Digital Preservation and Curation of Self-Tracking Data: A Position Paper

Frank Hopfgartner[1] and Joy Davidson[2]

[1] University of Sheffield - Sheffield S1 4DP, United Kingdom
f.hopfgartner@sheffield.ac.uk
[2] University of Glasgow - Glasgow G12 8QQ, United Kingdom
joy.davidson@glasgow.ac.uk

Abstract. Thanks to recent advances in the field of ubiquitous computing, an increasing number of users now rely on tools and apps that allow them to track specific aspects of their lives. An example are step counters and activity trackers that are promoted as unobtrusive tools to monitor our fitness levels. Interestingly, although significant research and development efforts went into improving the accuracy of these *self-tracking* devices, hardly any research is performed on the digital preservation of the data created. This position paper highlights challenges and opportunities arising from the digital preservation of self-tracking data.

1 Introduction

Increasingly, people rely on the services of Wearable devices and smartphone apps that unobtrusively capture various aspects of their lives. Popular examples include apps that keep track of users' steps, record running distances and estimate energy expenditures [12]. Informative statistics on users' performances and illustrative visualisations such as plotting running routes on a map provide valuable information that allow users' to better understand their own activities. In addition, features such as the ability to share these records via social media or app-specific features such as the use of leaderboards, points and badges are designed to keep users engaged with these apps [23]. Although studies have shown that users of self-tracking devices often give up on them after a few months already [15], some users rely on these services significantly longer. A few users (e.g., [6]) have even started to capture their life activities for years or even decades, creating a detailed lifelog consisting of automatically created digital records.

This paper elaborates on the hypothesis that self-tracking data of both short-term and long-term users are of high societal value since they depict snapshots of users' daily activities in the 21st century [16]. An illustrative example is given by the technology and wearable products company Jawbone who visualised averaged sleeping patterns of users of their activity trackers from different cities in the world [5]. Their sleep tracking data shows, for example, that customers in Tokyo (on average) go to bed later than the average customer of New York. Similarly, data of users of the running app Strava allows us to inspect popular running routes on a global scale, hence providing insights in peoples' recreation activities and their preferred locations [24].

While above examples showcase the value of aggregated data representing activities of a larger share of the overall population, individual data can evenly be as valuable to depict live in our age. Looking at this from a historic perspective, historians and museum curators crave for personal and detailed data that can help them to better reflect on or represent people's lives. For example, personal correspondence of the rich and famous (e.g., [7]) is often used as primary source to describe certain time periods. This does not help in drawing the full picture though as the lives of "the common people" is evenly as important if one wants to provide a good picture of the past. Setting rare exceptions (e.g., [11,19]) aside, their records are often not available as they not were created or did not survive the pass of time.

Self-tracking provides an opportunity to tackle this situation and archives and museums are well advised to invest in the digital preservation and curation of self-tracking data. Digital curation describes the process of storing, managing, protecting and sharing digital resources to "keep [them] authentic and re-usable for future users" [22]. In other words, it stands for the management and appraisal of digital data throughout its entire lifecycle. Although museums and archives are well aware of the challenges related to above steps, research to date on tackling these challenges is very limited. By using a basic digital curation lifecycle as a template, this position paper aims to contribute to the discussions by highlighting what these actions mean in the context of digital preservation of self-tracking data. The paper is structured as follows. In Section 2, we first survey adaptations of the digital curation lifecycle for different types of data. Section 3 then discusses digital curation actions in the context of self-tracking data. Section 4 concludes this position paper.

2 Related Work

As mentioned above, digital curation refers to the process of managing, storing, and preserving digital data for later use. Treating data as a digital entity that goes through various stages or cycles in its ”life”, Pannock [22] argues that these cycles need to be carefully planned in order to guarantee a feasible digital curation policy. This follows the argumentation of Humphrey [13] who describes lifecycle models as an ideal method to represent flow, representation, and transition of system components. Pennock puts forward three main arguments to support her statement. First of all, data in digital form is rather fragile and technical advances might result in issues related to access to this data. Moreover, she argues that activities (or lack thereof) can directly influence the digital curation of data. Thirdly, she highlights that re-use of curated data is only possible if the data's authenticity and integrity is guaranteed.

