Centre on Regulation in Europe

REPORT

June 2020

Jan Krämer
Pierre Senellart
Alexandre de Streel

# MAKING DATA PORTABILITY MORE EFFECTIVE FOR THE DIGITAL ECONOMY

## ECONOMIC IMPLICATIONS AND REGULATORY CHALLENGES

*The project, within the framework of which this report has been prepared, has received the support and/or input of the following organisations: Facebook and OFCOM.*

*As provided for in CERRE's by-laws and in the procedural rules from its "Transparency & Independence Policy", this report has been prepared in strict academic independence. At all times during the development process, the research's authors, the Joint Academic Directors and the Director General remain the sole decision-makers concerning all content in the report.*

*The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE, or to any sponsor or to members of CERRE.*

MAKING DATA PORTABILITY MORE EFFECTIVE FOR THE DIGITAL ECONOMY
Jan Krämer, Pierre Senellart, Alexandre de Streel
June 2020

info@cerre.eu
www.cerre.eu

# Table of contents

## About CERRE

Providing top quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) promotes robust and consistent regulation in Europe’s network and digital industries. CERRE’s members are regulatory authorities and operators in those industries as well as universities.
CERRE’s added value is based on:

- its original, multidisciplinary and cross-sector approach;
- the widely acknowledged academic credentials and policy experience of its team and associated staff members;
- its scientific independence and impartiality;
- the direct relevance and timeliness of its contributions to the policy and regulatory development process applicable to network industries and the markets for their services.

CERRE's activities include contributions to the development of norms, standards and policy recommendations related to the regulation of service providers, to the specification of market rules and to improvements in the management of infrastructure in a changing political, economic, technological and social environment. CERRE’s work also aims at clarifying the respective roles of market operators, governments and regulatory authorities, as well as at strengthening the expertise of the latter, since in many Member States, regulators are part of a relatively recent profession.

## About the authors

**Jan Krämer** is a CERRE Joint Academic Director and Professor at the University of Passau, Germany, where he holds the chair of Internet & Telecommunications Business. He holds a Ph.D. in Economics from the Karlsruhe Institute of Technology, where he also headed a research group on telecommunications markets.

His current research interests include the regulation of telecommunications and Internet markets, as well as digital ecosystems and data-driven business models.

**Pierre Senellart** is a CERRE Research Fellow, and a Professor in the Computer Science Department of the École Normale Supérieure in Paris. Pierre is an alumnus of ENS and obtained his Ph.D. (2007) in computer science from Université Paris-Sud. Before joining ENS, he was an Associate Professor (2008–2013) then a Professor (2013–2016) at Télécom ParisTech.

His research interests focus around practical and theoretical aspects of Web data management, including Web crawling and archiving, Web information extraction, uncertainty management, Web mining, and intensional data management.

**Alexandre de Streel** is a CERRE Joint Academic Director and Professor of European Law at the University of Namur where he manages the Research Centre for Information, Law and Society (CRIDS). His current research interests include the regulation and competition policy in the digital industries as well as the legal implication and the regulation of Artificial Intelligence.

He regularly advises international organisations and national regulatory authorities on regulatory and competition issues. He holds a Ph.D. in Law from the European University Institute.

# Executive summary

This CERRE report scrutinises the economic aspects of data portability in the context of the digital economy, against a background of legal and technical considerations. In particular, the report looks beyond the current requirements of data portability, as provided for by the GDPR. It critically assesses whether additional legal requirements, such as a right to continuous access to personal data, would provide complementary tools for making the right to portability more effective and for establishing an innovative, competitive and fair data ecosystem that benefits all players.

## I. The legal dimension: EU regulatory framework on data portability

The current EU legal framework contains a number of rules that encourage or impose the portability and the sharing of personal and non-personal data. **Some of these rules are horizontal. For personal data, they are covered by the *General Data Protection Regulation* (GDPR) and *competition law.* For non-personal data, they are covered by the *Digital Content Directive* applicable in a B2C relationship and the *Free Flow of Data Regulation* applicable in a B2B relationship as well as *competition law*. The others are sectoral and impose data sharing or portability**. In particular, this includes (i) the financial sector, with the *Second Payment Service Directive* (PSD2), which imposes access to payment account data (and which has been completed in the UK through the Open Banking Programme); (ii) the automotive sector, with the new *Motor Vehicle Regulation* imposing access to some vehicle data; and (iii) the energy sector, with the new *Electricity Directive* imposing access to some customer data.

The European Data Protection Board (EDBP) notes that the right to portability, according to **GDPR Article 20, should be interpreted broadly and should cover both volunteered data (actively and knowingly provided) and observed data, but not inferred or derived data.** However, it remains to be seen whether EU judges will embrace such a broad interpretation. If it is followed, web tracking and clickstream data should also be covered by the right to portability. However, currently these are not routinely included in the data sets that consumers can download pursuant to exercising their right to data portability.

**Tensions can emerge within the GDPR, as the right to portability of personal data promotes the exchange and reuse of data, while its principles - purpose limitation and data minimisation – tend to limit data sharing.** In practice, this means that these principles need to be considered when implementing the right to data portability and need to be articulated prior to the porting of the data. Indeed, the EDBP recommends that the data seeker should inform the data subjects on the purposes for which the ported data will be processed and on the categories of personal data that are adequate, relevant and necessary for these purposes. This will help prevent a breach of these purpose limitation and data minimisation principles. Moreover, if the data seeker realises that more data than necessary were ported for the required purpose, they will have to delete this excess data as quickly as possible to avoid any liability issue.

Some industry stakeholders have raised concerns that data portability may create a **liability issue if the data is misused by the recipient**. The EDBP has also indicated that, insofar as the data giver responds to the request for portability, it acts on behalf of the data subject and should not be responsible for any later infringement potentially committed by the data recipient. Nevertheless, according to the EDBP, the data giver should still establish certain safeguards, such as internal

procedures to ensure that the data actually transmitted matches that whose portability was requested, in respect of the purpose limitation and data minimisation principles.

A more contentious issue around data portability arises with requests relating to **personal data from other data subjects**. Article 20(4) of the GDPR provides that the portability right should not affect the rights of others. Indeed, while the data subject originating the portability request may have given his or her consent to the data seeker, or has concluded a contract with them, this is not the case for the other data subjects whose data could be ported as a result of the exercise of this right. Therefore, the consent of the other data subjects would be required in order to be able to port such data, which significantly complicates consent management.

From the sector-specific provisions on data portability, the PSD2 is the most relevant and interesting in the present context. This is because it complements and extends the B2B portability right under Article 20.2 of the GDPR by compelling banks (the original controllers) to allow direct transmission of the data subjects' personal banking information to third party providers (payment initiation services or account information services). PSD2 goes further than the GDPR because, on the one hand it forces the banks to ensure the technical feasibility of this B2B financial account data portability, while on the other it makes this portability continuous, as data subjects can request personal data at each transaction, facilitated by APIs.

## II. The technical dimension

From a technical perspective, we highlight the **various data models and formats** commonly used in the digital economy. These formats can be roughly categorised as structured, semi-structured and unstructured data. In both the structured and semi-structured cases, file formats only specify a syntactic layer on how information is represented. To make sense of it, it is necessary to know the schema of the data, i.e. what fields and data attributes exist, and what constraints on the data values should be respected. Beyond the syntax (provided by the file format), the schema and the constraints (given by the schema annotations, when available), data needs to be interpreted with respect to specific semantics, which give meaning to data fields and attributes. When data is exchanged between two data controllers using different schemas, it is necessary to transform it from one schema to the other, using schema mappings from the source to the destination. These schema mappings are, most of the time, hand written by data engineers, although there is sometimes the possibility of automated learning from examples.

