4/25/2022

Ivan Ivanov

2672214

What is the current state of energy consumption of Web sites?

[Document subtitle]

Contents

[Abbreviations 3](#_Toc105084204)

[Abstract 3](#_Toc105084205)

[1. Introduction 3](#_Toc105084206)

[Thesis Structure 4](#_Toc105084207)

[2. Theoretical Framework 4](#_Toc105084208)

[3. Research Method 5](#_Toc105084209)

[Method 5](#_Toc105084210)

[Software 6](#_Toc105084211)

[4. Calculations/API Description 6](#_Toc105084212)

[Energy per visit in kWh (E): 6](#_Toc105084213)

[Emissions per visit in grams CO2e (C): 6](#_Toc105084214)

[Annual energy in kWh (AE): 7](#_Toc105084215)

[Annual emissions in grams CO2e (AC): 7](#_Toc105084216)

[Annual Segment Energy: 7](#_Toc105084217)

[Annual Segment Emissions: 7](#_Toc105084218)

[Limitations 7](#_Toc105084219)

[5. Data Analysis & Results 7](#_Toc105084220)

[General Analysis 7](#_Toc105084221)

[Overview 7](#_Toc105084222)

[Top 1000 9](#_Toc105084223)

[Bottom 1000 9](#_Toc105084224)

[Hosting Type 9](#_Toc105084225)

[Per Domain 9](#_Toc105084226)

[Original Domains 9](#_Toc105084227)

[Domain Popularity (Top 20) 9](#_Toc105084228)

[Other Domains 9](#_Toc105084229)

[Per Regional Domain 9](#_Toc105084230)

[European Union and United Kingdom 9](#_Toc105084231)

[North America 9](#_Toc105084232)

[South America 9](#_Toc105084233)

[Asia 9](#_Toc105084234)

[Middle East 9](#_Toc105084235)

[Oceania 9](#_Toc105084236)

[Africa 9](#_Toc105084237)

[Per Company 9](#_Toc105084238)

[6. Discussion 10](#_Toc105084239)

[7. Conclusion 10](#_Toc105084240)

[References 10](#_Toc105084241)

[Appendices 10](#_Toc105084242)

# Abbreviations

|  |  |
| --- | --- |
| TWh | Terawatt-hour |
| MMT | Million Metric Tonnes |
| IEA | International Energy Agency |
| EPA | (US) Environmental Protection Agency |
| WRI | World Resources Institute |
| KWg | Kilowatt-hour per GB |
| TLD | Top-level Domain |

# Abstract

[abstract body]

# Introduction

Awareness towards the threat of climate change has increased in the in the last years[[ref](https://climatecommunication.yale.edu/publications/climate-change-in-the-american-mind-december-2018/%20)]. Sustainability, a practice concerned with the recognition of the finite nature of resources, and their conscious usage with regards to future generations has also experienced a similar increase in recognition[[ref](https://www.epa.gov/sustainability/learn-about-sustainability)]. As a result of this growing interest, there has been a larger amount of CO2-related scientific publications published in the past years[[Fabbrizzi et al. 2016](https://www.researchgate.net/publication/295841100_Sustainability_and_Food_A_Text_Analysis_of_the_Scientific_Literature)].

Along with the rise in climate change related concerns, there has also been a growth in global Internet data traffic. In the year 2020 alone ([PCH ref](https://www.pch.net/ixp/summary_growth_by_country)) web traffic exchange has expanded with an average of 40% in size, and has been previously predicted to experience further expansion with close to five and a half billion internet users expected by 2023. ([Cisco 2018](https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html)) This explosion in traffic growth is driven partially by the Coronavirus 2020 pandemic but also by the consistent increases in usage of Internet-connected devices from 2011 onward ([Ericsson 2021](https://www.ericsson.com/4ad7e9/assets/local/reports-papers/mobility-report/documents/2021/ericsson-mobility-report-november-2021.pdf), p. 3), and it is mostly caused by video streaming, conferencing, online gaming and social networking([IEA 2021](https://www.iea.org/reports/data-centres-and-data-transmission-networks)). Device adoption rates vary widely per region though it is estimated that in 2018 the average United States citizen owned about 10 Internet connection-capable devices and consumes 140 gigabytes of data per month ([The Shift Project 2019, 61](https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf))

