Toward a Solid Acceptance of the Decentralized Web of Personal Data: Societal and Technological Convergence

Pop Stefanija, Ana; Buelens, Bart; Goesaert, Elfi; Lenaerts, Tom; Pierson, Jo; Van Den Bussche, Jan

Published in:
Communications of the ACM

DOI:
10.1145/3624555

Publication date:
2023

License:
Unspecified

Document Version:
Accepted author manuscript

[submitted version]

Towards a “solid” acceptance of the decentralized web of personal data: social and software challenges

Ana Pop Stefanija, Bart Buelens, Elfi Goesaert,

Tom Lenaerts, Jo Pierson, Jan Van den Bussche

Citizens using common online services such as email, social media, or health tracking, effectively hand over control of their personal data to the service providers. The services used for processing personal data are also where the data must reside. This situation is problematic, as has been recognized for some time [1]: competition and innovation are stifled; data is duplicated; and citizens are in a weak position to enforce legal rights such as access, rectification, or erasure. One suggestion to address this problem has been to ascertain that citizens can access and update, with every possible service partner, the personal data that partner holds of them [2]. The question whether such an approach is realistic remains up for debate. Recently, however, various community efforts are taking a different approach, turning things around.

The central tenet of this complementary approach is that citizens should regain control of their personal data. Once in control, they can decide which partners they want to share data with, and if so, exactly which part of their data. Moreover, they can revisit these decisions at any time. This is the societal vision put forward by the MyData movement in Nordic countries since 2012 [mydata.org].

Software systems and applications for operating such a vision were already envisaged in the previous decade. In the field of data management and information systems, personal information management systems have been studied [3]. In the fields of security, privacy and usability, scholars have discussed Personal Data Services as an alternative aggregating platform under control of the end user [4]. In the area of genetics, personal data lockers have been proposed [5]. Last but not least, in the web area, we see the Solid project [6, solidproject.org] gaining momentum. Solid lets individuals store their data in personal data vaults called Pods: secure personal web servers for data.

The purpose of this note is to discuss challenges in making the decentralized web of personal data a reality. Some challenges are of a societal, others are of a technological nature, and we tackle both. Indeed, we will disclose some interesting social studies done in the context of Flanders, Belgium, and Europe.

The data ecosystem

Putting individuals in control of their personal data can stimulate a new economy, where providers compete to deliver useful and innovative services. Consider, for example, your daily exercise data. You may prefer one platform for producing insightful data analytics, another for appealing data visualization, and yet another for communicating reports to

friends. Contrast this to the current situation, where data needs to be duplicated across different platforms, and where migrating to different platforms requires painful data export and import procedures.

Such a citizen-centric data ecosystem raises pertinent questions. Which companies do we want to do business with? Which (non-)commercial organizations or partners do we want to voluntarily share data with? In studies conducted by KennisCentrum [7] and Itsme [8], 85% of the respondents indicate that they value privacy as important, and 75% to 78% worry that companies will misuse data they collect. These figures suggest that citizens will indeed be selective in deciding whom to share their data with.

Willingness to share data with a partner will likely depend on the nature of the data, as well as on the purpose [9,10]. Data considered less private, such as name or birthdate, is more likely to be shared in comparison with highly private data, such as health information, credit history, or current location. Furthermore, people tend to be more willing to share if serving the public good, especially in a medical context, or when improving their own health [11]. Following as a close second is the use of data for advancing academic knowledge in particular areas. Generally, data is seen as useful if it helps keep people safe, followed by usage by governments to improve public services.

Another factor is the identity of the partner [9]. Willingness to share data with an organization is higher if people are familiar with it. More generally, they must be able to trust the organization. In Europe, governments, health institutions and banks are seen as the most trustworthy kinds of organizations for sharing of personal data. The more commercial a sector, the less safe it is perceived. Trustworthy organizations may be perceived to have a long history in the protection of personal data with tried-and-tested solutions and are more likely to implement well-developed standards for data handling; they are more amenable to governmental oversight and complying with regulations. The trustworthiness of an organization may also depend on the nature of the data and its use. For example, one study reports reluctance to share health data for insurance purposes, which stands in contrast with the trust banks receive concerning financial data [11].

Data security: Pod providers and aggregators

Just like most individuals do not run their own web servers, they will likely not run their own Pods. There is an important role for companies, institutions or intermediaries that provide the service of hosting Pods. These *Pod providers* may be commercial companies, or public and not-for-profit institutions (e.g., civil society organizations). Some countries, such as Nordic countries involved in the MyData movement, and the Flanders region of Belgium, are preparing to set up public Pod providers as an integral component of their e-government infrastructure [12].

Pod providers carry a huge responsibility. They must keep data safe, be resilient to attacks, and guarantee quality of service. Their implementation of access rights to Pods must be watertight. Laws and regulations may appear for holding Pod providers accountable.

