

WEB DATA SCRAPING FOR DIGITAL PUBLIC RELATIONS ANALYSIS BASED ON THE EXAMPLE OF COMPANIES INSTALLING PHOTOVOLTAIC SYSTEMS

Dariusz ZDONEK

Silesian University of Technology in Gliwice, Faculty of Organisation and Management;
dariusz.zdonek@polsl.pl, ORCID: 0000-0002-6190-9643

Purpose: The first objective of this article was an attempt at identifying the major differences between such terms as public relations (PR), digital public relations (DPR) and digital marketing (DM). The second objective was to employ selected web data scraping techniques to analyse DPR of service providers installing photovoltaic systems.

Design/methodology/approach: The first objective of this article was achieved by analysing reference works. To achieve the second objective, the author used MS Excel, web scraping and proprietary computer scripts in R and Python. In this way, selected details were obtained from the companies catalogue at panoramafirm.pl and Google search engine, and then the received results were compared and analysed. What is more, the results from Google search engine were obtained and analysed for 964 towns and cities entered in the engine with the “photovoltaics” phrase.

Findings: 50 thousand URLs were obtained and 1,755 unique website domain addresses were extracted. Analysing the content of websites at the obtained Internet domains, 6 major categories of websites were identified, which appeared in the first 10 search results for the photovoltaic-related queries. These are: Company Websites (CW), Blog Websites (BW), Announcement Services (AS), SEO Landing Pages (SLP), Public Announcement Pages (PAP) and Social Media Page (SMP). Each of these categories is characterised briefly and a few examples are provided for each of them.

Research limitations/implications: The limitations of this article include the focus on one companies catalogue, i.e., panoramafirm.pl, and the results from Google search engine solely for the Polish language. Moreover, only the results of the first 10 links from Google engine for the single “photovoltaics” phrase and town/city name were taken into consideration.

Originality/value: This article has a theoretical and practical value. The analysis allowed to identify six categories of websites, which may be analysed with respect to digital public relations in the area of photovoltaic system installation. The most important of them are the websites belonging to the Company Website (CW) and Social Media Page (SMP) types. This article is addressed to anyone interested in obtaining data from the Internet using web scraping technique and data analysis in the area of digital public relations (DPR).

Keywords: digital public relations, Poland, cities, photovoltaics, web scraping.

Category of the paper: Research paper.

1. Introduction

Technology development, globalisation and continuous development of companies result in growing competition on the product and service market. This may affect, in particular, micro-enterprises, small and medium-sized ones (MSMEs) operating on local markets. For this reason, it is necessary to create a favourable image of the company and its products among current and prospective customers, employees and business partners. This is why the companies are more and more active in the widely understood public relations and marketing activities. This range can be further reduced to the Internet. Then, the studies will comprise the digital public relations and digital marketing.

One of recently developed business activity areas in Poland has been the market of services relating to replacing central heating boilers and installing photovoltaic systems. This service market has been developing dynamically for two years. This is proven by reports and statistical data (Raport_IOE, 2021). This is why, considering the continuous change of the theory and practice of public relations (Huang et al., 2017), (Arief & Saputra, 2019), (VanDyke & Lee, 2020) and the development of services related to photovoltaic systems installation, it is necessary to study current trends of using digital public relations in the operations of small and medium-sized service companies. To do it, analysis data must be obtained from websites and social media profiles of service providers installing photovoltaic systems. However, the problem is that there are a lot of such companies in Poland. It is infeasible to analyse the multitude of companies in terms of DPR alone. This is why it is necessary to look for and apply techniques and tools supporting it. The first step is obtaining current website addresses of installation companies in Poland.

The objectives of this article include (1) explaining major differences between the basic terms, such as public relations, digital public relations and digital marketing, as well as (2) obtaining selected web data for DPR analysis.

The article consists of 4 parts. The first part presents important differences between public relations, digital public relations and digital marketing. The second part shows the technique of using MS Excel and web scraping to obtain selected web data. MS Excel was used to retrieve collective data concerning all Polish cities from polskawliczbach.pl. Web scraping served to obtain address details of service providers installing photovoltaic panels from panoramafirm.pl and Google search engine data. The third part presents the results obtained. The data taken from the websites catalogue at panoramafirm.pl was compared to the one from Google search engine for the key phrase company name + company address. What is more, the results of analysing data obtained from Google search engine for the phrase of the city/town name and photovoltaics are presented. The final part contains discussion and conclusions.

2. Public relations, digital public relations and digital marketing

Advertising, the marketing communication tool used most frequently in commercial companies, has been losing its power of effective communication as its recipients are overloaded with advertisements and become resistant to traditional advertising stimuli. Advertising, as a one-way mass media tool, has not been as effective as it used to. This is why, it becomes necessary to change the structure of traditional marketing communications to more customised and interactive formats. A technique to improve the marketing communication efficiency is to shift the leading role from advertising to public relations. The public relations tools used in practice may ensure superior transparency in internal and external communications of the organisation and, at the same time, offer more effective marketing communication (Naumovska & Blazeska, 2016).

