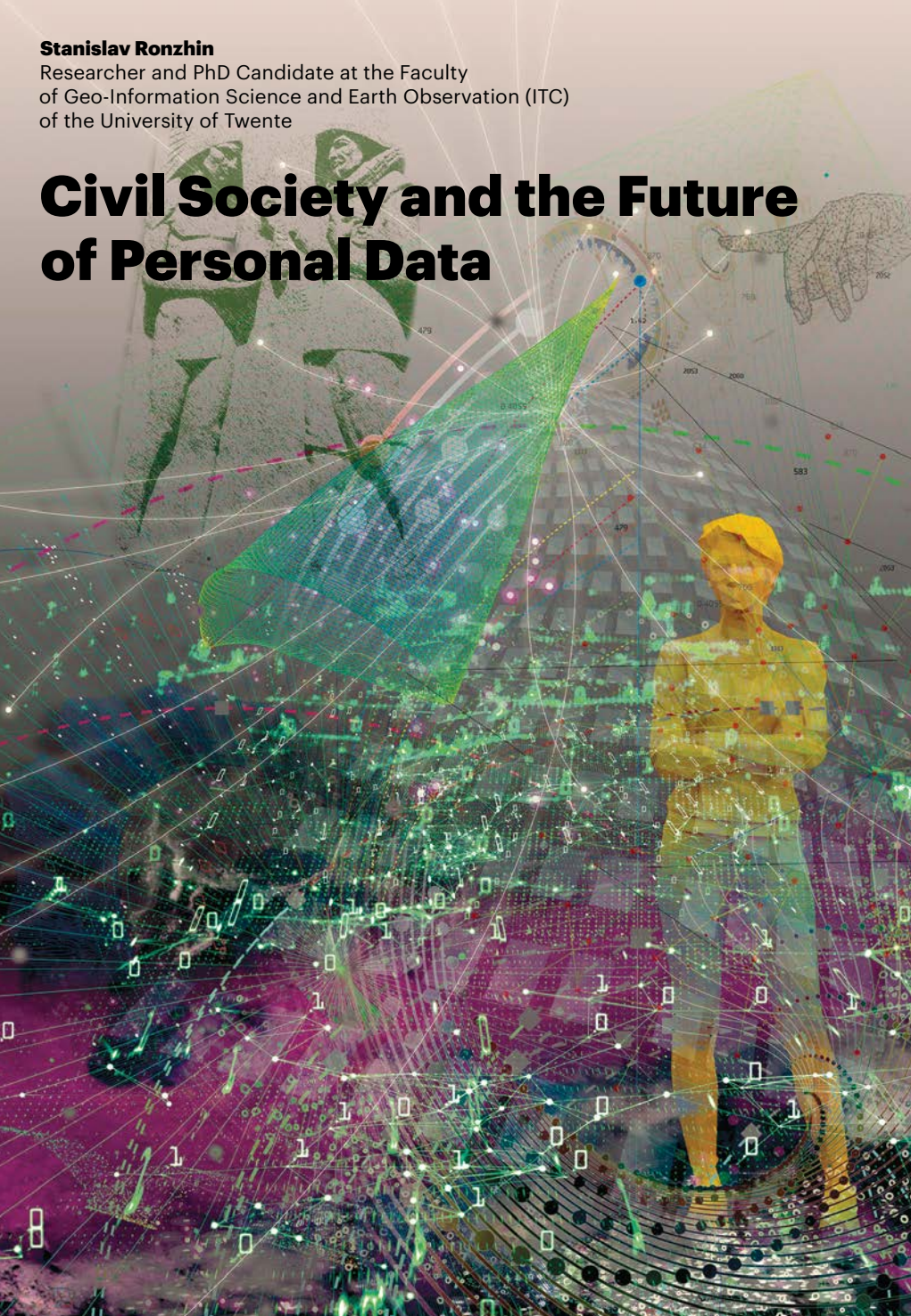
Stanislav Ronzhin
Researcher and PhD Candidate at the Faculty
of Geo-Information Science and Earth Observation (ITC)
of the University of Twente