Various digital curation models have been introduced that follow the idea of data lifecycles. An early example is the DCC Digital Curation Lifecycle [8] that focuses on research data. The model distinguishes between *full lifecycle actions*, *sequential actions* and *occasional actions* as its key elements. Full lifecycle actions include the description and representation of information, preservation planning, community watch & participation, and curation & preservation. Sequential actions include conceptualisation, creation or receiving of data, appraisal and selection, ingestion, preservation action, storage, providing access, use and reuse, and transformation. Occasional actions include activities such as disposal of data, reappraise, and data migration. A similar lifecycle model have been introduced by the US Library of Congress [17]. More recently, the UK Data Service introduced a more general lifecycle model [26] that can be applied to a wide range of different data types. Kowalczyk [14] identifies this as a limitation as some features unique to specific data might require a much more fine-grained approach to guarantee digital curation.

Considering this, it comes with no surprise that various domain-specific digital curation lifecycles have been introduced throughout the years. More recent examples include work by Emsley and De Roure [9] on the digital curation of Docker containers and by Yoon et al. [29] who focus on citizen-generated data. Probably the most relevant model in our context is introduced by Wallis et al. [27] who argue for the digital curation of ecological sensing data. Although sensing data might share similarities to self-tracking data that has been created using sensor platforms, we argue that there are specific differences and challenges in other steps of this data's lifecycle. A first discussion on these challenges is provided in the next section.

3 Towards a Digital Curation Lifecycle for Self Tracking Data

Pennock [22] argues that a basic digital curation lifecycle consists of six major actions that are required for the curation and preservation of data. These actions include creation of data, active use, appraisal & selection, transfer, storage & preservation, and access & re-use. In the remainder of this section, we look at these actions more in detail.

3.1 Creation of data

The first action of the lifecycle is the actual creation of data. A wide range of different self-tracking devices exist that create a manifold of different data [28]. An important step of data creation is also the additional creation of administrative, structural and technical metadata that describes this data as it can help us in better understanding its meaning. In a self-tracking scenario, this step could be rather challenging.

Although more and more self-tracking apps and devices are made available, the nature of accompanying metadata created is unknown.

3.2 Active Use

There can be different reasons of why users decide to rely on self-tracking devices. Lupton [18] identifies five main reasons, or modes, for self-tracking, including personal, communal, pushed, imposed and exploited. Although the initial motivation of these groups differs significantly, in all cases, self-tracking data is used to quantify aspects of the self-tracker's life, which eventually might lead to behaviour change [21].

Another case for the active use of self-tracking data is presented by Musakwa and Selala who analysed aggregated data of cycling records in the city of Johannesburg to further study cycling trends [20]. While their particular showcases benefits for city planners, transportation managers, and other stakeholders, similar analyses can be thought of in other contexts that would allow us to represent various aspects of life in the 21st century.

3.3 Appraisal & Selection

This action includes the evaluation of data and selection for long-term curation and preservation. Best practice guidelines (e.g., [4]) generally suggest to preserve raw data (and accompanying metadata) for future use. In a self-tracking context, such raw data would be sensor data, e.g., recorded by accelerometers. Raw data is only accessible to the tracking service provider though since users often only get to see the output of an additional data analysis step. For example, step counter apps do not visualise the actual raw that was captured by accelerometers or other sensors but instead interpret this data using undisclosed algorithms. Given the unregulated nature of the self-tracking market with a multitude of apps and devices made available, it is not clear what appraisal and selection actions are performed by the self-tracking service providers.

