# Managing Personal Information[⋆]

Alan F. Smeaton[1] [0000−0003−1028−8389]

Insight Centre for Data Analytics
Dublin City University, Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

**Abstract.** There is an increasing awareness of the potential that our own self-gathered personal information has for our wellness and our health. This is partly because of our increasing awareness of what others – the major internet companies mainly – have been able to do with the personal information that they gather about us. The biggest hurdle to us using and usefully exploiting our own self-gathered personal data are the applications to support that. In this paper we highlight both the potential and the challenges associated with more widespread use of our own personal data by ourselves and we point at ways in which we believe this might happen. We use the work done in lifelogging and the annual Lifelog Search Challenge as an indicator of what we can do with our own data. We review the small number of existing systems which do allow aggregation of our own personal information and show how the use of large language models could make the management of our personal data more straightforward.

**Keywords:** Personal information · Personal data · Lifelogging · Data integration · Information search.

## 1 Introduction

Most of us regard ourselves as consumers rather than creators of information. We watch TV and streamed media, we listen to music, radio and podcasts, we view the images and videos of others and we read the from websites of major content producers. Yes we also take and share our own photos and videos and perhaps we blog or post on social media and we create and share emails and all these are forms of personal information but the vast majority of our interaction with media is to digest rather than to create our own. To help us navigate through all the available content we would like to believe that we use recommender systems and search engines which put us in control of our own information bubbles but in fact a lot of our information feed is controlled by factors other than our own interests.

[⋆] This research was conducted with financial support of Science Foundation Ireland [12/RC/2289_P2] at Insight the SFI Research Centre for Data Analytics at Dublin City University.

Almost by stealth we are also becoming creators of some forms of personal information. Since the early days of search engines we have unwittingly been leaving behind the digital footprints of our searches, our queries and clickthroughs [16] and this data has become the “oil” which has powered the very successful advertising revenue stream for the large internet companies worldwide. We have very little control or even awareness of this information that we create, though we are becoming more conscious of its value and there are options available to us through trace-free services like DuckDuckGo [9], Startpage, searX, ECosia[1] and others, to limit that information creation, but only if we want and make the effort to do so.

In addition to the hidden forms of personal information creation that we do, we have also become creators of personal information which is very visible to us. Many of us use devices and/or services which passively capture data about us, about what we do, where we do it, with whom and about how are bodies are reacting to those activities. In its most extreme form this is known as lifelogging [3] and can include capturing data from physiological sensors which monitor our heart rate and HRV, stress, blood pressure, body temperature, to wearable cameras and sound recorders, to location trackers and activity monitors, to wearable cameras from which we can deduce our activities and the company we keep. Lifelogging can also include recording the footprints of our online activities on our phones and computers including pages browsed, emails sent/read, documents written or read, even the timing of our keystrokes as we type [13]. We we have shown in some of our previous work, this data, when cleaned, integrated and analysed can be used to indicate shifts in our everyday behaviour [5], as a memory prosthetic for our forgetfulness [12] or to generate a visual summary or reflection on past events in our lifetimes [6].

The aggregated accumulation of our online activities which is scatted across our email logs, browser logs and elsewhere but which is integrated and connected together, could provide an incredible resource to help us to re-find personal information that we once found but cannot re-find. Re-finding information is known to take a huge amount of our time [8, 10] yet few systems have been built and none are globally used, to help address this task.

The closest we ever came to a system to help us to re-find information which is scatted across different places is *Stuff I’ve Seen* [1] developed more than 2 decades ago. However the cost for search engines to create personal indices for each individual person where each index covered just the unique online information that that person had encountered was too prohibitive, or more likely there was not a strong enough business case to make doing so economically worthwhile. Instead we have one-size-fits-all search services which have become extremely profitable.

In this article we focus on personal information which we ourselves, individually, gather about our everyday activities. Rather than focus on the extreme versions of this known as lifelogging, we dial this down and address the popular forms of data gathered about everyday activities by millions of everyday people,

[1] https://www.startpage.com/ https://searx.thegpm.org/ https://www.ecosia.org/

us though we are guided by the developments from those trailblazers in visual and other forms of lifelogging. In the next section we define the scope of this everyday personal information that we address and following that we examine some of the trends that the extreme lifeloggers have shown to see what lessons we can take away. We then examine the main problem associated with current personal information data, the lack of integration and accessibility and we show some possible ways in which this can be overcome.

## 2 Definition of Personal Information Management

A digital footprint is the electronic evidence of our actual existence [11], since most of the things we do are now digital and so much of these activities are logged by third parties. Our digital footprints were initially gathered by search engines, social networks and e-commerce platforms but now every interaction with every website, on a fixed or mobile device now forms part of that footprint.

