Health data leaks to third parties in web-based health services

Sampsa Rauti[1],*, Robin Carlsson[1], Samuli Laato[2], Timi Heino[1], Panu Puhtila[1] and Ville Leppänen[1]

[1] University of Turku, Vesilinnantie 5, 20500 Turku, Finland
[2] Tampere University, Kalevantie 4, 33100 Tampere, Finland

Abstract

Today, users may share sensitive health data on web-based health services. We rely on these services to keep our data safe and secured, but this is not always the case. Therefore, this study investigates the privacy of a snapshot of 10 Finnish web-based health services, providing an analysis of health data leaks. We show that all analyzed services leaked at least some kind of personal data to third parties – from topics of visited pages to details on appointment bookings. While the situation has improved after we have notified the health service providers about this issue, the study serves as a reminder of the ongoing challenges in protecting user privacy in online health services and highlights the pressing need to address these issues.

Keywords

Medical websites, data leaks, data concerning health, web privacy, third-party services

1. Introduction

Web-based health services have become a vital part of essential electronic services [1]. Booking appointments, viewing personal health information and test results, and searching for health-related information can be conveniently carried out online. Many web-based healthcare services, such as medical centers' websites, process sensitive personal information concerning health. Due to the sensitivity of this data, it is critical to ensure it remains confidential and does not leak to third parties [2].

However, previous research has demonstrated that across websites and services, regardless of sensitivity requirements, numerous third-party services and components, such as web analytics, are often used [3, 4]. Using such services makes monitoring business goals and improving user experience more convenient, but at the same time, there is a risk that sensitive information is leaked through these third party services. This typically happens without users' knowledge, and also unbeknownst to website developers and maintainers.

This study conducts an in-depth examination of the privacy of 10 web-based health services. We present an overview of health data leaks, an issue that an even larger group of web-based health services is likely to have. Our study specifically focuses on the privacy and confidentiality of Finnish web-based health services. Hence, in this study we address the following research question: *Do web-based healthcare services leak sensitive data related to an individual user's health status?* This paper serves as an analysis and discussion on the privacy threats associated with integrating third-party services in web-based health services.

The rest of the paper is organized as follows. Section 2 reviews related work on the privacy of medical websites. Section 3 outlines the study setting and the method, describing how the studied websites were selected and how the network traffic analysis was performed. Section 4 discusses the results of our network traffic analysis and explores the found data leaks. Section 5 presents a discussion on our key findings and their implications. Section 6 concludes the paper.

2. Related work

In recent years, a number of papers pertinent to our research have been published. Huo et al. [5] analyzed 459 health-related web portals and found that Google Analytics was used in 14% of them. Sensitive health data leaks were present on 9 websites, and details on e.g. prescribed medicines and laboratory results were transferred to third parties. Libert [6] investigates the problem of leaking health data contained in URL addresses to third parties. Zheutlin et al. [7] studied user data tracking through third-party cookies on USA-based government, non-profit, and commercial health-related websites, but did not go into detail about what personal data is sent to third parties.

Friedman et al. [8] discussed the risks of third-party tracking technologies in hospital websites, highlighting potential legal liabilities. Yu et al. [9] conducted a large-scale automated survey on hospital websites around the world, revealing that 53.5% of them employed tracking tools that collected user data. Friedman et al. [10] examined the prevalence of third-party tracking tools in abortion clinic websites and concluded that the majority (99.1%) used some form of tracking tool leaking user data to third parties. Surani et al. [11] found clear deficiencies in privacy policies of web-based health services.

Huesch [12] reminds that searching and accessing free health-related information online raises concerns about privacy and the potential for information on a user's health to be used for profiling and targeted advertising. Wesselkamp et al. [13] studied 385 medical websites in the EU area. They found that 62% used tracking tools before user consent for data collection and 15% tracked the user even after consent rejection. Kes et al. [14] argue that collecting of users' health data on websites, despite privacy concerns, can lead to an improved user experience akin to a personalized customer relationship. Still, the actual benefits are debatable, and transferring health data to third parties to improve targeted advertising is very problematic in the light of the GDPR.

