Surveying the .NZ Top Level Domain: Cookie Policy

Liv Fletcher

***Abstract*—Web cookies are a pivotal asset for businesses and organisations alike. Their diverse utility ranges from offering personalised user experiences to enabling marketers to ascertain user engagement metrics such as page visits, page visit downtime, and on-site interactions. Such data can grant the website a personalised user experience, which in turn can develop into monetary gain. Yet, as websites continue to employ varied and more complex tracking-based cookies, so do the concerns surrounding their ethics and legality.**

This research aims to contextualise this discussion in a New Zealand context, analysing a list of all .NZ domains including all related subdomains such as .org.nz, .govt.nz, .co.nz, and more. To achieve this, we developed a script capable of scraping web cookies. This method enabled us to extract and analyse cookie data from nearly 250,000 NZ websites.

The results have unveiled intriguing patterns regarding the types of cookies New Zealand businesses employ.

These findings can act as a reference point, not just for businesses in New Zealand, but their corresponding user bases. This research not only reveals the current landscape of cookie usage in New Zealand, but can serve as a foundational study for future digital ethics inquiries, policy making, and for businesses seeking to align their practices with known standards while respecting user privacy.

I. Introduction

IN today's digital age, cookies serve as an essential tool for businesses to gain insights into user behaviour to better understand their customer base and to be able to deliver more targeted advertisements to propel their financial success. In this report, we aim to discover the various types of cookies employed by all Top-Level-Domains (TLD) and the correlating sub-domains associated with .nz (New Zealand). This includes but is not limited to organizations (.org.nz), government agencies (.govt.nz), educational entities (.ac.nz) etc. Due to handling such a large data-set, we will be facing some problems we need to consider and thus, we will explore the various options available that can assist us in completing this project. Our primary aim for this project was to thoroughly examine the prevalence and nature of cookie usage across all New Zealand TLD domains, assessing their overall adherence to standards stipulated by The New Zealand Privacy Act 2020 [1].

In the process of refining our script, it was of up-most importance to ensure its efficiency and its overall ability to capture all available cookie types, not just a select few. We have faced specific challenges, that we have had to adapt our project to take into account, such as:

- We needed to counteract potential security risks when accessing malicious domains.
- We needed to take into account domains that may contain blocking measures to ensure no 'bot' activity.
- We needed to design the script to handle domains that are no longer hosted or, currently defunct.
- We needed to set-up a time-out functionality, to ensure no domains overuse resources.
- We need to ensure the scraping measure itself does not require too much processing, taking more energy than required.

Addressing these issues is crucial, given the scale of the domains list and corresponding cookie output, and also in terms of emphasizing sustainable consumption and production patterns. Navigating through the myriad of domains naturally demands substantial computational resources and, consequently, energy use. For our project to critically align with the U.N's Sustainable Development Goal 12, we conscientiously implemented strategies to minimize our energy consumption and promote reasonable usage of computational resources.

After data-collection of the New Zealand top-level domains, the task of organising and analysing the data proved substantial. As a considerable amount of cookies contain a unique string identifier to the base cookie, this further complicates our ability to quickly and efficiently search by a given cookie name.

This project was supervised by Arman Khouzani.

II. Related Work

In this chapter, we will review similar articles and research papers that have helped develop our understanding in our projects field and provide relevant information that can propel our projects goals.

A. Can I Opt Out Yet? [2]

Understanding and analysing data privacy practices, particularly as they relate to online tracking, is of paramount importance. The research paper titled "Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control" provides a valuable perspective into this dynamic landscape. The team, which consists of 7 researchers ,spearheaded this comprehensive analysis of how the European Union's General Data Protection Regulation (GDPR), has impacted cookie-based tracking on the web.

Historically, before the introduction of the GDPR, data privacy norms had been ambiguous and implemented unevenly across websites. The eCommerce and digital marketing industry heavily relies on targeted advertisements to turn profits.

The report highlights the huge-scale profit margin (est 220 billion in 2018) is due to the marketing and analytics data from website cookie data.

The researchers took a deep dive into the behaviour of 2,000 high traffic websites, both within and outside the EU. Their primary objective was not merely to see if these sites complied with the GDPR, but to gauge the broader influence of the GDPR on the privacy norms on the internet globally.

This study documents the state of cookie tracking post-GDPR, and contrasts this with the scenario that existed prior to the legislation.

Findings

The research underscores the pervasive nature of web tracking via cookies. An overwhelming 92% majority of websites within their list were found to indulge in some form of tracking even before notifying the user via cookie consent notices or similar. The paper highlights that the findings could hint towards a lack of awareness in regards to the GDPR stipulations.

The report findings can be categorised as the following:

- *Illusion of Control.* A significant highlight of the study is the disparity between the illusion of control and the actual practices. Merely 4% of the analysed websites offered a clear opt-out option on the cookie consent notice. However, the effectiveness of this option is questionable as even after opting out, a significant proportion of websites either maintained or even increased their tracking output. This is a significant finding, and indicates the challenges users face in genuinely exerting control over their online data.
- *Question of Consent.* The research brings to light the trend of 'tacit consent', where users' continued navigation on a given website is deemed as an agreement to the setting of cookies. This disregards the GDPR's guidelines, and detracts from the principle of requiring explicit and informed consent.
- *Regional Influence.* A notable observation was made from the research was the differential adherence and influence of the GDPR across the globe. While websites in the EU exhibited discernible effects of the regulation, USA websites were also shown to be influenced by it. Although the rest of the globe are not required to include cookie consent notices, it appears that the GDPR has influenced this addition to outside the EU.
- *Cookie Longevity.* Highlighted concern is the relation to the lifespan of cookies where it was discovered that approximately 90% of websites generate cookies with a life span of more than a year. The question arises regarding the necessity and justification for such long-lasting trackers, especially since the GDPR has a mandate for minimal data storage durations.
- *Third Party Involvement.* Another finding was the role of third-party services in cookie notices and user settings. It was noted that a small percentage of websites employ third-party control over these functions. This raises a discussion around the trustworthiness of employing third-party entities for cookie management.

Strengths

- *Diverse Data.* This study's data set encompasses many different kinds of popular websites from around the world, shedding light onto the GDPR's far-reaching impact. This global approach has made the paper significant as it portrays a comprehensive picture of the GDPRs direct influence.
- *Rigorous Methodology.* The report showcases a strong methodology. Using a custom browser plugin, a similar approach to our employment of WebDrivers, the research team analysed the cookies set by these websites both before and after opt-out attempts, considering the provided user information, privacy policies, and available privacy controls.
- *Analytical Depth.* The report showcases detailed observations of the gathered data. The researchers found that the GDPR effects are palpable across the globe. Demonstrating a connection even to popular US-based websites. The US-based websites exhibited behaviour akin to their EU counterparts. This indicates a significant indirect influence of the regulation even in regions where it isn't legally binding.

Similarities

While the aforementioned study provides a comprehensive analysis of GDPR's global influence, our research focuses specifically on cookie collection across New Zealand TLD domains. By narrowing down our geographical focus, our project will build upon this research and delve deeper into the nuances of regional behaviour, regulation, and practices, providing stakeholders with more tailored and actionable insights.

Furthermore, our study aims to assess adherence to New Zealand, and the relevant privacy commissions such as The New Zealand Privacy Act 2020, thus filling the gap left by the global GDPR-centric focus of the research paper.

The research manually employed a custom browser plugin to assess cookie behaviour on websites, while our research utilised automated tools and scripts designed to scrape and analyse cookies across a vast array of NZ websites. This paper has provided an alternative view on how this project could be completed, and the vast amount of output data that could be collected in these means.

B. A Snoop at Privacy Issues on the Internet in New Zealand [3]

The conversation surrounding how New Zealand citizens data is handled on the internet is too far and in between and the conversation needs to be had, so we can catch up with the world's current technological ecosystem. We have comparatives available, such as the principles within the GDPR providing a great example for how organisations should handle their user data. The research paper by 'Winnie Chung' from the University of Auckland, delves into this issue and provides a perspective in how data in how New Zealand was handled in the early 2000's.