One approach to guarantee appraisal and selection could be to hand this action to the consumer or other trusted parties. However, considering that the success of self-tracking apps depends on the accuracy of their algorithms and that the release of raw data would open the gates for reverse engineering efforts, it is unlikely that raw data will be made available to the consumer. Consequently, pre-processed data might be the only alternative type of data available that could be considered in the appraisal and selection step.

3.4 Transfer

The transfer action, also referred to as ingestion, refers to the transfer of data to an archive, repository or data centre. Currently, self-tracking data is often transferred from the users' smartphone or Wearable device to the cloud servers of the self-tracking service provider. Here, they are then analysed further. Additional data transfer actions remains unclear but as shown in [2], self-tracking data is often used for further research, which would suggest that the data is transferred further. In case the digital preservation and curation of self-tracking data would be performed by the costumer or the general public, data would have to be transferred to their hands or to the hands of a trusted custodian. However, as will be discussed below, users might have transferred ownership of the data to the service provider by accepting their Terms and Conditions, and it is up to the service provider to decide whether they want to provide access to this data. When the popular tracking app Moves stopped their service, customers had a short time window during which they could download their (processed) data [10]. Although this example illustrate that there might be good will on the service providers' side, the implementation is not always satisfactory. In the case of the Moves app, giving short notice to customers resulted in complete loss of data for customers who had missed the time window.

3.5 Storage & Preservation

Storage refers to the need to secure data in a secure manner adhering to relevant standards. As discussed above, most self-tracking data is currently stored by the providers of self-tracking services. The state of data storage and use of standards is unknown. Considering that self-tracking data contains very personal and potentially even sensitive data, a special emphasis needs to be put on the security of the data. The recent introduction of the strict General Data Protection Regulations (GDPR) by the European Union which focuses on the protection of personal data highlights the importance of this step even further. A discussion on the impact of GDPR on businesses is provided in [25].

Data preservation includes all steps required to ensure long-term preservation and retention of the data. This means that data remains authentic and reliable. Here, it also remains unclear how self-tracking data is currently treated. An issue related to this is that self-tracking services often cease to exist, e.g., when the service is not profitable. The destiny of the data collected remains unclear. This highlights the need to allow transfer of data to the customer or trusted parties who have the resources and credibility to guarantee secure and long-term data storage.

3.6 Access & Re-use

The access and re-use step aims to make sure that stakeholders involved can easily access the data on a day-to-day basis. Here, an important question to be asked is who are the stakeholders involved in self-tracking. One obvious stakeholder is the self-tracker who created all the data but it might not actually be him or her who owns this data. Dependent on the terms & conditions that users signed up for when using a specific self-tracking app or device, another stakeholder, and owner of the data, might be the provider of the app. For a further discussion on data ownership, the reader is referred to [3]. At the same time, with self-tracking data being used in legal cases, other stakeholders might emerge who might get the legal right to access self-tracking data. For example, the Lancaster county district attorney in Florida stated that “when we have technology like Fitbit we’re going to take advantage of it” [1].

4 Conclusion

Self-tracking devices are increasingly being used to quantify various aspects of our lives. While the majority of users might rely on self-tracking services to better understand their current lifestyles, we argue that the data created is worthy of preservation of future use. Using a very basic digital curation lifecycle as template, this position paper highlights a few of the core challenges that emerge from the digital preservation of self-tracking data. Learning from more domain-specific digital curation models that have been introduced in this paper, future work includes outlining more specific guidelines for the digital preservation of self-tracking data.