There are numerous definitions explaining what public relations are. In 1976, Rex Harlow analysed 472 definitions of public relations (Theaker, 2020, p. 3). He defined public relations as a specific management function, which enables to establish and maintain two-way lines of communication, understanding, acceptance and cooperation between the organisation and the general public (social environment), covering problem or issue management, helping the managers to notify the general public on an ongoing basis and respond to it, defining and emphasising managers' responsibility to serve public interest, helping managers to follow changes and use them efficiently, also acting as an early alarm system, which helps to foresee trends (Harlow, 1976; Agee et al., 1992, p. 3; Theaker, 2020, p. 3). Public relations are defined as a reputation as well, i.e. the results of what you do, what you say and what others say about you (Theaker, 2020, p. 5). Public relations comprise all types of communication between the organisation and its social environment (Mahendra, 2020). Marketing is popularly associated with any activities in the company, which are to promote product or service sale.

Some researchers base the difference between marketing and public relations on two words, i.e., consumer and profit. The traditional marketing role is to understand the consumer and produce products or services, which will satisfy their needs with a benefit for the supplier (Theaker, 2020, p. 6). For public relations, this is a favourable image of the company and its products, i.e., **reputation**.

Digital public relations and digital marketing are based mostly on websites and digital platforms. Traditional media, i.e., printed press, business cards, hard-copy catalogues etc., are not entirely gone, but are used to a decreasing degree. Noticeably, more and more organisations have been using different digital platforms to carry out public relations activities. They created new dynamic channels to sell products, create consumers' groups, increase website traffic and the awareness of existing concerns (Alexander, 2016). Digital public relations are a foundation of effective competitive strategy now (Gifford, 2010). Using digital

public relations enables to create trust between the service provider, the product, the idea and the consumers (Mohamed, 2022).

Digital marketing and digital public relations are two separate notions. They may use different techniques, tools, impact area and results. Examples of differences between the digital marketing and digital public relations described in reference works and scientific works are presented in Table 1.

Table 1.

Examples of differences between the digital marketing and digital public relations

| | Digital marketing | Digital public relations |
|--------------------------------|--|---|
| General | The act of selling or promoting company products and services. The activities include Internet advertising as well. | Maintaining a favourable public image of the company, organisation or products. |
| Examples of experts' tasks | Buying advertising space on digital platforms; Planning advertising campaigns of products and services; Developing promotional materials; Developing forms of sale promotion and special offers. | Developing press releases concerning new organisational initiatives, products or services; Updating and managing company releases; Developing topics for discussion to involve Internet community in discussions concerning the company and its products; Participating in charity and social campaigns on the Internet. |
| Examples of success indicators | Has the sale of products or services increased and how much? Was it possible to achieve a higher ROI (Return on Investment) and how much higher was it? Has the number of visits to product or service pages increased and how much? | Was it possible to arouse the interest in the company, products or services among the Internet users? Was it possible to create favourable publications or responses concerning the company, products and services and how much? Has the number of favourable opinions and comments increased and how much? |

Source: own work based on (Website startup, 2021).

However, in practice marketing and public relations in commercial companies have much in common. They comprise activities aimed at building awareness of the company brand and products, as well as protecting its reputation on the market. Thanks to it, the company becomes more recognisable and competitive. Although they may be separate, the combination of marketing and public relations, as well as the other way round, often makes the end objectives even more visible for the company.

Public relations (PR) have evolved from the traditional to the digital era. At present, we can speak of PR 1.0, 2.0, 3.0 or even 4.0 (Arief & Saputra, 2019; Permatasari et al., 2021). Public relations 1.0 is a traditional PR era. In it, printed media were the basic information distribution channel. This meant the prevailing role of one-way communication from a single source to many targets, i.e., one-to-many communication. Public Relations 2.0 was the era when Internet media were born. It was characterised by many-to-many communication. This era witnessed the earliest days of printed media transfer to digital platforms. Public Relations 3.0 is the era when social media were born. It was when social, professional and even employee journalism appeared. Public Relations 4.0 is the era of artificial intelligence and big data.

This is when artificial intelligence solutions are introduced into public relations. Human public relations cooperate with robots, which are not only able to monitor the network and write publications, but also schedule digital content uploading and respond quickly to any discussions and opinions posted online (Arief & Saputra, 2019). This is why skilful, automated web data scraping may soon become a highly important area of organisation operations relating to digital public relations.

3. Materials and methods

When studying web data scraping to analyse digital public relations of service providers installing photovoltaic panels, answers to the following study questions were sought:

1. What are the techniques, advantages and disadvantages of scraping web data concerning companies installing photovoltaic systems?
2. What share of companies installing photovoltaic panels in panoramafirm.pl catalogue has their own websites? Are they up-to-date? Are address details from website catalogues (e.g., from panoramafirm.pl) sufficient and reliable?
3. How many websites relating to photovoltaic panel installation in Poland can be obtained by analysing the first 10 Google search results for the key phrase town/city name and photovoltaics? Are they solely websites of companies installing photovoltaic panels in the given town/city?

The studies were carried out from December 2021 to April 2022. Their major focus was analysing websites relating to photovoltaic system installation in Poland.

The study plan was composed of 4 stages repeated in several iterations. The study diagram is presented in Figure 1.