Civil Society and the Future of Personal Data



"Technology presumes there's just one right way to do things and there never is"

— Robert M. Pirsig[1]

How we lost our data

In 2019, the World Wide Web celebrated its 30th birthday. By then, the story of how the Web came about had become almost mythological. It is thought that the foundation of the Web was laid down in a proposal[2] written by Tim Berners-Lee, addressing "the management of general information about accelerators and experiments at CERN." This novel information management system was designed to deal with the information explosion that had already started affecting data-intensive fields of research such as high energy physics. The original code and specifications defining the Web were published in 1991, and the first website explaining how to set up one's own web server and start publishing documents on the Web was up and running in 1993.

In the following 30 years, the decentralized and distributed nature of the Web has been the main driver of its unconstrained growth. People have been free to publish any documents without registering them in any centralized catalogue; and once published, documents can be immediately accessed by any user with a web browser. The main goal of the initial proposal was achieved: the Web connected people regardless of borders and hierarchies.

However, despite the success story, many people are increasingly of the opinion that the Web has failed in serving the humanity. "But for all the good we've achieved, the web has evolved into an engine of inequity and division; swayed by powerful forces who use it for their own agendas," Tim Berners-Lee reported in his anniversary speech[3] addressing the state of the Web in March 2018. This is because the original concept of the Web relied on it being managed benevolently, that is with a desire to help others and benefit

people rather than to make a profit. This assumption may seem naive nowadays, but the truth is that 30 years ago CERN researchers simply could not imagine that the technology aimed at sharing data with a person next door would be used for "state-sponsored hacking and attacks, criminal behavior, and harassment."

Arguably all this existed well before the Web and it simply found its way online. What is more interesting is that the Web has created an ecosystem of technologies where novel ad-based revenue models have become possible. In such models, profit is dependent on the number of people visiting a page. This triggers a positive feedback loop: more users mean more profit. The downside is that it is all about the numbers; value for is not considered. In fact, Facebook did not originally aim at collecting our data; rather, it wanted more people to spend more time on the platform by making Facebook available on an ever-growing number of user devices. The problem is, as Tim Berners-Lee, put it: "It's amazing how clever people can be, but when you build a new system it is very, very hard to imagine the ways in which it can be attacked." This is how we inadvertently became involved in the unfolding battle for user data – a key asset in the 21st century. The outcomes of this battle will shape not only our digital lives but will directly influence our physical being (e.g., because healthcare is becoming extensively data-intensive).

The aim of this article is to discuss different ideas about how personal data is organized on the Web and to highlight those which are most promising for the development of civil society. Despite the ways the internet has been misused in the past, the future is still much greater than the past. Therefore, we will try to imagine what kind of data management regimes would benefit civil society and what steps should be taken by civil society actors to promote such a vision.

Data pyramid

Data has little value in and of itself, unless it is presented

in a form which can be understood. Once the data is understood and interpreted it becomes information; in turn, information has more value than raw data because it allows us to learn what is going on. Combining several sources of information and structuring and enriching them with context creates knowledge. Knowledge answers questions about how something happened and why. Wisdom is the top tier of the pyramid representing integrated knowledge which makes it possible to determine what is best to do. Fig. 7 illustrates this hierarchy. In short, data enables the generation of new knowledge, which in turn makes it possible to reason about the future.

The model presented in Fig. 7 is something of a holy grail for what is known as data science. The reason is simple: it makes it possible to explain to businesspeople why they need to hire a data scientist. The only difference is that in business language this would be something like "data-driven insights and actionable business intelligence to drive innovation for sustainable future." This would be likely followed by a sentence including concepts "big data" and "artificial intelligence." The idea that data-driven innovations are vital for keeping up with the market is at the heart of the ongoing fourth industrial revolution and the economy it creates.