In almost all cases, **the data needs to be (very) efficiently accessible upon request**. This is true whenever data may be used by the data controller in real-time applications, e.g. for display when a web page is accessed. This means any data item of interest needs to be retrievable with a latency in the order of one second or less. Although traditional SQL systems remain by far the dominant data storage mode, most large technology companies have switched from traditional relational database systems to **NoSQL systems**, which focus on performance, reflecting their extreme needs in terms of latency, data volume, or query throughput. In addition to the core data storage system, there is also often an additional caching layer that stores responses in the main memory in order to react more quickly to common queries.

A **web service, or API**, is a technical interface for accessing data meant to be used by programmes - in particular by third-party software - to introduce novel applications of the data. Although there is no requirement to offer such APIs, they are already commonplace, as they allow data controllers to specify what type of access third-party software can have. They even offer the

possibility of monetising richer forms of data access. In order to be used for accessing personal, potentially private, data, API use needs to be combined with an *access delegation* protocol. This verifies that the call to the API has been authorised by the user whose data is being accessed. The most-commonly used protocol is OAuth 2.0. The output of APIs is usually in the format of JSON files, in a wide variety of schemas, with little to no standardisation between companies.

With respect to data transfer, **Personal Management Information Systems (PIMSs)** act as a separate data controller, with direct exchanges from external data controllers to the PIMS. A PIMS may also offer the possibility of pushing the data to other data controllers, in this case acting as a third party between the source and destination data controllers. The PIMS can initiate API calls, control access tokens and implement schema mappings. It is therefore crucial that the user fully trusts the PIMS.

**Technical solutions for standardised data exchange remain in their infancy**. Noteworthy projects include Solid and the Data Transfer Project (DTP). The DTP is a technical initiative launched in 2018 and is supported by - among others - Apple, Facebook, Google, Microsoft and Twitter. The main aim of this initiative is the development of a specification of an open-source platform for data transfer. Although these five companies are nominally involved, the project builds on Google's former *Data Liberation Front*, and Google is by far the main contributor to the DTP platform. When compared to other successful open source projects, both Solid and DTP are still at an early stage of development and have progressed little recently.

We claim that, **in general, there are no strong technical challenges to providing continuous pull- or push-based data exports, with limited delay,** as long as specific solutions are implemented for large, unstructured data sets. The fact that large data controllers provide similar (but incomplete) features through APIs means there are no particular obstacles to implementing them. However, in order to better exploit exported data, data controllers should aim for greater standardisation of data models (e.g. using common RDF schemas). Currently, data exchange capabilities are impeded by the problem of schema heterogeneity. However, assuming this problem can be resolved (either by standardising the data export models or by manually compiled schema mappings), they represent a manageable technical challenge. Data exchange through a trusted third-party (as in a PIMS or the DTP where the hosting entity is on a trusted external host) has the advantage that there is no need to provide access tokens to the original data controllers.

## III. The economic dimension: Impacts of data portability on innovation

From an economic perspective, although data consumption is non-rival, **observed user data collection (as opposed to volunteered user data) is rival.** This is because for key services (such as search, or social networking) the market is concentrated with only a few firms able to track user activity across the web. Thus, observed data is not ubiquitously available, and it is also usually neither feasible nor socially desirable to duplicate the collection of the same observed data. This would mean that users would have to conduct the same search, the same post or the same purchase on several platforms, leading to even more web trackers being built into the websites that we visit. Thus, although rivalry in data collection is not a problem per se, it does provide a strong rational for sharing data.

The more prevalent sharing of 'raw' user data will likely render the market for data intermediaries - which simply acquire and sell raw data, but do not offer further advanced analytics on it - more competitive and possibly unprofitable. However, this does not destroy the incentives to compete on

the basis of data-derived insights. Indeed, **as raw data becomes more prevalent, the focus of competition is likely to move from collection to analytics, which is more likely to stimulate, rather than stifle, innovation**. Furthermore, as data collection is highly concentrated and the services through which (observed) data is collected usually exhibit strong network effects, stronger competition at a data analytics level seems much more feasible and desirable than at the data collection level.

Having access to greater quantities of data (e.g. both volunteered and observed data) will, in many applications, yield a better quality of inferred data (i.e. the actionable knowledge) and thus offer higher profit opportunities for firms. Therefore, the application scope of data portability - i.e. whether restricted to volunteered data or also encompassing observed data - is also crucial from an economic perspective.

Data portability lowers switching costs and facilitates multi-homing. Moreover, widespread data portability, particularly if it occurs on a continuous basis and includes observed data, can facilitate algorithmic learning outside of the organisation where the data was created. The advantage of data portability is that personally identifiable data can also be transferred, and thus there is no trade-off between competition and privacy, which is inherent to access requests that are not user-initiated. At the same time, however, it is unlikely that all users will initiate a transfer of their data. Thus, the data set that is ported under data portability is likely to be more detailed on specific data subjects, but less representative of the user base as a whole. Whether such a data set is useful for a competing or complementing firm, is context specific and depends on the extent to which consumers make use of data portability. However, data portability does not alleviate consumer lock-in due to network effects; this would require some form of interoperability of services.

Irrespective of the extent and mode of data portability, **we do not think that data portability will lead to greater or less competition and innovation in established digital markets per se. It may, however, spur innovation in complementary and new digital markets**. Widespread data portability could make it possible that innovation at the service level and innovation at the analytics level occur independently, i.e. within different organisations. Thanks to the non-rivalry of data, this would not mean that the current data controllers will lose access to the data, and can thus continue to be innovative at both the service *and* the analytics level. This lends itself to the hypothesis that user-induced data portability may increase the innovativeness of digital markets, rather than stifle it. However, although there is some tentative empirical evidence from Open Banking in the UK, currently there is a lack of empirical studies testing this hypothesis or other economic effects specific to data portability.

## IV. The economics of Personal Information Management Systems (PIMSs)

The central premise of a **PIMS for users is that it offers a centralised dashboard** that seamlessly integrates with the various services that they are using, offering key functionalities such as identity management, permission management and data transfers. This **requires a common set of de facto standards and high-performance APIs**, through which a PIMS would be able to access the various services and users' data. To date, however, such common standards are lacking.

Furthermore, even if we look beyond the need for standards and API access to connect the various data sources of a user in a centralised PIMS, the question arises of **how the business model of a privately-financed 'neutral' data broker can ever be made sustainable**. We find that common business models that seek to generate revenues from: i) data markets by selling users' data, ii) users directly, via a subscription model, or iii) data controllers by offering a compliance service are either not feasible or are unlikely to see widespread adoption. Specifically, a number of PIMSs that set out to monetise personal data on behalf of their users have failed in the recent past. Paying users for their data also gives rise to an ethical issue, as such a practice would quickly reveal that the data of some users is more valuable than others.

## V. Increasing the effectiveness of data portability

To date, there is limited evidence that data portability is widely used in digital markets, and thus there is scope to make it more effective. To this end, we have developed policy recommendations in three areas.

1. The first set of recommendations entails **more effective enforcement and legal certainty on existing legal frameworks for data portability**. Here, a first priority for policymakers is to **increase the legal certainty on the scope and the limits of data portability** under Article 20 of the GDPR in the context of digital markets. In particular, it should be clarified to what extent observed data - including tracking and clickstream data - is to be included. It should also clarify whether there is an obligation to ask consumers for consent regarding the transfer of other data subjects that may concern them.