Compared to many earlier studies, the current study conducts a more in-depth examination of types of personal data that web-based health services leak to third parties in different scenarios. We show that the issue of third-party analytics being present in web-based health services re-

TKTP 2024: The Annual Symposium of Computer Science, June 10-11, 2024, Vaasa, Finland

*Corresponding author.

sjprau@utu.fi (S. Rauti); crcarl@utu.fi (R. Carlsson); samuli.laato@tuni.fi (S. Laato); tdhein@utu.fi (T. Heino); papuht@utu.fi (P. Puhtila); ville.leppanen@utu.fi (V. Leppänen)
0000-0002-1891-2353 (S. Rauti); 0009-0003-7255-0239 (R. Carlsson); 0000-0003-4285-0073 (S. Laato); 0009-0008-4798-5261 (T. Heino); 0009-0004-6418-1063 (P. Puhtila); 0000-0001-5296-677X (V. Leppänen)

mains a significant problem despite having been addressed in research well over ten years ago [15].

3. Study Setting and Method

We selected 10 Finnish web-based health services for closer inspection in this study. We chose the websites of several important healthcare providers in Finland, such as medical centers, therapy houses, and laboratories. We searched healthcare providers using the Google search engine, with keywords "lääkärikeskus" (medical center), "terapia" (therapy) and "laboratorio" (laboratory). Instead of analyzing a large number of health services, our study examines the network traffic of these services more thoroughly. It includes various usage scenarios where sensitive health data web services process can leak to third parties. We examined the data leaks in the chosen services two times, first in December 2022 and then again in February 2024 after the service providers had been informed of the issue.

It is important to note that we aim to address privacy challenges at a general level and avoid singling out the affected health service providers in a negative light. To adhere to ethical research practices, the chosen web services are not referred to by their actual names but are denoted by abbreviations WS1–WS10.

In our test sequence, the browser cache was first cleared, cookies were deleted, and then the front page of the health service under examination was opened. On the front page, all cookies and data collection were accepted. When using the health service, all network traffic was recorded using Google Chrome browser developer tools (DevTools). The network traffic recordings were saved as HAR files (HTTP Archive) for more detailed analysis. We manually examined the log files, searching through the HTTP request payloads and documented all instances of personal data meticulously. Here we considered two distinct categories of personal data:

- *Identifying data*, capable of uniquely identifying the website user, such as IP addresses, User-Agent strings, and device-specific identifiers. Identification may also happen with a combination of technical details, including operating system or browser details, window size, etc.
- *Sensitive contextual data*, for example an URL address containing a sensitive search term used on a medical website, or details on a booked appointment. Although this kind of sensitive contextual data is often contained in URL addresses sent to a third party, it may also be elsewhere in the HTTP request payload.

What makes data leaks dangerous is the combination of these two categories: identifying a user by e.g. their IP address and then combining this to sensitive contextual data such as details on doctor's appointment. This enables third parties to infer user's potential medical conditions, for example. It is also worth noting that while the identifying personal data such as an IP address cannot always be immediately combined to a person's identity (real name), large technology companies such and Google and Meta often have the capability to fully identify the user, as users may use the same device to login to the other services run by these companies.

Four common usage scenarios where the leakage of health data to third parties is possible were recorded while using the health services. The chosen scenarios were key functionalities of the web-based health services that involved processing of sensitive personal data, and the scenarios varied based on the tested service. Network traffic was recorded when 1) booking an appointment, 2) viewing personal information, 3) using the search function, and 4) accessing information pages.

For the appointment booking scenario, network traffic was recorded from clicking the appointment link on the front page to the final stage of making the appointment. In other words, the test was concluded before the final confirmation of the appointment. In the appointment scenario, an appointment was scheduled with a specific specialist (such as a doctor or therapist). We also conducted a separate test for booking an appointment for a specific procedure or service (e.g. a COVID-19 test or influenza vaccination) if such an option was available in the tested health service.

The second scenario, viewing personal information, refers to the section behind the authentication of the web service. In this section of the web service, users can usually review their own prescriptions, test results, vaccinations, or previous appointments. In this scenario, we investigated whether data leaks occur when the user displays different types of personal information. For example, information about laboratory results and previous appointments could potentially be disclosed to third parties.