Value of personal data

Scientists were the first to experience this revolution at the end of the 1980s. For business and industry, it took another 15 years to find out that management, decision making and marketing must be data-driven just to keep afloat. When it comes to policymakers and the general public, it seems that they are yet to find their way around this new world. User data greatly benefit tech giants, while users are getting tailored advertising in return. This does not sound like a fair trade. Only recently politicians have started taking actions to improve (or, more accurately,

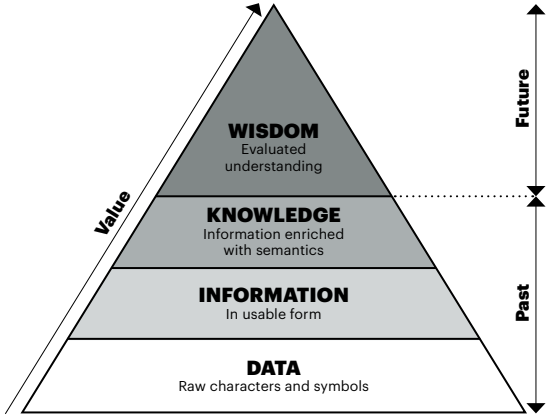


Fig. 7. The Data Information Knowledge Wisdom hierarchy pyramid (Adopted from Rowley, 2007[4])

create) data-related policies such as GDPR. However, so far, all legal efforts have mainly addressed privacy issues and misuse of information. This is undoubtedly important, since it provides grounds for sueing Facebook in court in the case of the next Cambridge Analytica scandal. Still, generals are always fighting the last war, therefore it is not realistic to expect governments to step in to harness the current technology revolution for the benefit of all. Policy-makers will always lag behind industry, and industry will always benefit from the existence of the grey zone. What makes it even worse is the fact that existing legal initiatives do not provide technical frameworks to back them up. As a result, real-life implementation is put back in the hands of app developers. We already know that being driven by the ad-based revenue model these hands will keep extending user agreements so they can get more and more data.

Data as a social divider. Data science has changed many aspects of our lives, but up until recently these changes have not touched on public health. This is rapidly changing since artificial intelligence has become sufficiently mature to be used in medical care systems. When this happens, the quality of the data used for diagnosis and treatment has a direct impact on an individual's wellbeing. Money can already provide access to better healthcare. In

the future, you will need data as well.

"Unfortunately, I believe that the class divide of the future will be data. And if you are not careful those who have access to data have better health than those who don't have access to data," argues Naveen Rao[5], Vice President and General Manager of Intel's AI division. The striking part here is that this was said by a representative of a giant private company whose core business is selling technological solutions. This means that they already know how to use this social divider for their own benefit, and clearly this has little to do with making data accessible to everyone. As such, the quote could have been put this way: "Fortunately, I know that the class divider of the future will be data. And if you are careful enough you will be able to benefit from it for at least 15 years, the time needed for policymakers to understand the problem and start reacting."

Data as an economic asset. The Web became participatory in the mid-2000s. From that moment on, users have become *creators* which means that they have started creating and consuming online content at the same time. This was a breakthrough point because the ability to post, comment, share and like allowed users to become first-class citizens of the Web. The logic was simple: if a comment section on a webpage is the main reason for visiting the page, then commenters would deserve a share of the profit made from displaying ads on that page. In reality, this did not happen because online platforms were not interested in doing so for obvious reasons; additionally, the average revenue per user (ARPU) was too low for taking it seriously.

For instance, Facebook was earning just around 10$ on every user in 2011.[6] However, this number has grown 10-fold since then and there is no reason to think that Facebook couldn't double it again in the next few years. Ad budgets will keep growing as long as Facebook keeps improving targeting algorithms. Therefore, it is plausible that an average user would be able to generate revenue approaching the user's annual income in a not-so-distant future. If this happens, will it be an argument to recon-

sider the terms and conditions of user agreements?