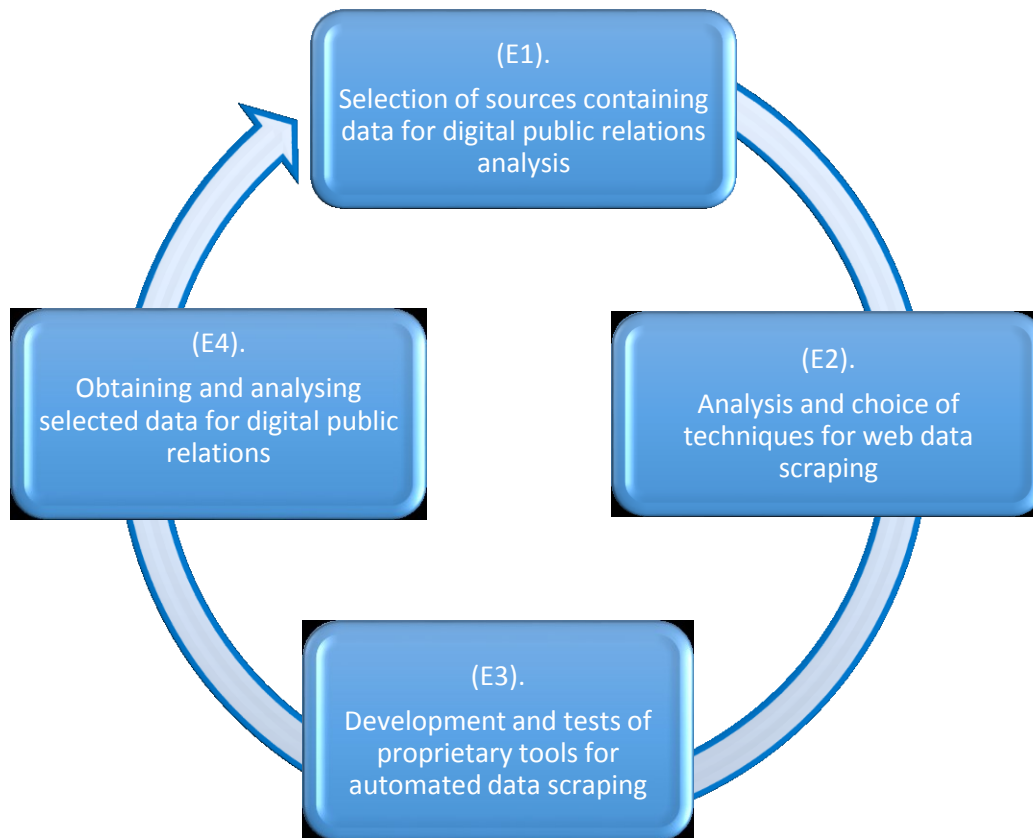


Figure 1. Study stages.

The first stage (E1) consisted in selecting sources containing data for digital public relations analysis. In this case, it was decided to analyse the content of a companies catalogue of the popular Polish portal (panoramafirm.pl). This website is available on the Internet for free. It contains address and website details of companies, e.g., those installing photovoltaic systems. Initially, when scraping the data, the “copy-paste” technique was used. However, this is highly time-consuming. It is not useful when analysing big data, i.e., the website address catalogues. This is why, it was decided to look for techniques to automate that operation, which can be used in practice.

The second stage (E2) consisted in analysing and selecting techniques to automate web data scraping. The techniques included: importing data from the Internet in MS Excel; using R language; using Python language, using libraries, such as Pandas, BeautifulSoup, HTMLSession, os, requests (Dogucu & Çetinkaya-Rundel, 2021; de las Heras-Pedrosa et al., 2020; BeautifulSoup Doc., 2021; GitHub Facebook Scraper, 2021), using API (Application Programming Interface), using `get_posts` module from the `facebook_scraper` package (GitHub Facebook Scraper, 2021) etc. It was decided to test various techniques and employ them in practice to retrieve data from panoramafirm.pl. The selected techniques included MS Excel, as well as R and Python languages with additional libraries.

The third stage (E3) consisted in the development and tests of proprietary tools for automated web data scraping. That was the most time-consuming stage of the study. It required getting technical expertise and applying it in practice.

The fourth stage (E4) consisted in obtaining and analysing data from the website catalogue (panoramafirm.pl). The data analysis carried out required another iteration of the presented research diagram. Consequently, several further iterations were carried out according to the diagram presented in Figure 1. The subsequent iterations brought web data for all Polish towns/cities and Internet domain addresses for websites relating to photovoltaics from Google search engines. The phrase used was town/city name + photovoltaics. The results were analysed and assumptions for the next iteration of studies were developed. Each iteration was followed by the analysis of advantages and disadvantages of the techniques and tools employed.

4. Results

During the first iteration, data was scraped from panoramafirm.pl. Once “photovoltaics” was entered, the list of more than 25 pages with company addresses was displayed. Due to the high number of entities, retrieving the data manually and analysing it would be very time-consuming and monotonous. This is why work was automated. Proprietary scripts in Python and R were developed. Thanks to a loop, all pages were read out and data of 568 companies was retrieved. The data was saved in an xlsx file. The saved data included a company name, address, URL of a website, e-mail, phone number and service category. Once duplicates and poor-quality data were removed, 442 companies were obtained. The most frequent categories, in which the companies in this database operated, are listed in Table 2.