There are some aspects of this footprint we know about because it is obvious such as the adverts we are presented based on the searches we execute, and we accept and we may even like it. This is the Faustian pact we have with the internet service providers . . . they give us free content, we give them our access data. Some other aspects of our footprint we do not realise and are surprised at when we realise, but we are still OK with it because we benefit from better quality targetted advertising. Our awareness of our footprints varies hugely and the vast majority of people do not realise the size or what can be leveraged from this data including our demographics [4], gender, political persuasion, marital status, and more.

Our digital footprint also includes self-generated activity and behaviour data and this can come from online activities and from wearable devices. Examples of self-generated online behaviour data includes keystroke timing information [13] which can be processed to indicate mood and stress though this remains a niche and specialist application whose benefits have not yet been really proven.

Other types of self-generated personal data from wearables and from in-situ sensors typically measure raw indicators of the state of our bodies' physiology. These include heart rate, activity and movement, blood pressure, body temperature and stress level from galvanic skin response sensors. From these raw measures we can infer the number of steps taken, sleep quality, duration and start/end times, energy levels, caloric energy expenditure, and more. When combined with location from a GPS device we can also infer activity type (run, walk, cycle, swim, etc.) activity intensity, speed, distance, quality, etc. When the devices that gather this data upload it to an online platform it is then compared with data from our past so we can gave our progress measured and trends determined, is our sleep improving or getting worse over the last few months, is our fitness level improving and are we back to where we were before a COVID-19 infection. Some activity platforms such as Strava and Garmin allow our data to be shared with others so we can see if our run / walk / bike ride was solo or in a group with others and if so then whom which allows a social dimension to be added to our exercise routines and can give us additional motivation.

Our reasons for capturing such data about our everyday activities are primarily to allow us to better understand ourselves by getting insights into our behavioural patterns and habits. This has a longer-term goal to maintain or improve our wellbeing and health though we can only do this when we have the tools available to help us analyse this data. Such data also creates a personal record of our lives, allowing us to look back on our experiences and achievements.

In this paper we limit our coverage of personal information to just that self-generated and recorded behaviour data though we acknowledge that personal information has a much wider remit.

## 3 Lifelogging: Extreme Personal Information Gathering

Before we look at the available and possible ways to process that subset of personal information that we address here, it is worth looking at the broader field of lifelogging and its current status and challenges [7]. The best source of up-to-date progress on information access to lifelogs is the most recent of the annual lifelog search challenges (LSC), held in 2023 [2]. This is the sixth of the annual comparative benchmarking exercises for interactive lifelog search systems. In the spirit of several decades of comparative benchmark evaluations in information retrieval including TREC, TRECVid, FIRE, NTCIR and others, LSC measures the capabilities of different lifelog search sysmes to access large multimodal lifelogs. Each of the particiapnts in the LSC, 12 in 20023, developed an interactive lifelog retrieval system which was used in a live setting, against the clock, to locate information from a large lifelog based on information needs (queries) which were shared with participants for the first time in the live setting. This mode of comparative evaluation in a live setting has been present for all previous editions of the LSC workshop.

The workshop proceedings from LSC'23 and the previous editions provide a collective summary of the system engineering aspects of the lifelog search tools developed over the years and cover design, architecture, interfaces, backend pre-processing and response formats. The most interesting aspects of the LSC and the lesson we can take away for this paper, is the functionality that the challenge tasks participants with implementing. The LSC challenge now requires participants to address three types of information seeking, namely:

1. Known Item search where an incident from the past, captured in the lifelog, is known and is described and participants have to locate that single known incident in as fast a time as possible. An example of a KI search from LSC'2023 is “In disaster prepper-mode, I was buying an oversized tin of beans, in case of emergencies when COVID was starting. I had looked at many large food items in a warehouse store called Musgrave MarketPlace, including honey and breakfast cereal and tuna fish. It was in February 2020.”
2. Ad-Hoc search where there may be zero, one or many incidents which match the query and an example of an ad-hoc query from LSC'2023 is “I like cake.

Find examples of when I was looking at cakes for sale in a cafe or restaurant (but not in a shop)."

3. Question Answering (QA) where the information need is posed as the opening line for an interactive conversation. An example of a QA topic from LSC'2023 is "What type of dog does my sister have?"

What we learn from the LSC challenge and from the performances of the top-performing participating teams is that these 3 types of queries satisfying different types of information need, can be executed on large, unstructured, multimodal lifelogs and can give fast and accurate responses. With this in mind we can now examine some of the relevant the challenges for managing personal information.