The question of why we still do not have an infrastructure to monetize our own digital footprints has to do with an ambiguity of how we should treat our data. On the one hand, people should be owners of information about themselves, and, as owners of this property, should be in full control over it. Moreover, treating data as property would stimulate the development of the data market. On the other hand, it is also clear that data is an intimate part of an individual's identity or being; it needs to be treated with care. Therefore, to be on the safe side, policymakers originally preferred to focus on creating laws which would provide a tort remedy for invasion of data privacy. Ironically, the approach that was believed to serve as a rule of thumb has led to the situation where actual owners of data don't have the means to monetize it while all the parties involved in the value creation chain (see Fig. 7) are making money.

Data as a political asset. Technically speaking, there is no difference in what's promoted on social media, so we see political ads alongside all others. This comes with tracking of engagement and conversion rates. Therefore, in this sense politicians are no different from other salespeople. They also heavily rely on for steering campaigns efficiently.[7] Political parties are busy organizing internal data infrastructures and processes to be able to climb the pyramid from Fig. 7. As a result, data has become an asset that adds weight to potential candidates.

Despite all the speculation around Cambridge Analytica, there is still no clear evidence that big data collection can be used for predicting and manipulating future outcomes. However, it is certain that it gives a better understanding of the current situation. therefore politicians will be in favor of obtaining more details about people's lives, e.g., data about the location,[8] since political campaigns always have a clear geographical extent. If data centralization empowers those who are responsible for policymaking, it is very unlikely that they will easily move towards giving this power away. The situation when

personal data is centralized in the hands of the few is known as data oligarchy – a rule of a tiny, privileged circle that occupies the top of the pyramid in Fig. 7.

To conclude, data ownership rights and the ability to have a fair share of the cake baked by the new data-centric economy will not simply be given away and they must be taken instead. However, there is a huge imbalance of power between individuals who want to protect their data and those who want to use it for their own gain.

It is going to be extremely difficult to centralize data ownership.

Weak signals

The inability of users to realize the value of their own data belongs to the category of *wicked* problems – those that don't have a single true-or-false solution. Instead, the potential numbers of solutions to a wicked problem is infinite and they can only be evaluated in terms of comparison. The following section overviews and discusses recent promising developments that are aimed at tackling the problem in question. These are weak signals which can help in imagining the trajectory of the problem in the future.

The Social Linked Data (Solid) project is a new endeavor led by Tim Berners-Lee, the inventor of the WWW and the Semantic Web. The project proposes a set of conventions and tools for building decentralized social apps based on Linked Data principles.

Solid implies that people store their data in personal databases called pods. In Fig. 8, these pods are shown as circles. Apps (dark blue shapes) access as many pods as needed instead of working with a single database. Users control which apps can read or write data from/in their pods.

The project aims to disrupt the ad-based revenue model by creating an infrastructure which would allow separating data from apps. Data always stays in a data pod and can be potentially reused by any other application. This will pro-

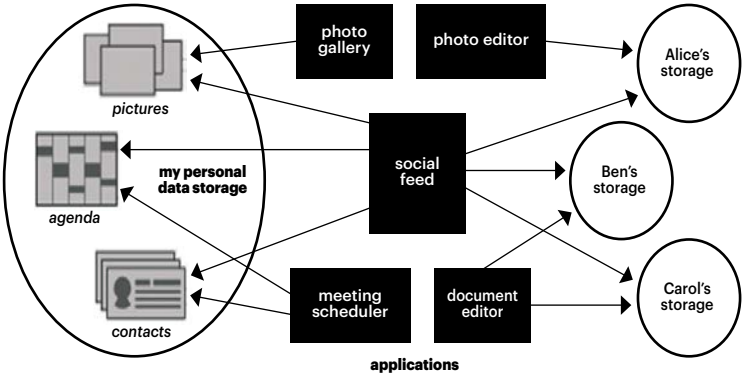


Fig. 8. Model of decentralized personal data storage and distributed applications. Source: Berners Lee & Verborgh, 2018.[9]