Table 2

Service category data obtained using web scraping from panoramafirm website for the photovoltaics phrase

| Category | Category (Polish name) | Number |
|---|----------------------------------|--------|
| Renewable energy | Energia odnawialna | 135 |
| Electric power distribution | Dystrybucja energii elektrycznej | 107 |
| Electrical installation works | Elektroinstalatorstwo | 98 |
| Electrical heating | Ogrzewanie elektryczne | 35 |
| Heating installation and maintenance | Instalacja i serwis ogrzewania | 26 |
| Turnkey building and finishing works | Budowa i wykończenia pod klucz | 16 |
| No classification | Brak klasyfikacji | 13 |
| Air-conditioning maintenance and installation | Serwis i instalacja klimatyzacji | 9 |
| Renewable energy generation | Wytwarzanie energii odnawialnej | 9 |
| Renewable heat sources | Odnawialne źródła ciepła | 8 |
| Heating devices | Urządzenia grzewcze | 6 |

Source: own work.

There were 240 URLs provided for the companies (54%). Next, it was checked how many of those addresses were active and could be opened in a browser. It turned out that 212 out of 240 addresses were active (88%). For 28 (12%) of addresses, the browser reported security issues or website absence.

In the second iteration, it was decided to verify the consistency of URLs in panoramafirm.pl and the ones obtained in Google search engine. For that purpose, a company name and town/city were entered (1). Next, data was retrieved from the Google business profile (2-5) and the search results (6-8). All numbers are plotted on Figure 2.

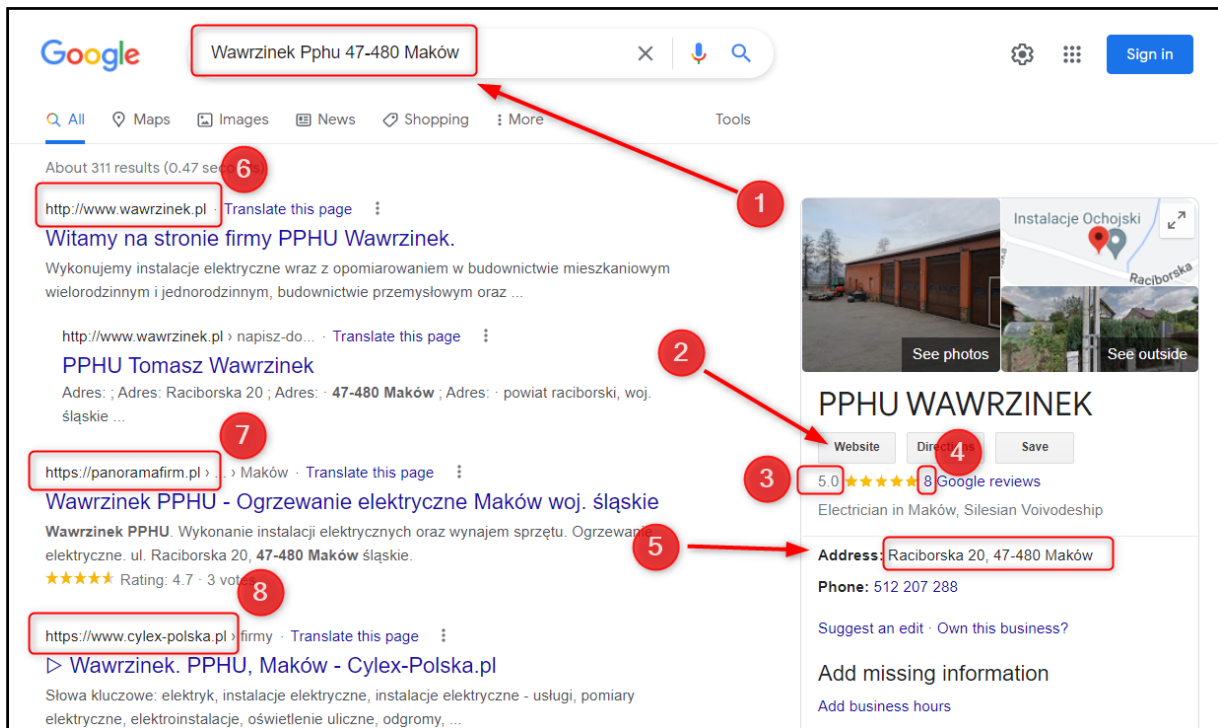


Figure 2. Data obtained from Google search engine for the phrase consisting of company name and town/city.

To automate the process, a script was written in Python using an automated test platform called Selenium. The data was collected in a separate worksheet. Next, it was unified to consist solely of main Internet domains and https:// protocol. The obtained results were compared and the consistency of the results from panoramafirm.pl companies catalogue and Google search engine was verified.

It turned out that websites' URLs were consistent in 151 cases. The results are presented in Table 3.

Table 3.

Comparison of URLs of companies installing photovoltaic panels at panoramafirm.pl and in Google search results (business profile – Google Button Website button)

| | Data from panoramafirm.pl (PF) | Data from Google Button Website (GBW) search engine |
|--|--------------------------------|---|
| Number of companies | 442 | 442 |
| Number of URLs | 240 | 265 |
| Number of Internet domain addresses consistent in PF and GBW | 151 | 151 |

Note: PF – panoramafirm.pl, GBW – Google Button Website.

Source: own work.

What is more, it turned out that many of the first Google search results showed addresses of various popular Polish companies catalogues (Table 4).

Table 4.