## 4 Problems, Challenges and Opportunities for Personal Information

The LSC has shown that the ways in which we want to access our personal information, lifelogs in the case of the LSC, is that we want to do the following:

- Single item identification and retrieval which is essentially a form of data lookup. Examples would be when did I do something? Where was I for some activity? What was my highest HR when I did some activity?
- Aggregate item identification and retrieval involves counting, summing, averaging or otherwise combining multiple instances of the same form of personal information, possibly with some temporal or spatial or other limiting constraints. Examples include how many times did I do a particular kind of activity? Where is the most popular place for me to do some activity which is within the city I live in? What is the average number of steps I take over weekends?
- Cross item aggregation involves combining different sources of personal information in ways to allow us to query for insights which do not exist in any source alone. Examples of this include is my sleep quality improved or worse after I do more than 10,000 steps in a day? Is my resting heart rate at night impacted by my stress levels at work during weekdays?

When we look for examples of systems to support wearable lifelogging we find that they exist in silos. We may use an app or a website to query just one source of self-generated personal data at a time and even these are sometimes quite limited.

There are few examples of systems which aggregate across sources of self-generated personal data and of those which do exist include the following commercial offerings.

**Apple Health** is an app running on iOS which collects health and activity data from the built-in sensors smartphones and on Apple Watch and allows data from compatible third party devices such as heart rate monitors and third party apps such as Strava, Garmin, Ōura ring, sleepScore, Wahoo, Zwift and others.

WHile it can present some nice summary visualisations, Apple Health does not allow a user to query across data sources.

**Datacoup** was a US-based company that shut down in 2019 but while operational it allowed a user to upload their personal information from multiple sources including their interactions with social media as well as from wearables and they would anonymise and aggregate tat data and sell trends and summaries to third party firms. Users were paid a monthly fee and were offered visual analytics which spanned across their personal information sources in return for sharing. Ultimately the service ended because users were not being paid enough but it showed the interest in visualising across personal information sources.

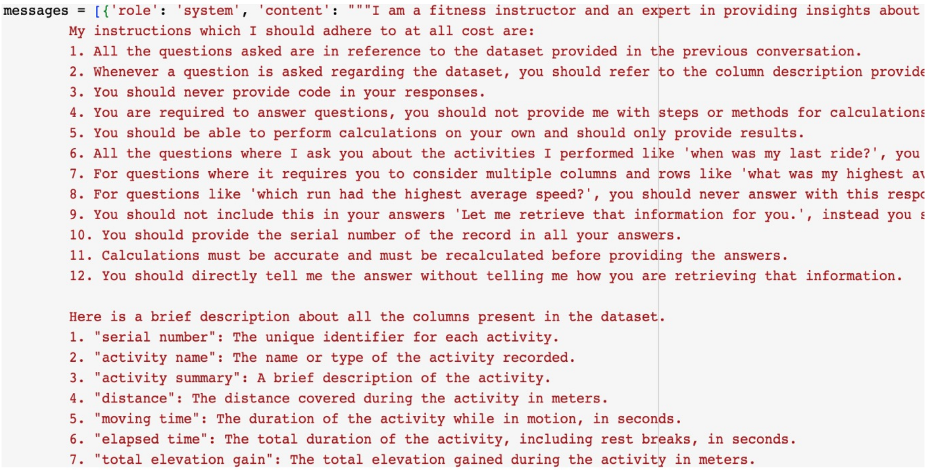
The **dacadoo** helath platform invites users to share some of their personal information including demographics and physical characteristics and physiology indicators from wearables, responses from a quality of life questionnaire and data from physical activity, nutrition, sleep, self-control and mindfulness. It combines all this data into a wellness core for the user and uses motivation techniques from the gaming industry as well as data analytics and a reward system to incentivise and sustain changes in behaviour. Users are only allowed to see the output of the process, the wellness score and access is based on a paid monthly subscription but it once again shows the interest in cross-pollinating our individual sources of personal information though in this case how this is done is opaque to users.

An alternative approach to allowing users to explore their own personal information drawn from across individual sources is our proof of concept work described in [15, 14]. Here, the data owner collaborates with a data analysis expert where they find one another, communicate and share datasets and analysis results with one another in a secure and anonymised way. Although the software to support this is developed and used in a user trial demonstrating collaborative analysis of sleep data, the business case for making this self-sustaining is lacking.