Findings

The research findings begin with providing a direct definition and categorisation of what privacy is, and what that looks

like in a human-digital context. Lim, 2000 [4], defines the specific instance of invasion of privacy, relating that to when an individual does not and is not able to continue, maintaining control over the usage and sharing of their personal details. As technology continues to advance, with the addition of data collection means such as web cookie data, the control an individual has over their details becomes more and more blurred.

The report discloses a survey taken of 750 New Zealanders in 2001 by UMR Research LTD shows that 86% of surveyors were concerned, and 76% were very concerned of the idea that an online business monitors, and collects data on consumers browsing the internet without their consent.

Another component of the paper's focus is in regards to this information privacy aspect, where individuals have raised concern over the control of the usage, release and circulation of their personal data. The main concerns discussed within the report are these key points that; secret tracking occurs on all website visits, unauthorised capture and usage of personal data for marketing purposes, sale of personal information to third parties without permission, and lastly, the theft of personal credentials such as credit card information. A defining factor of these concerns is in relation to website cookies, and how they are being employed.

The paper elaborates on how cookies, often used for user identification on websites, can be a concern due to the implications of how and what data is being collected from these cookies. The user is unaware of the data being stored, and what that entails. From a business perspective, cookies provide a tool for generating profit by using third party marketing companies that relate customer profiles over other visited webpages, offering more targeted advertisements. Another discussed point of concern was the usage of web bugs, which can range from simply tracking user behaviour across many domains, to directly executing scripts for various objectives. Such bugs can expose a user's device to malicious actors that manipulate the bug due to lack of security, and could steal files/data from the user's hard drive without their knowledge.

The other side of the debate are arguments against these raised concerns. One such counter-argument is that the level of data extraction gained as technology develops is essentially a scare tactic. The comparison would be that of an employee in a physical shop monitoring its customers, assessing what they may be interested in to potentially help put a sale through. The sentiment here is that people accept the in-store counterpart of surveillance without much concern, and should be considered the same when it comes to online shopping. The cookie counter-argument was that it was fear-based bait where cookies can't specifically extract personal details such as names, addresses etc unless they have been willingly provided by the user. A comparison was made with that of mail order catalogues where consumers voluntarily provide personal information. However worth noting, as this paper was released in 2002, modern cookie capabilities have since changed, and are more complex than when this research was endeavoured upon.

Strengths

The overall strengths of the report can be categorised as such:

- *Comprehensive Overview.* The report provides a thorough examination of privacy concerns, weaving together a historical context, and a background for our projects development.
- *Balanced Perspective.* By presenting arguments both for and against internet privacy concerns, the report ensures an unbiased and balanced view, encouraging readers to critically engage with the content and form their own opinions.
- *Legal Framework.* The discussion and relation to The New Zealand Privacy Act 1993 version - offers readers a solid understanding of the legal backdrop against which privacy issues could unfold. This contextualises the technological and ethical concerns within a regulatory framework.

Similarities

The paper delves into the correlation of user privacy on the internet, and directly relates it to the most recent New Zealand Privacy Act 1993 as per the paper's publishing date. The section refers to the implications of what organisations in New Zealand are legally allowed to do with user data and refers to two specific principles in relation to this. The aim of our project is to not only scrape NZ TLD web cookie data, but to draw the overall comparison of what cookies New Zealand as a whole employs, and what that means in accordance with the Privacy Act 2020.

Another similarity lies in regards to the research and understanding that cookies hold a significant role in regards to user privacy, and how organisations are allowed to handle user data. Within the reports section 'Technological Solutions' they discuss on how advancements in technology are being harnessed to address privacy concerns (cited from Lim 2000) [4]. Furthermore, the report touches on the updates at the time, to browsers such as Internet Explorer which introduced enhanced privacy controls for user data protection measures, such as privacy controls for cookies, providing details for the purpose of each cookie and differentiating between first-and third party cookies.

In conclusion, the report, despite being produced 20 years ago, is still relevant today, and has helped provide a background into the history of user data protection online, and what the implications of organisations mishandling user data looks like.

III. Laws and Regulations

New Zealand, like many countries, has its unique set of laws and regulations governing digital privacy and user data protection. These laws are crafted to ensure that businesses and websites operate within the bounds of legality, and to safeguard the rights and privacy of New Zealand Citizens. Within this section, we will discuss the direct laws and relate them with our project's research, and also delve into an international perspective to build a better understanding of how modern laws and regulations currently stand with online privacy.

A. Historical Context

The main compliance factors in regards to user internet data usage is disclosed within The New Zealand Privacy Act [1]. The most recent revised version is the 2020 document, where beforehand it was the 1993 version that handled the guidance for organisations handling user data, whether digital or physical [5]. The long-awaited updated legislation provides a modernised approach, bringing about several changes to better reflect the realities of our developing digital age. The main comparisons between the two versions are as follow:

Mandatory Reporting of Privacy Breaches A significant addition to the 2020 version is the requirement for an organisation to report any and all serious privacy breaches to those directly affected. In the case that the breach poses a harm or risk to an individual or group of individuals (identify theft, financial loss, etc), the involved organisations must notify both the affected individuals and the Privacy Commissioner [1, Part 6, sections 112-118].

Cross-border Data Flow The 2020 Act contains new provisions regarding the transfer of personal data out of New Zealand. This section states that organisations can only send data overseas in the situation that the recipient country contains comparable privacy protections. [1, Principle 12, section 22].

Compliance Notices Organisations are now required to abide by The Privacy Commissioner's requests and must follow notices to do something, stop doing something, in order to comply with the Act [1, section 123].

Right to Personal Data If an individual requests the personal information that is stored from a given organisation and the request is denied, the Privacy Commissioner has the authority to issue the individual's right to access their own data/information [1, Section 91].

Overseas Organisations The Act explicitly states that international organisations that are doing business within New Zealand, are strictly subject and must abide by the Act.

Additional Offences The 2020 version introduces new criminal offences, such that it is considered a crime to mislead organisations in a way that affects the personal data of another individual, or to mislead them into destroying or requesting the data of another individual.

Strengthening Privacy Principles The 2020 revision retains the original 12 privacy principles, but provides further clarification and emphasis on certain points such as delving deeper into the responsibilities of businesses and organisations concerning the collection of personal data, ensuring it is done for lawful purposes and is in a manner that doesn't intrude upon the individuals affairs.

Targeted Marketing and Online Privacy The Act provides clearer rules and guidance for businesses around the use of personal information for targeted advertisements, especially unsolicited electronic messages (emails etc) [1, Principle 7, sections 18 and 19].

In summary, while the New Zealand Privacy Act 1993 laid the foundation for privacy regulations in the country, the Privacy Act 2020 strengthens the protections to better correlate to the modernisation in today's digital age. The revised version emphasises accountability, particularly in the areas of data breaches and cross-border data transfers, and provides the Privacy Commissioner with enforcement capabilities.

B. Cookiebot: In Compliance with the New Zealand Privacy Act

[6] Although New Zealand does not have specific laws in place for how and what Cookies are allowed to be hosted within New Zealand websites, there are some existing guidelines that provide an overall basis of connection to legitimate principles within jurisdiction such as, The New Zealand Privacy Act 2020. Cookiebot offers CMP - Consent Management Platform solutions that provide websites with the ability to display cookie consent banners, and available cookies in compliance with regulations such as those set by the Act.

Cookiebot also provides a scanning tool that checks if a given website is currently running within global standards (GDPR, ePR). Although most cookies are essential and employed for general website functionality and ease of access for the user base, we will draw our focus to specific cookies that arguably are seen as unnecessary, and for many, a breach of privacy.

Cookiebot Results

As Cookiebot scanning results can be time consuming, I have categorised and scanned New Zealand's top 3 most used websites.