An additional role that can be played by Pod providers, and by other parties, is that of an *aggregator* — a trusted party that can collect and aggregate the totality of data of many individuals. Think, for example, of a research institute conducting population health studies on health characteristics, health-impacting behavior, eating and exercising habits etc. Through the aggregator, each health study could request access to specific types of data

directly from citizens who are willing to participate. The aggregator can subsequently make this data, or parts thereof, available to selected health studies, on the condition that they meet criteria of credibility, quality, confidentiality, ethics and regulation. Additionally, Pod technology may make it easier for the real stakeholders to get involved through trustworthy organizations. Clearly, all these types of aggregators will be subject to prevailing regulations including the GDPR and the Data Governance Act.

Web agents, schema mappings, and mediators

As we have seen, the promise of Solid (or any similar enabling technology) is to give citizens the autonomy and the agency to share personal data with partners whom they trust. Recent research by some of the authors [13] shows that to gain a sense of agency and trust, people have three related requirements. They should be *able to act* regarding the entire data cycle; they require *transparency* concerning the use, purposes or goals; and they should be able to meaningfully *understand* the outputs of the data processing and how they affect them.

At the same time, it is evident that people need support in exercising their autonomy. How do they keep track of what was shared, and with whom? To fill this need, we envisage a *web agent* as an autonomous piece of software between a Pod and the world. This web agent can assign unique client identifiers to different partners and allow them access to data they were authorized to.

Such an arrangement, however, requires quite some care. Each partner will expect data to be organized in a certain form. Facebook, for example, may expect your posts, reactions, and comments, to be arranged chronologically, so that it can efficiently query for your latest information. A blogging website may have similar expectations but may require a different structure for metadata or textual content. At the same time, your data is originally represented, in your Pod, in yet another manner. Think for example of someone who frequently publishes short pieces and wants to share them both on Facebook and a blogging platform.

This situation is familiar in the field of information integration [14], where solutions have been developed based on the notion of *schema mapping*. S*chema* refers to a structuring of data in a certain form. A schema mapping is a transformation of data over some source schema into some target schema. In our setting, the source schema is the schema of the Pod, and the target schema is the one expected by the partner. A query formulated by the partner over the target schema can be automatically *rewritten* to an equivalent query over the source schema; the rewritten query can be answered by the Pod, and the results can be handed over to the partner. The schema mapping serves a double purpose: it serves to convert from one data format to another, and it is the formal mechanism by which the data to be shared is specified precisely. Managing the schema mappings for the different partners will be the task of the web agent. The process of verifying a partner's identity, consulting the appropriate schema mapping, receiving and rewriting a query and delivering the answer is known as a *mediator*. Care should be taken that such use of AI proxies leads to beneficial outcomes [15].

Software challenges

Schema mappings can be written using standard query languages. The Web community is working towards a recommended schema mapping language (RML, rml.io). However, *who* will write the schema mappings? It is clearly unrealistic that individual citizens will write

their own. Instead, we expect that trusted partners, in collaboration with Pod providers and aggregators, will offer generic schema-mapping packages that could be used out of the box.

We must keep in mind the principle of the human-in-control. It is therefore essential that we develop creative solutions that allow people to modify, configure, and understand schema mappings. We see exciting opportunities for innovative software development. Tools are needed for visualization of mappings, for learning, explaining, and verifying mappings for correctness. These tools need to be usable by people without computing expertise.

A related research direction is that of expressive schema languages for RDF data (the standard data model for open linked data on the web). The W3C has developed the Shapes Constraint Language "SHACL". More research is needed to see if SHACL suffices to express requirements on schemas for exchange and sharing of personal data. SHACL schemas could be learned from example data, and queries over RDF could be typechecked on the input or output side for conformance to SHACL schemas.

Conclusion

In a decentralized web of personal data, social, economical, legal, and technological factors come into play. We highlighted some important issues and pointed to new potential ways to overcome them. Building human agency-enabling technological solutions that remedy complex socio-technological issues is not an easy or a quick job, and will require the joint work of many disciplines. A thorough discussion of regulation is outside the scope of this article, but we can point at relevant developments such as the European Health Data Space [16], and guidelines such as the Datasheets for Datasets [17]. In this article, we have emphasized that social trust and advances in software will have to go hand in hand to make the vision of decentralized personal data management into a reality.