10 most popular websites taking initial positions in Google search results after the company name and address was entered from panoramafirm.pl websites catalogue

| URL | Number |
|----------------------------|--------|
| https://panoramafirm.pl/ | 174 |
| https://aleo.com/ | 48 |
| https://biznesfinder.pl/ | 24 |
| https://krs.infoveriti.pl/ | 15 |
| https://krs-online.com.pl/ | 13 |
| https://dnb.com/ | 12 |
| https://mapa.targeo.pl/ | 10 |
| https://krs-pobierz.pl/ | 9 |
| https://pkt.pl/ | 4 |
| https://aktualnyodpis.pl/ | 3 |

Source: own work.

As not all companies had their URLs in the websites catalogue (panoramafirm.pl), and when their names were entered in the browser solely 265 addresses were obtained (Google Button Website), it was decided to expand data scraping by entering the phrase consisting of a town/city name + photovoltaics. This is why another iteration was carried out based on the study diagram (Figure 1).

During the study, data for all 964 towns/cities in Poland was obtained from polskawliczbach.pl/miasta. Data presented there come from Statistics Poland (GUS_cities, 2021) and display the situation as at 31 December 2020 (Figure 3).



| | RANKINGI | WOJEWÓDZTWA | POWIATY | GMINY | MIASTA | WSIE |
|----|-------------|--------------------|---------|--------------------|---------|------|
| 5 | Poznań | powiat Poznań | | wielkopolskie | 532 048 | |
| 6 | Gdańsk | powiat Gdańsk | | pomorskie | 470 805 | |
| 7 | Szczecin | powiat Szczecin | | zachodniopomorskie | 398 255 | |
| 8 | Bydgoszcz | powiat Bydgoszcz | | kujawsko-pomorskie | 344 091 | |
| 9 | Lublin | powiat Lublin | | lubelskie | 338 586 | |
| 10 | Białystok | powiat Białystok | | podlaskie | 296 958 | |
| 11 | Katowice | powiat Katowice | | śląskie | 290 553 | |
| 12 | Gdynia | powiat Gdynia | | pomorskie | 244 969 | |
| 13 | Częstochowa | powiat Częstochowa | | śląskie | 217 530 | |
| 14 | Radom | powiat Radom | | mazowieckie | 209 296 | |

Figure 3. Database of Polish towns and cities.

It was impossible to extract the data using “copy+paste” technique to MS Excel, as the data did not save correctly. It would be too time-consuming to copy individual items. For this reason, the technique of importing data from the Web was used in MS Excel (MS Excel > Data > Get Data from Other Sources > From Web). After the URL <https://www.polskawliczbach.pl/Miasta> was entered, the entire content of the database was retrieved and saved in a cities.csv file.

For the subsequent iteration, it was decided to enter the phrase consisting of the town/city name + photovoltaics in the search engine and then to retrieve the resultant URLs (Figure 4).

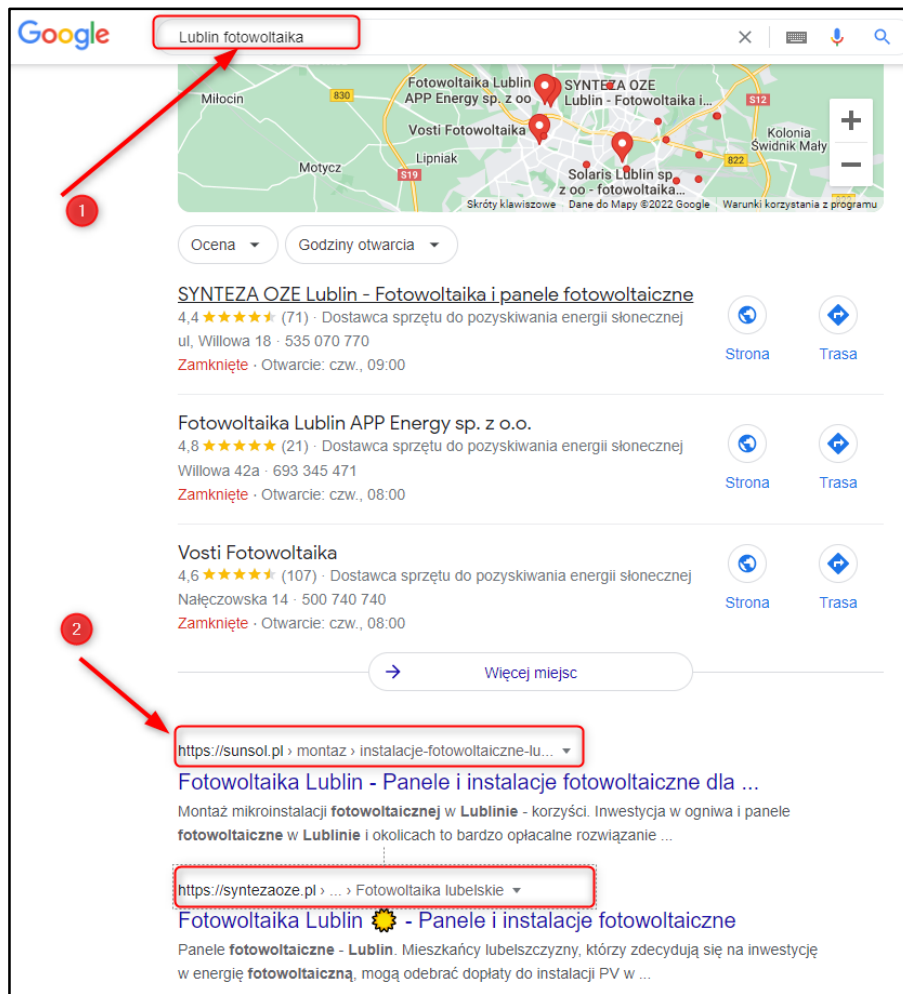


Figure 4. Obtaining data from the Google search engine for the town/city name and photovoltaics.