Domain	Cookies	GDPR Compliant	NZ Compliant
Stuff.co.nz	241	No	Yes
Nzherald.co.nz	252	No	Yes
Trademe.co.nz	61	No	Yes

1) **Stuff Website Notes**
   - The scan results show 30 necessary, 4 preferences, 55 statistics, 123 marketing and 29 unknown.
   - The list contains extensive details regarding each cookie, and their general purpose.
   - The majority of the cookies are collected from the initial page loading, however the standards set by The New Zealand Privacy Act 2020 do not require that a domain preemptively receives user consent before doing do.
2) **NzHerald.co.nz**
   - The scan results show 252 cookies were identified, with 25 necessary, 2 preferences, 65 statistics, 125 marketing and 35 unknown.
   - A majority of the marketing/statistics cookies are collected upon page loading, signifying that the majority of the tracking happens before the user's awareness of such tracking.
3) **TradeMe.co.nz**
   - The scan results show 61 cookies were identified, with 8 necessary, 9 statistics, 17 marketing and 27 unknown.

- Majority of marketing cookies send the data to the United States. Similar to above, most cookies are upon initial page-loading.

However, although these websites are considered non-compliant as per GDPR and ePR standards, it is not entirely applicable in a New Zealand context. The New Zealand standard is set by The New Zealand Privacy Act 2020 of which states that the NZ Privacy Principles are of the responsibility of the website owner and operator to ensure the adequate information is provided to users who request it. The main section within the Act that directly correlates to website owners, is Principle 3 - 'What to Tell an Individual'.

The New Zealand Privacy Act 2020

Principle 3 requires a website owner to ensure the following:

- Users are aware of the data collection methods
- The reason of the data collection
- The agency the data is being shared with (e.g third party)
- Where the data is stored and for how long.

A note in regards to the above considerations, is that the domain is required to inform the users in regards to these actions before any data has been collected on an individual. Personal data contains vast amounts of information regarding an individual, not just the obvious 'name, age, location etc'. Personal information a website can gather on you can include many more aspects such as; social security numbers, signature, passport numbers, race, political views, religious beliefs, sexual orientation, health and genetic information and on the technical side, IP addresses, unique user ID's (Google Analytics etc), search and browser history, device information, purchase and shopping history and many more.

In a New Zealand context, this has been integrated into a form of cookie consent notices, however, due to not having any specific guidelines, have been a subject of discussion in regards to the readability of these consent notices.

Many websites have adopted cookie consent notices in order to abide by the domains local compliance standards (GDPR, ePD, CCPA etc), however, there has been discussion in relation to the way some domains utilise their CMP (Cookie Management Platform) notice to deceive a user into accepting an agreement they may not otherwise, given the circumstances. A research paper titled '"Okay, whatever": An Evaluation of Cookie Consent Interfaces' [7] discusses the misuse of cookie consent notices, by employing immoral means, specifically known as 'dark patterns' [add the ref]. Dark patterns is a design technique that some domains employ, that essentially nudges a user towards decisions that may compromise their privacy.

Another research paper titled '(Un)informed Consent: Studying GDPR Consent Noticed in the Field' [8], provides another perspective into this issue. The papers focus on how users on a day-to-day basis behave when interacting with cookie consent notices, and that the content displayed on the notice, can easily affect the users decision to accept/decline. The research discovers that from an examination of over 80,000 participants, the influence of various design aspects of these notices, including their placement, choice type and content framing heavily attributed to the uninformed consent. Users did not want to necessarily consent to specific cookies, but felt fatigued by the extensive list on the notice, and accepted anyway to 'get rid of it'. The paper concludes with the need to establish clearer regulatory guidance on what and how websites display their consent notices.

In a New Zealand context, there is no provided guide for how a given website presents their privacy policy, or cookie consent notices, as long as they provide the relevant data as per the Act. However, many of the principles hold significant value, some could be misconstrued, or seen as a 'grey area'.

- **Principle 1.** A website is only allowed to collect personal information, if it is in direct connection with the functions of the web page itself. These points could be skirted around, for example a News website is collecting data such as general information, and in addition, sexual orientation and religious views. The argument could align with why does the website require this type of collection? However the argument could be that it is in relation to analytics, and marketing purposes. [9]
- **Principle 3.** A website is required to notify its users about why the data is collected, whether the data collection is compulsory or not, and who the data is shared with. A large portion of New Zealand domains employ cookie consent notices as means to abide by this principle, but most just include this information within their subpage that includes their privacy policy. However, this could be viewed as inadequate, as most users don't bother with reading the contents of a cookie consent notice, or navigating to a given domain's privacy policy page. [9]
- **Principle 4.** A website is required to only collect personal information that is considered legal and fair. This includes not coercing or misleading users to give out their personal data. The argument here is the data is being requested, and the wording used on the consent notice themselves. Some believe that consent notices are somewhat misleading, and difficult to manoeuvre, and users in frustration just accept the notice 'to get rid of it'. [9]
- **Principle 5.** A website is required to have proper safeguards in place around the collected user data to prevent loss or misuse of personal information. What would be considered misuse? And how is that defined? [9]
- **Principle 6.** A website is required to allow for a given user to request access to all collected information about themselves. However, in some cases the provided output data can be difficult to navigate/understand. [9] This has also been discussed in the paper "Okay, whatever" released April 2022 [7].
- **Principle 7 and 8.** A website is required to allow for the correction and accuracy of user information. [9]
- **Principle 9.** Websites are not allowed to store and use the collected personal data for longer than intended. What is the maximum length? How is that decided? This principle does not seem entirely specific, and could be looped around. [9]
- **Principle 10 and 11.** Websites are only allowed to use the collected data for its original purpose and if used for other means, must require notifying the user. However, for example, as if the original purpose was of means such

as marketing, the data could be used for different aspects of marketing. [9]

- **Principle 13.** Websites are only allowed to assign unique user identification when deemed necessary. However, what would be considered necessary? This could be potentially argued. [9]

C. The 5 Tikanga

When reviewing cookies and associated cookie policies of top New Zealand websites, it's essential to consider not only the New Zealand Privacy Act 2020 [1], but also the foundational Māori principles such as the framework offered within the Ngā Tikanga Paihere [10]. Incorporating the Ngā Tikanga Paihere - The 5 Tikanga principles to our research not only provides a culturally informed perspective, but also emphasises the importance of aligning modern digital practices with Māori values. These principles provide an essential cultural context that can guide and inform our understanding of data protection from a Māori perspective.

The 5 Tikanga Principles and their relation to online data policies can be categorised/considered as so:

Principle 1

- *Pūkenga - Skills.* Website owners should invest in a team that is skillful and have the required expertise to implement robust and informed data protection measures.
- *Whakapapa - Genealogy.* Every aspect of a website's inner workings from its backend processes to the front-facing user interface, has a connection to its users.

Principle 2

- *Pono - True to the principles of culture.* This principle relates to ensuring website owners should be accountable and to provide sufficient reasons for data collection and to ensure users fully understand the implications. Websites should be innovative in how they approach data privacy, ensuring they meet both legislative and cultural standards.
- *Tika - Value for all.* This principle can coincide with ensuring that cookie policies should remain clear, and transparent, enabling users to have full insight into how their data is being used.

Principle 3

- *Wānanga - Organisations.* This Tikanga is to guide organisations to ensure that they have established systems in place to appropriately handle data and offer guidance when necessary.
- *Kaitiaki - Guardians.* A given website is the guardian of its users' data. To ensure this principle requirements are sufficiently met, the website owner/s must ensure the user data is protected and used responsibly.

Principle 4

- *Wairua - Spirit or soul of a person.* This principle relates to the spririt/soul of an individual and the power it holds. Website owners must ensure respect is maintained, and to properly ensure no harm will arise from the way the data has been used.
- *Mauri - life principle or force.* This Tikanga relates to everyone's right to exist in a space equally, working together to create an ecosystem of which everyone takes part in contributing to. This relates to website owners making sure their part in addressing privacy issues with their users is vital for a natural and mutually beneficial and respectful environment.