References

1. Alejando Alba. Police, attorneys are using fitness trackers as court evidence, 2016.
2. Stephen Armstrong. What happens to data gathered by health and wellness apps? *British Medical Journal*, 353, 2016.
3. Ann Babe. Wearable fitness devices: Who owns your data?, 2016. Accessed in July 2018.
4. Elizabeth T. Borer, Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. Some simple guidelines for effective data management. *The Bulletin of the Ecological Society of America*, 90:205–214, 2009.
5. Brian Wilt. In the city that we love. the jawbone blog, 2014. Accessed in July 2015.
6. Niamh Caprani, Paulina Piasek, Cathal Gurrin, Noel E. O’Connor, Kate Irving, and Alan Smeaton. Life-long collections: Motivations and the implications for lifelogging with mobile devices. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 6(1):15–36, 2014.
7. Derek J. de Solla Price. The correspondence of isaac newton. vol. 1. 1661-1675. h. w. turnbull, ed. cambridge university press, new york, 1959. xxxviii + 468 pp. illus. + plates. *Science*, 131(3408):1202–1202, 1960.

8. Digital Curation Centre. Dcc digital curation lifecycle, 2008. Accessed in July 2018.
9. Iain Emsley and David de Roure. A framework for the preservation of docker containers. *International Journal of Digital Curation*, 12(2):125–135, 2017.
10. Facebook. Hello. tbh, we're moving on, 2018. Accessed in July 2018.
11. David A. Gerber. *Authors of Their Lives: The Personal Correspondence of British Immigrants to North America in the Nineteenth Century*. New York University Press, 2006.
12. John P. Higgins. Smartphone applications for patients' health and fitness. *The American Journal of Medicine*, 129(1):11–19, 2016.
13. Chuck Humphrey. e-Science and the life cycle of research. Technical report, University of Alberta, 2006.
14. Stacy T. Kowalczyk. Modelling the digital research data lifecycle. *International Journal of Digital Curation*, 12(2):331–361, 2017.
15. Dan Ledger and Daniel McCaffrey. Inside wearables. how the science of human behavior change. Technical report, Endevour Partners, 2014.
16. Na Li and Frank Hopfgartner. To log or not to log? swot analysis of self-tracking. In *Lifelogging*, pages 305–325. Springer Verlag, 2015.
17. Library of Congress. Preserving our digital heritage: The national digital information infrastructure and preservation program 2010 report. Technical report, Library of Congress, 2011.
18. Deborah Lupton. Self-tracking modes: Reflexive self-monitoring and data practices. In *Proceedings of the Imminent Citizenships: Personhood and Identity Politics in the Informatic Age workshop*, 2014.
19. Martyn Lyons. French soldiers and their correspondence: Towards a history of writing practices in the first world war. *French History*, 17(1):79–95, 2003.
20. Walter Musakwa and Kadibetso M. Selala. Mapping cycling patterns and trends using strava metro data in the city of johannesburg, south africa. *Data in Brief*, 9:898–905, 2016.
21. Mitesh S. Patel, David A. Asch, and Kevin G. Volpp. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, 313(5):459–460, 2015.
22. Maureen Pennock. Digital curation: A lifecycle approach to managing and preserving usable digital information. *Library & Archives Journal*, 1, 2007.
23. Amon Rapp. Gamification for self-tracking: From world of warcraft to the design of personal informatics systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 80:1–80:15, New York, NY, USA, 2018. ACM.
24. Strava. The global heatmap, 2017. Accessed in July 2018.
25. Colin Tankard. What GDPR means for businesses. *Network Security*, 2016(6):5–8, 2016.
26. UK Data Service. Research data lifecycle, 2018. Accessed in July 2018.
27. Julian C. Wallis, Christine L. Borgman, Matthew S. Mayernik, and Alberto Pepe. Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative research field. *International Journal of Digital Curation*, 3(1):114–126, 2008.
28. Joshua H. West, P. Cougar Hall, Carl L. Hanson, Michael D. Barnes, Christophe Giraud-Carrier, and James Barrett. There?s an app for that: Content analysis of paid health and fitness apps. *J Med Internet Res.*, 14(3), 2012.
29. Ayoung Yoon, Lydia Spotts, and Andrea Copeland. Data curation for a community science project: CHIME pilot study. *International Journal of Digital Curation*, 12(2):220–231, 2017.