Personal Data Transfers to Non-EEA Domains: A Tool for Citizens and An Analysis on Italian Public Administration Websites

Lorenzo Laudadio
Politecnico di Torino
Turin, TO, Italy
lorenzo.laudadio@polito.it

Antonio Vetrò
Politecnico di Torino
Turin, TO, Italy
antonio.vetro@polito.it

Riccardo Coppola
Politecnico di Torino
Turin, TO, Italy
riccardo.coppola@polito.it

Juan Carlos De Martin
Politecnico di Torino
Turin, TO, Italy
demartin@polito.it

Marco Torchiano
Politecnico di Torino
Turin, TO, Italy
marco.torchiano@polito.it

ABSTRACT

Six years after the entry into force of the GDPR, European companies and organizations still have difficulties complying with it: the amount of fines issued by the European data protection authorities is continuously increasing. Personal data transfers are no exception. In this work we analyse the personal data transfers from more than 20000 Italian Public Administration (PA) entities to third countries. We developed "Minos", a user-friendly application which allows to navigate the web while recording HTTP requests. Then, we used the back-end of Minos to automate the analysis. We found that about 14% of the PAs websites transferred data out of the European Economic Area (EEA). This number is an underestimation because only visits to the home pages were object of the analysis. The top 3 destinations of the data transfers are Amazon, Google and Fonticons, accounting for about the 70% of the bad requests. The most recurrent services which are the object of the requests are cloud computing services and content delivery networks (CDNs). Our results highlight that, in Italy, a relevant portion of Public Administration websites transfers personal data to non EEA countries. In terms of technology policy, these results stress the need for further incentives to improve the PA digital infrastructures. Finally, while working on refinements of Minos, the version here described is openly available on Zenodo: it can be helpful to a variety of actors (citizens, researchers, activists, policy makers) to increase awareness and enlarge the investigation.

ACM Reference Format:
Lorenzo Laudadio, Antonio Vetrò, Riccardo Coppola, Juan Carlos De Martin, and Marco Torchiano. 2024. Personal Data Transfers to Non-EEA Domains: A Tool for Citizens and An Analysis on Italian Public Administration Websites. In *International Conference on Information Technology for Social Good (GoodIT '24), September 04–06, 2024, Bremen, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3677525.3678632

GoodIT '24, September 04–06, 2024, Bremen, Germany

ACM ISBN 979-8-4007-1094-0/24/09
https://doi.org/10.1145/3677525.3678632

1 INTRODUCTION

In today's digital and connected world, personal data can be considered a new currency. Commercial data exchanges have a huge impact on the economy, with many companies having built their business around extracting and selling such data [12]. These pieces of information tell much about our lives and contribute to forming what is called our *digital identity*. In the last few decades, several episodes have been reported in which personal data have been collected and exploited without consent from their owners. In 2013, Edward Snowden disclosed the mass surveillance programs conducted by the US government under the name of PRISM [6]. In 2018, Facebook users discovered that their personal data were being collected by the British company Cambridge Analytica and exploited to influence their political opinions [7]. More recently, in 2023, the use of facial recognition systems by Israel to set up biometric surveillance systems targeting Palestinians has been reported by Amnesty International [8]. As a consequence, more and more people seem to have developed concerns about their privacy. At the same time companies and institutions are not performing very well when it comes to complying with privacy regulations [1].

In this work, we address the problem of personal data transfers from the EEA to third countries with a focus on data transfers operated by the Italian Public Administration (PA) entities: we provide the justification of our choice in Section 2 together with other contextual information and related work. In Section 3 we report on methodology and instruments of our analysis, presenting Minos: a software tool which can detect HTTP requests directed to third countries. We report and discuss the results we obtained by running a mass analysis on the Italian PA entities in Section 4, and identify limitations of the analysis in Section 4.4. We summarize the contributions and future work in Section 5. The extended pre-print version of this article, accepted to the conference, is available on arXiv[2].

2 BACKGROUND AND RELATED WORK

The General Data Protection Regulation (GDPR) is the European regulation which addresses data protection and privacy problems

[1] See, for example, the highest fines issued for General Data Protection Regulation (GDPR) violations as of January 2024 https://www.statista.com/statistics/1133337/largest-fines-issued-gdpr/ , last visited on 15 May 2024
[2] https://arxiv.org/abs/2407.13467

in the EU and EEA, and the transfers of personal data outside the EU and EEA.

