In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press, 90-94. DOI 10.5117/9789463722124/CH15.

Citizens in data land

*Arjen P. de Vries**

Abstract

How can people escape the hidden profiling omnipresent in our online world? Ten years back, Profiling the European Citizen argued for legal and technological tools, where we have seen much more progress in the former than in the latter. Two design principles for information services and tools should complement the new regulations: keep data where it originates, and decentralise information services.

Keywords: Profiling, web and social media search, distributed architectures, personal web archiving, decentralisation.

Introduction

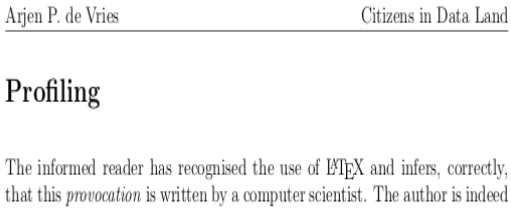
My provocation in the panel on *Legal and political theory in data driven environments* at the workshop '10 Years of Profiling the European Citizen' started with a quote from the closing chapter of Profiling the European Citizen (Hildebrandt and Gutwirth 2008):

> *For individual citizens to regain some kind of control over the way they live their lives, access is needed to the profiles applied to them. This will require both legal (rights to transparency) and technological tools (the means to exercise such rights).*

Looking at progress with respect to these two requirements, European citizens have been successful in creating a legal framework that gives people the power to claim substantial rights in their personal data. Even if we have not yet gained much experience with the law being tested on its practical usefulness, serious restrictions have been imposed upon the parties that control the processing of personal data (e.g., data minimisation, data portability). Switching our perspective to the technological tools however, I am much less optimistic. Wouldn't it be so much easier to exercise our right on e.g. data portability if we actually knew who has our data, in what form, on what server, and how to access and manipulate that data – and not merely transfer this data from one service that we do not control to yet another one?

Profiling

Take a look at the original rendering of my provocation for the online workshop proceedings:



As you read in the Figure already, the informed reader would recognise immediately the use of the LaTeX typesetting system and infer, correctly, that this provocation is written by a computer scientist.[1] The author is indeed trained as computer scientist and the first thing he had to do upon receiving the invitation to join the workshop with a provocation was to look-up the meaning of that term, using a search engine (I might as well share my ignorance with you, the reader, given that *I shared this information already with one of the largest tech companies in the world*). The title of the panel revealed more gaps in my background knowledge, because my immediate association with "political theory" is the title of a Coldplay song. Wikipedia came to the rescue, although I would tell my students not to simply rely on the information in the online encyclopaedia when it concerns my area of expertise... At this point in my provocation, you know most of the information about me that you would have learned also from my bio on one of the various social media sites where I have an account.[2]

Now, the simple fact that you can find this personal information about me via a web search by name (you need to include the middle initial) is no issue of concern; the bio is a public self-description I

contributed voluntarily to the online world, as a 'citizen of data land', advertising why to connect to me. What does (and should) raise objections is the detailed information that I gave away implicitly, mostly unaware, through usage of online services such as the search engine. And it is not easy to escape hidden forms of profiling if I want to stay a 'citizen of data land'; a recent analysis of the CommonCrawl 2012 corpus found that the majority of sites contain trackers, even if websites with highly privacy-critical content are less likely to do so (60% vs 90% for other websites) (Schelter and Kunegis 2018). I learned from an independent blogger that her commissioning parties demand Google Analytics based statistics: to generate any income as an online writer, sharing visit data from your blogging site with Google has become a *de facto* prerequisite, even if you keep your site free from advertisements. The way the Web has evolved, accessing online information implies being profiled.

Civic responsibility in 'data land'

Will the new legal rights (transparency and control) help enforce a new balance? We should not sit back and expect the GDPR to save our privacy from organisations' hunger for data. If only 'citizens of data land' had the means to take control of their data, including the traces they leave online; alas, we have seen less progress with regard to the technological tools necessary to exercise our new rights.

The current situation is that 'we the people' give those who run online services a *carte blanche* to collect our data. The legal framework will make this collection more transparent (we hope), but it cannot change the status quo if we do not act ourselves. It is – to a large extent – our own personal choice (if not to say mistake) that we let a few, very large and omnipresent organisations build their business model on harvesting personal data *en masse*.

If we do not modify our online behaviour, the GDPR creates an improved legal context, sure; but the balance of power between individual citizens and the (public and private) organisations they deal with online shifts back just a tiny fraction of how it could shift back to the citizen, if only we were more responsible in taking care of our data.

Our data, our devices

We have been seduced to give up, voluntarily, the control over our personal data, in exchange for convenience: the convenience of having services managed for us, in the cloud, seemingly for free. We give away our data without much consideration of their value, or the long-term consequences of doing so. We might try to claim back our data with the re-gained legal rights, or at least exercise control over the ways our data is used – but would it not be so much easier to "simply" keep our data for ourselves?

We create our personal data ourselves, and, at least initially, on our own devices.

Instead of handing over that data to an external organisation that runs an information service for us, I put my cards on two design principles to help establish a renewed, better balance, where the people who create the data exercise a significantly larger degree of ownership over their data.