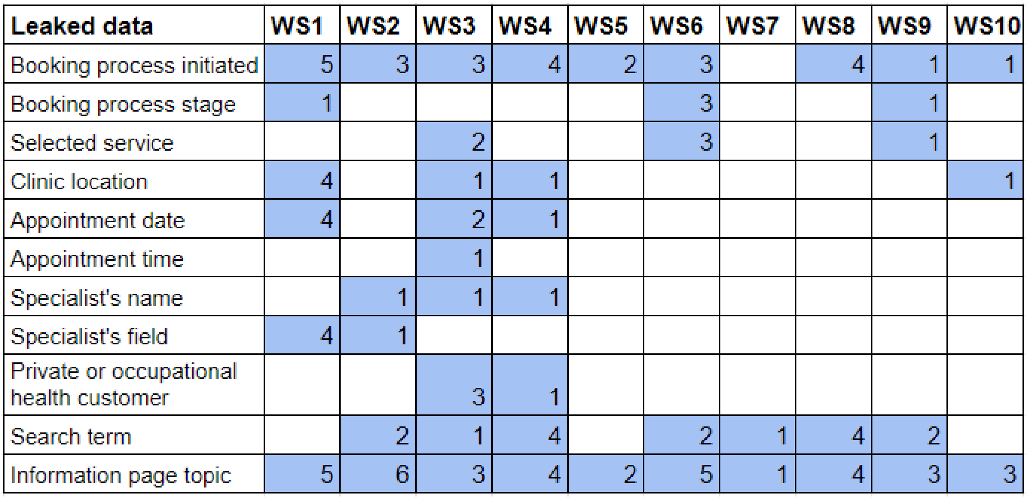
We also examined the possible leaks when using the search functions of the studied web services. The leakage of search terms to third parties can be particularly dangerous, because users may input highly sensitive terms, such as the name of a specific disease or symptom. If user-defined search terms are transmitted to third parties, these external actors can possibly build a detailed profile of the user's assumed health status and medical history.

The fourth usage scenario was related to information pages within web services, often containing information about specific diseases. It can be problematic if information about the pages a user browses is sent to third parties, as users can be profiled based on this. This can be especially effective over a longer time period.

4. Results

Figure 1 displays information leaked to third parties on the studied websites (December 2022). Each cell in Figure 1 indicates a leak of specific information type in a specific health service. The numbers indicate how many third parties the information was leaked to. For example, information about initiating an appointment booking was leaked to 5 different third parties in WS1.

A common data leak pertained to the use of the appointment booking function. Even though the appointment booking process was not completed in this study, the information about initiating this process indicates the user's intention to make a booking. In all services except for one (WS7), information about initiating the appointment booking process leaked to at least one third party. In three services, details about entering specific stages of the appointment booking process (e.g., selecting a time for the appointment, entering personal information) also leaked. Leaking any information about the appointment booking process is a problem because it strongly indicates a relationship between the patient and health provider. This kind of relationship must be kept confidential according to the Finnish Deputy



Leaked data	WS1	WS2	WS3	WS4	WS5	WS6	WS7	WS8	WS9	WS10
Booking process initiated	5	3	3	4	2	3		4	1	1
Booking process stage	1					3			1	
Selected service			2			3			1	
Clinic location	4		1	1						1
Appointment date	4		2	1						
Appointment time			1							
Specialist's name		1	1	1						
Specialist's field	4	1								
Private or occupational health customer			3	1						
Search term		2	1	4		2	1	4	2	
Information page topic	5	6	3	4	2	5	1	4	3	3

Figure 1: Data leaked in the web-based health services in December 2022.

Ombudsman[1].

Seven of the studied web services leaked additional information about appointments to third parties. These included the selected clinic location (3 web services), appointment date (3), appointment time (1), the name of the specialist (e.g., doctor) (3), the specialist's field of expertise (2), and whether the appointment was made as a private or occupational health customer (2). The selected service (e.g., influenza vaccination, COVID-19 test, or STD test) also leaked on three of the studied websites. In one case (WS10), the specific region (e.g., Central Finland) leaked instead of the exact clinic location.

The information transmitted to the third party about the initiation of the appointment is problematic by itself, because it implies a relationship between a patient and a healthcare provider. Details about the reserved health service or the doctor's name reveal the nature of this relationship even more precisely. It is also important to understand that a third party can often track a specific individual's online activities over a long period of time. When multiple appointments accumulate, a clear picture of the patient's treatment measures and health status begins to emerge.