Article 4 defines personal data as any information which can lead to the identification of a natural person, including location data and online identifiers. The IP address of a host can be considered as personal data. Due to how the Internet works, whenever a personal device makes a request to a certain domain, its IP address is shared with that domain. This means that any HTTP request results in a personal data transfer.

Article 28 states that the data controller should choose only processors which provide sufficient guarantees to implement appropriate technical and organizational measures such that processing meets the GDPR requirements and ensures the protection of the rights of the data subject.

Chapter V of the GDPR regulates personal data transfers to third countries or international organizations. The European Commission (EC) has the power to decide whether a non-EEA country offers an adequate level of protection for personal data. Such a decision is called an *adequacy decision*. If such a decision is missing, data can still be transferred if appropriate safeguards are provided by the controller or processor. Ultimately, in the absence of an adequacy decision and appropriate safeguards, data transfer can take place based on specific derogations. Any data transfer which does not fall under one of these cases is not GDPR-compliant.

In our study, we focus on personal data transfers operated by the Italian Public Administration entities because of the highly sensitive nature of the data they operate on. In addition, to the best of our knowledge, no studies so far addressed the problem of third-countries data transfers concerning the PA ecosystem. A quantitative investigation on privacy compromising mechanisms on popular websites was conducted by Timothy Libert, with a focus on third-party tracking mechanisms [9]. He found that 88% of the analysed websites made requests to a third-party domain. Guamàn et al. [5] suggested a method to assess the GDPR compliance of data transfers in Android applications, highlighting the presence of ambiguities and inconsistencies within the mobile application environment. They also considered privacy policies for their analysis. They found that 66% of the analysed apps were found as ambiguous, inconsistent or omitting cross-border data transfer disclosures. Granata et al. [4] applied the NIST SP-800-53 security control framework to GDPR compliance assessment, introducing a mapping between the GDPR articles and the NIST SP-800-53 security controls. Loré et al. [10] proposed a new AI-based framework to detect potential violations of the GDPR in the context of Italian Public Administration, even though they did not address the analysis of data transfers.

3 METHODOLOGY

Our analysis is driven by the following research questions:

- **RQ1:** What is the distribution of Italian PA sources that transfer personal data to non-EEA countries?
- **RQ2:** Which are the most common destinations of the data transfers to non-EEA countries?
- **RQ3:** Which are the most common types of services requested in the context of the data transfers to non-EEA countries?

Data Sources. The list of Italian PA entities – along with other information such as the category or official website of each entity – is retrieved from the publicly available OpenData IPA database[3].

Two datasets have been used for the analysis: `enti.csv`, containing all the personal data of the entities considered for the analysis, and `categorie-enti.csv`, which contains data about the categories of the entities.

Some URLs were unusable for the analysis due to being inaccurate, incomplete or inconsistent. Three kind of badly formatted URLs were identified: empty ones, invalid URLs (e.g. `about:blank`) and those containing typos. Since there is no way to check programmatically whether a given URL is valid or not without contacting the corresponding host, the number of unreachable websites was taken as an estimate of the number of badly formatted URLs. The empty ones were removed before starting the analysis.

Instrumentation Our investigation required the collection and analysis of a large number of HTTP requests to identify the third countries involved in data transfers with the Italian PA entities.

For this purpose, we developed Minos, a software application that allows non-advanced users to navigate the web while recording their HTTP requests. The tool was developed with the Electron JavaScript framework and is available as open-source on Zenodo [2]. Minos relies upon an internal blacklist which was created by volunteers from the nonprofit community MonitoraPA [4], which operates a public observatory since 2022 and periodically reports about GDPR violations within the Italian Public Administration world. Such blacklist contains popular third-party domains from non-EEA countries[5]. These domains are divided in groups. For example, all the `youtube.com.*` domains belong to the `youtube` group. A further refinement has been operated on these domain groups in order to retrieve the companies these domains belong to and their relative countries. For example, both the domains of the `microsoft` group and the ones of the `azure` group belong to the US company Microsoft. A mix of different sources was used to trace the companies and countries behind the domain groups, namely: Wikipedia, the website *opencorporates.com* – an open database which contains data of corporations –, the *ipinfo.io* IP Geolocation API service, and the *webXray* domain owner list – a large hierarchical collection of third-party domains on the web and the owning companies [1] –.