As these devices become more and more of a central point in human lives the Internet usage grows rapidly as well. 92% of internet users nowadays access the web with mobile phones. ([Statista 2022](https://www.statista.com/statistics/1289755/internet-access-by-device-worldwide/)) Because of this growth the amount of site hits increases with each year and that, in turn leads to a collective increase in overall data consumption ([United Nations 2019, XV](https://unctad.org/system/files/official-document/der2019_en.pdf)).

Data centers are the backbone of website accessibility and they consume around 1% of global electricity use ([Mansanet 2020](https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf)). They are estimated to use around 200-250 TWh of electricity per year ([IEA 2021](https://www.iea.org/reports/data-centres-and-data-transmission-networks)). In general, the usage has fallen “by a factor of four” due to improvements in hardware efficiency.([Mansanet 2020](https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf)).

About the split of electricity usage between data centers and devices.

All of these have one thing in common: they generate CO2.

CO2 emissions have been growing rapidly for more than a hundred years. In the early 1900s they have been estimated to have been around 1000 MMT of carbon dioxide. In the 2010s these numbers reached almost 10,000 MMT ([EPA 2022](https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#:~:text=Global%20carbon%20emissions%20from%20fossil,increase%20from%201970%20to%202011.)). This growth has been driven by a variety of factors: increased globalization, fossil fuel usage, population increase and more ([Climate.gov](https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide#:~:text=Carbon%20dioxide%20concentrations%20are%20rising,people%20are%20burning%20for%20energy.)). Numerous attempts have been made to decrease these emissions, from the Paris Agreement to public awareness campaigns and government reforms([ref](https://www.gov.uk/government/publications/net-zero-strategy)) ([WRI 2019](https://www.wri.org/insights/which-countries-have-long-term-strategies-reduce-emissions) ref 2). The IT sector alone amounts to 1.4% of global emissions but it can be reduced to below 20% of that if a switch to renewables were to happen ([Ericsson 2020](https://www.ericsson.com/4907a4/assets/local/reports-papers/consumerlab/reports/2020/ericsson-true-or-false-report-screen.pdf)).

(Will need to add something about the type of energy used)

There has been research made about the IT sector’s emission generation but none of them have attempted to analyze the web’s CO2 footprint. In this thesis, an attempt will be made to create a clear overview of some of the web’s most popular websites and their CO2 generation footprint.

To do so, the following research question will be addressed:

*What is the current state of energy consumption of Web sites?*

## Thesis Structure

In the next chapter the [ ] is defined. Chapter 3 talks about [ ] and Chapter [ ] is about [ ]. In the last chapter an answer to the research questions will be provided and suggestions for further research as well.

There have been attempts in estimating the electricity consumption of Internet transmissions, defined as the “energy consumed per amount of data transmitted” ([Coroama et al. 2013, 2](https://files.ifi.uzh.ch/hilty/t/Literature_by_RQs/RQ%20100/2013_Coroama_Hilty_Heiri_Direct_Energy_Demand_of_Internet.pdf)) Each website hit generates [insert amount generated).

# Theoretical Framework

The following have been defined by Geringer et. al. (2020) as the main drivers on a website’s ecological footprint:

1. Performance: Optimizing a website decreases its size, which, in turn reduces the number of bytes that need to be loaded on every page refresh. The biggest factors for a decreased website performance are: images, videos and scripts.
2. Hosting: If a website receives energy from renewable sources then and only then it can be considered ‘green’.
3. Content Strategy & SEO: Good search-engine optimization means that a website will be found faster because the user is spending less time opening other websites to find the relevant information they need.
4. UX-Design: Appropriate user experience design means that the user can find the information they’re looking for faster.