Figure 1 also shows how users' searches were tracked. Notably, in all seven cases where a health service website had a search function, potentially sensitive search terms were transferred to at least one third party, and in the worst cases (WS4 and WS8), even up to four separate analytics services.

In all 10 examined health services, the URL addresses of information pages opened by the user were delivered to at least one third party. In the case of one service (WS2), the URL was sent to six third parties. Of course, viewing an information page about a specific illness does not necessarily imply that the visitor has that illness or even suspicion of it. However, the exposure of sensitive browsed pages to multiple third-party analytics services is not favorable.

Lastly, in our experiments we found no data leaks when viewing personal information such as laboratory results after logging in to the studied services. It seems these more sensitive sections of the health services have been implemented with the privacy-by-design approach in mind.

To sum up, the findings of Figure 1 are concerning: for each examined health service, information leaked to third parties either from the appointment booking page or search function, in most cases, both. These pieces of information – possibly combined with the pages the user browsed – can, in just one visit, give a third party an accurate picture of the user's current health.

Figure 2 shows the most common third parties (two instances or more) present in the studied health services in December 2022. Google Analytics and Meta Pixel were the most common ones, Google appearing in every single service and Meta in 8 services out of 10. The average number of third parties per health service was 5.2, which we consider a large number in websites processing such sensitive data. WS1 had a staggering 9 third parties, WS2 and WS6 following close behind with 8 third parties.

After discovering the data leaks in December 2022, the studied healthcare providers were informed about the issue. Figure 3 shows the updated status of data leaks in February 2024. The number of data leaks has decreased. For example, calculating the sum of all data leaks in Figure 1 yields 116, while this sum is 70 in Figure 3. However, this number is still very disappointing. Figure 3 shows clearly that revealing the initiation of the appointment booking process, and leaking viewed pages and search terms to third parties are still a significant issue in majority of the studied health services, although the number of leaks has gone down. It is also surprising that highly sensitive information such as the selected health service or the name of the specialist the patient is going to see is still being leaked. Only a single service, WS5, has completely removed third-party web analytics and eliminated data leaks.

5. Discussion

While the sensitivity of the data leaked by studied services ranged from visited information pages (not so sensitive) to details on booked appointments (highly sensitive), this data is still often directly related to the visitor's health status [6]. Also, even though the dataset we collected for the current study is not large in quantity, the finding that all of the analyzed web services leaked personal data to third parties cannot be simply dismissed. Although the situation has improved with time, web-based health services in Finland

[1] https://yle.fi/a/3-11213545

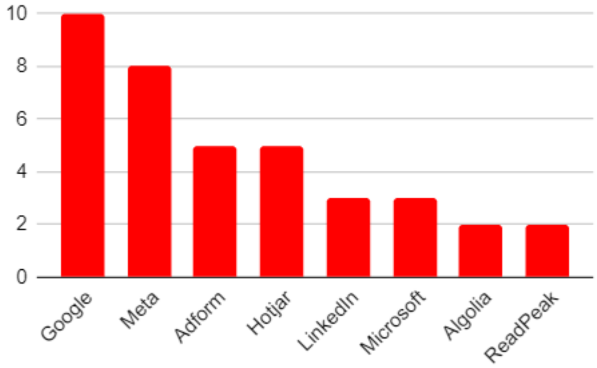
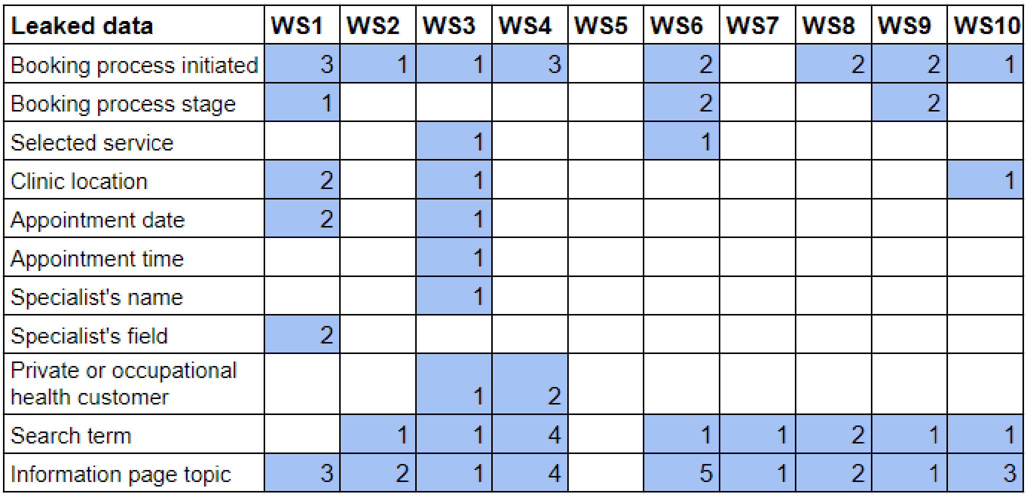