Principle 5

- *Tapu - Sacred, prohibited, or restricted.* This Tikanga concept relates to the responsibility of keeping private data private. This principle ensures that web owners sufficiently ensure that their user data is properly protected.
- *Noa - Ordinary, unrestricted, or normality.* This principle is defined to meaning that something free and open, should remain so. That data that is inherently beneficial, should be openly shared with users/communities.

In summary, integrating Māori principles into data protection not only aligns with New Zealand's cultural values, but also provides a robust framework that prioritises respect, transparency, and sovereignty over one's individual data. As we analyse and review New Zealand website cookies, it's important to critique them through the lens of the Ngā Tikanga Paihere as well emphasising the importance of weaving New Zealand cultural values into digital practices.

D. GDPR Web Cookie Compliance

[11] There are many websites and tools available that offer domain scanning for corresponding cookie output and compare their cookie usage against various compliances such as the GDPR. Although the GDPR is not directly relevant in a New Zealand context, due to the policies only relating to European Law and Order. New Zealand domain holders only need to lawfully comply with the GDPR in the case that the organisation operates any offices within any of the European Union associated countries, that the organisation's approach targets European users (website is in any given European language), sells goods or services to European individuals, including physical items or digital processing, and lastly, if the organisation handles the personal information of anyone living in the EU.

The reason the GDPR is so important is because it provides a basis for how websites should handle its users personal data, even if it's lawfully obtained and shared by the locations compliance standards. The GDPR takes a lead and has influenced how data is processed even outside of the EU, reaching across the world. When it comes to these standards and approaches in New Zealand, we have fallen behind and are yet to have specific laws for personal data protection and how user data is handled and processed within an organisation.

One tool we will delve into, relates to scanning a given domain, and reviews every cookie the organisation employs, and checks if it is compliant as per the GDPR [12]. For our research, I scanned a list of New Zealand's top used websites to compare the output against the GDPR policies and decipher if any NZ website is compliant.

GDPR Top 10 NZ Domain Results

An analysis was undertaken to further evaluate New Zealand's cookie compliance standards in context of the GDPR. The

list consists of New Zealand's top 10 websites as per usage statistics this year [13]. These domains encompass a range of online entities, from news outlets and ecommerce platforms to financial organisations and entertainment hubs.

The scan results generally indicate a trend that many of these high-traffic websites may not be fully GDPR compliant. General observations:

- **Pre-loading of cookies.** A common trend observed across the majority of these domains is the pre-loading of cookies before acquiring explicit user consent. GDPR guidelines specify that websites should not initiate non-essential cookies without obtaining clear consent from the user first. From a New Zealand legislative perspective, it is not necessary to do so, but is in good practice to do so.
- **Varied Cookie Purpose.** These domains tend to utilise a mix of essential and non-essential cookies. While essential cookies are vital for core website functionality, the non-essential ones range from extensive marketing cookies to statistics and analytics. Many of these marketing based cookies were from the online marketing giants such as Google, Facebook and Amazon. Some of these marketing specific cookies target user advertising profiles for serving better targeted advertisements.
- **International Data Transfer.** An area of potential concern is the significant amount of user data that gets transferred to international third-party servers, primarily located in the United States. GDPR emphasises data protection, and the transfer of user data outside of the EU without adequate protections is considered a significant breach of compliance.

In conclusion, while these observations indicate non-compliance as per GDPRs standards, this does not necessarily reflect the laws and legislations these domains must abide by in a New Zealand context. It's essential to recognise the dynamic nature online platforms take with utilised cookies, as it opens a door to better understanding the purposes and goals an online organisation has with their employed cookies. As GDPR has heavily influenced many online businesses, even outside of the EU, it continues to act as a marker for good user privacy to online business relationships.

IV. Design & Implementation

The implementation and initial planning phase was pivotal in transforming the initial concept into a practical and feasible solution capable of scraping and analysing the web cookies of nearly 250,000 NZ TLD websites. In this section, I will provide an in-depth description of the specific components utilised and the rationale behind the selection of each.

A. Webdriver

The first and most important component of the project, was selecting an appropriate webdriver to act as the browser for data collection. WebDrivers are a form of simulated browsers, where you can program it to imitate real user behaviour on a browser in order to extract data for valuable insights. We will utilise a WebDriver to simulate this experience in order to perform web cookie scraping for analysis. WebDriver browser simulation provide various benefits to this project, some aspects include:

- **Cookie Handling Support:** Due to mimicking the behaviour of a legitimate browser, WebDrivers contain the same functionality of cookie handling as regular browsers too. We will be able to retrieve the same available cookie data that would be on a regular browsing session.
- **Scalability:** Are programmed to handle large volumes of data, over multiple instances (multi-processing support). WebDrivers are also able to perform parallel processing, handling multiple browser instances simultaneously.
- **Cross-Platform Compatibility:** Like their browser-engine counterpart, are able to run on most, if not all, operating systems such as Linux, MacOS and Windows. This is important, due to utilising a mix of devices (University labs - Linux, Personal - Windows).

I narrowed it down to two potential candidates: *Pyppeteer* and *Selenium*. Pyppeteer is a Chromium-based webdriver that proved to be incredibly efficient in-terms of the domain process time, however, we faced some issues with the configuration not properly scraping full cookie data. Selenium is a Firefox-based webdriver that however took some more time to process, gathering vast amounts of more data than the Pyppeteer test run did. This could have been boiled down to the script but within the project's time constraints - could not build upon this further. Ultimately, due to its robustness in handling large-scale scraping tasks, and the level of collected cookie output, we settled on Selenium [14] for the webdriver of choice. This webdriver enabled real-time interaction with web content, mimicking the behaviour of an actual user navigating the site, allowing for full web cookie loading and collection.

Testing

From the initial steps of this project, I initiated simple scripts for both Selenium and Pyppeteer, focusing primarily on their fundamental scraping capabilities. As timing in this project is a significant component, I needed to ensure that not only am I prioritising the data scraping capabilities, I also needed to take the total processing time into account. To test on a smaller-scale, I tested each script against a list of 10 domains, and included a total processing time to compare the results of the two Webdrivers.

Pyppeteer

Figures 1 and 2 relate to the initial Pyppeteer test and the corresponding output data. As we can observe from the terminal window, the total processing of 10 domains took approximately 24 seconds. The output also unveils gaps in scraped cookie data, where some domains on the list were not processed. The output file itself only contained 72 lines of output cookie data, which is insignificant.

Selenium

Figures 3 and 4 relate to the initial Selenium test and the corresponding output data. For Selenium, the total processing of 10 domains spanned about 31 seconds, yielding substantial output with 410 lines of scraped cookie data.

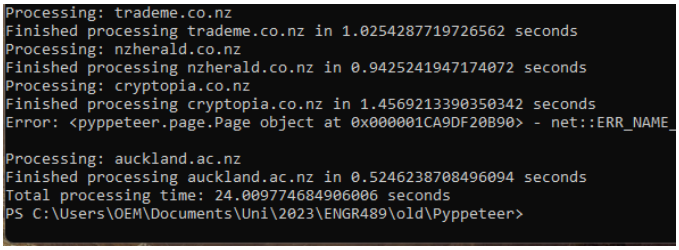

Fig. 1. Pyppeteer processing time against 10 domains

Fig. 2. Pyppeteer output cookie data, proving insignificant

Although the extra 10 seconds seems insignificant, when extrapolated to 220,000 domains, the processing time extends to approximately 7.9 days.

In summary, Selenium's robustness in handling large-scale scraping tasks and the extensive amount of output data we are able to retrieve, make it our Webdriver of choice for this project. This Webdrive enables genuine real-time interaction with web content, emulating a real user's site navigation, allowing for full domain loading and cookie acquisition.

B. Programming Language

The backbone of this project, and a large portion of this project's timeline heavily attributed to meticulously crafting and refining of the web scraper script. Thus, the decision on the programming language would be vital in ensuring the project's success. Python, with its rich features and versatility, emerged as an ideal candidate [15]. Specified below, are features offered by Python, that was a deciding factor in our decision process.

