Data Portability as a Tool for Audit

Zoe Zwiebelmann*
University of St Andrews
School of Computer Science
St Andrews, Fife, UK
zoe@zwiebelmann.de

Tristan Henderson
University of St Andrews
School of Computer Science
St Andrews, Fife, UK
tnhh@st-andrews.ac.uk

ABSTRACT

Pervasive systems are almost omnipresent in their collection and processing of personal data. Understanding what these systems are doing is essential for trust, and to ensure that data being collected are accurate. Auditing these systems can help to determine the accuracy of these data. Such audit may take place internally by systems designers, but external audit is important for accountability.

In this paper we explore whether users can conduct their own external audit of the systems with which they interact. In particular, we use the Right to Data Portability afforded to data subjects through the General Data Protection Regulation. Using fitness trackers, we collect and upload running data to a set of data controllers. By using data portability to then obtain a copy of our data, we compare the data held by the controllers with our ground-truth data. We find some inaccuracies in the data, but also that audit can be impeded by insufficient explanations from data controllers.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Social and professional topics** → **Technology audits**; • **Hardware** → *Sensor applications and deployments.*

KEYWORDS

GDPR; Data Portability; Activity-Tracking; Audit

ACM Reference Format:
Zoe Zwiebelmann and Tristan Henderson. 2021. Data Portability as a Tool for Audit. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460418.3479343

*The first author conducted this work as part of their MSc dissertation at the University of St Andrews.

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09…$15.00
https://doi.org/10.1145/3460418.3479343

1 INTRODUCTION

Pervasive computing has arrived: sensors, actuators and applications and devices using these are commonplace, collecting and processing data with or without the knowledge of those who are generating the data. Understanding what data are collected and whether those data are accurate is crucial for instilling trust in these systems. In other words, we need to be able to audit these systems.

In this paper we explore the use of the Right to Data Portability (RtDP), provided to data subjects as part of the EU General Data Protection Regulation (GDPR), to see whether it can be used to help audit pervasive computing systems. The RtDP allows data subjects to obtain their personal data in a machine-readable format. In particular, we examine the use case of a wearable fitness tracker. Using a tracker to collect fitness data locally to provide ground-truth data, we upload data to a variety of fitness-tracking services. We then exercise the RtDP to obtain the data that these services believe that they hold about us, and compare these data to our original ground-truth data.

We find that in many cases, we are able to use our RtDP requests to determine the accuracy of the data held, and find that some data are inaccurate. In other cases, we cannot successfully audit, because of insufficient data returned by the controllers (in one case, no data at all), or insufficient metadata or explanation to allow us to interpret the data. This raises interesting questions both for the use of data rights in general, but also for further study into the use of rights for auditing systems.

This paper is laid out as follows. Section 2 outlines background and relevant related work in both law and computer science. Section 3 describes the study, and we highlight some interesting results in Section 4. Finally in Section 5 we discuss implications of our findings and future directions.

2 BACKGROUND AND RELATED WORK

The auditing of algorithmic decision-making systems and other large-scale data collection mechanisms is increasingly recognised as an important means of accountability in such systems. Bandy surveys 62 studies where external agents audit public systems in order to provide public accountability [4]. Among the suggestions for future work is the need to establish baselines and metrics. Tools have also been created for the internal audit of systems by the system designers themselves [16].

Our study employs data subject rights, which are increasingly recognised as a research method [3], as a tool to obtain and audit data held by data controllers. Article 20 of the GDPR provides data subjects with the right to data portability: data subjects have

the right to receive personal data that they have provided to a controller, in a "structured, commonly used and machine-readable format" (Article 20(1) [9]). As the only new data subject right in the GDPR, Article 20 has been the subject of several empirical studies. Wong and Henderson make portability requests to 230 controllers and find variation in their ability to conform with the right's requirements [21]. Syrmoudis et al. [19] look at 182 controllers to specifically study the ability to exercise Article 20(2) — the right to port data from one controller to another. In a smaller-scale study, Li compares a set of portability requests with another right, the right to erasure [13]. Two recent studies are most similar to our work in that they focus on the Internet of Things [5, 20] — their aim, however, is more on examining whether portability can be exercised, rather than the accuracy of the data that we study here.

