimodal models. They are

foundation models that can be used in a variety of ways, and they can understand website code as well as screenshots. As development continues, it is becoming easier to use these models for automatic extraction, summarization, analysis, and categorization of municipality websites. As these models generate text, natural language analysis is essential.

3. Recent and Future Work

The field of our work in this project consists of two parts:

- The URL gathering consists of the following questions: Has the municipality a website, and if so, what is the URL? Furthermore, the retrieved URLs must be distinguished from non-municipality URLs to eliminate false positives.
- The website must be preprocessed (website segmentation, selection of relevant data, and removal of noisy data) and processed. The municipality website must be assessed based on the criteria defined by political scientists. A classifying model must be capable of detecting certain features if they exist on this website.

Assessing a website requires a semantic understanding of a website by the machine learning model used to process the Websites. Whether it is URL classification (specifically discerning municipality websites from others), topic modeling (classification of services), or e-service detection on web pages, a robust foundation in embedding is essential. In our previous work, we conducted not yet published experiments with general pre-trained webpage embedding models and developed a basic embedding method to effectively differentiate municipality websites from non-municipality ones. All methods yielded very good results, with the more complex ones resulting in slightly better results. However, it's crucial to acknowledge that basic embeddings demonstrated a faster processing speed than more complex models, a significant consideration given the vast number of websites in our study. We additionally evaluated different data sources concerning their completeness of data. The categories evaluated were search engines, encyclopedias, and blind requests with fabricated URLs based on certain patterns. The retrieved URLs partially consisted of wrong URLs that did not belong to the local government or municipality. Although the URL appeared to be correct in many of those cases, containing the municipality name, the content was of another topic such as tourism, airports, other official organizations in this municipality, or even completely unrelated content to the municipality. Thus, an automated distinction and classification by analyzing the website's content was required.

Furthermore, as mentioned in Section 1.4, there are many ways of measuring digitization. In a conference paper, we defined three key aspects of our analysis, which consisted of different indices. The categories are Service Maturity (measurement of provision of information, communication possibility, and transactions), Usability (evaluation of accessibility and convenience of use), and Technical Maturity (evaluation of security and privacy). This index was published in a conference paper [28]. We tested the index on a sample of municipality websites and are currently working on implementing and applying it to the whole data set. Looking ahead, our plan encompasses the application of webpage embedding techniques for e-form detection, including webpage segmentation and relevant information extraction. Further, we plan to

leverage large Language Models for topic modeling of webpages and webpage content. This approach aims to further automate the process of monitoring the digital transformation of European communities.

4. Acknowledgment

This work is supported by Grant No. GR 200839 of the Swiss National Science Foundation (SNF) and German Research Foundation (DFG) for the research project “Digital Transformation at the Local Tier of Government in Europe: Dynamics and Effects from a Cross-Countries and Over-Time Comparative Perspective (DIGILOG)”.

References

- [1] European Commission, The Digital Economy and Society Index (DESI), 2023. URL: <https://digital-strategy.ec.europa.eu/en/policies/desi>.
- [2] UN DESA, UN E-Government Survey 2022 - The Future of Digital Government, Technical Report, New York, 2022.
- [3] Initiative D21 and TUM, eGovernment Monitor 2023, Technical Report, 2023. URL: https://initiativesd21.de/uploads/03_Studien-Publikationen/eGovernment-MONITOR/2023/egovernment_monitor_23.pdf.
- [4] Kompetenzzentrum Öffentliche IT, Deutschland-Index der Digitalisierung, 2023. URL: <https://www.oeffentliche-it.de/deutschland-index>.
- [5] I.-M. García-Sánchez, L. Rodríguez-Domínguez, J.-V. Frias-Aceituno, Evolutions in e-governance: evidence from Spanish local governments, *Environmental Policy and Governance* 23 (2013) 323–340. Publisher: Wiley Online Library.
- [6] V. Pina, L. Torres, S. Royo, Are ICTs improving transparency and accountability in the EU regional and local governments? An empirical study, *Public administration* 85 (2007) 449–472. Publisher: Wiley Online Library.
- [7] F. Monnet, L. Pivodic, C. Dupont, R.-M. Dröes, L. Van den Block, Information on advance care planning on websites of dementia associations in Europe: A content analysis, *Aging & Mental Health* 27 (2023) 1821–1831. Publisher: Taylor & Francis.
- [8] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, *Information* 10 (2019) 150. Publisher: Multidisciplinary Digital Publishing Institute.
- [9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning-based text classification: a comprehensive review, *ACM computing surveys (CSUR)* 54 (2021) 1–40. Publisher: ACM New York, NY, USA.
- [10] M. Hashemi, Web page classification: a survey of perspectives, gaps, and future directions, *Multimedia Tools and Applications* 79 (2020) 11921–11945. Publisher: Springer.
- [11] M. Hashemi, M. Hall, Detecting and classifying online dark visual propaganda, *Image and Vision Computing* 89 (2019) 95–105. Publisher: Elsevier.
- [12] A. Akusok, Y. Miche, J. Karhunen, K.-M. Bjork, R. Nian, A. Lendasse, Arbitrary cate-