- **Extensibility.** Python's vast ecosystem, equipped with libraries tailored for a diverse range of tasks, provides a robust platform for web scraping tasks. For web-scraping tasks, libraries such as 'socket' and 'sqlite' for storage solutions, have provided a foundation for easing into our data extraction process. The rich assortment of various libraries ensures that no matter the complexity of the task, there was likely a Python library equipped to handle it.
- **Compatibility.** Python seamlessly integrates with other necessary components. Whether it was interfacing with SQL databases for storage purposes, or integrating with our chosen Webdriver, Selenium, Python ensures seamless data flow between components.

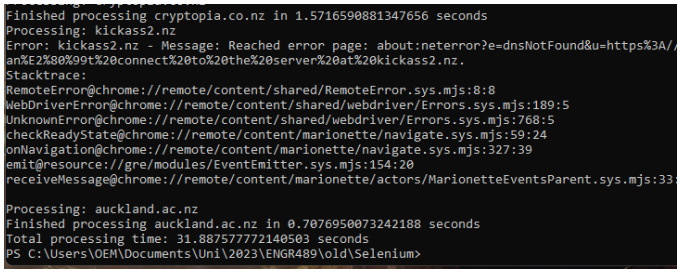

Fig. 3. Selenium processing time against 10 domains

Fig. 4. Selenium output cookie data, proving significant in comparison to Pyppeteer

- **Simplicity.** One of Python's most desired features lies with its general simplicity. This makes development more straight-forward and aids in quick prototyping, giving room for necessary tweaking. Python's simplicity and readability ensures that adaptations could be made promptly, keeping the project on schedule.
- **Web Capabilities.** Beyond the basics, Python can be configured to, and excels in web-interactions. The available functions allowed for simple configuration such as interacting with web elements, which in turn made the task of cookie scraping more straight-forward. Coupled with Selenium integration, Python offered a fool-proof solution to reaching our project's goals.
- **Versatility** After completing our scraping, the data required to be sufficiently processed and analysed. Python's ability to convert and re-organise data is an essential feature for this project. With Python, we would be able to ensure that data from post-scraping is efficiently parsed, stored, and structured, readying it for the subsequent analysis phase of the project.

In summary, Python was the final candidate for this project's development [15]. Python, with the necessary additions, was equipped with the tools ready to navigate the multifaceted processes to perform the various required tasks such as web scraping, data processing, and data storage.

C. Scripts

Throughout the development phase, the script's core structure has been refined and enhanced to address the evolving

needs and challenges of the project. In this section, is a provided deeper exploration into the script evolution.

Prototype - First Trimester

As mentioned in section Webdriver IV-A, the primary goal was to gauge the effectiveness of web scraping tasks using Selenium. The first prototype was straightforward in that the only functions present were as follows:

- Read domain names from a file
- Deploy Selenium's Webdriver and loop through, navigating to each domain on the list
- Extract the cookies from each loaded domain
- Store the data to a text file.

This provided a baseline understanding of Selenium's general capabilities and allowed for quick proof of concept.

selenium_scraper.py

C: > Users > OEM > Documents > Uni > 2023 > ENGR489 > old > Selenium > selenium_scraper.py > ...

```python
from selenium import webdriver
import time
# Read domain names from a text file
with open('first_10_domains.txt', 'r') as file:
    domain_list = file.read().splitlines()
# Set up Selenium WebDriver with Firefox
driver = webdriver.Firefox()
start_time = time.time()
for domain in domain_list:
    print(f"Processing: {domain}")
    domain_start_time = time.time()
    try:
        # Load the website
        driver.get(f"https://{domain}")
        # Retrieve the cookies
        cookies = driver.get_cookies()
        # Save the cookies to a text file
        with open('cookie_data.txt', 'a') as file:
            file.write(f"{domain}:\n")
            for cookie in cookies:
                file.write(f"\t{cookie['name']}: {cookie['value']}\n")
        domain_processed_time = time.time() - domain_start_time
        print(f"Finished processing {domain} in {domain_processed_time} seconds")
    # Add Error check
    except Exception as e:
        print(f"Error: {domain} - {str(e)}")
        with open('cookie_data.txt', 'a') as file:
            file.write(f"\tError: {str(e)}\n")
# this doesn't work, spits back last domain time, need fixing
total_time = time.time() - start_time
print(f"Total processing time: {total_time} seconds")
# Close the browser
driver.quit()
```

Fig. 5. Prototype selenium cookie scraper script, base scraping capabilities

Development - Mid Trimester

As the project was evolving, the need for more structured and efficient storage mechanisms became apparent. We implemented a storage solution by integrating SQL database functionality, for parsing and structuring the scraped data. The initial script for creating the database instance included two separate tables, one for instantiating the domain data with general error checking, and the other for specifically structuring the full cookie data for organisational purposes.

```
cursor.execute('''
    CREATE TABLE IF NOT EXISTS Domains (
        id INTEGER PRIMARY KEY,
        domain_name TEXT UNIQUE,
        is_valid INTEGER,
        is_timeout INTEGER,
        error_message TEXT,
        created_at TIMESTAMP
        DEFAULT CURRENT_TIMESTAMP
    )''')
```

At this point of development, both scripts were equipped with the following capabilities:

Database Script:

- Initialise SQL database, equipped with two tables for structuring purposes
- During the initialization function, the Domains table gets populated with the required entries (e.g IS_VALID = 1/0) which coincides with added filtering in the main script, as a form of error checking to additionally avoid unnecessary time delays.

Scraper Script:

- The main script now pulls the domains list from the IS_VALID = 1 entries in the SQL database Domains table, and now populates the Cookies table with the scraped output data.

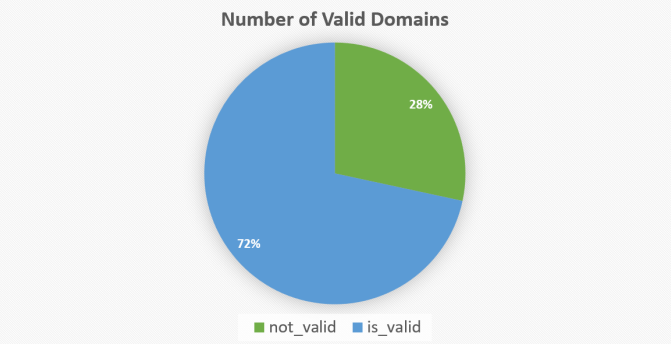

Fig. 6. Percentage of Valid Domains from the Domains List

Additional Scripts:

Created a separate script not included in final artefact, that is a version of the SQL database initialisation, but with an included malicious domain check, that is implemented from the Google Safe Browsing API. This was an addition to the database initialization script, where it also includes a check (similar to the check validity function) that the domain is either present (1) or is not present (0) in the API list. However there were some complications where the maximum amount of API requests allowed was set to 10,000. From my perspective, this was unfeasible in the given timeframe of this project, as that would require 22 days to process the list of 220,000 domains. However, a fellow student, and sister project by Tomas Borsje, debugged this problem and was able to process the list of domains in batches, of 100 to sufficiently fit within the maximum 10,000 API requests.

Final Script - Second Trimester

The final script implementation was multi-faceted, one aspect we included was the addition of a list of malicious domains and attached it onto the existing SQL database, and it was used as an additional filter, to ensure the security of the device performing the full scrape.

Another addition was implementing multiprocessing support, via a docker image categorised into 10 subsections for batch processing. This ensured that the web-scraping process itself required too much in our project's timeline, allowing for some leeway, in the situation we needed to further configure

the script itself. To summarise the final script included the following capabilities:

- Initialise the two databases, and populate the Domains table with is_valid and is_malicious
- Scraper script, reads in the domains table from the database file, and runs through, scraping each domain that is valid, and not malicious. The output cookie data is then populated within the Cookies table in the SQL database file
- Additional scripts include sample visualisation of the output cookie data, one for visualising the top errors into a pie graph, and another pie graph for the most frequent cookies
- The final scripts have been combined into one file, with a Python Docstring initialisation, allowing for command line execution via the specified function. E.g [python ./final_script.py -h] help function, displays the available function calls.

