

## Opinion

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

*Giving individuals more control of their personal data.*

CITIZENS USING COMMON online services such as social media, health tracking, or online shopping effectively hand over control of their personal data to the service providers—often large corporations. The services using and processing personal data are also holding the data. This situation is problematic, as has been recognized for some time:<sup>13</sup> 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. The approach to address this problem has been to ascertain that citizens can access and update, with every possible service provider, the personal data that providers hold of or about them<sup>6</sup>—the foundational view taken in the European General Data Protection Regulation (GDPR).

Recently, however, various societal, technological, and regulatory efforts are taking a very different approach, turning things around. The central tenet of this complementary view is that citizens should regain control of their personal data. Once in control, citizens can decide which providers they want to share data with, and if so, exactly which part of their data. Moreover, they can revisit these decisions anytime.



This is the societal vision put forward by the MyData movement in Nordic countries since 2012 (see <https://mydata.org>), which since has grown into a global movement including an award system for certified data operators. Another prominent example of individuals embracing decentralized solutions is the uptake of Mastodon. After Elon Musk's take-

over of Twitter in November 2022, the decentralized and open-source social media that runs using W3C's standards, drew millions of new users in a matter of days.

Apart from these initiatives, new European legislation is providing more rights to data subjects, ostensibly supporting this trend. The Data Governance Act (DGA) for example,

applicable since September 2023, provides roles for data intermediaries, mediating data access between data subjects and companies wishing to use their data for products and services. The data intermediaries as described in the DGA match the role of data operators in the MyData movement. Additionally, proposals for the Data Act (DA) and European Health Data Spaces (EHDS<sup>16</sup>) provide more access rights and data portability options for data subjects, again expanding on GDPR data subjects' rights in line with the MyData movement and a more citizen-centric approach.

Upcoming trends in Web technology—collectively referred to as Web3 or Web 3.0—include semantics, decentralization, and personalization. These come together in the specifications of the Solid project<sup>9</sup> (see <https://solidproject.org>). This new technology lets individuals store their data in personal data vaults called Pods: secure personal online Web servers for data. This citizen-centric approach in Solid, with decentralized management of personal data, is an important answer to the trend of upping the rights of data subjects, set forward in both the MyData movement and the new EU legislations mentioned here (DGA, DA and EHDS). Software systems and applications for operating such a vision, however, were already envisaged in the previous decade. In the field of data management and information systems, personal information-management systems have been studied.<sup>2</sup> 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.<sup>3</sup> In the area of genetics, personal data

**Upcoming trends in Web technology include semantics, decentralization, and personalization.**

lockers have been proposed.<sup>13</sup> Against this historical backdrop, Solid can be seen as an instrumental next step in this process. Under the auspices of the World Wide Web Consortium (W3C), the Solid Community Group (see <https://www.w3.org/community/solid/>) is working toward making the Solid specification a W3C standard.

Here, we connect these societal and technological trends, with recent regulatory initiatives following suit. We are now at a stage where data ecosystems of the future are being designed in a decentralized, personalized fashion. We elaborate on this view and discuss outstanding challenges.

### From Accessing to Sharing Data

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, daily exercise data stored in a citizen-centric fashion with the individual in control of which services they use. One service provider may offer insightful data analytics, another impressive data visualizations, and yet another may allow for communicating reports to friends. Contrast this to the current situation, where data needs to be duplicated across different service platforms, and where migrating to other platforms requires painful data export and import procedures—if at all possible. In addition, researchers from academia, government, or industry may want access to personal exercise data for health studies or other research. In a citizen-centric system, they can simply seek consent from citizens to share the relevant data they hold, for specific research purposes.