To these four, I will also consider here a fifth factor: type of the website. Some websites lend themselves to a higher footprint. Streaming services generally generate more CO2 because of the high amount of content a user consumes there and on the other side websites like Wikipedia.com are naturally less of an offender of this due to their textual nature. To add to that, independent of what kind of content a website serves is the fact that some sites are inherently structured to a high amount of page refreshes and/or content loading. Examples to that can be YouTube.com where a playlist of short and addicting clips can lead a user to stay on it for hours and Twitter.com which constantly fetches photos and videos on the user’s home page.

# Research Method

The purpose of this chapter is to explain the research methods used in this paper.

## Methodology

To write this paper, books, academic articles, journals, reports, and statistical analyses have been used. Additionally, the data used has been collected from several sources: Website Carbon’s API, [list other sites].

The data analysis will be strictly quantitative. The goal here is to analyze the impact of our internet surfing habits on a relatively large scale and to compare an innocuous habit with the real-world impact it has. Due to the nature of the service that the data is sourced from, the analysis will be done entirely on the homepage of a given website and because of that there will be no focus on analysis of other types of data transfer like streaming video from a particular streaming service, loading multiple pages from the same website or infinitely scrolling pages (e.g., Twitter, Facebook, NBCNews.com).

The main data has been collected by using the Website Carbon API. It is an online tool created by Wholegrain Digital that provides an estimate for the amount of CO2 a website generates. It awaits a query in the form of a URL and returns a JSON file containing several statistics about the website.

The dataset where all the statistics have been collected is a .csv file generated from the data retrieved by the aforementioned API. It contains information about the 50,034 most popular websites and it has the following features:

1. Website URL
2. Type of hosting – Depends on the energy source used by the data center. Saved as either ‘True’ for websites hosted by a service provider using green energy or ‘unknown’ for those whose green status could not be determined. The status is determined by the Green Web Foundation’s own API. ([ref](https://www.thegreenwebfoundation.org/)) There, any websites that is hosted by a ‘Green’ data center is shown as using Green energy (Note: Not all centers mentioned on the GWF website use Green energy, in some cases they use standard grid and the emissions are offset afterwards.).
3. Number of bytes – The amount transferred upon the initial page load, provided that the website hasn’t been visited before.
4. Number of bytes (adjusted) – An adjusted value which takes into consideration browser caching.
5. Percentage of sites it is cleaner than – A simple comparison between the amount of CO2 the currently tested website generates and the others in the database. Comparison is done at [ ].
6. Energy – The amount transferred upon a single page load, in Kilowatts per Gigabyte.
7. Grams and liters of CO2 transferred for the renewable grid
8. Grams and liters of CO2 for the national grid.

The website list we are using has been sourced from the [Tranco](https://tranco-list.eu/) list of 1 million most popular websites. Tranco is a list which uses averaged data from four other ranking providers (Alexa, Cisco Umbrella, Majestic and Quancast). The data has been [explain Tranco here] ([Pochat et. al. 2019](https://tranco-list.eu/assets/tranco-ndss19.pdf)). The original list was retrieved at 03/05/2022.

The dataset that was used for analysis has been sourced between the period of 04/05/2022 and 21/5/2022.

### Errors and Limitations

The set has been cleaned up of any accidentally repeated data and several manual edits have been made for the following issues:

1. Google redirects: Some URLs (like Google’s regional domains) redirect to the main Google.com domain. This created several hundred duplicate “https://www.google.[region]/” entries. In this case the redirected entries have been removed and only the first one from each region has been kept with the presumption that it is the original one.
2. General regional redirects: The same issue appeared for several other websites and has been dealt with in the same way.

Additionally, the data has been formatted to account for visual clarity and ease of use (decimal sign placement, general formatting).

An occurring issue encountered during this data collection process was also a large number of websites which could not be analyzed for different reasons. These reasons are as follows:

1. The website was offline.
2. The website was hosted by Cloudflare – In this case the API returned a response in an HTML form and not JSON. Each time that happens the corresponding website is skipped.