We realise that at some point, these questions can become so complex that a case-by-case analysis will be necessary. Here, it should be clear what the main interests of the trade-offs are and where organisations and consumers can find legal guidance on balancing those trade-offs in a timely manner. In these cases, providers that are willing to facilitate data portability for consumers should be able to receive specific guidance by the privacy regulator in a cooperative approach. In this context, it is also worthwhile to discuss the use of sandbox regulation, as is the case in Open Banking, in order to provide a safe harbour under which data portability can be developed further.

A second priority is for **greater transparency on the categories and extent of personal data** that firms in the digital economy hold on a certain data subject. This information should be readily available to users before any formal access request (Art 15(3) GDPR) or data portability request (Art. 20 GDPR) is initiated. Data subjects already have these rights under Art. 12 – Art. 15 GDPR, but currently there still seems, in some cases, to be a lack of transparency over the actual extent of data collection pertaining to each data subject (e.g. on the extent of tracking data).

A third priority is for **more effective monitoring and enforcement of the existing provisions on data portability** under GDPR. This requires that the scope and the limits of these provisions are clear in the context of the digital economy (primary priority), and that users are well aware about which data is available about them and can be ported (secondary priority).

2. Next, we argue for **investigating the need and feasibility of a new, proportionate rule that enables consumers to transfer their personal and non-personal data in a timely and frequent manner**, from their existing digital service provider to another one at any time. This is what we refer to as 'continuous data portability'. As there is a possibility that such a regulation amplifies the legal and economic risks and trade-offs inherent to data portability, it is vital that the previously raised legal uncertainties are thoroughly addressed in advance. The scope of data to be ported under such continuous data portability should match that under GDPR Article 20. Moreover, in accordance with the proportionality principle, the obligation to implement and enable continuous

data portability should only apply when the benefits are likely to outweigh the costs; it should not be overly burdensome for small and emerging digital service providers.

**Continuous data portability requires a dialogue and code of conduct on common data standards and APIs**. We believe that standardised APIs that enable continuous data portability are a prerequisite for encouraging more firms to import personal data, and for encouraging more consumers to initiate such transfers. Ultimately, this is likely to spur innovation and competition in digital markets, although it is unlikely to disrupt existing market structures.

This will echo ongoing policies in the UK and Australia, and we believe that the European Commission - in its Data Strategy - should follow suit. We therefore propose first attempting a participatory, adaptive and soft approach, similar to what was done in the Free Flow of Data Regulation. If there is insufficient progress made in establishing standards and operational interfaces within a specified period, it may require stronger governmental intervention or guidelines to ensure progress is made and that neutrality of interests are warranted, as was the case for PSD2 and Open Banking.

3. Finally, **in order to enable a centralised consent management through PIMS, additional standards need to be agreed above and beyond those needed for data transfers**. We think that the importance of this should not be underestimated, because it is crucial that consumers are aware of their given consents and are able to exercise their rights with little to no transaction costs, particularly if this is the basis on which data is being shared between firms.

We also expect that, if such standards are in place, there will be considerable **development in open-source communities**, providing decentralised, non-profit solutions. Given the potentially sensitive nature of the data being handled through PIMS, public oversight may still be necessary, such as through **privacy seals and certification**.

To achieve critical mass for PIMSs, one fruitful avenue may be to build a user base from existing (or developing) identity management solutions. In particular, the EU could be more active in **encouraging PIMSs by coupling development of its consent standards more closely** to its ongoing efforts for a joint European identity management solution.

01

# INTRODUCTION

# 1. Introduction and background[1]

Following a prolonged political tug-of-war, the new European General Data Protection Regulation (GDPR) came into force on 25 May 2018. It replaced the previously applicable European Data Protection Directive (95/46) from 1995. This originated in a time when personal data did not have its current economic and societal importance. Nowadays, of course, massive amounts of personal data are routinely collected, analysed and monetised - particularly in the digital economy – making use of advanced (big) data analytics.

From both a legal and economic perspective, there are important elements in the GDPR that strengthen the rights of data subjects. Of particular note is the "right to data portability" under Article 20 of the GDPR, which is intended to allow users to exercise greater control over their personal data. This is also supposed to enable users to counteract lock-in effects when using digital services and to facilitate switching to alternative content or service providers.

However, the GDPR was built on the premises of fundamental rights, but not on the premises of competition law, despite the fact that the right to data portability may have competition-enhancing effects. In fact, there has been little research to date on the actual economic effects that come about with the right to personal data portability. There are several reasons for believing that the economic effects of data portability are much more complex than simply a reduction in switching costs, as initially anticipated.

This CERRE report scrutinises these economic aspects of data portability against a background of legal and technical considerations. In particular, the report looks beyond the current requirements of data portability as provided by the GDPR. It critically assesses whether additional legal requirements, such as a right to continuous access to personal data, would be complementary in making the right to portability more effective. It also examines how to establish an innovative, competitive and fair data ecosystem that benefits all players.

For example, the GDPR and the right to data portability apply horizontally, independent of any measures of market power. Therefore, the right to data portability may, in some circumstances, reinforce the market power of some data-rich firms, while their smaller competitors are likely to suffer more from unique user data being ported to incumbents than vice versa. Moreover, the right to data portability does not encompass a continuous porting of data (facilitated, e.g. by Application Programming Interfaces or APIs). In addition, it is currently not entirely clear to what extent users can port their observed data (i.e. data implicitly given, such as via clicks and location). However, continued access to volunteered (i.e. data explicitly given) *and* observed data may be crucial in stimulating competition and the emerging landscape of Personal Information Management Systems (PIMS), whose vision is to offer users a centralised dashboard for monitoring and controlling the flow of their personal data.

The topic of how the porting of data from one provider to another would affect innovation incentives and stimulate entry is one that also creates controversy. Some argue that, with frictionless data portability, there is a concern that the quality of content and service offers (e.g. financed by advertising) will decline. This is because data portability will allow more companies

[1] The authors are grateful to (in alphabetical order) Malte Beyer-Katzenberger, Marc Bourreau, Richard Feasey, Claire-Marie Healy, Bertin Martens, and Thomas Tombal for their helpful comments and suggestions.

access to the same personal data, intensifying the competitive situation on the data and advertising market. In other words, third parties could act as free riders on the data market, which could ultimately also harm the customer. The contrary argument says that the free flow of personal data (with safeguards to mitigate privacy and security risks and with the users' consent) would stimulate innovation, as data would be freed from its existing silos and a much larger set of minds could gain. More generally, the OECD (2019)[2] noted that data access and sharing is estimated to generate social and economic benefits worth between 0.1-1.5% GDP - in the case of public-sector data - and between 1.0-2.5% of GDP (as high as 4.0% in some studies) if private-sector data is included.

We also investigate legal and technical questions arising from the context of more widespread data portability; for example, dealing with conflicts in fundamental rights of others if entire profiles are to be ported in social networks; or regarding the compatibility of data formats. From an economic perspective, we highlight some of the potential complex economic trade-offs arising from data portability. For example, committing to let the consumers port their data may make them more willing to send data in the first place, increasing consumer lock-in.

We also examine ongoing developments of technical solutions for facilitating data portability beyond the current legal standards. This includes the *Data Transfer Project* (DTP), which is particularly interesting, as it is supported by major, data-rich companies such as Facebook, Google, Microsoft and Twitter. The expressed aim of this project is to build "a common framework with open-source code that can connect any two online service providers, enabling a seamless, direct, user-initiated portability of data between the two platforms." The project intends to build on "existing APIs and authorisation mechanisms to access data" and then "uses service specific adapters to transfer that data into a common format, and then back into the new service's API." In other words, the platforms within the DTP become - to an extent - interoperable, allowing immediate and eventually continuous exchange of user data back and forth between platforms.