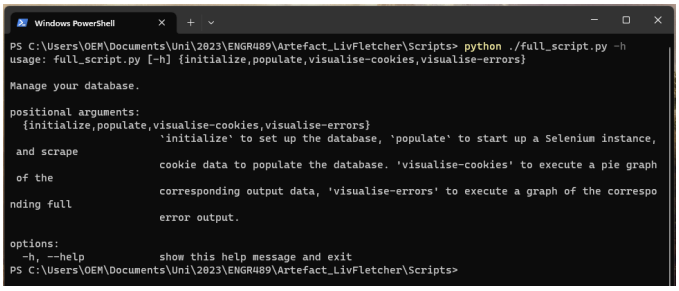

Fig. 7. Python Docstring

D. Database

As discussed in the previous section IV-C our projects required the proper storage facilities in order to provide a clean basis for later analysis. The inherent structure and organisation that SQL databases offer, ensures that the data remained is both accessible and analyzable. A rudimentary Python script was formulated to instantiate a base SQL database file, using sqlite [16] ensuring a smooth data transfer process from the data scraping phase, to the data storage phase. The reason for implementing SQL as our storage means, can be categories as so:

- **Structured and Organised.** SQL databases are inherently organised into tables, allowing for parsed in data to be segregated by type and relation. This ensures that our web cookie data was not just stored, but stored in a manner that made it straightforward to retrieve and analyse.
- **Data Integrity.** SQL databases rarely fail due to software defects, ensuring that data integrity is maintained.
- **Scalability.** Due to the size of the output data we intend to handle, SQL databases offer scalability, ensuring that as the data output grows, the database can handle it without digs to its overall performance.
- **Usability.** Combined with DB Browser SQL viewer [17], the interface was straight-forward and easy to use.

Once scraping and populating was complete, later on in the project, I opted to parse the cookie output data into Excel spreadsheets, for the built-in graphing functionality for visualising the output data.

E. Infrastructure

Given the magnitude of the scraping task at hand, a solid infrastructure was paramount, and the specifications offered by my personal device were not enough. There were no guarantees that my personal device won't falter, leading to inaccurate data or prolonged scraping durations. Thus, we utilised a dedicated server provided by the university department. This ensured that the script ran on a clean OS, free from potential background disruptions or conflicts. Additionally, to expedite the scraping process, the server was connected to the university's high-speed fibre Ethernet. This configuration allowed the script to run uninterrupted for approximately five days, ensuring the timely completion of our data collection phase. The general reasons for this decision are marked as below:

- **Performance and Reliability.** The Universities available servers offered a high-performance environment, crucial for a task as resource-intensive as web loading and subsequent data scraping. With a clean OS, there was minimal risk of background applications interfering with the scraping process.
- **High Speed Connectivity.** Web scraping relies heavily on network connection speed. With access to the University's high-speed fibre Ethernet, the script could navigate and load domains at a more efficient pace, than would it would on my personal device.
- **Uninterrupted Execution.** A dedicated, untouched environment meant the script could run without any interruptions. This was pivotal given that the scraping process was expected to run for several days.
- **Safety and Security.** Running the script on the university's infrastructure ensured a level of security. In the event of unexpected challenges or data breaches, the universities IT protocols and security could intervene, ensuring data integrity and project safety.

With the above considerations, our data scraping process was not just a possibility but a guarantee. The script ran (mostly) uninterrupted for approximately 3-4 days, leading to the successful and timely completion of our data collection phase.

V. Analysis

After completing our data collection phase and reorganising our output cookie data, we embarked on a comprehensive analysis journey. In this section, we will delve into the analysis phase of the project, and showcase the culmination of interesting trends we have discovered from the output data.

A. Visualisation

Dominance of Major Web Entities: A Look at Most Employed Cookies

Our cookie output data unveiled the dominance of certain companies, reflecting their widespread influence and presence across the web. We have illustrated our findings through a tree-map styled graph, providing the visual representation of cookie prevalence by their corresponding company.

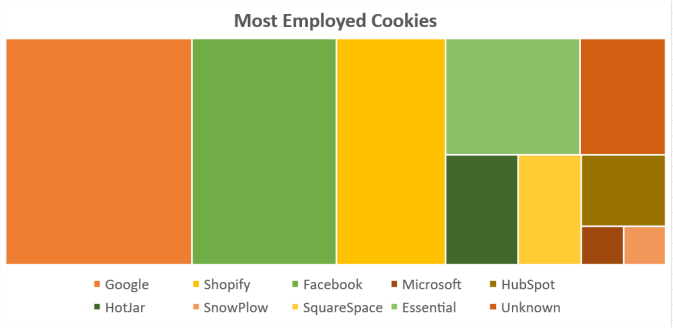

Fig. 8. Most Employed Cookies

The main findings are as follows:

- **Google's Dominance.** Topping the list was Google, with an impressive amount of over 60,000 cookies - approx 30% of total cookie output. This statistic isn't only about specific numbers, but as a testament to Google's overall omnipresence online. This verifies Google dominance on the online market, whether it be through search, analytics, advertising, or its myriad of services, Google's footprint is present everywhere on the web.
- **Facebook's Reach.** Facebook interestingly continues to exert itself amongst our collected data, despite many of the coinciding websites with such cookies, are not in relation to the social media site itself, but for marketing purposes. With a presence of over 45,000 cookies, Facebook exerts itself amongst the group of most employed cookies in New Zealand.
- **Rise of E-Commerce.** Shopify with just over 35,000 cookies, showcases the power and reach that ECommerce has online. These numbers signify the boom in online shopping and the website's success in simplifying e-commerce for businesses.
- **The Enigma of Unknown.** The unknown category has some interesting entries, that with research in comparing known cookie data bases, was still unknown. One such cookie is specifically named 'crumb' with no additional identifiers. The presence of this cookie appeared random amongst the employed domains, and no comparisons could be made. The crumb cookie itself, made up just under 10,000 of the 14,000 unknown cookies, proving to be quite the enigma. This was categorised by manually uploading the cookie details to various online database tools, namely *Cookiedatabase.org [18]* and Github repository *Open Cookie Datbase [19]* to finalise this categorisation.

Cookie Categories: Understanding their Roles in Digital Interactions

Beyond identifying the companies' cookies that dominate the digital space online, it's crucial to understand the purpose these cookies serve and why. By categorising them based on their primary functions, provided below is a bar graph of a visual representation of the cookie categorisation by their overarching function.

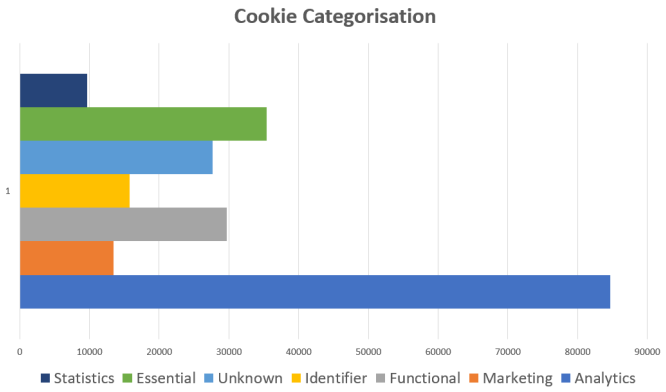

Fig. 9. Cookie Categorisation

- **Analytics.** Dominating the chart with just under 85,000 cookies, analytic cookies provide website owners insights into user behaviour when navigating their page. Cookies classified as analytics can provide web owners user data such as; time spent on page, and time spent on a particular section of a page, scroll/click and mouse-hover data. Analytic cookies can also provide a broad-overview of their whole user-base such as number of users, session statistics, approximate location, and browser/device information. This kind of data can help a web owner better understand its user base, and potentially offer insights into restructuring the website for better user-interaction.
- **Marketing's.** With just over 13,000 cookies, the marketing cookies are objectively speaking, the most important aspect of this project's goals. These cookies play a vital role in delivering targeted advertisements to users, ensuring that marketing campaigns resonate with the intended audience. Their presence indicates New Zealand's approach to employing personalised online advertisements.
- **Functional's.** These cookies' primary use is ensuring websites operate as intended, providing baseline website functionality, such as previous session details (saved logins, user preferences etc) and other settings, overall enhancing the user experience.
- **Identifier's.** Serving as digital identification, these cookies are for recognising returning users, ensuring a cohesive experience across sessions and site visits.
- **Unknown's.** A considerable number of cookies, just over 27,000 were unable to be categorised. These cookies could serve a myriad of functions. The large amount of Unknown highlights the nature of many cookies not having proper identifying information available online. This could be due to the website hosting their own cookies, and not providing the relevant information online.
- **Essential's.** As the name suggests, these cookies are essential for the basic functionality of the website. Their presence ensures websites function as intended, by providing session management to security protocols and other various functions.