We choose to focus on fitness trackers, which are commonly associated with self-tracking or the quantified self [18]. Lupton [14] situates the quantified self in "audit culture", whereby technologies such as the fitness trackers studied in this paper can be used by institutions for surveillance and accountability of individuals, but can also be used by individuals for their own self-audit. Our work attempts to operationalise this self-audit. Schreiber [17] proposes a model for checking the provenance of quantified self data. Provenance in this model is concerned with what data were generated, which activities generated the data, and which parties were involved or have access to the data. Accuracy is not considered.

Audit and the GDPR have also been studied elsewhere, e.g. Arfelt et al's system for data controllers to automatically audit their compliance with the GDPR [1]. While some data subject rights such as Article 15 (right to access) are formalised and audited in this system, Article 20 is not.

3 METHODOLOGY

Fitness trackers and services pervasively collect vast amounts and various types of personal and sensitive data [10]. As such they provided an interesting case study for exercising the RtDP. Our study's methodology can be split into: data collection and storage; data transmission; exercising of rights; and data analysis.

To collect data, we used an Amazfit Bip fitness tracker, and connected this to the open-source application Gadgetbridge running on a Samsung Galaxy A20e Android phone. Gadgetbridge allows the use of a selected number of fitness tracking devices without the vendor's closed source application, thus eliminating the need to create an account and transmitting personal data to the vendor's server [11]. This allowed us to collect GPS files from the Amazfit Bip after running sessions, and store these locally in a NoSQL database to keep as ground-truth data.

To transmit data to data controllers, we first chose an appropriate set of service providers based on popularity (Table 1). All apps selected for assessment offered tracking of basic running metrics such as pace, heart rate, burned calories, and distance. We signed up for each service using a fake profile that was similar but not identical to one of the authors. Most service providers required the submission of personal data such as gender, date of birth, height and weight (justifying this, for instance, with the need to accurately calculate the number of calories burned during a run).

Table 1: Popular Running Apps (State: June 2020) [12]

App	*Company*	*Play Store Downloads*
Adidas Running by Runtastic	Adidas	50m+
Strava	Strava	10m+
MapMyRun	Under Armour	10m+
Endomondo	UnderArmour	10m+
Nike Run Club	Nike	10m+
Runkeeper	ASICS	10m+
Komoot	komoot	5m+

Table 2: Data Collection Details

Collection Start Date	5th May 2020
Collection End Date	2nd July 2020
Number of Runs	39
Number of Towns	3
Minimum distance per run	2km
Maximum distance per run	15km

Table 3: Data Transmission Details

Data controller	*Transmission method*
Adidas Running	GPX import
Strava	Amazfit app integration
MapMyRun	GPX import onto web
Endomondo	GPX import onto web
Nike Run Club	None of the methods supported
Runkeeper	GPX import onto web
Komoot	GPX import into app
Amazfit	Amazfit app integration

Having created accounts, we then carried out a number of runs (Table 2) and recorded the data in Gadgetbridge.

Data were submitted to the service providers in three ways: through the device vendor, the Amazfit app, that was connected to the Amazfit Bip fitness tracker; through GPX (GPS Exchange Format) import functions on the service providers' websites; or through directly importing the raw GPX file from Gadgetbridge into the Komoot app (Table 3). We selected one service provider (Nike Run Club) as a destination for Article 20(2) transmissions since it could not interact with our chosen Amazfit device, nor did it provide any option to import raw GPX files. We uploaded data to the other six chosen applications and to the device provider Amazfit.

To exercise the RtDP and obtain a copy of the uploaded data, we examined the privacy notices for each service provider to find contact details for the Data Protection Officer. We sent requests via e-mail using a template (similar to Wong and Henderson [21]) to four providers (Adidas Running, Strava, Komoot, Amazfit) and three providers (Endomondo, Runkeeper, MapMyRun) were contacted using an online form. We asked both for a copy of our data (as per Article 20(1)) and to port data to another controller (Article 20(2)).

Table 4: Overview of file formats of responses sent by data controllers.