mote the creation of the data market and will democratize app development. Users will be able to monetize their own data and move between apps freely based on their functionality. For developers, this would open an opportunity to innovate at the app level since the user data would not be locked down by an app anymore. Moreover, decentralization of storage will return control of privacy back into users' hands. Trading of user data to third parties will not be possible since data exists only in a single place – the user's database – and is never copied. This principle is known as *data-at-the-source* in enterprise system architecture.

Adoption of legal frameworks such as GDPR will create problems not only for organizations with questionable or malicious intent, but also for everyone who deals with personal data. This is often the case for non-governmental and non-profit organizations. For example, a group of skilled volunteers wants to help a non-profit organization with a data project. This requires copying data, which is illegal without additional user consent. The solution to the problem is to provide citizens with personal data pods, so that all their public and private data remains in one place. Instead of moving data between organizations, they individually ask for permission to view relevant parts of the data only. This way data does not have to be moved around, and GDPR compliance

can be assessed automatically for every single data request.

Solid is not the only development that attempts to deploy distributed social networks at a Web-scale. For instance, Diaspora and Indie Web are working examples; however, the difference is that Solid is supported by a company called Inrupt whose main mission is to foster the development of apps and communities around Solid.

Federated learning is an emerging trend in machine learning that does not require centralization of data for model training. In contrast to the traditional centralized approach, federated training takes place on a personal device using local data. The data is never sent to a central server and only model parameters are exchanged instead. This enables the development of machine learning algorithms without sharing and exchanging personal data.

What will the future look like?

This section addresses the possible trajectories of the problem in the next 20 years.

The future I would like to live in

In 2040, Web apps will not be able to copy and store user information. Instead, every time a user accesses a Web resource, the ad provider will have to request the data needed for content personalization from the user's personal storage. However, unfortunately for the ad provider, the user has already set a price (let's say 10 cents) for each of the data calls that ask about their preferences. If the ad provider agrees, then the user receive 10 cents in their bank account and the page shows personalized content. Otherwise, the user gets default ads. In a similar way, users can monetize any tracking.

However, even if the provider personalizes their content to the user's needs, they never get access to the underlying data used to calculate the preferences. The user uses another app to make it upfront and only exposes

the final results for the ad personalization service. It is likely that modern social media, e.g., Facebook, would become such services, since they would not survive the competition with an avalanche of new social networks leveraging the benefits of decentralized personal data storage. Instead, Facebook would serve as an ad-broker and proxy between companies and users. Moreover, these new social apps would be ad-free because the very existence of the data market would trigger a competition between app developers and, consequently, they would be forced to seek other forms of monetization.

Potential for civil society: independent identity provider. The presence of the state in people's lives will decrease. Governments will still be responsible for the maintenance of the official national registers, but data will be stored and controlled by users. Government bodies will request access to personal data on an individual basis. In this context, government organizations will be nothing but another service to keep your precious data up to date. Personal data storage, apps and access control will be provided by different parties. Therefore, users will be free to choose independent service providers at each of these levels. This is exactly where NGOs and non-profits can step in. The deployment and maintenance of independent communal infrastructures plays an important role and provides a great opportunity to strengthen civil society. This is especially important when it comes to identity provision. Similarly to using a Facebook account to log into other services, people will be able use identity services run by the community.

Potential for civil society: data unions. Individuals will use their data as a means of democracy. If a user does not share the political agenda of a certain political party or candidate, then the user can either deny access to their data or set a price for such access. In this context, support would mean granting access to certain data to the candidate. In a similar way, the ability to donate digital footprints will boost citizen science projects. Neighborhood level traffic optimization will come together with strengthening

the local community. Scaled up on spatial and conceptual levels, this may serve as a catalyst for developing global coordination.