- gory classification of websites based on image content, *IEEE Computational Intelligence Magazine* 10 (2015) 30–41. Publisher: IEEE.
- [13] L. Espinosa-Leal, A. Akusok, A. Lendasse, K.-M. Björk, Website classification from webpage renders, in: *Proceedings of ELM2019 9*, Springer, 2021, pp. 41–50.
- [14] D. López-Sánchez, J. M. Corchado, A. G. Arrieta, A CBR system for image-based webpage classification: case representation with convolutional neural networks, in: *The Thirtieth International Flairs Conference*, 2017.
- [15] V. K. Bhalla, N. Kumar, An efficient scheme for automatic web pages categorization using the support vector machine, *New Review of Hypermedia and Multimedia* 22 (2016) 223–242. Publisher: Taylor & Francis.
- [16] G. Matošević, J. Dobša, D. Mladenčić, Using machine learning for web page classification in search engine optimization, *Future Internet* 13 (2021) 9.
- [17] E. Buber, B. Diri, Web page classification using RNN, *Procedia Computer Science* 154 (2019) 62–72. Publisher: Elsevier.
- [18] B. Y. Lin, Y. Sheng, N. Vo, S. Tata, Freedom: A transferable neural architecture for structured information extraction on web documents, in: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1092–1102.
- [19] A. K. Nandanwar, J. Choudhary, Semantic features with contextual knowledge-based web page categorization using the GloVe model and stacked BiLSTM, *Symmetry* 13 (2021) 1772.
- [20] Y. Zhou, Y. Sheng, N. Vo, N. Edmonds, S. Tata, Simplified dom trees for transferable attribute extraction from the web, *arXiv preprint arXiv:2101.02415* (2021).
- [21] X. Chen, Z. Zhao, L. Chen, D. Zhang, J. Ji, A. Luo, Y. Xiong, K. Yu, WebSRC: a dataset for web-based structural reading comprehension, *arXiv preprint arXiv:2101.09465* (2021).
- [22] A. Gupta, R. Bhatia, Ensemble approach for web page classification, *Multimedia Tools and Applications* 80 (2021) 25219–25240. Publisher: Springer.
- [23] J. Li, Y. Xu, L. Cui, F. Wei, MarkupLM: Pre-training of Text and Markup Language for Visually-rich Document Understanding, 2022. URL: <http://arxiv.org/abs/2110.08518>, arXiv:2110.08518 [cs].
- [24] A. K. Nandanwar, J. Choudhary, Contextual Embeddings-Based Web Page Categorization Using the Fine-Tune BERT Model, *Symmetry* 15 (2023) 395. URL: <https://www.mdpi.com/2073-8994/15/2/395>. doi:10.3390/sym15020395, number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- [25] R. Bruni, G. Bianchi, Website categorization: A formal approach and robustness analysis in the case of e-commerce detection, *Expert Systems with Applications* 142 (2020) 113001. Publisher: Elsevier.
- [26] S. Lugeon, T. Piccardi, R. West, Homepage2Vec: Language-Agnostic Website Embedding and Classification, *Proceedings of the International AAAI Conference on Web and Social Media* 16 (2022) 1285–1291. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/19380>. doi:10.1609/icwsm.v16i1.19380.
- [27] J. Kiesel, L. Meyer, F. Kneist, B. Stein, M. Potthast, An Empirical Comparison of Web Page Segmentation Algorithms, 2021, pp. 62–74. doi:10.1007/978-3-030-72240-1_5.
- [28] J. Marquardt, J. Gerber, J. Machljankin, C. Kaiser, & R. Steiner, Applying web crawling for data collection in the social sciences - Opportunities and limits using the example of digital transformation in European local governments, Zagreb, Croatia, 2023.