Personal web archives

The first principle is to build systems for online information interactions such that they **keep data where it originates**: in your own device.

As a proof of concept, consider the personal web archive and search system called WASP,[3] that archives and indexes all your interactions with the Web and enables effective re-finding (Kiesel et al. 2018). Those searches remain completely local (and therefore private). While WASP did not yet address the case of a user managing multiple devices (like a smartphone and a desktop computer), this is resolved with Prizm, a small personal device that acts as a gatekeeper between your edge devices and the outside world (Lin et al. 2016).

A more radical version of the design principle (of keeping all your personal Web interactions local) would be to expand those interactions, as a seed to a personal crawl that captures also the information for highly likely future interactions, while also storing a significant fraction of the Web as a snapshot local to your device, instead of in your favourite search engine's data centres.

Practical implementation of this idea raises many interesting technical questions (exciting for the computer scientist in me), where I imagine a role for commercial and/or non-profit organisations too.

They could, for instance, package recent web crawls for distribution, sliced per topic of interest.[4] People could then subscribe to regular updates of their own personal search engine index without the need to crawl the Web themselves; the GDPR helps us trust those organisations to keep subscription information private and secure.

Decentralised social Media

Obviously, whenever we want to share information with others, we cannot keep that data on our own infrastructure. The second design principle would therefore be to **decentralise online services** (or, better, to *re*-decentralise the Web).

The recent rise of decentralised alternatives to existing centralised social media services is especially promising. ActivityPub[5] is a W3C standard that has been granted the status of 'recommendation' (since January 23rd, 2018) and has already been implemented in an increasing number of open source projects. For example, Mastodon is essentially a 'decentralised version of Twitter' where ActivityPub facilitates the communication among thousands of Mastodon instances that together host over 1 million registered users. Other community projects have created decentralised alternatives for Instagram (PixelFed), YouTube (PeerTube), and Medium (Plume).

This cooperation of decentralised online services that exchange social information using ActivityPub has been called the Fediverse (a partial blend of federated and universe). Members of the Fediverse interact freely with each other, even if their accounts reside on different so-called 'instances'. This enables communities to organise themselves, independent from large corporations that would like to collect this data in a huge centralised database. Examples of Mastodon instances that serve a community include the recent Mastodon instance created for 'all people with an email address from University of Twente', an MIT instance, and, an instance I created myself, aiming to be a new online home for the Information Retrieval community.[6]

Closing statement

The directions in which I seek a solution for better technological support are still a long way from empowering the 'citizens of data land'.

A hurdle to take is how to get these new solutions in a state so that 'data land' ends up under 'the rule of the people'. Managing your own personal data is a '21st century skill' that the 'citizens in data land' will have to master. If we do not pay attention, we end up replacing one 'aristocracy', of an elite of large tech corporations, by another one, consisting of tech savvy people who know how to operate their own data infrastructure, thus excluding others from exercising the same level of control over their data.

The exciting technological developments that underpin the two principles of data ownership and decentralisation create an opportunity to exercise a higher level of control over the decision as to who gains access to our data. However, we need to pay for this control in the form of an investment in personal computer infrastructure and the effort to acquire the skills to manage this infrastructure.

Are we, the people, willing to make that effort? Paraphrasing Hildebrandt and Gutwirth (2008, 365):

> *Citizenship, participation in the creation of the common good and personal freedom cannot be taken for granted, they presume that citizens 'acquire the competences to exercise control over what is known about them and by whom'.*

Notes

* Arjen P. de Vries is professor of information retrieval at Radboud University Nijmegen, the Netherlands. His research aims to resolve the question how users and systems may cooperate to improve information access, with a specific focus on the value of a combination of structured and unstructured information representations. Homepage: http://www.cs.ru.nl/~arjen/.

[1] The format of the text in the Figure is another, more subtle hint that the author might be a computer scientist.

[2] 'Computer scientist and entrepreneur. Information access & integration of IR and DB. And Indie music'.

[3] https://github.com/webis-de/wasp/.

[4] Consider a new service provided by The Common Crawl Foundation, http://commoncrawl.org/, or, alternatively, a new community service provided via public libraries.

[5] ActivityPub, https://www.w3.org/TR/activitypub/.
[6] Visit https://idf.social/ or https://mastodon.utwente.nl/ for more information

References

Hildebrandt, Mireille, and Serge Gutwirth. 2008. "Concise conclusions: Citizens out of control." In Profiling the European Citizen: Cross-Disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth, 17-45. Dordrecht: Springer.

Kiesel, Johannes, Arjen P. de Vries, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. “WASP: Web Archiving and Search Personalized.” In Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018. CEUR Workshop Proceedings 2167: 16-21.

Lin, Jimmy, Zhucheng Tu, Michael Rose, and Patrick White. 2016. “Prizm: A Wireless Access Point for Proxy-Based Web Lifelogging.” In Proceedings of the first Workshop on Lifelogging Tools and Applications (LTA '16). ACM, New York, USA: 19-25.

Schelter, Sebastian, and Jérôme Kunegis. 2018. “On the ubiquity of web tracking: Insights from a billion-page web crawl.” The Journal of Web Science 4(4): 53–66.