Technical and legal aspects of personal data management are frequently far too complex to be managed by individuals themselves. For example, consider a scenario. Alice keeps her personal browsing data in a data pod. Bob is a researcher at a university who would like to use Alice's data. Alice is a member of a data union, a non-profit community-driven organization that helps its members to manage their data permissions. The data union agrees that Bob can be trusted with Alice's data for his research. Data unions can be of different geographical extent (local-global) and application domain.

Such a future can be possible for two reasons. First, by 2020, tech giants of the Western world had started losing competition for Asian media platforms, especially those in China. On the one hand, GDPR resulted in a number of devastating lawsuits against Facebook and Google. On the other hand, the ever growing middle-classes in Africa and Asia choose platforms using their languages. Together, these reasons have raised awareness and are considered a threat to the security of the Western world. As a result, in order to disrupt these novel platforms, tech giants have initiated deployment of decentralized data storage to protect personal data of their users, thus creating a competitive advantage. Second, such steps were backed up with social mobilization that have given rise to anew generation of non-profits such as data unions and independent identity providers.

Undesirable future

Governments will utilize the problem of personal data protection as a reason to increase state control over all personal data. The most efficient way to do this is to tightly couple hardware, software and "dataware" on our devices. In 2020, smartphones were fitted dedicated AI chips. By 2040 devices will receive another dedicated chip and come preinstalled with software as part of obligatory

national certification. This chip will be permanently busy with maintaining the official digital twin of the device. The digital twin is a digital representation of everything which is happing with and on the device, including detected surroundings and data gathered from them.

Data is an asset whose management will become far too difficult for individuals to mage themselves. Therefore, governments will centralize all personal data under the flag of a national data service running digital twins and protecting individuals and their networks from criminal activity. Alternative storage solutions will be considered to be an attack on the state and will be blocked since they give alternative identities. Intelligence services will be able to access all the information at once. Endless data breaches will feed the black market of personal data. Anonymity will become impossible because patterns of online behavior will unambiguously identify an individual, similar to the way fingerprints are currently used. Access to digital fingerprints will be a question of state security, which is another reason for governments to enforce centralized personal data storage. Moreover, national data protection laws will enforce control over cross-border data transfer and access. It will accelerate the process of fragmentation of the Web into national sub-networks.

Wildcards and early warnings. The undesirable future stems from a fear of losing cyberwars in the future. This is a natural response to the overwhelming complexity of future war scenarios. The military instinctively acts overprotectively, therefore any potential threats add to this fear. This works like a pressure-cooker, and the increased pressure on security issues will fan the flames of the undesirable future. In this context, global climate change will only add to this pressure by forcing people to move away from regions with unbearable climate conditions. The sheer number of climate refugees will trigger social unrest towards newcomers, thus propagating the adoption of global surveillance.

The recent allegations that Huawei is providing the Chinese government with a backdoor to their citizens'

devices is an early warning. In terms of technological and economic development the US government can afford to ban Huawei products from being sold to US telecom companies. But what happens in other, less developed parts of the world? Huawei is difficult to beat pricewise. This makes their products far more difficult to be rejected by many users in the Global South.

Recent news about the release of a smartphone from ByteDance, the Chinese company behind TikTok, is another similar warning. The company has already been accused of cooperating with the Chinese government and violating children's data use policy. The use of proprietary hardware will ensure that the company will be able to harvest user data regardless of the software they use.

The future we already have (the most likely scenario)

"The future is already here – it's just not very evenly distributed" goes the famous quote from William Gibson. Life in New York will likely look very futuristic in comparison with rural Mongolia. In other words, developed countries will be the first to start reclaiming personal data. There is already a growing number of cases of individuals applying their right to obtain their personal data from Facebook and Microsoft. European GDPR will be the model and target for many countries in the next 20 years.