However, DTP is not the only project of its kind seeking to facilitate the exchange and transfer of data between data holders. The *Solid* project, for example, which was founded by Tim Berners-Lee, envisions a new internet built on "linked data", in much the same sense as the internet was envisioned by Berners-Lee as being "linked documents", so-called hypertext. However, despite the importance of the subject, all of these projects are still relatively small in size and relevance, as we highlight in the report.

Ultimately, the goal of this CERRE project is to identify *how personal data portability can be made more effective*, in order to truly empower consumers in the use of their data and to use the services they want.

[2] OECD (2019). Enhancing Access to and Sharing of Data. Reconciling Risks and Benefits for Data Re-Use Across Societies. OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en.

**The remainder of the report is organised as follows:**

- In Section 2, we summarise the EU legal framework on data portability, and more generally on data access and data exchange, focussing particularly on personal data.
- In Section 3, we discuss technical aspects of data portability and data exchange, including standards, Personal Information Management Systems (PIMSs) and other ongoing software projects, such as the Data Transfer Project.
- In Section 4, we highlight the relationship between data portability and the competitiveness and innovativeness of digital markets.
- In Section 5, we focus on the economic aspects of PIMSs.
- In Section 6, we draw conclusions from our legal, technical and economic discussion, developing policy recommendations to help make the portability right of the GDPR more effective and to truly empower users in a future European personal data ecosystem.

02

# THE EU LEGAL FRAMEWORK ON DATA PORTABILITY

# 2. The EU legal framework on data portability[3]

Although this report is focused on the portability of personal data, we also mention in this section the rules on non-personal data because the distinction between personal and non-personal data is not always easy to draw in practice and some predict that it will become increasingly difficult in the future with the advancement of big data analytics and the increased technical ability to re-identify a person through the available data points.[4] The EU legal framework contains several rules imposing or encouraging the portability and the sharing of personal and non-personal data. Some rules are horizontal and are mainly composed of:

- for personal data, the *General Data Protection Regulation* (GDPR)[5] and *competition law*;
- for non-personal data, the *Digital Content Directive* (DCD)[6] applicable in a B2C relationship and the *Free Flow of Data Regulation* (FFDR)[7] applicable in a B2B relationship as well as *competition law*.

Other rules are sectoral and impose data sharing or portability in particular in:

- the financial sector with *Second Payment Service Directive* (PSD2) imposing access to payment account data, which has been completed in the UK with the Open Banking Programme;[8]
- the automotive sector with the new *Motor Vehicle Regulation* imposing access to some vehicle data;[9]
- the energy sector with the new *Electricity Directive* imposing access to some customers data.[10]

Those rules and incentives to share and port data may increase in the future. The Commission has clearly indicated in its new Data Strategy adopted in February 2020 that it wants to promote the sharing of non-personal data within EU common data spaces.[11]

[3] This section is partly based on I. Graef, T. Tombal and A. de Streel, *Limits and Enablers of Data Sharing : An Analytical Framework for EU Competition, Data Protection and Consumer Law*, TILEC Discussion Paper 2019-024, November 2019.

[4] According to GDPR, art. 4(1): "personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." On the distinction between personal and non-personal data, see Commission Guidance of 29 May 2019 on the Regulation on a framework for the free flow of non-personal data in the European Union, COM (2019) 250, p.4-5

[5] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46 (General Data Protection Regulation), OJ [2016] L 199/1.

[6] Directive 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services, OJ [2019] L 136/1.

[7] Regulation 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, OJ [2018] L 303/59.

[8] Directive 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market, OJ [2015] L 337/35, arts.66-67.

[9] Regulation 2018/858 of the European Parliament and of the Council of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, OJ [2018] L 151/1, arts.61-66.

[10] Directive 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity [2019] OJ L 158/125, art.23.

[11] Communication from the Commission of 19 February 2020, A European strategy for data, COM(2020) 66, p.13.

**Table 1: EU legal framework for data portability and sharing**

| | Personal data | Non-personal data |
|---|---|---|
| **Horizontal** | - GDPR (2016)<br>- Competition law | - DCD (2019) in B2C<br>- FFDR (2018) in B2B<br>- Competition law |
| **Sector-Specific** | - Financial: PSD2 (2015) and UK Open Banking (2016)<br>- Automotive: Motor Vehicle Regulation (2018)<br>- Energy: Electricity Directive (2019) | |

## 2.1 *GDPR*

***Article 20 - Right to data portability***

*1.The data subject shall have the* ***right to receive the personal data*** *concerning him or her, which he or she has provided to a controller,* ***in a structured, commonly used and machine-readable format*** *and have the* ***right to transmit*** *those data to another controller without hindrance from the controller to which the personal data have been provided, where:*
*(a) the processing is based on consent pursuant to point (a) of Article 6(1) or point (a) of Article 9(2) or on a contract pursuant to point (b) of Article 6(1);*
*and (b) the processing is carried out by automated means.*

*2.In exercising his or her right to data portability pursuant to paragraph 1, the data subject shall have the* ***right to have the personal data transmitted directly*** *from one controller to another,* ***where technically feasible****.*

*3.The exercise of the right referred to in paragraph 1 of this Article shall be without prejudice to Article 17. That right shall not apply to processing necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.*

*4.The right referred to in paragraph 1 shall not adversely affect the rights and freedoms of others.*

### *2.1.1 The two data portability rights*

Data portability aims to strengthen the data subject empowerment, i.e. the power of control that the data subjects have on their own personal data and to re-balance the relationship between data subjects and data controllers.[12] To do that, two specific rights are given to the data subjects:

- First, a data subject has the right to receive the personal data concerning him which he has provided to a controller (the data giver) and to transmit those data to another controller (the data seeker) in a **B2C2B relationship** (art. 20(1) GDPR). For instance, a data subject can receive his current playlist from a music streaming service to find out how many times he listened to specific tracks or to check which music he wants to purchase and to port it to another platform to listen music from there.[13]
- Second, a data subject has also the right to have his personal data transmitted directly from one controller to another in a more direct **B2B relationship** (art. 20(2) GDPR). In essence,

[12] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, WP242 rev.01, p. 4.
[13] *Ibid.*, p. 5.

this means that a data seeker can import data directly from the data giver with the consent of the data subject.

The first portability right (B2C2B) is the strongest as it should be exercised without hindrance from the data giver. According to the European Data Protection Board (EDBP), such hindrance could be 'fees asked for delivering data, lack of interoperability or access to a data format or API or the provided format, excessive delay or complexity to retrieve the full dataset, deliberate obfuscation of the dataset, or specific and undue or excessive sectorial standardisation or accreditation demands'.[14] The second portability right (B2B) is weaker as it can only be exercised when technically feasible, which is assessed on a case-by-case basis. In this regard, the *Data Transfer Project* aims to create an open source platform allowing the direct portability of data between the participating data controllers (see Section **0**).

Those two (new) portability rights complement than the (old) **data access right** given by Article 15(3) of the GDPR which provides that:

> *The controller shall provide a copy of the personal data undergoing processing. For any further copies requested by the data subject, the controller may charge a reasonable fee based on administrative costs. Where the data subject makes the request by electronic means, and unless otherwise requested by the data subject, the information shall be provided in a commonly used electronic form*.