- **Statistics's.** Associated closely with analytics cookies, statistical cookies provide a more generic understanding of website performance, gathering data on page views and other various engagement metrics. The statistical cookies gather data on users as a whole, less-so than the individual behaviour like analytic cookies.

The bar plot in Figure 8 provides a general distinct categorisation of the cookies, offers a glimpse into New Zealand's approach to employing cookies on their websites.

B. Manual Cookie Analysis

StorbieAnon

Storbie anon appears to be used on various NZ websites that host products. StorbieAnon cookie is created by the New Zealand-based ecommerce platform Storbie. Cookie tracks user data such as;

- User preferences
- Shopping cart contents
- Anonymous user data.

StorbieAnon cookie is only hosted on websites via HTTP-ONLY, which could raise some security concerns for the domains hosting such cookies. Not only is it third-party associated which directly suggests that user data is sent through the server to the Storbie domain - which could suggest the use of collecting/creating user data profiles across Storbie-associated domains.

Shopify [20]

Shopify cookies are directly associated with the Shopify ecommerce platform. Shopify hosts many different kinds of cookies such as:

- Session management
- Authentication
- Store preferences, language, currency, location etc
- Analytics, user data collection.

Directly, the Shopify_S cookie refers to the Shopify analytical cookie which is a part of Shopify's user tracking system which gathers anonymous data about user behaviour for analytical and reporting purposes.

The above webpage notes that Shopify shares/sells user data for advertising purposes to make "shown advertisements more relevant to you". Around 300 out of the 10k list contain the shopify_sa_t token, which is this token that is used for advertising purposes, and is allowed to share/sell your data to third party advertisers.

VI. Evaluation

As with any Engineering project, the post-implementation phase offers an opportunity to reflect, evaluate, and derive lessons for future undertakings. As expected, we faced many unique challenges. The project was a complex amalgamation of various tools, technologies, and strategies, each chosen after careful consideration. Several factors influenced our general decisions thus far.

- *Feasibility:* The availability of resources at our disposal, such as the university server, played a role in our implementation choices.
- *Cost-Efficiency:* Leveraging open-source and widely available tools and university resources ensured that the project remained cost-effective.
- *Ethical Consumption:* By utilizing multiprocessing in our final script, we have sufficiently addressed our need to ensure we have aligned with the U'N's Sustainable Development Goal 12, by reducing our overall energy consumption for this project.
- *Sustainability:* Our approach, particularly the timeout for each domain, ensured that the project did not overly exhaust resources and time, keeping our practices efficient and sustainable.

A. Selenium

Utilising a simulated browser was vital to this project's development, as there are not many well-known means available that offer the same or similar capabilities. Web scraping presents a unique set of challenges, particularly when it comes to extracting data from dynamic web pages. The implementation of a Webdriver such as Selenium was essential, due to some challenges we faced early on.

Python Shortcomings

Early on in our research efforts, we began with testing out Python's Request library for script implementation. Although it's a powerful tool for making HTTP requests, its capabilities are limited when it comes to simulating user intersections on dynamic websites, due to only getting HTTP page contents, not being able to load modern Javascript websites. The full-loading of Javascript is vital for generating a website's cookie data, and we quickly found out this was not possible with just Python's Requests library implementation. This is where simulated browsers such as Selenium come into play. Selenium, combined with Python, allowed for the full page loading, and cookie extraction, in one step. If revisiting this project in the future, we may explore other web scraping libraries or tools that offer a blend of the simplicity of Requests and the dynamic capabilities offered by Selenium.

B. Python

The programming language of choice forms the backbone of any software project. For our endeavour, we considered the following points:

- **Versatility.** Python's extensive capabilities and available libraries, with coinciding tools, made it an ideal choice for this project's goals. Combined with Selenium, it provided us the abilities to scrape, store, organise and analyse the expected large-scale data output.
- **Integration.** Python's ability to integrate seamlessly with databases, such as SQL, ensured a smooth transition of data from the initial scraping stage, to the parsing data stage.
- **Readability.** Python's syntax is well-known for its general readability and clarity.

While Python proved to be an excellent choice for this project, exploring and comparing with other languages could provide insights into potential optimizations or alternative solutions.

C. Feasibility

Cookie Consent Notices

For this scope of this project, an important component I must consider is the concept and use of 'Cookie Consent Notices' and what that means for our data collection measures and reporting. Checking available website cookies before and after accepting the cookie notice is an important component for assessing a given website's privacy practices while also ensuring compliance with data protection regulations such as The New Zealand Privacy Act 2020 and the GDPR. Here's why this is significant in how this can affect our overall results if we do decide to include the comparable data.

- **Consent Compliance.** In a New Zealand context, web hosting falls under The New Zealand Privacy Act 2020 and for general practice, the GDPR. Websites in New Zealand, are required by law to obtain informed consent from the users that visit their domain on the initial data collection, and the type of data being collected. By including a scan for the presence of Cookie Consent Notices, we can check the active cookies on a given website before and after consenting, to assess whether they are up to standard.
- **Data Protection.** Cookies have the ability to collect various types of data, and build user profiles by assessing user behaviour on a given webpage, or other webpages for profitability to better target advertisements at users. We can review the types of cookies being employed.
- **Cookie Policies.** Domains have an expectation to provide and maintain accurate and up-to-date cookie policies that provide details on the types of cookies used and their corresponding purposes. By checking the cookies before and after consenting, we can check if a given domain's policy accurately reflects the notice provided.

However, in terms of utilising this type of scan, by gathering data before and after clicking consent on the notice, we face a significant problem. The job of just collecting the data from 220,000 domains is such a monumental task that adding the consent notice comparator factor, doubles this task. I have created a simple script, whose job is to only use the Python Requests library, which logic of only collecting one type of data within the pages HTML, whether or not a consent notice is present, and to organise this data as 1 - true, or 0 - false per domain into an SQL database file.

This problem can be broken down into four main phases:

1) Creating/sourcing a script for scanning specifically for checking the presence of cookie consent notices per domain.
2) Running the script against our domain list of 220,000 domains to 'sort' the domains that contain the notice, vs not containing the notice. This process online has been estimated to take (on my own machine) two full days despite only using the HTML request functionality.
3) Once we have sorted the list of 0/1's of domains, we need to add this data to the original SQL database that contains all the other data entries we require (is_third_party etc etc). Once the data has been integrated, we start the full domain scan of the original data sets, this scan will ignore the 0/1's data points of the consent notice entry and scan every non-malicious domain.
4) After we have finished our scan on all domains, we can start the scan, only on websites that contain a consent notice. This is where the difficulty skyrockets, as we will need to create another script to handle this functionality, where we create a comparator table of before and after consenting.

The next step involved into this process is to initialise the first scans to develop further into the next stage of organising and developing into researching the validated cookie data and to get back onto the track.