Data was collected solely from the first results page. No paid advertisements and business profiles on Google maps were considered. Analysing and saving such a high number of pages would not be possible if it were not for using scripts automating the data saving operations. In order not to overload servers with excessive number of queries, high delay was set for queries and the code was set to save to text files to enable its analysis on a later date. The obtained results were analysed. They are presented in Tables 5 and 6.

Table 5.

Aggregate number of data retrieved from the search engine for the query of a town/city name and photovoltaics (for 964 towns/cities in Poland)

| Variable | Number |
|--|--------|
| Total URLs retrieved from the search engine. | 53020 |
| 1 Internet domain instance | 1271 |
| < 4 Internet domain instances | 287 |
| < 10 Internet domain instances | 117 |
| < 100 Internet domain instances | 63 |
| >= 100 Internet domain instances | 17 |
| Total unique Internet domain addresses | 1755 |

Source: Own study.

Table 6.

List of 15 Internet domain addresses appearing most frequently in search results for the key phrase of a town/city name and photovoltaics (for 964 towns/cities in Poland)

| Internet domain address | Results number |
|---|----------------|
| https://oze.net.pl | 760 |
| https://fotowo.pl | 668 |
| https://farmy.pl | 531 |
| https://ongeo.pl | 514 |
| https://www.oferteo.pl | 478 |
| https://www.fotowoltaika.org.pl | 351 |
| https://syntezaoze.pl | 247 |
| https://miastoslonca.pl | 232 |
| https://sunsol.pl | 224 |
| https://sitarsky.pl | 217 |
| https://hymon.pl | 209 |
| https://kalkulatormocy.pl | 206 |
| https://mkfotowoltaika.pl | 147 |
| https://www.olx.pl | 137 |
| https://www.brewa.pl | 125 |

Source: Own study.

Having saved the unique Internet domain addresses and analysed the content of 100 selected websites they redirected to, a division into 6 main website categories was proposed:

- Company Website (CW) – company websites offering products and services relating to photovoltaic systems.
- Blog Website (BW) – blogs, guidebooks and useful tools relating to photovoltaic system installation.
- Announcement Services (AS) – companies catalogues and announcement services (AS).
- SEO Landing Page (SLP) – websites, which are positioned and developed by specialist companies to reach the most prospective customers. Those are usually simple websites with main key phrases and a contact phone number, aimed at achieving high positions in search engines in response to specific key words.
- Public Announcement Page (PAP) – announcement websites of government agencies and local government authorities informing about e.g., programmes, investments and grants relating to photovoltaic system installation.
- Social Media Page (SMP) – public websites and groups connected with photovoltaic system installation in popular social media.

Table 7 presents examples of website addresses included in particular groups.

Table 7.

Categories and examples of websites in the area of photovoltaic panel installation

| ID | Category | Examples of websites |
|----|--------------------------------|--|
| 1 | Company Website (CW) | https://www.brewa.pl ; https://www.soltechenergy.pl ; https://www.esoleo.pl ; https://www.brewa.pl ; https://eeef.pl ; https://sundaypolska.pl ; https://jbenergia.pl ; https://zatokaenergii.pl |
| 2 | Blog Website (BW) | https://fotowoltaikaonline.pl ; https://ongeo.pl ; https://enerad.pl/ |
| 3 | Announcement Services (AS) | https://www.olx.pl ; https://www.oferteo.pl ; https://fixly.pl ; https://panoramafirm.pl/fotowoltaika |
| 4 | SEO Landing Page (SLP) | https://oze.net.pl ; https://fotowo.pl ; https://farmy.pl ; https://kalkulatormocy.pl ; https://mkfotowoltaika.pl ; https://www.solar2biznes.pl ; https://instalacje-solarne.pl ; https://fotowoltaicznainstalacje.pl |
| 5 | Public Announcement Page (PAP) | https://samorzad.gov.pl/web/gmina-bransk/projekt-oze-2 ; https://samorzad.gov.pl/web/gmina-kalwaria-zebrzydowska/odnawialne-zrodla-energii-przewodnik ; http://bip.wolin.pl/strony/13976.dhtml ; https://www.pruszkow.pl/srodowisko/dofinansowanie/moj-prad/ |
| 6 | Social Media Page (SMP) | https://pl-pl.facebook.com/FotowoltaikaStargard/ ; https://www.facebook.com/fotowoltaikastarachowicemarcingalka/ ; https://pl-pl.facebook.com/protonenergia/ |

Source: Own study.

The subsequent stage would be an automatic classification of every website into specific categories. Depending on the website category, it would require proposing new analysis tools and techniques. For a Company Website, it would be obtaining information concerning, e.g., social media and opinions in a Google business profile. While for a Social Media Page, it would be, e.g., obtaining information on the number of news, comments etc. Every additional information requires new script development, testing and practical application. This means that this is a highly time-consuming process, though required for a comprehensive analysis of DPR data sets.