The objectives of those different rights are distinct and complement each other: while the portability right aims at facilitating the technical re-use of the data and preventing user lock-in, the data access right aims at empowering the data subject by allowing him to understand what is done concretely with his data. Moreover, as explained in the following sub-sections, the scope of those different rights is also different: while the portability right is limited to personal data provided by the data subject, the right to data access applies to all personal data.[15]

### 2.1.2 Data covered

The scope of the portability right is limited to certain categories of personal data. The GDPR mentioned the data provided by the data subject. In its interpretative Guidelines, the EDPB mentions three categories of data:[16]

- The *data actively and knowingly provided* by the data subject such as name, age, email address;
- The *observed data* provided by the data subject by virtue of the use of the service or the device, such as search history, traffic and localisation data, the heartbeat tracked by a wearable device;
- The *inferred data and derived data* created by the data controller on the basis of the data provided by the data subject such as the outcome of an assessment regarding the health of a user or the profile created in the context of risk management and financial regulations to assign a credit score.

[14] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p.15.

[15] On those differences, see T. Tombal, "Les droits de la personne concernée dans le RGPD", In *Le règlement général sur la protection des données (RGPD/GDPR): analyse approfondie*, Larcier, 2018, p.514-515.

[16] *Ibidem*, p.10.

The EDPB notes that the portability right should be interpreted broadly and should cover the first two categories, i.e. the data that have been actively provided by the data subject but also the observed data, and only the third category (the inferred data) should not be covered. However, it remains to be seen whether the EU judges will follow such broad interpretation.

If this interpretation is followed, web tracking data should be covered by the portability right as this is observed data. For such data, the data controller is the legal person that determines the purposes and the means of the processing of the web tracking data which could be the website or the third-party company depending of the specific circumstances of the case. The e-Privacy Directive is complementary to the GDPR as it imposes prior information obligation by the website, which could then be used by the data subject to exercise his portability right.

The scope of the portability right is also limited by the type of processing and covers only personal data whose **processing is based on consent or a contract**. There is thus no general right to data portability as it does not apply to processing operations necessary for the performance of a task in the public interest vested in the controller, nor to processing operations necessary for the compliance with a legal obligation to which the controller is subject. For instance, a financial institution has no obligation to respond to a portability request relating to personal data that has been collected in the context of the compliance with its legal obligation to fight money laundering.[17]

Finally, the right to data portability only applies if the data **processing is carried out by automated means**, and therefore does not cover most paper files.

### *2.1.3 Conditions for the data transfer*

Article 20 of the GDPR imposes that the data have to be provided in a **structured, commonly used and machine-readable format**.[18] Recital 68 of the GDPR clarifies further that data controllers are encouraged to develop interoperable formats that enable data portability. According to the EDBP, "the terms structured, commonly used and machine-readable are a set of minimal requirements that should facilitate the interoperability of the data format provided by the data controller. In that way, 'structured, commonly used and machine readable' are specifications for the means, whereas interoperability is the desired outcome". However, such interoperability goal should not go as far as imposing technical compatibility, as it is clarified by Recital 68 of the GDPR.

According to the EDPB, 'the most appropriate format will differ across sectors and adequate formats may already exist, and should always be chosen to achieve the purpose of being interpretable and affording the data subject with a large degree of data portability. As such, formats that are subject to costly licensing constraints would not be considered an adequate approach' and 'where no formats are in common use for a given industry or given context, data controllers should provide personal data using commonly used open formats (e.g. XML, JSON,

[17] *Ibidem*, p. 8.
[18] Machine-readable format is not defined in the GDPR but is defined in the Open Data Directive as 'a file format structured so that software applications can easily identify, recognise and extract specific data, including individual statements of fact, and their internal structure': Art.2(13) of the Directive 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, OJ [2019] L 172/56.

CSV,…) along with useful metadata at the best possible level of granularity, while maintaining a high level of abstraction.'[19]

The EDBP also encourages cooperation between industry stakeholders and trade associations to work together on a common set of interoperable standards and formats to deliver the requirements of the right to data portability as is done by the European Interoperability Framework (EIF)[20] which creates an agreed approach to interoperability for organisations that wish to jointly deliver public services.[21]

In addition, Article 12(3) of the GDPR requires that the data giver provides information on action taken to the data subject **without undue delay and in any event within one month** of receipt of the request. This one-month period can be extended to a maximum of three months for complex cases if that the data subject has been informed about the reasons for such delay within one month of the original request.

Article 12(5) of the GDPR provides that data should be ported **free of charge**, unless the data controller can demonstrate that the requests are manifestly unfounded or excessive, in particular because of their repetitive character. In this case, the controller may either charge a reasonable fee taking into account the administrative costs of porting the data or refuse to port the data.

The EDPB notes that: 'for information society services that specialise in automated processing of personal data, implementing automated systems such as Application Programming Interfaces (APIs) can facilitate the exchanges with the data subject, hence lessen the potential burden resulting from repetitive requests. Therefore, there should be very few cases where the data controller would be able to justify a refusal to deliver the requested information, even regarding multiple data portability requests'.[22] The EDBP also specifies that 'the overall system implementation costs should neither be charged to the data subjects, nor be used to justify a refusal to answer portability requests'.[23]

### *2.1.4 Limits and safeguards*

Tensions can appear within the GDPR, as the personal data portability right promotes the exchange and reuse of data, while its principles of purpose limitation and data minimisation limit data

[19] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p.17-18.
[20] https://ec.europa.eu/isa2/eif_en
[21] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p.18.
[22] *Ibidem*, p.15.
[23] Moreover, some commentators noted that: Given that the right to data portability should not be considered solely as a "one-shot" - as it does not entail an obligation for the data holder to erase the ported data (Art. 20.3) - the "repetitive character" of the request should not be interpreted too strictly, otherwise this right will be deprived of its effectiveness. Thus, the mere fact of renewing the request to port the same data a second time, for example to forward them to a third controller, should not be sufficient to conclude that the request is "repetitive". In doing so, it will be necessary to make a case-by-case assessment of the repetitive or non-repetitive nature of the request, and therefore of the possibility for the controller to claim payment. Similarly, if the data subject exercises his right to portability again, towards the same controller, in order to port the data that has been updated since the first request (new data, modified data, etc.), this request should not be considered as repetitive as, by assumption, the data concerned will be different from the data ported the first time: T. Tombal, "Les droits de la personne concernée dans le RGPD", In *Le règlement général sur la protection des données (RGPD/GDPR): analyse approfondie*, Bruxelles, Larcier, 2018, p.504.

sharing.[24] In practice, this means that these two principles have to be considered when implementing the data portability right. This articulation has to be done prior to the porting of the data.

Indeed, the EDBP recommends that the data seeker should inform the data subjects about the purposes for which the ported data will be processed and about the categories of personal data that are adequate, relevant and necessary for these purposes, in order to prevent a breach of these purpose limitation and data minimisation principles.[25] Moreover, if the data seeker realises that more than necessary data were ported for the purpose pursued, he will have to delete this excessive data as soon as possible, in order to avoid any liability issue.