Local Storage Data

As discussed in my preliminary report, initially at the beginning of the project I considered the aspect of additionally scraping for local storage data, on top of the base web cookie data, to expand upon our potential data output. However, within our set project time-frame, this was unattainable and would require further, more complex research due to the limited documentation on specifically scraping data from local storage. The findings as stated in the preliminary report, is that even well-known scraping tools with many capabilities, did not contain the functionality of such data collection measures. The discussion surrounding local storage data is insignificant, and we decided it was not critically necessary for this project scope.

Changes

The current script now implements the use of a time-out functionality where if a website takes longer than 10 seconds to scrape all cookie data, to skip the domain. This is due some websites taking up to a full minute on their own as they contain complex and many varied cookies. This time-out functionality also works as an error check, when a domain is unable to be loaded for reasons that are not known (currently down, no longer hosted etc) to skip the scraping efforts. We have also included the ability to skip websites that are within the Google Safe Browsing API list, to ensure we do not accidentally scrape data from malicious webpages and potentially infect the system.

VII. Future Work

This research into the realm of web cookies and their prevalence across New Zealand's digital landscape aims to influence further in-depth explorations and discourse in the domain of online privacy.

A. Advocacy for Robust Cookie Policies

While the GDPR serves as a benchmark in establishing rigorous standards for user data protection and privacy in a global perspective, the dialogues surrounding cookie consent remain somewhat dispersed and unheard of in New Zealand. The aim of this research is to instigate a conversation surrounding user data privacy, and the importance of data sovereignty in New Zealand, and to encourage proper and sufficient policies in this field. A closer look into New Zealand's' Privacy Act 2020 and potentially augmenting it. Better mirroring the user-centric

principles of the GDPR, could pave the way for a safer digital environment for New Zealand citizens, and better abide by cultural ramifications as per the Ngā Tikanga Paihere.

B. Extension to Cookie Consent Notices

Given the expansiveness and depth of cookie-related data and their applications, a natural progression of this project would be to explore into 'Cookie Consent Notices'. Future research could align towards analysing and understanding the variance between collected cookie data prior to, and post user consent. This could unveil insightful patterns regarding actual user awareness in general cookie usage, and whether or not websites are transparently honest and ethically acting as per their own consent notice details. This conversation has arisen in a global perspective, as outlined in the paper 'Can I Opt Out Yet?' as referenced in sectionII-A. Possibly due to New Zealand's legislation's not requiring a website owner to include cookie consent notices, it isn't entirely expected, but from my endeavour upon this project, have noticed many New Zealand websites including notices despite this, and the corresponding scraped output data could still provide interesting insights into the general honesty of websites in New Zealand.

C. Investigating Local Storage Data

Another interesting avenue for potential future works is the exploration into Local Storage Data. While this topic hasn't had mainstream influence in discussions relating to online user privacy, it warrants further analysis. Local Storage allows for websites to store data persistently within a user's browser, which can include a myriad of information such as user preferences, session information, and other related to the user session. This data is utilised by the website to provide a better user-experience, however, if used to store sensitive user data, could face negative implications. As discussed in a blog post by Nurhaliza Binte Sapari [21], there is potential for attackers to infiltrate the data stored in a users Local Storage. A deeper dive into how websites utilize Local Storage, the types of data stored, and the implications on user privacy would propel our projects findings and provide a deeper view of data management practices employed by websites, in either a global standards, or in a New Zealand context.

Concluding the frameworks, data, and findings from this project lay down a foundation that can be utilised and expanded upon with further research and discussions. This projects goal was to serve as a reference point for opening a discourse that influences better safeguards and digital rights for New Zealand citizens.

VIII. Conclusion

Throughout this projects journey we have explored various aspects of the world of cookies, particularly within a New Zealand context and related the output data to relevant laws and regulations. We have uncovered an array of interesting insights into New Zealand's regulatory practices for user privacy and cultural ramifications.

Addressing challenges along the way served as a learning experience in developing a robust and efficient cookie scraping script. Through our methodology that held up against the practical and ethical considerations, but also opened doors into the possibility of applying this research into a larger, more diverse digital field.

The exploration into New Zealand's legislation's and privacy frameworks unveiled data about the current digital practice in New Zealand, but also to serve as a pivot in how future-policy making could evolve, to provide better guidance and privacy practices in New Zealand.

References

[1] (Version as at 1 July 2023) The New Zealand Privacy Act 2020. [Online]. Available: https://www.legislation.govt.nz/act/public/2020/0031/latest/whole.html#LMS23322

[2] Sanchez-Rola, Iskander and Dell'Amico, Matteo and Kotzias, Platon and Balzarotti, Davide and Bilge, Leyla and Vervier, Pierre-Antoine and Santos, Igor, "Can I opt out yet? GDPR and the global illusion of cookie control," in *Proceedings of the 2019 ACM Asia conference on computer and communications security*, 2019, pp. 340–351. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3321705.3329806

[3] Chung, Winnie, "A Snoop at Privacy iIssues on the iInternet in New Zealand," *University of Auckland, Business Review*, vol. 4, no. 3, pp. 2–15, 2002. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b46d1ca20e7a8513f97be38bcfd35333ae006dab

[4] Lim E, "Electronic Commerce and the Law," *Unpublished BCom(Hons) dissertation, Department of Management Science and Information Systems, The University of Auckland*, 2000. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b46d1ca20e7a8513f97be38bcfd35333ae006dab

[5] (Version as at 1 July 2023) The New Zealand Privacy Act 1993. [Online]. Available: https://www.legislation.govt.nz/act/public/1993/0028/latest/DLM296639.html

[6] Cookiebot. (2021) The New Zealand Privacy Act 2020: In Compliance with Cookiebot. [Online]. Available: https://www.cookiebot.com/en/new-zealand/

[7] Hannah H, Megan L, Ellie Y, Lorrie C, ""Okay, whatever": An Evaluation of Cookie Consent Interfaces," pp. 1–27, 2022. [Online]. Available: https://doi.org/10.1145/3491102.350198

[8] Christine U, Martin D, Sascha F, Florian S, Thorston H, "(Un)informed Consent: Studying GDPR Consent Notices in the Field," pp. 973–990, 2019. [Online]. Available: https://doi.org/10.1145/3491102.3501985

[9] New Zealand Privacy Comissioner. (2020) Privacy Act 2020 and the Privacy Principles. [Online]. Available: https://www.privacy.org.nz/privacy-act-2020/privacy-principles/

[10] (2020) Ngā Tikanga Paihere. [Online]. Available: https://data.govt.nz/assets/data-ethics/Nga-Tikanga/Nga-Tikanga-Paihere-Guidelines-December-2020.pdf

[11] (Version as at 1 July 2023) General Data Protection Regulation. [Online]. Available: https://gdpr-info.eu/

[12] (2016) General Data Protection Regulation Scanner. [Online]. Available: https://2gdpr.com

[13] (2023) Top new zealand-based and used websites 2023. [Online]. Available: https://www.similarweb.com/top-websites/new-zealand/

[14] Jason Huggins. (2004) Selenium. [Online]. Available: https://www.selenium.dev/

[15] Guido Van Rossum. (1991) Python v3.8.0. [Online]. Available: https://www.python.org/

[16] Dwayne Richard Hipp. (2000) Sqlite. [Online]. Available: https://www.sqlite.org/index.html

[17] Pete Morgan. (2012) Db browser for sqlite. [Online]. Available: https://sqlitebrowser.org/

[18] Cookiedatabase. (2012) Understanding tracking with cookiedatabase.org. [Online]. Available: https://cookiedatabase.org/#about

[19] Github User Jkwakman. (2023) Open cookie database. [Online]. Available: https://github.com/jkwakman/Open-Cookie-Database/blob/master/open-cookie-database.csv

[20] Shopify. (2023) Shopify Cookie Policy. [Online]. Available: https://www.shopify.com/nz/legal/cookies

[21] Nurhaliza Binte Sapari. (2022) Your browser's local storage can be misused. here's what you should know. [Online]. Available: https://www.horangi.com/blog/misuse-of-local-storage