Facebook makes most of its profits from its Western audiences – individuals with democratic values and the open market. Therefore, an opportunity to make money from the users' own data will likely drive the development of improved protection of data ownership rights.

China will bring technologies of state surveillance to the developing world to create the second-highest data pyramid in the world after Google. Local governments will harvest data from their citizens, while China will harvest data from all of them. The use of personal data for political activism will be strictly controlled.

What we don't currently know is the potential impact of 5G technology[10] and the Internet of Things. The latter refers

to the idea that any electrical device can be augmented with a Web interface and, consequently, can be connected to other devices via the internet. Therefore, together they will increase the volume of personal data by several magnitudes. The volume of data generated by endless interconnected devices will require cutting-edge computing power, which means that the data will not be sent to data centers for storage and processing. Instead, it will be stored and processed close to the location where it is needed. This may also retransform the infrastructure into a more decentralized one.

Conclusions

The Web will celebrate its 51st anniversary in 2040. We can hope that by that time, Tim Berners-Lee will have come round to the view that the Web has not failed humanity but that it has, in fact, empowered individuals and given them the ability to leverage the power of their own data.

GDPR is a reality and there are growing numbers of cases of individuals reaching for it to request their data from apps. However, at the same time, the legislation creates difficulties for everyone working with personal data. In this context, personal data storage controlled by individual users is a way of solve the legal complexities of moving personal data around.

Once you have data, you can use it for your own or collective goals. Granting or withdrawing access to data is a new type of collective action which is not yet available. Combining personal data to build an even higher volume of data will foster and develop horizontal links within communities. Last but not the least, the monetization of personal data can create a source of basic income. All in all, once individuals are owners of their own data, civil society has the potential to become much stronger.

doi: 10.24412/cl-35945-2021-1-112-127

Endnotes

1 Pirsig M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values* (Phaedrus #1). NYC: William Morrow & Inc.

2 Berners-Lee T. (1990). Information Management: A Proposal. *CERN*. URL: https://www.w3.org/History/1989/proposal.html (retrieval date 22.08.2020).

3 Berners-Lee T. (2018). One Small Step for the Web.... Medium.com. URL: https://medium.com/@timberners_lee/one-small-step-for-the-web-87f92217d085 (retrieval date 22.08.2020).

4 Rowley J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*. no. 33(2). C. 163–180. URL: https://doi.org/10.1177/0165551506070706.

5 Time (2019) "Unfortunately I believe the class divide in the future will be data," says Dr. Naveen Rao at the #TIME100Health Summit. "And if you're not careful, those who have access to data will have better health than those who don't have access to data." Twitter. 17 October 2019. URL: https://twitter.com/TIME/status/1184908321666547712?s=17 (retrieval date 22.08.2020)

6 Facebook.com (2018). *Reports Fourth Quarter and Full Year 2018 Results*. URL: https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q4/Q4-2018-Earnings-Release.pdf (retrieval date 22.08.2020).

7 Bashyakarla V., Hankey S., Macintyre A., Rennó R., Wright G. (2019). Personal Data: Political Persuasion. Inside the Influence Industry. How it works. *Tactical Tech's Data and Politics team*. URL: https://cdn.ttc.io/s/tacticaltech.org/Personal-Data-Political-Persuasion-How-it-works.pdf (retrieval date 22.08.2020).

8 Tactical Technology Collective (2018). *Geotargeting: The Political Value of Your Location*. URL: https://ourdataourselves.tacticaltech.org/posts/geotargeting/ (retrieval date 22.08.2020).

9 Verborgh R. (2018). Decentralizing the Semantic Web through incentivized collaboration. Ruben Verborgh blog. URL: https://ruben.verborgh.org/articles/incentivized-collaboration/ (retrieval date 22.08.2020).

10 5G is the next generation of wireless networks that will allow bandwidth of up to two gigabits.