To automate the analysis of the PA entities, a slightly different version of Minos has been created, called Minos-cli. This version consists of a set of JavaScript and Python scripts to collect, analyse and visualize the data. Minos-cli relies on a running instance of a Chromium-based web browser in order to exploit the Chrome Debugging Protocol (CDP) which allows analysing HTTP requests. A JavaScript module, called Chrome HAR Capturer (CHC)[6], has been modified and used to read data from the fields of the HTTP requests and store them in a csv file.

Data collection The data collection took place according to the following process: (1) the Minos-cli scripts read the list of entities' URLs; (2) a running instance of a Chromium-based web browser is

[3] https://indicepa.gov.it/ipa-dati/
[4] https://monitora-pa.it/
[5] The blacklist can be found at https://github.com/MonitoraPA/Minos/blob/main/hosts.json
[6] https://github.com/cyrus-and/chrome-har-capturer

instrumented, in order to contact the website and save the HTTP request in an internal HAR object, through to the Chrome Debugging Protocol; (3a) if an HTTP request's `entry.request.url` field matches one of the blacklisted domains, a line containing all the data about the request is appended to the `bad-requests.csv` file, and the IPA Code of the entity being considered is logged to the `done.csv` file; (3b) if the website cannot be contacted due to some error, the IPA Code of the entity is written to the `done.csv` file, along with an error message. To detect matches between the request's URL and the blacklisted domains the longest prefix matching is adopted. For example, even if the URL `https://www.youtube.com/` matches against both `youtube.co` and `youtube.com`, only the latter is taken as valid.

Entities making at least one request to blacklisted domains have been classified as *bad entities*, and their requests as *bad requests*. After data collection, the files `bad-requests.csv` and `done.csv` are processed by scripts that compute statistics on the number of bad requests, bad entities, bad entities by category and bad requests by domain, domain group and company.

4 RESULTS AND DISCUSSION

Our analysis targeted all the 22890 PA websites present in the OpenData IPA datasets. 14.83% of the websites were not reachable due to errors or missing URLs. The most common error encountered during the analysis was the `net::ERR_NAME_NOT_RESOLVED` error, which is a symptom of a badly formatted or misspelt URL.

4.1 RQ1: Transfer of personal data to non-EEA

We found that 15% of the entities send data out of EEA countries: since we analyze only the request made to the home page only, we expect this number to be higher in a real navigation context.

Figure 1 displays the top 10 PA categories the bad entities belong to. The category *Municipalities* ranks first: this is consistent with several investigations made in Italy that reported a general trend towards widespread violations of the GDPR by Italian municipalities in recent years. A 2019 report by Federprivacy revealed that at the time 47% of the Italian municipalities did not provide `https` access to their websites, while 36% did not specify the contact details of their Data Protection Officer, although the regulation has been in force since 2018 [3]. Moreover, between 2020 and 2022, the Italian Data Protection Authority sanctioned several municipalities for unlawful processing or diffusion of citizens' personal data. In general, Italian municipalities seem to have difficulty complying with the GDPR. The second place is held by the category *National teaching institutes*. Several sources have reported an increase in the use by Italian educational institutes of services offered by the US big tech companies in recent years, in particular after the COVID-19 global pandemic. Paolo Monella, in his 2021 article «Education and GAFAM: from awareness to responsibility», reported that since the outburst of the COVID-19 pandemic, schools and colleges based their teaching almost exclusively on big tech infrastructures and platforms [11].

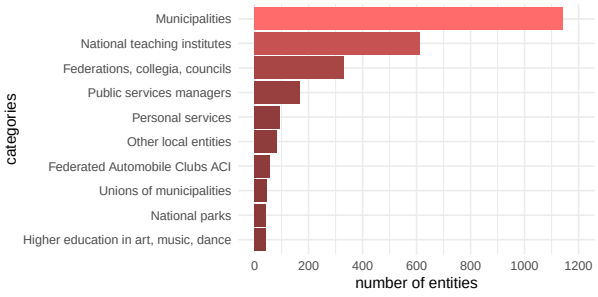

Figure 1: Distribution of bad entities per category

4.2 RQ2: destinations of personal data transfers

Figure 2 and Figure 3 visualize the number of bad requests, respectively grouped by target domain group and company.