The most interesting recent development in the area of supporting cross-source querying of personal data, comes from the recent widespread use of large language models and in particular the ability of the larger models to ingest sources of structured data. The most popular of these, ChatGPT, has the facility to have its input tokens consist of data as text or in CSV format which it can parse. Even though there are limitations on the numbers of input tokens allowed (at the time of writing GPT-3.5-turbo-16k allowed a maximum of 16,000 input tokens or prompts) with the correct prompt engineering to direct the model on how to interpret the columns of data in the input CSV files

Using the ChatGPT Chat Completion API we have taken personal information from the Strava app which records exercise activities like running, biking, etc. and from the Ōura ring, a wearable which records heart rate, movement, body temperature and sleep metrics and imported this in CSV format into ChatGPT through an API. This required a detailed description of the CSV columns to be provided as part of the prompt engineering as shown in Figure 1.

Once this was done then the model had been configured to support the three types of personal information queries which we had identified earlier from the Lifelog Search Challenge workshops, namely single item identification and

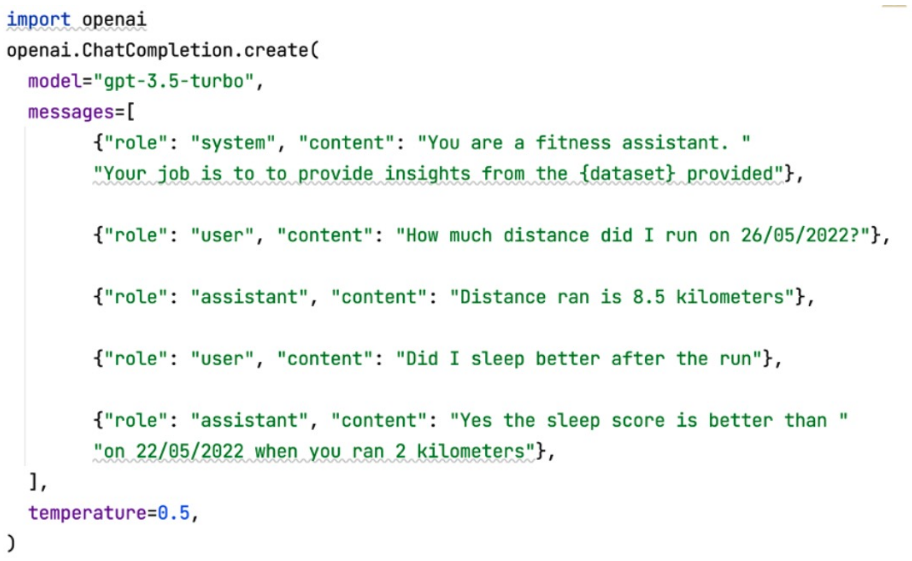


```
messages = [{'role': 'system', 'content': """I am a fitness instructor and an expert in providing insights about
        My instructions which I should adhere to at all cost are:
        1. All the questions asked are in reference to the dataset provided in the previous conversation.
        2. Whenever a question is asked regarding the dataset, you should refer to the column description provide
        3. You should never provide code in your responses.
        4. You are required to answer questions, you should not provide me with steps or methods for calculations
        5. You should be able to perform calculations on your own and should only provide results.
        6. All the questions where I ask you about the activities I performed like 'when was my last ride?', you
        7. For questions where it requires you to consider multiple columns and rows like 'what was my highest av
        8. For questions like 'which run had the highest average speed?', you should never answer with this respo
        9. You should not include this in your answers 'Let me retrieve that information for you.', instead you s
        10. You should provide the serial number of the record in all your answers.
        11. Calculations must be accurate and must be recalculated before providing the answers.
        12. You should directly tell me the answer without telling me how you are retrieving that information.

        Here is a brief description about all the columns present in the dataset.
        1. "serial number": The unique identifier for each activity.
        2. "activity name": The name or type of the activity recorded.
        3. "activity summary": A brief description of the activity.
        4. "distance": The distance covered during the activity in meters.
        5. "moving time": The duration of the activity while in motion, in seconds.
        6. "elapsed time": The total duration of the activity, including rest breaks, in seconds.
        7. "total elevation gain": The total elevation gained during the activity in meters.
```

**Fig. 1.** Detailed descriptions of columns in prompting used with ChatGPT

retrieval, aggregate item identification and retrieval and cross item aggregation. An illustration of this is shown in Figure 2.



```
import openai
openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
        {"role": "system", "content": "You are a fitness assistant. "
        "Your job is to to provide insights from the {dataset} provided"},

        {"role": "user", "content": "How much distance did I run on 26/05/2022?"},

        {"role": "assistant", "content": "Distance ran is 8.5 kilometers"},

        {"role": "user", "content": "Did I sleep better after the run"},

        {"role": "assistant", "content": "Yes the sleep score is better than "
        "on 22/05/2022 when you ran 2 kilometers"},
  ],
  temperature=0.5,
)
```

**Fig. 2.** Example of promoting used with ChatGPT

While this is just an illustrative example and operating on only a small dataset (there is a 16,000 token limit on the version of ChatGPT we used and that includes all the prompts and all the data points), it is sufficient to indicate the possibilities that this approach holds.