References

1. A. Poikola et al. MyData: A Nordic model for human-centered personal data management and processing. Finnish Ministry of Transport and Communications, 2012. See also later white papers, 2020, 2022, at https://mydata.org.
2. Y. Gurevich, E. Hudis, J.M. Wing. Inverse Privacy. *Communications of the ACM*, 59(7):38–42, 2016.
3. S. Abiteboul, B. André, D. Kaplan. Managing your digital life. *Communications of the ACM*, 58(5):32–35, 2015.
4. A. Acquisti, C. Bettini, R. Böhme, C. Castelluccia, T. Dimitriou, F. Dürr, R.K. Ganti, J. Grossklags, D. Estrin, M. Friedewald, R. Mayrhofer, D. Phillips, K. Rannenberg, N. Sadeh, M. Scipioni. Personal Data Service: Accessing and aggregating personal data (Chapter 4.1). A. Acquisti, I. Krontiris, M. Langheinrich and M.A. Sasse (Eds.) *'My Life, Shared': Trust and Privacy in the Age of Ubiquitous Experience Sharing (Dagstuhl Seminar 13312)*, Dagstuhl Reports, 3:7, Wadern: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 87-92, 2013. https://doi.org/10.4230/DagRep.3.7.74
5. Overkleeft, R., Tommel, J., Evers, A. W., den Dunnen, J. T., Roos, M., Hoefmans, M. J., Schrader, W. E., Swen, J.J. ,Numans, M. E., Houwink, E. J. (2020). Using personal genomic data within primary care: a bioinformatics approach to pharmacogenomics. Genes, 11(12), 1443.

6. E. Mansour, A.V. Sambra, and others. A demonstration of the Solid platform for social web applications. In *WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web*, pages 223–226. ACM, 2016.
7. M. Martens, R. De Wolf and T. Evens. Algoritmes en Artificiële Intelligentie in een medische context: een studie naar de perceptie, mening en houding van Vlaamse burgers. *Kenniscentrum Data & Maatschapij*, December 2020. (In Dutch.) https://data-en-maatschappij.ai/publicaties/algoritmes-en-artifici%C3%ABle-intelligentie-in-een-medische-context-een-studie-naar-de-perceptie-mening-en-houding-van-vlaamse-burgers
8. Itsme. The power of digital ID — Survey 2020, Belgians and digitalization https://www.itsme-id.com/files/itsme-Survey-2020.pdf
9. Open Data Institute. Who do we trust with personal data. 2018. https://theodi.org/article/who-do-we-trust-with-personal-data-odi-commissioned-survey-reveals-most-and-least-trusted-sectors-across-europe/
10. Cavestany, M., van den Dam, R. and Fox, B. The trust factor in the cognitive era. *IBM Institute for Business Value*, 2017. https://www.ibm.com/thought-leadership/institute-business-value/report/digitaltrust#
11. P. Raeymaekers. Zorg voor je data. Koning Boudewijnstichting, January 2022. (In Dutch.) kbs-frb.be
12. NN. Who owns the web's data? The fightback against Big Tech's feudal lords has begun. *The Economist* (Business section – Schumpeter "Free the data serfs!"), October 22, 2020.
13. A. Pop Stefanija and J. Pierson. How to be algorithmically governed like that—data- and algorithmic agency from user perspective. *AoIR Selected Papers of Internet Research*, 2021. https://doi.org/10.5210/spir.v2021i0.12227
14. K. Aberer. *Peer-to-Peer Data Management.* In *Synthesis Lectures on Data Management,* lecture 15. Morgan & Claypool, 2011.
15. E. Fernández Domingos, I. Terrucha, R. Suchon, J. Grujić, J.C. Burguillo, F.C. Santos, T. Lenaerts. Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports,* 12(1), 1–12, 2022.
16. M. Shabani. Will the European Health Data Space change data sharing rules? *Science,* 375(6587):1357–1359, 2022.
17. T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, and others. Datasheets for Datasets. *Communications of the ACM,* 64(12):86–92, 2021.

Authors

Ana Pop Stefanija (ana.pop.stefanija@vub.be) is a PhD researcher at imec-SMIT, research group at Vrije Universiteit Brussel, Belgium.

Bart Buelens (bart.buelens@vito.be) is Head of Data Science at VITO, the Flemish Institute for Technological Research in Belgium.

Elfi Goesaert (elfi.goesaert@vito.be) is researcher and project leader Data Science at VITO, the Flemish Institute for Technological Research in Belgium.

Tom Lenaerts (Tom.Lenaerts@ulb.be) is professor at the Department of Computer Science in the Faculty of Sciences of the Université Libre de Bruxelles (machine learning group) and the Department of Computer Science in the Faculty of Sciences of the Vrije Universiteit Brussel (artificial intelligence group), both located in Belgium.

Jo Pierson (jo.pierson@vub.be) is professor of Responsible Digitalisation in the new School of Social Sciences at Hasselt University (research group R4D), and professor of Media and Communication Studies in the Faculty of Social Sciences & Solvay Business School at the Vrije Universiteit Brussel (research group imec-SMIT), Belgium.

Jan Van den Bussche (jan.vandenbussche@uhasselt.be) is professor of databases and theoretical computer science, and member of the Data Science Institute, at Hasselt University, Belgium.

(max 1800 words)