Such a citizen-centric data ecosystem raises pertinent questions. Which companies do people want to do business with? Which organizations, commercial or not, or other partners, do people want to voluntarily share data with? In studies conducted by the Knowledge Centre Data & Society<sup>10</sup> and Itsme,<sup>8</sup> 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 citizens will indeed be selective in deciding

**Putting individuals in control of their personal data can stimulate a new economy, where providers compete to deliver useful and innovative services.**

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.<sup>4,14</sup> 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.<sup>14</sup> Following as a close second is the use of data to advance 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.<sup>4</sup> 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 the E.U., governments, health institutions, and banks are seen as the most trustworthy kinds of organizations for sharing 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 to 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.<sup>14</sup>

### The New Data Holders

Just like most individuals do not run their own Web servers, they will likely not run their own Solid 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 (for example, civil society organizations). For example, the company Inrupt, co-founded by Sir Tim Berners-Lee works globally on putting the Solid principles in practice for companies. Then in the Flanders region in Belgium, recent government-supported initiatives include the establishing of the Data Utility Company (see <http://www.data-nutsbedrijf.be>), which will operate as a Pod provider and at the same time put in place infrastructure and policies to support decentralization of personal data. Funds have been allocated to support Solid-based activities for data innovation based on Solid (see <http://www.solidlab.be>). There is a strong community of developers gathered around the Solid Community Flanders (see <https://solidcommunity.be>). In addition, a project to develop a Pod-providing platform for health data, and the governance of it, started earlier in 2022 (see <http://www.we-are-health.be>). These recent initiatives are likely among the first of their kind to develop decentralized data architectures at state level with government buy-in. Moreover, new Flemish companies are active in the Solid sphere as well.

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. While for the moment, there is no regulation targeting specifically decentralized technologies, the task of mapping the existing principles (see DGA, DA, EHDS) to the Solid context, while a challenge, is feasible. For example, while DGA does not

target a specific type of technology, it does propose regulations for the role of data intermediaries and data cooperatives, who mediate between data subjects and companies. To the extent that Pod providers serve this intermediary role, DGA could provide the first set of regulations that apply to decentralized technologies.

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. Through the aggregator, other parties can 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 studies, on the condition that they meet criteria of credibility, quality, confidentiality, ethics, and regulation. As previously mentioned, and just like any other data processors, aggregators can be well aligned with prevailing regulations. Yet, again, mapping the principles therein to new roles played in the decentralized Web is an important task to be investigated by law researchers and legislators.

### Web Agent Technology for Humans

As we have seen, the promise of technology such as Solid 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<sup>16</sup> 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 to access data they were authorized to.

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

This situation is familiar in the field of information integration,<sup>1</sup> where solutions have been developed based on the notion of *schema mapping*. *Schema* 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

**The trustworthiness of an organization may also depend on the nature of the data and its use.**



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.<sup>5</sup>

### Software Challenges

Schema mappings can be written using standard query languages. The Web community is working toward a recommended schema mapping language (RML, see <https://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 type-checked on the input or output side for conformance to SHACL schemas.

### Conclusion

We have identified three well-aligned trends indicating a convergence toward a decentralized Web of data. These trends meet recent and upcoming regulations, as well as societal and ethical demands for personal data handling (much more than many present-day practices). Each trend has its roots in the years 2000 and started largely independent of the others.


First, the Solid movement originated from the view that the power in

## Tools are needed for visualization of mappings for learning, explaining, and verifying mappings for correctness.

the Web must be redistributed. While the Web was conceived as a decentralized system, platforms operated by Big Tech turned it into a practically centralized architecture. Solid as a protocol and specification is being developed actively under auspices of the W3 to turn it into a standard web technology.

Second, there is a clear societal trend, first initiated by privacy activists, and increasingly being pushed by data ethicists.<sup>7</sup> This trend opposes practices where people lose control of their personal data, and must obtain it from the data holders—businesses, typically—as opposed to the businesses obtaining data from the people. MyData emerged as an umbrella movement of this societal view, as is laid down in their Declaration.

Third, regulators reacted to formal consumer complaints about personal data handling practices which in Europe led to adoption of the GDPR. More recently it was recognized that this may stifle the data economy and the Data Governance Act and Data Act were initiated. It is safe to say societal personal data trends have influenced policymakers and legislators, to align the permissible with the desirable.