Most of the requests are addressed to the `aws` group. Amazon Web Services (AWS) is a subsidiary of Amazon which offers on-demand cloud computing services. Various online sources report AWS as the leading provider of cloud services globally. The website *european-alternatives.eu*[7] lists several cloud service providers from the EU. However, organizations and companies still seem to prefer Amazon for its simplicity, where the self-hosted deployment instead would require advanced knowledge and skills, as well as high deployment costs in the early stages of the operation. Right after, we find Google, another popular big tech giant which offers a wide range of services: maps, web search, website development tools, multimedia hosting, etc. The Google services being more

[7] https://www.european-alternatives.eu

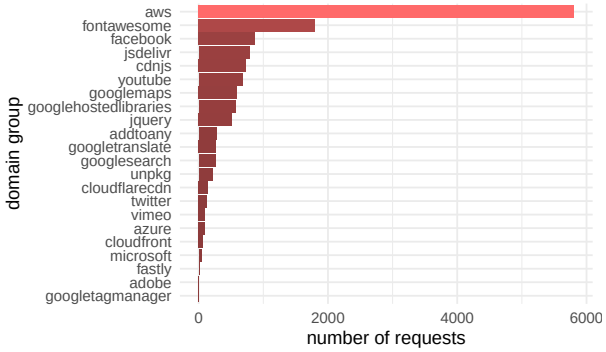

Figure 2: Bad requests per domain group

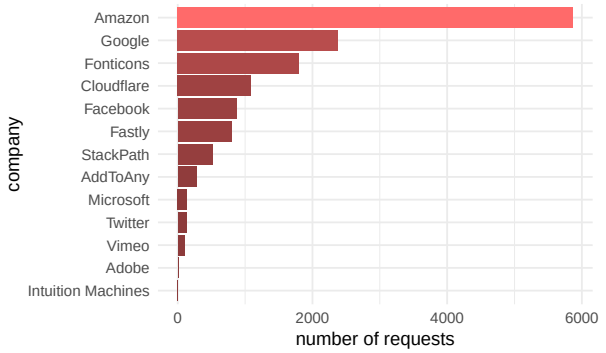

Figure 3: Bad requests per company

requested are the ones from YouTube, Google Maps and Google Hosted Library (Google's CDN). In third place, we have Fonticons, the company behind Fontawesome, a popular icon library used in many websites. Requests directed to Amazon are more than 40%, and more than 70% of requests are directed to the first three domains (Amazon + Google + Fontawesome). Other companies offer mostly content delivery networks (e.g. Fastly), social networking and multimedia (e.g. Facebook) or other services (e.g. Adobe).

4.3 RQ3: Distribution type of services

Figure 4 shows the number of bad requests per type of service requested. The results highlight a prevalence of essential services being requested, such as cloud computing services and content delivery networks. Together, the requests for these two types of services represent almost 70% of the overall number of requests. The distribution of requests for cloud computing services is consistent with the global distribution of requests per group: Aws cloud services are the most requested (98%) followed by Azure (2%), Microsoft's cloud services. For CDNs, instead, the distribution is much more uniform: in first place, there is Fontawesome (22%), followed by JSDelivr (21%) and CDNjs (20%). Cloud services and CDNs require a considerable amount of computational power and large IT infrastructures. This is in line with the categories of PA entities identified in RQ1 as most frequently transferring personal data out of EEA: municipalities and teaching institutions, which are - on average - small-sized and, in the absence of a centralized infrastructure provided by the Italian State, do not have the proper know-how and funding to build their own services.

4.4 Threats to validity

The adopted data collection approach produces underestimates of the number of bad requests, since only home pages are loaded by the Chromium browser and the session is closed immediately. Therefore, any requests originating by further navigation are not detected. We are working to improve this aspect and minimize the number of false negatives by waiting for an additional timeout after the page has been loaded and triggering random navigation events with the help of a browser automation tool. Anyway, we believe that the relative number of bad requests might not be significantly influenced by including navigation to pages other than the considered home pages.