5. Discussion and conclusions

Analysing a single, specific Internet activity, we can perceive it in the context of digital marketing and digital public relations. The difference may be slight for commercial companies, but an attempt can be made to identify it. The studies carried out in Europe in 2020 among 2,324 managers revealed that PR practitioners faced more moral challenges in their everyday work than 8 years earlier. There are growing concerns relating to bots, big data, influencers in social media and sponsored content (Hagelstein et al., 2021). It is a difficult task to create online content able to push its way through the digital chaos and appeal directly to consumers. There is a problem of competing with a multitude of information, choosing the suitable

platform, tools and forms of content distribution (Myers, 2016, pp. 40-55), as well as techniques and tools for web data scraping and analysis.

This article focuses on the market of photovoltaic panel installation services in Poland. It presents techniques of obtaining address details using a spreadsheet and scripts in Python and R. Automated address details scraping to analyse digital public relations is an introduction to a comprehensive analysis. The first step was to obtain complete, up-to-date and reliable data concerning website addresses containing information important for a DPR analysis. It was highly time-consuming to prepare suitable tools to scrap and cleanse data. Attention had to be paid to tool testing and error handling. For those applications, web scraping technique was employed, as well as R and Python scripts were developed. This allowed to obtain 1,755 unique Internet domains for websites relating to photovoltaic panel installation in 964 Polish towns and cities based on ten first Google search results for the phrase consisting of the town/city name and photovoltaics. Web scraping, as a technique of obtaining web data, is used by a growing number of companies and organisations, as well as practitioners and scientists. It is used, e.g., to collect data concerning prices of listed stock on an ongoing basis, with the frequency higher than standard (Lin & Yang, 2022) and to collect and compile prices of products with high variability frequency in different locations (Benedetti et al., 2022). Web scraping is also used to collect and analyse employers' needs concerning competences of future employees from the most popular advertising websites (Wykwarski, 2021) and to collect and analyse Facebook data concerning criminal activity relating to illegal trading in prohibited products (Xu et al., 2020). This technique is sometimes employed to collect data to detect abusive comments on social media, including Twitter, YouTube and Facebook (Vrysis et al., 2021), and to collect for sentiment analysis during COVID-19 (Bhagat et al., 2021). There are numerous publications devoted to collecting data from social media, i.e. Reddit (Higgins et al., 2021), Twitter (Hernandez-Suarez et al., 2018; Dongo et al., 2021), Facebook (Mancosu & Vegetti, 2020). There are even manuals concerning scraping data from the web and social media and using it in analytics (Russell, 2013). The Internet has become a vast source of information, which may support decision-making processes both in companies and in public entities. Analytical departments may obtain data from the market and social media and then analyse it.

The analysis of websites relating to photovoltaic system installation based on URLs obtained in this study reveals that they may be divided into 6 major categories. They include Company Websites (CW), Blog Websites (BW), Announcement Services (AS), SEO Landing Pages (SLP), Public Announcement Pages (PAP) and Social Media Pages (SMP). Only a small fraction of them can be ascribed to more than one category.

For this reason, when creating a database of entities for DPR analysis, it is necessary to consider different categories of websites and prepare suitable analytical tools and techniques for them. This should be kept in mind when planning any operations and data analysis in this area, with respect to digital public relations and digital marketing. The website categories identified in the article are the best example. Business profiles should be analysed with respect

to DPR, different than the government and local government websites or social media. For this reason, it is important to consider the category of websites used by the organisation and serving for its comparison to others. Thanks to analytics, managers are able to understand the environment where their companies operate better and adapt their activities to the rapidly changing market.

The Internet becomes a more and more important environment of promotion and building business relations with others. This is why DPR activity and the ability to obtain suitable data from the Internet in this respect and analysing is may become a more and more important component of the operations of contemporary public organisations and commercial companies.

References

1. Agee, W.K., Ault, P.H., & Wilcox, D.L. (1992). *Public Relations: Strategies and Tactics*. Harper Collins.
2. Alexander, D. (2016). What digital skills are required by future public relations practitioners and can the academy deliver them? *PRism*, 13(1), 1-13.
3. Arief, N.N., & Saputra, M.A.A. (2019). Kompetensi baru public relations (PR) pada era artificial intelligence [Public Relations in The Era of Artificial Intelligence]. *Jurnal Sistem Cerdas*, 2(1), 1-12. <https://doi.org/10.37396/jsc.v2i1.19>.
4. Beautiful Soup Doc (2021). *Beautiful Soup Documentation*. Available online: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>, 25.01.2021.
5. Benedetti, I., Laureti, T., Palumbo, L., & Rose, B.M. (2022). Computation of High-Frequency Sub-National Spatial Consumer Price Indexes Using Web Scraping Techniques. *Economies*, 10(4), 95. <https://doi.org/10.3390/economies10040095>.
6. Bhagat, K.K., Mishra, S., Dixit, A., & Chang, C.-Y. (2021). Public Opinions about Online Learning during COVID-19: A Sentiment Analysis Approach. *Sustainability*, 13(6), 3346. <https://doi.org/10.3390/su13063346>.
7. de las Heras-Pedrosa, C., Sánchez-Núñez, P., & Peláez, J.I. (2020). Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems. *International Journal of Environmental Research and Public Health*, 17(15), 5542. <https://doi.org/10.3390/ijerph17155542>.
8. Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education*, 29(sup1), S112-S122. <https://doi.org/10.1080/10691898.2020.1787116>.
9. Dongo, I., Cardinale, Y., Aguilera, A., Martinez, F., Quintero, Y., Robayo, G., & Cabeza, D. (2021). A qualitative and quantitative comparison between Web scraping and API