Data Controller	*Description*	*File Formats*	*Files*
Adidas Running	YES	JSON, GPX, JPG	279
Strava	YES	CSV, GPX, JPG	82
MapMyRun	NO	XLSX, CSV	2
Endomondo	YES	JSON, HTML, TCX, JPG	43
Runkeeper	NO	GPX, CSV	41
Amazfit	NO	CSV	6
Komoot	N.A.	N.A.	N.A.

4 RESULTS

Article 12(3) states that data controllers shall provide information in response to a request within one month, with potential extension of a further two months depending on the complexity and number of the requests. Our seven data controllers responded within 1 to 23 days. One controller replied to say that fulfilling the request would take longer than one month, and eventually provided the data after 32 days. Two controllers asked for further identity verification. After responding to the request, four controllers pointed to an existing "Download my data" tool, while two sent data by e-mail. One controller, Komoot, stated that they were unable to provide any data at all, and did not provide further explanation.

Article 20(1) states that data controllers shall provide data in a structured, commonly used and machine-readable format. Data were returned in a variety of formats and number (Table 4). Data were uploaded to controllers in GPX format, and some controllers returned data in this same format. Endomondo returned the GPS data in TCX format, while MapMyRun returned an Excel spreadsheet containing GPS traces. The JPG files returned were those images that were uploaded during the submission of runs. Table 4 also shows that only half of the responses contained a description of the dataset or any metadata. A lack of description and conversion of formats between upload and download may hinder the use of portability for audit.

We considered the accuracy of the data in three parts: the user profile data, the running statistics and the GPS data. Only one data controller returned all of the user profile data in a complete and accurate state. Three controllers returned partial information (e.g. missing date of birth, gender, or name). Two data controllers returned inaccurate profile information (weight, height and date of birth). All other profile data elements were returned accurately and in the same format that were uploaded (excluding minor changes such as date string formats such as 01.01.1996 being uploaded and 1996/01/01 being returned).

Running statistics (duration, distance, average pace, calories and so forth) were calculated by the data controllers on the basis of the submitted GPX files. Such processed or inferred data are excluded from the right to data portability [2], but they were returned by five controllers (all apart from Endomondo and Komoot). We were able to calculate these statistics using our ground-truth data and compare. Table 5 shows that only one data controller provided complete and accurate statistics. Completeness was measured by the extent to which the data that was provided was returned by data controllers, and accuracy was measured by the extent to which the returned data were correct and coincided with the transmitted data. We saw that Strava's *duration* measure was wrong for more than half of the submitted runs. Amazfit returned data in metrics that we could not understand (e.g. distance = 1.6649797 for a 10km run, average pace = 218) and no information was provided on how to convert these.

Table 5: Assessment of completeness and accuracy of returned running statistics (starting time, distance, duration) per run by data controller. The number of inaccurate runs (out of a total of 39 runs) per statistic is indicated in the respective column.

Data Controller	*Complete*	*Accurate*	*Time*	*Distance*	*Duration*
Adidas Running	YES	NO	1	0	0
Strava	YES	NO	0	0	21
MapMyRun	NO	YES	0	0	0
Endomondo	N.A.	N.A.	N.A.	N.A.	N.A.
Runkeeper	YES	YES	0	0	0
Amazfit	YES	NO	0	39	39
Komoot	N.A.	N.A.	N.A.	N.A.	N.A.

Table 6: Number of data points in each GPS-data set, further separated into total number of track points (TP), geolocation, elevation, time, and heart rate data.

Source	*Total TP*	*Geolocation*	*Elevation*	*Time*	*HR*
Raw Data	95,037	95,037	95,037	95,037	94,666
Adidas Running	95,037	95,037	95,037	95,037	94,669
Strava	100,393	94,146	93,223	100,393	43,654
MapMyRun	N.A.	N.A.	N.A.	N.A.	N.A.
Endomondo	95,037	95,037	95,037	95,037	95,037
Runkeeper	95,037	95,037	95,037	95,037	0
Amazfit	N.A.	N.A.	N.A.	N.A.	N.A.
Komoot	N.A.	N.A.	N.A.	N.A.	N.A.