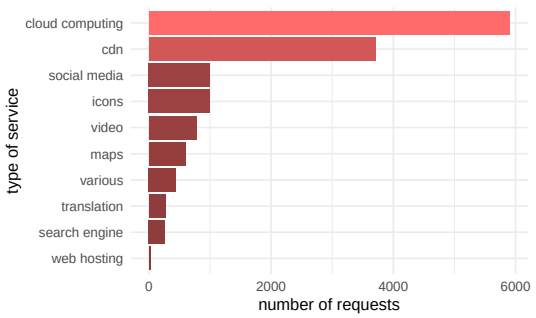

Figure 4: Bad requests per type of service

We stress that the present study considered mainly US destinations of requests (83% of the companies involved in the analysis are from the US). Our goal was to investigate the reliance of Italian PAs on big tech corporations and foreign domains. The majority of domains are from the US simply because many big techs are located in the US.

5 CONCLUSION AND FUTURE WORK

In this work, we presented the results of an ongoing analysis of personal data transfers by Italian PA websites to countries outside of the EEA. We leveraged a command line request analyzer, Minos-cli, to run a massive analysis on Italian PA entities from public datasets. Our analysis identified a large number of requests coming from these entities. All of them were directed to US domains, mostly to Amazon domains. The PA entities with the most violations reported were municipalities and schools, requesting most of the time cloud computing services and content delivery networks. Even if some alternatives exist in the EU, numbers are still low, and the self-hosting does not seem a popular alternative, since it requires advanced knowledge and, at least at the beginning, larger investments. In terms of technology policy, these results highlight the need for incentives for the PA's organizations to access IT services within the EEA.

ACKNOWLEDGMENTS

We thank Giacomo Tesio and Massimo Maria Ghisalberti, who have followed the development of Minos, and Marco Ciurcina, whose contribution has been crucial in understanding the legal concepts.

REFERENCES

[1] Reuben Binns. 2018. webXray Domain Owner List. https://github.com/RDBinns/webXray_Domain_Owner_List.
[2] ebmaj7 and Giacomo Tesio. 2024. *ebmaj7/Minos: MINOS_v1.0.0.* https://doi.org/10.5281/zenodo.11384690
[3] Federprivacy. 2019. Privacy, il 47% dei siti dei comuni italiani è a rischio hacker. https://www.federprivacy.org/informazione/societa/privacy-il-47-dei-siti-dei-comuni-italiani-e-a-rischio-hacker.
[4] Daniele Granata, Michele Mastroianni, Massimiliano Rak, Pasquale Cantiello, and Giovanni Salzillo. 2024. GDPR compliance through standard security controls: An automated approach. *Journal of High Speed Networks* 30, 2 (Jan. 2024). https://doi.org/10.3233/JHS-230080 Publisher: IOS Press.
[5] Danny S. Guamán, Jose M. Del Alamo, and Julio C. Caiza. 2021. GDPR Compliance Assessment for Cross-Border Personal Data Transfers in Android Apps. *IEEE Access* 9 (2021), 15961–15982. https://doi.org/10.1109/ACCESS.2021.3053130
[6] The Guardian. 2013. Edward Snowden: the whistleblower behind the NSA surveillance revelations. https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance.
[7] The Guardian. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.
[8] Amnesty International. 2023. Israel and Occupied Palestinian Territories: Automated Apartheid: How facial recognition fragments, segregates and controls Palestinians in the OPT. https://www.amnesty.org/en/documents/mde15/6701/2023/en/.
[9] Timothy Libert. 2015. Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites. https://doi.org/10.48550/arXiv.1511.00619 arXiv:1511.00619 [cs].
[10] Filippo Lorè, Pierpaolo Basile, Annalisa Appice, Marco de Gemmis, Donato Malerba, and Giovanni Semeraro. 2023. An AI framework to support decisions on GDPR compliance. *Journal of Intelligent Information Systems* 61, 2 (Oct. 2023). https://doi.org/10.1007/s10844-023-00782-4
[11] Paolo Monella. 2021. Education and GAFAM: from awareness to responsibility. *Umanistica Digitale* 5, 11 (Jan. 2021), 27–45. https://doi.org/10.6092/issn.2532-8816/13685
[12] Shoshana Zuboff. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st ed.).