- methods for Twitter credibility analysis. *International Journal of Web Information Systems*, 17(6), 580-606. <https://doi.org/10.1108/IJWIS-03-2021-0037>.
10. Gifford, J. (2010). Digital Public Relations: E-Marketing's Big Secret. *Continuing Higher Education Review*, 74, 62-72.
 11. GitHub Facebook Scraper (2021). *Facebook Scraper*. Available online: <https://github.com/kevinzg/facebook-scraper>, 25.01.2021.
 12. GUS_cities (2021). *Główny Urząd Statystyczny—Statistics Poland—Cities, Voivodship*. stat.gov.pl. <https://stat.gov.pl/en/topics/other-studies/cities-voivodship/>.
 13. Hagelstein, J., Einwiller, S., & Zeffass, A. (2021). The ethical dimension of public relations in Europe: Digital channels, moral challenges, resources, and training. *Public Relations Review*, 47(4), 102063. <https://doi.org/10.1016/j.pubrev.2021.102063>.
 14. Harlow, R.F. (1976). Building a public relations definition. *Public relations review*, 2(4), 34-42.
 15. Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). *A Web Scraping Methodology for Bypassing Twitter API Restrictions*. arXiv:1803.09875 [cs]. <http://arxiv.org/abs/1803.09875>.
 16. Higgins, J.A., Lands, M., Valley, T.M., Carpenter, E., & Jacques, L. (2021). Real-Time Effects of Payer Restrictions on Reproductive Healthcare: A Qualitative Analysis of Cost-Related Barriers and Their Consequences among U.S. Abortion Seekers on Reddit. *International Journal of Environmental Research and Public Health*, 18(17), 9013. <https://doi.org/10.3390/ijerph18179013>.
 17. Huang, Y.-H. C., Wu, F., & Huang, Q. (2017). Does research on digital public relations indicate a paradigm shift? An analysis and critique of recent trends. *Telematics and informatics*, 34(7), 1364-1376. <https://doi.org/10.1016/j.tele.2016.08.012>.
 18. Lin, C.-F., & Yang, S.-C. (2022). Taiwan Stock Tape Reading Periodically Using Web Scraping Technology with GUI. *Applied System Innovation*, 5(1), 28. <https://doi.org/10.3390/asi5010028>
 19. Mahendra, A.I. (2020). Simplification of the role of public relations. *Jurnal ASPIKOM*, 5(1), 102-114.
 20. Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, 6(3), 205630512094070. <https://doi.org/10.1177/2056305120940703>.
 21. Mohamed, F.A. (2022). The Role of Social Responsibility in the Digital Public Relations Age. In: A.M.A. Musleh Al-Sartawi (ed.), *Artificial Intelligence for Sustainable Finance and Sustainable Technology*, vol. (pp. 480-492). Springer International Publishing. https://doi.org/10.1007/978-3-030-93464-4_47.
 22. Myers, C. (2016). Disclosure in Online Promotions: The Effect of FTC Guidelines on Digital Public Relations and Advertising. *Social Media for Government*. Routledge, 40-54.

23. Naumovska, L., & Blazeska, D. (2016). Public relation based model of integrated marketing communications. *UTMS Journal of Economics*, 7(2), 175-186.
24. Permatasari, A.N., Soelistiyowati, E., Suastami, I.G.A.P.P., & Johan, R.A. (2021). Digital Public Relations: Trend and Required Skills. *Jurnal ASPIKOM*, 6(2), 373-386.
25. Raport_IOE (2021). *Rynek fotowoltaiki w Polsce*. <https://aviasolar.pl/wp-content/uploads/2021/05/Raport-Rynek-Fotowoltaiki-w-Polsce-2021.pdf>.
26. Russell, M.A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O'Reilly Media, Inc.
27. Theaker, A. (2020). *The public relations handbook*. Routledge.
28. VanDyke, M.S., & Lee, N.M. (2020). Science public relations: The parallel, interwoven, and contrasting trajectories of public relations and science communication theory and practice. *Public Relations Review*, 46(4), 101953.
29. Vrysis, L., Vryzas, N., Kotsakis, R., Saridou, T., Matsiola, M., Veglis, A., Arcila-Calderón, C., & Dimoulas, C. (2021). A Web Interface for Analyzing Hate Speech. *Future Internet*, 13(3), 80. <https://doi.org/10.3390/fi13030080>.
30. Website startup (2021). *Różnica między marketingiem cyfrowym a PR*. <https://startup.info/the-difference-between-digital-marketing-and-pr/>.
31. Wyskwarski, M. (2021). Requirements for Project Managers—What Do Job Advertisements Say? *Sustainability*, 13(23), 12999. <https://doi.org/10.3390/su132312999>.
32. Xu, Q., Cai, M., & Mackey, T.K. (2020). The illegal wildlife digital market: An analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation*, 47(3), 206-212. <https://doi.org/10.1017/S0376892920000235>.