The GPX data provided to each controller included location data (latitude, longitude, elevation) recorded periodically in track points (TP), and heart rate (HR) data. Such health data are sensitive (special category under the GDPR) and with potential uses such as insurance, their accuracy is of importance. Four out of the seven data controllers returned GPS data for each of the 39 runs (Table 6). The disparity between HR and TP can be explained by the fact that heart rate was not measured at the beginning of each run. The most interesting disparities are between Strava and the raw data. Closer examination of the data showed that Strava structures its data into TPs occurring every two seconds of a run, while the uploaded data contained TPs every three to four seconds. Interpolation appears to have led to an increased number of TPs, and yet fewer points for elevation and HR.

We further examined the HR data since this is health and more sensitive. Table 6 shows that Runkeeper failed to return the HR data with the GPS traces. Instead it provided average HR per run; we were able to calculate these averages from our ground-truth

data and they were all accurate. Adidas and Endomondo provided differing numbers of HR points to the raw data. Some of these appear to be errors (e.g. 368 Adidas TPs had no HR data yet 263 of these did have HR data in the upload). Endomondo appears to have interpolated the HR data to make up for the missing number of points compared to TP. Strava's differing TP structure made it difficult to directly compare with our ground-truth. Instead we used a two-sample Kolmogorov-Smirnov test to compare the distributions of the two datasets - these showed a significant difference ($p < 0.01$).

Finally, out of the six data controllers, only one respected the Article 20(2) request to port the data to another data controller. Unfortunately the latter data controller, Nike Run Club, interpreted the incoming transmission as an Article 17 (erasure) request and asked for confirmation that the data subject wished to have all of their data erased. This reiterates the findings of others [19, 20] that Article 20(2) is not widely understood.

5 CONCLUSION AND FUTURE WORK

In this paper we have presented an initial exploration into the use of subject access rights for auditing systems. We have shown that data portability, when combined with local collection of ground-truth data, can act as an instrument for understanding the accuracy of the data held by data controllers. We find that we were able to highlight inaccuracies in the data held by data controllers, but also that insufficient descriptions and metadata provided by data controllers might impede our ability to audit. In one case, the complete lack of data returned by a controller made it impossible to audit. Differing formats, and differing ways in which the data were interpreted by data controllers, also raise challenges to audit.

This study raises many additional possible avenues for future work. One question is whether other rights could also help to audit. Article 15 (the right to access) is more expansive than Article 20, but also introduces potential problems in that requests do not have to be fulfilled in a machine-readable format. For simplicity, therefore, we focused on Article 20 in this study. Article 16 (the right to rectification) or Article 17 (the right to erasure) could be used to correct inaccurate data, such as those that we found for user profiles. One could imagine an ongoing monitoring loop of Article 20 requests to check data followed by Article 16 to correct data, although the GDPR also places limitations on the number of requests that can be made (Article 12(5)'s conditions for "manifestly unfounded or excessive" requests). Even without these restrictions, the 32-day response time by one controller would make such a loop challenging.

We have examined a portability request from a single data subject. This only provides one viewpoint into the system being audited. We aim to explore tools for crowdsourcing and aggregating these individual self-audits, thus extending from a single citizen's audit to a societal audit. Note that "it is important to acknowledge that the rights under Regulation (EU) 2016/679 can only be exercised by each individual and cannot be conferred or delegated to a data cooperative" [8], so we envisage such crowdsourcing as being made up of individual requests carried out by individual data subjects, which could then later be aggregated. This could build on Mahieu and Ausloos' [15] concept of the ecology of transparency to create collective empowerment, thus re-balancing the power between data subjects and controllers.

While this initial study is small, it does indicate some limitations to this approach. We were forced to use three different methods of uploading data to our chosen service providers, which potentially creates inconsistency. This was due to technical constraints: Amazfit, Strava and Komoot did not allow us to import raw GPX files from Gadgetbridge. To upload data to Strava we had to use the Amazfit integration, and this may have led to preprocessing by Amazfit before transmission (e.g. rounding). Exploring solutions to overcoming these limitations, by collecting more data in this sector (fitness) but also exploring other sectors, is another challenge for future work.