A few other reasons, as mentioned by Wholegrain Digital on their website. Those are:

1. The website can be accessed by the public through a standard web browser.
2. It does not require login.
3. It allows search engines.
4. Contains unique content aimed at human visitors – this excludes holding pages, error pages, server notification pages, demo pages or pages that are generally useless (this is highly subjective).
5. Is free from illegal or explicit content.

The original goal of the project was to parse the entire one million website list. Unfortunately, another limitation was encountered at the data collection process: the API used has a daily limit of 25,000 hits available to it. The API is also shared with other users and because of that the number of sites that could be parsed per day was no more than two to three thousand.

Overall, the first 65,600 websites from the Tranco list were parsed. 52,431 of those were parseable and after the removal of any duplicates there were 50,034 usable websites. The loss from the parsed to parseable is 21.1% and from parseable to usable is an additional 4.6%.

## Software

The research question and the scope of the paper were defined in the previous sections. Here, the tools used for the data collection process are described.

There were two main methods used:

* An HTTP Request/Response program (async\_alternate.py) which main function is to send out asynchronous requests to the WebsiteCarbon API and receives JSON formatted responses in return. It has been written with Python 3.10.4 64bit.

Additionally, these libraries were used:

1. ASYNCIO ([Python.org](https://docs.python.org/3/library/asyncio.html)). Part of the Python Standard Library. Utilizes concurrent programming concepts to make asynchronous programming possible. It is used because asynchronous requests fare multiple times faster than sequentially programmed ones, depending on the implementation and the receiving server’s limitations. [detailed explanation about concurrency)
2. AIOHTTP ([Documentation](https://docs.aiohttp.org/en/stable/)) written by Nikolay Kim and Andrew Svetlov. A client/server library that utilizes ASYNCIO for making asynchronous HTTP GET requests targeted towards a public endpoint from the WebsiteCarbon API.

* The WebsiteCarbon API which provides the data to the Python program.

[fill out when you get access to the API source]

# Calculations/API Description

The amount of emissions generated by a webpage are calculated with the following factors:

1. Data transfer over the wire – the amount of data that is transferred from the server to the user upon a page load.
2. Energy intensity of web data – The amount of energy used to load a page.
3. Energy source used by the data center – can be either True or False. True represents the energy used by the data center coming from green sources, False means standard energy. This is checked through the Green Web Foundation database.
4. Carbon intensity of electricity – Carbon intensity is the amount of carbon dioxide generated to create a unit of electricity.( <https://www.nationalgrideso.com/future-energy/net-zero-explained/what-carbon-intensity>) Usually measured in grams of CO2 per kilowatt-hour. Here it is based on the international average for grid electricity.
5. Website traffic – The number of page views on a website.

The API uses a number of formulas to make the exact calculations. They work as follows [(Sustainable Web Design 2022](https://sustainablewebdesign.org/calculating-digital-emissions/)):

### Energy per visit in kWh (E):

E = [Data Transfer per Visit (new visitors) in GB x 0.81 kWh/GB x 0.75] + [Data Transfer per Visit (returning visitors) in GB x 0.81 kWh/GB x 0.25 x 0.02]

### Emissions per visit in grams CO2e (C):

C = E x 442 g/kWh (or alternative/region-specific carbon factor)

### Annual energy in kWh (AE):

AE = E x Monthly Visitors x 12

### Annual emissions in grams CO2e (AC):

AC = C x Monthly Visitors x 12

### Annual Segment Energy:

Consumer device energy = AE x 0.52

Network energy = AE x 0.14

Data center energy = AE x 0.15

Production energy = AE x 0.19

### Annual Segment Emissions:

Consumer device emissions = AC x 0.52

Network emissions = AC x 0.14

Data center emission = AC x 0.15

Production emission = AC x 0.19

# Data Analysis & Results

It can start with tables showing the amount of data I’m working with.

Then there will be a structured analysis, on a few major categories and subcategories.