This clarifies one of the uncertainties regarding the **liability faced by the data givers**, namely whether there is a risk that they might be found liable for the unlawful processing of the ported data made by the data seeker because of a breach of these purpose limitation and data minimisation principles. Such a concern has been raised, among others, by Facebook in its White Paper on Data Portability and Privacy.[26] This uncertainty stems from the fact that the GDPR does not tackle this issue. The EDBP has indicated that insofar as the data giver responds to the request for portability, it should not be held liable as a result of the processing carried out on the data by the data seeker.[27] Indeed, the data giver acts on behalf of the data subject and should not be responsible for any later infringement potentially committed by the data seeker. Nevertheless, according to the EDBP, the data giver should still set up certain safeguards, such as internal procedures to ensure that the data that is actually transmitted matches the data whose portability is requested, in light of the purpose limitation and data minimisation principles.[28]

These two principles will also have to be considered in order to limit the **porting of personal data from other data subjects** than the one exercising his data portability right. Article 20(4) of the GDPR provides that portability shall not affect the rights and freedoms of others. Accordingly, when a data subject exercises his right to data portability, it is necessary to ensure that the personal data of other data subjects, who have not given their consent to such portability, are not transmitted if the data seeker is likely to process the personal data of such third parties.[29] Indeed, while the data subject at the origin of the portability request has given his consent to the data seeker or has concluded a contract with him, this is not the case for the other data subjects whose data could be ported as a result of the exercise of this right.[30]

Given that the third parties in question have not consented to the transfer of their data to the data seeker, this transfer can only take place if the purpose for which the transfer is made is compatible with the data giver's initial purpose of processing.[31] If this is not the case, the data seeker has to

[24] Resp. Article 5(1b) and (1c) of the GDPR.
[25] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p. 13.
[26] See Question 5 in Facebook (2019). Charting a Way Forward: Data Portability and Privacy. White Paper. Available at: https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf
[27] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p. 6.
[28] Ibidem.
[29] *Ibidem*, p. 11.
[30] See also Question 3 in Facebook (2019). Charting a Way Forward: Data Portability and Privacy. White Paper. Available at: https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf
[31] Article 5 (1b) of the GDPR.

rely on a new lawful basis for the processing of these third parties' personal data, such as the legitimate interests basis.[32]

In order to avoid such an issue, the EDBP suggests that the processing of these other data subjects' personal data should be authorised only insofar as these data remain under the sole control of the data subject requesting the portability, and that they should only be processed for the purposes determined by this data subject.[33] The data seeker could therefore not process these third parties' data for purposes that he has defined himself, such as prospecting purposes. Moreover, the data seeker could not process these data for purposes that are not compatible with the purposes of the data giver. While being appealing in theory, this suggestion is nevertheless extremely restrictive and provides little interest for the data seeker, whose margin of manoeuvre will be severely limited.

However, the EDBP makes another suggestion that is more interesting. It invites both the data giver and data seeker to implement technical tools allowing the data subject to select the personal data he wishes to port, while excluding, where possible, the personal data of other data subjects.[34] This makes it possible to avoid, upstream, a potential infringement of the rights of these third parties. However, this is not sufficient in itself, as some personal data of third parties might necessarily have to be ported. Accordingly, in addition to these technical tools, it must also be reflected on the implementation of consent mechanisms for these other data subjects, in order to facilitate data portability.[35] Once again, the difficulty is the practical implementing of such a mechanism. For example, in the banking sector, it would be nearly impossible to obtain the consent of all the persons appearing in a list of banking transactions that a data subject would like to port to another bank.

[32] Article 6(1.f) of the GDPR.
[33] Guidelines of 13 April 2017 of Working Party 29 on the right to data portability, p. 12.
[34] Ibidem.
[35] Ibidem.

## 2.2 *Digital Content Directive*

> ***Article 16 - Obligations of the trader in the event of termination***
>
> *3. The trader shall refrain from using any content other than personal data, which was provided or created by the consumer when using the digital content or digital service supplied by the trader, except where such content:*
> *(a) has no utility outside the context of the digital content or digital service supplied by the trader; (b) only relates to the consumer's activity when using the digital content or digital service supplied by the trader;*
> *(c) has been aggregated with other data by the trader and cannot be disaggregated or only with disproportionate efforts; or*
> *(d) has been generated jointly by the consumer and others, and other consumers are able to continue to make use of the content.*
>
> *4. Except in the situations referred to in point (a), (b) or (c) of paragraph 3, the trader shall, at the request of the consumer, **make available to the consumer any content other than personal data**, which was provided or created by the consumer when using the digital content or digital service supplied by the trader.*
>
> *The consumer shall be **entitled to retrieve that digital content free of charge**, without hindrance from the trader, within a reasonable time and **in a commonly used and machine-readable format**.*
>
> *5. The trader may prevent any further use of the digital content or digital service by the consumer, in particular by making the digital content or digital service inaccessible to the consumer or disabling the user account of the consumer, without prejudice to paragraph 4.*

### *2.2.1 Data retrieval right*

Like personal data protection law, consumer law also enables data portability, notably through Article 16 of the Digital Content Directive (DCD). The Directive applies to all suppliers of digital content or services (i.e. virtually any firm in the digital economy that collects data) when dealing with a consumer (i.e. any natural person who is acting for purposes which are outside that person's trade, business, craft, or profession).[36] The DCD grants a **form of portability right for the non-personal data provided or created by the consumer**. However, this right for consumers does not apply in a number of situations when the content is of little practical use to the consumer, who therefore has a limited interest in the portability of such data, particularly in view of the fact that requiring such a mechanism is costly for the trader.[37]

The DCD solely provides the consumer with a right to retrieve some of its non-personal data but does not allow the direct transmission of data between two traders. Nevertheless, the underlying idea of the DCD is to allow the consumers to retrieve their data in order to then share this data with other traders. This new right ensures that consumers can easily switch content providers, by reducing legal, technical and practical obstacles, such as the inability to recover all the data that the consumer has produced or generated through his use of digital content.[38]

Unlike the GDPR, the DCD provides that, when the consumer terminates the contract, the trader must refrain from using the non-personal data provided or created by the consumer.[39] The fate of

[36] Art.2(6) DCD.
[37] Recital 71 of the DCD.
[38] Recital 70 of the DCD.
[39] Article 16(3) of the DCD. The only exceptions are if the data has no use outside the context of the content or service; if the data only relates to the consumer's activity when using the content or service; if the data has been aggregated with other data by the trader and cannot be disaggregated or can only be disaggregated with disproportionate effort; or if the data has been generated jointly by the consumer and other persons who continue to use them (Article 16(3) of the DCD).

the data held by the original controller/trader therefore differs in the two regimes, as the GDPR does not prevent the original controller from continuing to use the ported data, while the DCD provides that the trader must refrain from using the data in the future unless it has been generated jointly by the consumer and others, and other consumers are able to continue to make use of the content.[40] This difference can be explained by the fact that data can be ported at any time under the GDPR, while data portability is only made possible after the termination of the contract by the consumer in the DCD.

### *2.2.2 Data covered*

While the GDPR applies to personal data that has been provided by or observed on the data subject, the DCD applies to any **content other than personal data, which was provided or created by the consumer** when using the digital content or digital service supplied by the trader. The scope of application of the DCD is thus complementary to that of the GDPR.[41] This is welcome as the distinction between personal and non-personal data might be difficult to draw in practice.

Indeed, given the GDPR's broad definition of personal data and the technological progress in big data and AI for identification, the vast majority of the data provided or created by the consumer are likely to be considered as personal data. In any case, it should be underlined that the "inferred and derived" personal data, which are not considered as data "provided" by the data subject, are neither covered by the GDPR nor by the DCD, and thus cannot be ported.[42]

### *2.2.3 Conditions for the data transfer*

Using similar language than the GDPR, the DCD provides that the data must be returned to the consumer in a **commonly used and machine-readable format**. Regarding the deadline to reply to the request, the DCD only provides that the data should be given to the consumer within a **reasonable time** after the termination of the contract.

While the DCD does not provide any further information as to how these terms must be interpreted, the deadline of one month provided for in the GDPR could arguably be used to assess this reasonable character. Finally, similar to the GDPR, the DCD provides that the consumer shall be entitled to retrieve the data **free of charge**.[43]

[40] Article 16(3d) of the DCD.
[41] This is explicitly stated in art. 16(2) of the DCD, which provides that the trader remains bound by the obligations of the GDPR, which prevails over the DCD in case of conflict (art.3(8) of the DCD).
[42] M. Ledger and T. Tombal, Le droit à la portabilité dans les textes européens: droits distincts ou mécanisme multi-facettes ?, RDTI, N°72, p. 25-44.
[43] Article 16.4 of the DCD.