Finally, going beyond the GDPR, there are indications that the right to data portability may become more widely applied, e.g. for interoperability [6]. Gatekeepers will have obligations to "provide effective portability of data generated through the activity of a business user or end user" in the forthcoming EU Digital Markets Act [7]. So our methods may be able to be applied elsewhere.

REFERENCES

[1] Emma Arfelt, David Basin, and Søren Debois. 2019. Monitoring the GDPR. In *Computer Security – ESORICS 2019*, Kazue Sako, Steve Schneider, and Peter Y. A. Ryan (Eds.). Springer International Publishing, Cham, 681–699. https://doi.org/10.1007/978-3-030-29959-0_33

[2] Article 29 Data Protection Working Party. 2017. Guidelines on the right to data portability. https://ec.europa.eu/newsroom/article29/items/611233/en

[3] Jef Ausloos and Michael Veale. 2021. Researching with Data Rights. *Technology and Regulation* (Jan. 2021), 136–157. https://techreg.org/index.php/techreg/article/view/61

[4] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 74 (April 2021), 34 pages. https://doi.org/10.1145/3449148

[5] Marlene Barth. 2021. A Case Study on Data Portability. *Datenschutz und Datensicherheit — DuD* 45, 3 (2021), 190–197. https://doi.org/10.1007/s11623-021-1416-3

[6] Ian Brown. 2020. *Interoperability as a tool for competition regulation.* https://doi.org/10.31228/osf.io/fbvxd

[7] European Commission. 2020. Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:842:FIN

[8] European Commission. 2020. Proposal for a Regulation of the European Parliament and of the Council on on European data governance (Data Governance Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

[9] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119 (4 May 2016), 1–88. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

[10] Grace Fox and Regina Connolly. 2018. Mobile health technology adoption across generations: Narrowing the digital divide. *Information Systems Journal* 28, 6 (2018), 995–1019. https://doi.org/10.1111/isj.12179

[11] Gadgetbridge [n.d.]. Gadgetbridge for Android. https://gadgetbridge.org/

[12] Google. 2020. Google Play Store. https://play.google.com/store

[13] Wenlong Li. 2018. A tale of two rights: exploring the potential conflict between right to data portability and right to be forgotten under the General Data Protection Regulation. *International Data Privacy Law* 8, 4 (Nov. 2018), 309–317. https://doi.org/10.1093/idpl/ipy007

[14] Deborah Lupton. 2016. The diverse domains of quantified selves: self-tracking modes and dataveillance. *Economy and Society* 45, 1 (Jan. 2016), 101–122. https://doi.org/10.1080/03085147.2016.1143726

[15] René L. P. Mahieu and Jef Ausloos. 2020. Harnessing the collective potential of GDPR access rights: towards an ecology of transparency. https://policyreview.info/articles/news/harnessing-collective-potential-gdpr-access-rights-towards-ecology-transparency/1487

[16] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker

Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[17] Andreas Schreiber. 2016. A Provenance Model for Quantified Self Data. In *Lecture Notes in Computer Science*. Springer International Publishing, 382–393. https://doi.org/10.1007/978-3-319-40250-5_37

[18] Melanie Swan. 2013. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1, 2 (June 2013), 85–99. https://doi.org/10.1089/big.2012.0002

[19] Emmanuel Syrmoudis, Stefan Mager, Sophie Kuebler-Wachendorff, Paul Pizzinini, Jens Grossklags, and Johann Kranz. 2021. Data Portability between Online Services: An Empirical Analysis on the Effectiveness of GDPR Article 20. *Proceedings on Privacy Enhancing Technologies* 2021, 3 (2021), 351–372. https://doi.org/10.2478/popets-2021-0051

[20] Sarah Turner, July Galindo Quintero, Simon Turner, Jessica Lis, and Leonie Maria Tanczer. 2020. The exercisability of the right to data portability in the emerging Internet of Things (IoT) environment. *New Media & Society* (July 2020), 146144482093403. https://doi.org/10.1177/1461444820934033

[21] Janis Wong and Tristan Henderson. 2019. The Right to Data Portability in Practice: Exploring the Implications of the Technologically Neutral GDPR. *International Data Privacy Law* 9, 3 (Aug. 2019), 173–191. https://doi.org/10.1093/idpl/ipz008