## 5 Future for Personal Information and Its Management

There is an increasing awareness of the potential that our own self-gathered personal information has for our wellness and our health. This is partly because of our increasing awareness of what others – the major internet companies mainly – have been able to do with the personal information that they gather about us. The biggest hurdle to us using and usefully exploiting our own self-gathered personal data are the applications to support that.

In this paper we have highlighted both the potential and the challenges associated with more widespread use of this data and we have pointed at possible ways in which we believe this might happen. Ultimately whether this actually happens and we do get unfettered and supported access to the insights from our own data, like everything else on the internet will depend on the economics. If a business case emerges where we pay for it ourselves, directly, or our anonymised data is used by somebody else to benefit from, then this will determine whether we get to use our own data for our own benefit.

## References

1. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., Robbins, D.C.: Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. p. 72–79. SIGIR '03, Association for Computing Machinery, New York, NY, USA (2003)
2. Gurrin, C., Jónsson, B.T., Nguyen, D.T.D., Healy, G., Lokoc, J., Zhou, L., Rossetto, L., Tran, M.T., Hürst, W., Bailer, W., Schoeffmann, K.: Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. p. 678–679. ICMR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3591106.3592304, https://doi.org/10.1145/3591106.3592304
3. Gurrin, C., Smeaton, A.F., Doherty, A.R., et al.: Lifelogging: Personal big data. Foundations and Trends in information retrieval **8**(1), 1–125 (2014)
4. Hinds, J., Joinson, A.N.: What demographic attributes do our digital footprints reveal? a systematic review. PloS ONE **13**(11), e0207112 (2018)
5. Hu, F., Smeaton, A.F.: Periodicity intensity for indicating behaviour shifts from lifelog data. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 970–977. IEEE (2016)
6. Hu, F., Smeaton, A.F.: Image aesthetics and content in selecting memorable keyframes from lifelogs. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24. pp. 608–619. Springer (2018)
7. Ksibi, A., Alluhaidan, A.S.D., Salhi, A., El-Rahman, S.A.: Overview of lifelogging: current challenges and advances. IEEE Access **9**, 62630–62641 (2021)

8. Meier, F., Elsweiler, D.: Going back in time: An investigation of social media re-finding. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 355–364 (2016)
9. Parsania, V.S., Kalyani, F., Kamani, K.: A comparative analysis: DuckDuckGo vs. Google search engine. GRD Journals-Global Research and Development Journal for Engineering **2**(1), 12–17 (2016)
10. Sappelli, M., Verberne, S., Kraaij, W.: Evaluation of context-aware recommendation systems for information re-finding. Journal of the Association for Information Science and Technology **68**(4), 895–910 (2017)
11. Sjöberg, M., Chen, H.H., Floréen, P., Koskela, M., Kuikkaniemi, K., Lehtiniemi, T., Peltonen, J.: Digital me: Controlling and making sense of my digital footprint. In: Symbiotic Interaction: 5th International Workshop, Symbiotic 2016, Padua, Italy, September 29–30, 2016, Revised Selected Papers 5. pp. 155–167. Springer International Publishing (2017)
12. Smeaton, A.F.: Lifelogging as a Memory Prosthetic. In: Proceedings of the 4th Annual on Lifelog Search Challenge. p. 1. LSC '21, Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3463948.3469271
13. Smeaton, A.F., Krishnamurthy, N.G., Suryanarayana, A.H.: Keystroke dynamics as part of lifelogging. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. pp. 183–195. Springer (2021)
14. Tuovinen, L., Smeaton, A.F.: Remote collaborative knowledge discovery for better understanding of self-tracking data. In: 25th Conference of Open Innovations Association (FRUCT). pp. 324–332. IEEE (2019)
15. Tuovinen, L., Smeaton, A.F.: Privacy-aware sharing and collaborative analysis of personal wellness data: Process model, domain ontology, software system and user trial. Plos ONE **17**(4), e0265997 (2022)
16. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., Gai, K.: Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1059–1068 (2018)