## 2.3 *Free Flow of Data Regulation*

> ***Article 6 - Porting of data***
>
> *1.The Commission shall encourage and facilitate the development of **self-regulatory codes of conduct at Union level** ('codes of conduct'), in order to contribute to a competitive data economy, based on the principles of transparency and interoperability and taking due account of open standards, covering, inter alia, the following aspects:*
>
> *(a) best practices for facilitating the switching of service providers and the **porting of data** in a structured, commonly used and machine-readable format including open standard formats where required or requested by the service provider receiving the data;*
>
> *(b) **minimum information requirements** to ensure that professional users are provided, before a contract for data processing is concluded, with sufficiently detailed, clear and transparent information regarding the processes, technical requirements, timeframes and charges that apply in case a professional user wants to switch to another service provider or port data back to its own IT systems;*
>
> *(c) approaches to **certification schemes that facilitate the comparison** of data processing products and services for professional users, taking into account established national or international norms, to facilitate the comparability of those products and services. Such approaches may include, inter alia, quality management, information security management, business continuity management and environmental management;*
>
> *(d) communication roadmaps taking a multi-disciplinary approach to **raise awareness** of the codes of conduct among relevant stakeholders.*

The Free-Flow of Data Regulation (FFDR) applies to the porting of non-personal data in B2B relationships. The Regulation instructs the Commission to contribute to the development of **EU Codes of conduct to facilitate the porting of (non-personal) data in a structured, commonly used and machine-readable format including open standard formats**.

On that basis, SWIPO (Switching cloud service providers and Porting Data), which is one of the Digital Single Market (DSM) Cloud Stakeholders Working Groups gathering more than 100 stakeholders, adopted in November 2019 two drafts Code of conduct, whose relevant parts are reproduced in Annex I of this Report.

- one on the **Infrastructure as a Service** (IaaS) market,
- and another on the **Software as a Service** (SaaS) market.[44]

Those Code of conducts will be assessed by the Commission by the end of 2022.[45] In particular, the Commission will focus on: "(i) the impact on the free flow of data in Europe; (ii) the application of the Free Flow of Data Regulation, especially to mixed datasets; (iii) the extent to which the Member States have effectively repealed existing unjustified data localisation restrictions; and (iv) the market effectiveness of codes of conduct in the area of porting of data and switching between cloud service providers."[46]

[44] https://ec.europa.eu/digital-single-market/en/news/presentation-codes-conduct-cloud-switching-and-data-portability

[45] FFDR, art. 8.

[46] Commission Guidance of 29 May 2019 on the Regulation on a framework for the free flow of non-personal data in the European Union, COM (2019) 250, p. 18.

The Commission also expects that the codes of conduct will be complemented by model contractual clauses allowing "sufficient technical and legal specificity in the practical implementation and application of the codes of conduct, which will be of particular importance for SMEs."[47]

## 2.4 Competition law: compulsory access to essential data

***Conditions of essential facilities***[48]

*78. The concept of refusal to supply covers a broad range of practices, such as a refusal to supply products to existing or new customers, refusal to license intellectual property rights, including when the licence is necessary to provide interface information, or refusal to grant access to an essential facility or a network*

*81. The Commission will consider these practices as an enforcement priority if all the following circumstances are present:*
*- the refusal relates to a product or service that is* ***objectively necessary*** *to be able to compete effectively on a downstream market,*
*- the refusal is likely to lead to the* ***elimination of effective competition on the downstream*** *market, and*
*- the refusal is likely to lead to* ***consumer harm****.*

### *2.4.1 Conditions of essential facilities and application to data*

The three main conditions of the essential facilities doctrine, which is the legal standard to impose access in EU competition law, are summarised in the Commission Guidance on the application of Article 102 TFEU to exclusionary abuses of dominant position. When applying those conditions to data, it is important to distinguish at which level of the data value chain access is requested. As explained by Gal and Rubinfeld (2019),[49] this value chain consists of several refinement steps (see also Sections 4.1.1 and 4.1.2 below):

i. first, raw personal and non-personal data are collected directly or bought on a secondary data market;
ii. second, data are structured and turned into information;
iii. third, those structured data are analysed by algorithms and information is turned into knowledge, such as a prediction;
iv. finally, the analysis of the structured data leads to an action such as improving products or offerings.

Going down the value chain, the efforts and investments by the data owner increase and may even be protected by intellectual property rights. Hence, the interest in protecting property and investment incentives, which is one part of the balance when deciding to impose access, increases. The place in the value chain will determine the upstream market on which indispensability is assessed and the downstream market on which the elimination of competition and the emergence of a new product are assessed. It is also important to apply those conditions in the light of the characteristics of data as explained in Cremer et al (2019, pp.98-108).

[47] Ibidem, p. 17 and FFDR, recital 30.
[48] Guidance of 3 December 2008 on the Commission's Enforcement Priorities in Applying Articles [102 TFUE] to Abusive Exclusionary Conduct by Dominant Undertakings O.J. [2009] C 45/7, paras78 and 81.
[49] Gal M.S. and D.L. Rubinfeld (2019), 'Data Standardization', *NYU Law Review* 94.

*(i) Condition 1: Indispensability of data*

When access to raw data is requested, assessment of the indispensability condition implies an enquiry as to whether an alternative raw dataset is available or could be collected by a firm having the same size as the data owner (e.g. assessed by market share in the consumer market). Obviously, this is an empirical analysis that should be examined on a case-by-case basis. The wide availability and the non-rivalry of data often do not make them indispensable as the Commission has concluded in several past merger cases. However, in some cases, data collection may be subject to legal, technical and economic barriers which may make them indispensable (see, e.g. our discussion in Section 4.1.3). In addition, the fact that the requested data have not already been traded, which is very often the case in practice, should not be an obstacle to imposing sharing as it suffices that there is demand and that such demand can legally and practically be met.

When access is about data structure, the assessment of the indispensability condition implies an enquiry as to whether the same information (not necessarily derived from the same raw data sets) is available or could be built by a firm having the same size as the data structure owner. Again, this is an empirical issue, but data structuring may show important network effects and become a *de facto* industry standard (see IMS Health explained below). Access may also be about structured data, i.e. the collected data and the structure. In this case, both assessments indicated are required.

*(ii) Condition 2: Elimination of effective competition in the downstream market*

The assessment of the elimination of downstream competition is very complex in case of data. First, the downstream market is not always known, as one of the main features of big data and AI is to experiment, crunch a lot of data without knowing in advance what information or knowledge will be found and what action might be taken. Therefore, the refusal to share data may lead to the possible elimination of a competitor on a not-yet-defined and future market. This requires a more dynamic analysis, better in line with market realities, but is more difficult to do and possibly increasing the risks of antitrust errors.

Second, the data owner is often not (yet) active on the downstream market because, as explained by Drexl (2016:49): "a typical feature of the data economy is that data is collected for one purpose but may turn out to be interesting for very different purposes pursued by other firms of very different sectors."[50] The evolution of digital industries is quick and uncertain, and many firms are 'paranoid' about the next disruptive innovation.[51] Thus, a data owner may refuse to share data with a firm that is not (yet) a competitor either because it plans to enter in the downstream market (future offensive leverage) or because it fears that the data seeker will disrupt its business (defensive leverage). In short, given the characteristics of the data economy, refusal to deal while not being active on the downstream market may be an anti-competitive exclusionary conduct.

[50] J. Drexl, *Designing Competitive Markets for Industrial Data - Between Propertisation and Access*, Max Planck Institute for Innovation and Competition Research Paper 16-13, 2016.