Figure 2: The most common third-party services present in the web-based health services in December 2022. Each third-party has only been counted once for each web service.



Leaked data	WS1	WS2	WS3	WS4	WS5	WS6	WS7	WS8	WS9	WS10
Booking process initiated	3	1	1	3		2		2	2	1
Booking process stage	1					2			2	
Selected service			1			1				
Clinic location	2		1							1
Appointment date	2		1							
Appointment time			1							
Specialist's name			1							
Specialist's field	2									
Private or occupational health customer			1	2						
Search term		1	1	4		1	1	2	1	1
Information page topic	3	2	1	4		5	1	2	1	3

Figure 3: Data leaked in the web-based health services in February 2024.

still appear to have many privacy challenges. Regrettably, it is highly likely that these issues extend well beyond the scope of the websites we examined.

Compared to many other studies (e.g. [5]), we found a high number of data leaks and observed these data leaks were widespread among the services we studied. One reason for this is likely to be different data collection methods. While many previous studies use automatic collection methods, we analyzed the network traffic and data leaks manually. Also, the other studies may not consider all the same data items our study does. Our goal was to consider all contextual data items that may relate to the user's health status. Some previous studies may only include the most sensitive data leaks like leaking laboratory results and medications and possibly exclude appointment booking related information, for example. Therefore, our set of studied data items and included use scenarios was more extensive than in most studies, which affects the numbers of found data leaks.

The use of third-party analytics is very difficult to justify on web-based health services. While we strongly believe the studied web-based services have not leaked sensitive personal data intentionally and while the third parties may not abuse it, the fact this data is sent to third parties remains a concern. There are multiple precautionary measures web developers and website maintainers should adopt to prevent such leaks.

A convincing argument can be made that third-party web analytics do not belong to websites processing sensitive health data. A straightforward alternative would be eliminating third-party analytics entirely. In the cases web analytics are necessary, locally hosted services like Matomo [16, 17] should be used. With the use of such self-hosted analytics, the health service provider now has full control over the collected data and there is no need to transfer it to a third party.

If third-party services really are necessary, chosen services should be thoroughly assessed and their use should be carefully justified. Of course, there are some well-justified use cases for trusted third-party services such as chat services or appointment booking systems that are vital for the functionality of the web-based health service. On the other hand, third-party analytics cannot be deemed essential for the functionality of web-based health services to the same extent.

During the software testing phase, a careful assessment of data leakages to third parties should be conducted, similar to the approach taken in the current study. In this examination of outgoing network traffic, special attention should be paid to pages that handle sensitive data, such as appointment bookings pages. Analyzing network traffic gives developers an accurate understanding of the data third parties collect. This analysis also helps website administrators in deciding which third-party services should be excluded from the service altogether. It is worth noting developers may unknowingly incorporate third-party analytics into websites, as off-the-shelf platforms commonly offer easy integration options or include them by default. This is why a network traffic analysis is essential.

A good understanding of the application area, such as the healthcare sector, holds great significance. The development team should aim to gain knowledge about the privacy regulations governing this particular industry. Effective communication with stakeholders is important in order to understand the requirements for protecting sensitive health data. When talking about essential online services such as medical center websites, the implemented service should also undergo an external privacy audit.

6. Conclusion

Our alarming discoveries should urge software developers and data protection officers overseeing web-based healthcare services to carefully assess the used third-party services and adopt a privacy-by-design approach. Developers and administrators of web services have to acknowledge their responsibility in protecting sensitive customer data and following fair data processing practices. The nature of processed personal data and the involved third parties have to be transparently communicated to users. When it comes to web-based medical services, it is unreasonable to rely on external services that may collect sensitive data. Failing to address serious data leaks, such as the ones presented in this study, increases the vulnerability of specific user groups online, especially in terms of privacy. Users of web-based health services should be able to see these websites as trustworthy and confidential equivalents to traditional onsite healthcare.