The MyData Declaration, the GDPR, as well as the DGA, are technology-agnostic. Nevertheless, the movement of decentralizing the Web matches them very well and provides a technology to actually achieve what is legally required or soon will be, and what is increasingly ethically desired from the public. We now see convergence of these trends, and from a technology perspective, readiness for the uptake of decentralized Web technologies. 

### References

1. Aberer, K. Peer-to-Peer Data Management. *Synthesis Lectures on Data Management, Lecture 15*. Morgan & Claypool, 2011.
2. Abiteboul, S., André, B., and Kaplan, D. Managing your digital life. *Commun. ACM* 58, 5 (2015), 32–35.
3. Acquisti, A. et al. Personal Data Service: Accessing and aggregating personal data (Chapter 4.1). In *'My Life, Shared': Trust and Privacy in the Age of Ubiquitous Experience Sharing (Dagstuhl Seminar 13312)*, Dagstuhl Reports, 3:7, A. Acquisti et al., Eds. Wadern: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, (2013); <https://doi.org/10.4230/DagRep.3.7.74>
4. Cavestany, M., van den Dam, R., and Fox, B. The trust factor in the cognitive era. *IBM Institute for Business Value*, (2017); <https://bit.ly/46TgXKw>.
5. Domingos, E.F. Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports* 12, 1 (2022), 1–12.
6. Gurevich, Y., Hudis, E., and Wing, J.M. Inverse privacy. *Commun. ACM* 59, 7 (2016), 38–42.
7. Hicks, T. Who owns the Web's data? The fight back against Big Tech's feudal lords has begun. *The Economist* (Oct. 22, 2020).
8. Itsme. The power of digital ID — Survey 2020, Belgians and digitalization. (2020); <https://bit.ly/3tEoyGn>
9. Mansour, E. et al. A demonstration of the Solid platform for social Web applications. In *WWW '16 Companion: Proceedings of the 25th Intern. Conf. Companion on World Wide Web*, (2016), 223–226.
10. Martens, M., De Wolf, R., and Evens, T. Algoritmes en Artificiële Intelligentie in een medische context: een studie naar de perceptie, mening en houding van Vlaamse burgers. *Kenniscentrum Data & Maatschappij (In Dutch)*. (Dec. 2020); <https://bit.ly/45vTiGF>
11. Open Data Institute. Who do we trust with personal data. (2018); <https://bit.ly/3S2Tom0>
12. Overkleeft, R. et al. Using personal genomic data within primary care: A bioinformatics approach to pharmacogenomics. *Genes* 11, (2020), 12.
13. Poikola, A. et al. MyData: A Nordic model for human-centered personal data management and processing. *Finnish Ministry of Transport and Communications*, (2012); <https://mydata.org>.
14. Raeymaekers, P. Zorg voor je data. *Koning Boudewijnsstichting, (In Dutch)*. (Jan. 2022); <https://www.kbs-frb.be>
15. Shabani, M. Will the European health data space change data sharing rules? *Science* 375, (2022), 6587.
16. Stefanija, A.P. and Pierson, J. Algorithmic governmentality, digital sovereignty, and agency affordances: Extending the possible fields of action. *Weizenbaum Journal of the Digital Society* 3, 2 (2023); <https://doi.org/10.34669/WI.WJDS/3.2.2>

**Ana Pop Stefanija** ([ana.pop.stefanija@vub.be](mailto:ana.pop.stefanija@vub.be)) is a Ph.D. researcher at imec-SMIT, research group at Vrije Universiteit Brussel, Belgium.

**Bart Buelens** ([bart.buelens@vito.be](mailto: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](mailto: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](mailto: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 lab), both located in Belgium.

**Jo Pierson** ([jo.pierson@vub.be](mailto:jo.pierson@vub.be)) is professor of Responsible Digitalisation in the 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](mailto:jan.vandenbussche@uhasselt.be)) is professor of databases and theoretical computer science, and member of the Data Science Institute, at Hasselt University, Belgium.