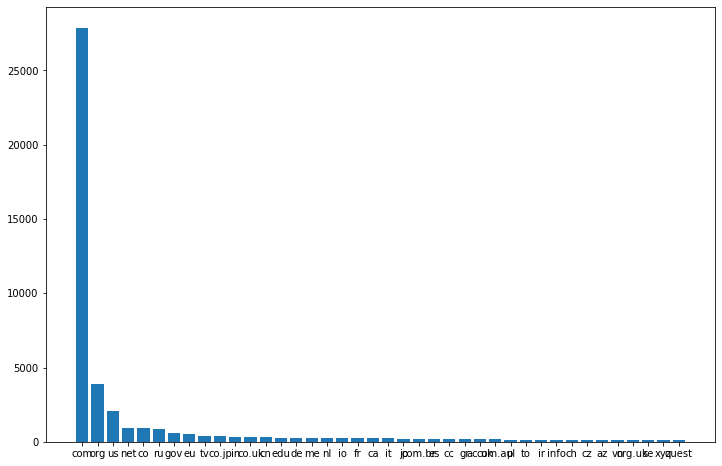
## General Analysis

### Overview

#### Dataset Distribution

Top-Level Domains: There are 50,034 websites in the final database. From those, 27,873 have the “.com” top-level domain, 3868 have “.org“ and 2065 have “.net“ as the domain. Full distribution follows in the table snippet below and the full table is located in the appendices [app index here]. Overall, there are 716 different top-level domains. Keeping only those which occur more than 100 times leaves us with exactly 40 domains. The distribution is heavily skewed towards the first 3 TLD’s which account for 67.5% of the entire dataset.

Figure 1: TLD Distribution



Chart, bar chart

Description automatically generatedWebsite Sizes: A website’s size affects the time it will take to load it and the amount of CO2 it will generate. The sites collected vary greatly in range. The smallest one is [] bytes and the largest one is <https://www.lematin.ma/> at 304MB. These are only outliers though as the mean size is only 3.69MB and the unaffected by outliers median stands at 2.23MB. Overall, all the websites take 184.93GB of space.

A webpage, though, has two different sizes registered under it. One, under the column ‘Bytes’ contains the amount the server transfers upon the initial page load, but the second one, under the ‘Statistics: CO2’ columns is the amount transferred on a second load, with caching taken into account.

#### CO2 Distribution

On average, a site generates 0.562 grams of CO2 per first page load. The minimum amount registered is 0.000042 but the largest one is much bigger, at 76 grams per load.

### Top 1000

### Bottom 1000

For these two we will look at the general trends, in a similar way to the overview section but we will compare the differences between the entire dataset and the top and bottom, on average. Also will compare the difference between the top/bottom.

### Hosting Type

We will look at how much of a difference the type of hosting used makes on the emissions. Things that can be looked at are the differences in averages between the two categories here and the proportion of big sites that have green vs unknown hosting. The amount of overly big sites (those that are much bigger than the average).

## Per Domain

### Original Domains

The original domains are looked at in terms of size, amount of co2 that they contribute to the overall dataset, and it will be shown how many unknown hosting sites every domain has.

### Domain Popularity (Top 20)

Sourced from <https://w3techs.com/technologies/overview/top_level_domain>

Will every category have the same general data in it?

### Other Domains

## Per Regional Domain

### European Union and United Kingdom

### North America

### South America

### Asia

### Middle East

### Oceania

### Africa

For these there will be a few standard measurements shown and they will be compared between each other.

## Per Company

The different companies will be analyzed in a manner similar to the Regional section. Look at which companies say are green friendly and check if their websites are so as well.

Other Ideas: Per category of site (requires paid API)

# Discussion

# Conclusion

# References

Adams, C. et al. (2022) Calculating Digital Emissions, In *Sustainable Web Design,* (Accessed: April, 12 20221) <https://sustainablewebdesign.org/calculating-digital-emissions/>

# Appendices

[notes]

Mention any issues, topics and things to look into for future research