Acknowledgments

This research has been funded by Academy of Finland project 327397, IDA – Intimacy in Data-Driven Culture.

References

[1] P. Wang, Z. Ding, C. Jiang, M. Zhou, Design and implementation of a web-service-based public-oriented personalized health care platform, IEEE Transactions on Systems, Man, and Cybernetics: Systems 43 (2013) 941–957.

[2] S. Saha, C. Chowdhury, S. Neogy, A novel two phase data sensitivity based access control framework for healthcare data, Multimedia Tools and Applications 83 (2024) 8867–8892.

[3] R. Carlsson, S. Rauti, S. Laato, T. Heino, V. Leppänen, Privacy in popular children's mobile applications: A network traffic analysis, in: 2023 46th MIPRO ICT and Electronics Convention (MIPRO), IEEE, 2023, pp. 1213–1218.

[4] S. Rauti, R. Carlsson, S. Mickelsson, T. Mäkilä, T. Heino, E. Pirjatanniemi, V. Leppänen, Analyzing third-party data leaks on online pharmacy websites, Health and Technology (2024) 1–18.

[5] M. Huo, M. Bland, K. Levchenko, All eyes on me: Inside third party trackers' exfiltration of phi from healthcare providers' online systems, in: Proceedings of the 21st Workshop on Privacy in the Electronic Society, WPES'22, Association for Computing Machinery, New York, NY, USA, 2022, p. 197–211.

[6] T. Libert, Privacy implications of health information seeking on the web, Communications of the ACM 58 (2015) 68–77.

[7] A. R. Zheutlin, J. D. Niforatos, J. B. Sussman, Data-tracking on government, non-profit, and commercial health-related websites, Journal of general internal medicine (2021) 1–3.

[8] A. B. Friedman, R. M. Merchant, A. Maley, K. Farhat, K. Smith, J. Felkins, R. E. Gonzales, L. Bauer, M. S. McCoy, Widespread third-party tracking on hospital websites poses privacy risks for patients and legal liability for hospitals, Health Affairs 42 (2023) 508–515.

[9] X. Yu, N. Samarasinghe, M. Mannan, A. Youssef, Got sick and tracked: Privacy analysis of hospital websites, in: 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2022, pp. 278–286.

[10] A. B. Friedman, L. Bauer, R. Gonzales, M. S. McCoy, Prevalence of third-party tracking on abortion clinic web pages, JAMA Internal Medicine 182 (2022) 1221–1222.

[11] A. Surani, A. Bawaked, M. Wheeler, B. Kelsey, N. Roberts, D. Vincent, S. Das, Security and privacy of digital mental health: An analysis of web services and mobile apps, in: Conference on Data and Applications Security and Privacy, 2023.

[12] M. D. Huesch, Privacy threats when seeking online health information, JAMA Internal Medicine 173 (2013) 1838–1840.

[13] V. Wesselkamp, I. Fouad, C. Santos, Y. Boussad, N. Bielova, A. Legout, In-depth technical and legal analysis of tracking on health related websites with ernie extension, in: Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society, WPES '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 151–166.

[14] I. Kes, D. Heinrich, D. M. Woisetschlager, Behavioral targeting in health care marketing: Uncovering the sunny side of tracking consumers online, in: Let's Get Engaged! Crossing the Threshold of Marketing's Engagement Era: Proceedings of the 2014 Academy of Marketing Science (AMS) Annual Conference, Springer, 2016, pp. 297–297.

[15] K. Masters, The gathering of user data by national medical association websites, The Internet Journal of Medical Informatics 6 (2012).

[16] J. Gamalielsson, B. Lundell, S. Butler, C. Brax, T. Persson, A. Mattsson, T. Gustavsson, J. Feist, E. Lönroth, Towards open government through open source software for web analytics: The case of matomo, JeDEM-eJournal of eDemocracy and Open Government 13 (2021) 133–153.

[17] D. Quintel, R. Wilson, Analytics and privacy, Information Technology and Libraries 39 (2020).