[51] Andy Grove, the iconic founder of Intel, wrote in 1999 a book that he famously titled: Only the paranoid will survive. On disruptive innovation, see Gans (2016).

*(iii) Condition 3: New product and consumer harm*

The interpretation of this condition is not very clear. The European Courts link this condition to the protection of the facility by an intellectual property right but have applied it more strictly in some cases than in others. The Commission integrates this condition into a more general consumer harm assessment. Taking the Courts' interpretation, the first issue is thus to determine whether the data to which access is required are protected by intellectual property (IP) rights. That will depend, among other things, on the level in the value chain at which access is required. If there is IP protection, the next issue is whether the product that the access seeker aims to bring on the downstream market is sufficiently new or, at least improved, compared to the data owner's products. Drexl (2016:52) is doubtful that this will often be the case as he considers that the generation of new information due to data sharing is often not sufficiently innovative to justify the compulsory licensing of the intellectual property right.

However, more fundamentally, the assessment of this condition faces the same two difficulties analysed previously for the second condition, i.e. the product to be offered by the access seeker is often still unknown and the facility owner is often not (yet) providing a competing product. Therefore, the more general consumer harm approach proposed by the Commission is more appropriate to the characteristics of the data economy. Thus, the competition authority will have to examine whether, for consumers, the likely negative consequences of the refusal to share data outweigh, over time, the negative consequences of imposing data sharing.

### *2.4.2 Relevant cases up to now*

Among the main essential facilities cases decided by the Court of Justice of the European Union, three are information-related. In *Magill*, the Court of Justice validated the compulsory access to programme listings, data for which there was a legal barrier (the copyright) and which was a by-product of the main activities of the broadcasters. In *IMS*, the Court of Justice set the conditions to impose access to a structure for data which was a *de facto* industry standard. In *Microsoft*, the General Court validated the compulsory access to interoperability information which were also close to *de facto* industry standard.

Specifically in *IMS*,[52] IMS-Health collected pharmaceutical sales information from wholesalers in Germany, structured them with the so-called 1860 brick structure linked to the German postal codes and developed with pharmaceutical companies and then provided sales reports to those pharmaceutical firms. IMS-Health had an intellectual property right on the 1860 brick structure and refused to licence it to NDC-Health which wanted to compete on the downstream pharma sales reports. Upon complaint by NDC-Health, the Commission had ordered interim measures forcing IMS to licence its brick structure that was found indispensable to carrying on business in the downstream market.[53] In the meantime, a litigation took place before a German Court which made a preliminary reference to the Court of Justice.

In its reply, the Court of Justice decided that the refusal to licence an intellectual property right constitutes an abuse of a dominant position where the following conditions are fulfilled:

[52] Case C-418/01, *IMS Health v. NDC Health*, ECLI:EU:C:2004:257.
[53] Case 38 044.

- *'the refusal is such as to reserve to the owner of the intellectual property, the right to market for the supply of data on sales of pharmaceutical products in the Member State concerned by eliminating all competition on that market;*
- *the undertaking which requested the licence intends to offer, on the market for the supply of the data in question, new products or services not offered by the owner of the intellectual property right and for which there is a potential consumer demand;*
- *the refusal is not justified by objective considerations*.'[54]

Moreover, the Court of Justice decided that: '*the degree of participation by users in the development of that structure and the outlay, particularly in terms of cost, on the part of potential users in order to purchase studies on regional sales of pharmaceutical products presented on the basis of an alternative structure are factors which must be taken into consideration in order to determine whether the protected structure is indispensable to the marketing of studies of that kind*.'[55]

Next to those EU cases, two non-digital national cases, which are very similar, are interesting. In both cases, a firm uses a customer list developed when it enjoyed a legal monopoly to promote a new service allowing it to compete unfairly through data cross-subsidisation which 'un-levels' the playing field between the former monopolist and the new entrants.

The first case was decided by the French competition authority against the previous gas monopolist *Gaz de France* (now Engie) which was using its customer list to promote a new gas service. In an interim decision, the authority forced Gaz de France to share the list with its competitors on the gas market as such a database was developed under a legal monopoly and was not easily reproducible by new entrants.[56] In the final decision, the authority imposed a fine of € 100 million on GDF.[57]

The second case was decided by the Belgian competition authority against the National Lottery which was using its customer list to send a one-off promotional email to launch its new sports betting product.[58] Given its nature and size, the authority concluded that the contact details could not have been reproduced by competitors in the market under reasonable financial conditions and within a reasonable period of time.[59]

In the digital sector, two American cases are also interesting. In both cases, a small firm was relying on the data of bigger digital platform to provide data analytics services and then, at some point, was cut off from the access to that data. In the first case, *PeopleBrowsr* analysed Twitter data to sell information about customer reactions to products or about Twitter influencers in certain communities. At some point, Twitter decided that its data will not anymore be accessible directly,

[54] Para 52 of the Case with a re-ordering of the conditions.

[55] Para 30 of the Case.

[56] Decision 14-MC-02 of 9 September 2014 of the French Competition Authority, *Direct Energie and UFC Que Choisir v. Engie*. This decision is based on the opinion that the French competition authority had adopted in 2010: Opinion 10-A-13 of the French Competition Authority of 14 June 2010 on cross-use of customers database.

[57] Decision 17-D-06 of 31 March 2017 of the French Competition Authority, *Direct Energie and UFC Que Choisir v. Engie*.

[58] Decision 2015-P/K-27 of 22 September 2015 of the Belgian Competition Authority, *Stanleybet Belgium/Stanley International Betting and Sagevas/World Football Association/Samenwerkende Nevenmaatschappij Belgische PMU v. Nationale Loterij*.

[59] *Ibidem*, par. 69-70.

but should be bought from certified data resellers. Following a complaint by PeopleBrowsr, a Californian Court ordered, with interim measures, that Twitter had to continue to provide its data directly. Then the parties settled the case deciding that after a transition period, PeopleBrowser will get the data from the certified data resellers.[60]

In the second case, *hiQ* analysed LinkedIn public available data to provide information to business about their workforces. At some point, LinkedIn limited access to this data by legal and technical means, because it wanted to provide similar services itself. Following a complaint by hiQ, a US federal district judge ordered LinkedIn to resume the supply of its data.[61]

## 2.5 *Sector-specific regimes*

### *2.5.1 Financial sector: Access to payment account data*

To stimulate competition and innovation in financial services, the **Second Payment Service Directive** of November 2015 (PSD2) establishes a framework for new FinTech services to access the payment account data[62] for free in a secure way and after having obtained the consent of their customers.

With regards **payment initiation services**,[63] Article 66(4) of the PSD2 provides that:

> *The account servicing payment service provider shall:*
>
> *(a)* ***communicate securely*** *with payment initiation service providers in accordance with [the common and secure open standards for communication imposed by the Commission];*
>
> *(b) immediately after receipt of the payment order from a payment initiation service provider,* ***provide*** *or make available* ***all information on the initiation of the payment transaction*** *and all information accessible to the account servicing payment service provider regarding the execution of the payment transaction to the payment initiation service provider;*
>
> *(c) treat payment orders transmitted through the services of a payment initiation service provider* ***without any discrimination*** *other than for objective reasons, in particular in terms of timing, priority or charges vis-à-vis payment orders transmitted directly by the payer*

With regard to **account information services**,[64] Article 67(3) of the PSD2 provides that:

> *In relation to payment accounts, the account servicing payment service provider shall:*

[60] http://blog.peoplebrowsr.com/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access/ and http://blog.peoplebrowsr.com/2013/04/peoplebrowsr-and-twitter-settle-firehose-dispute/

[61] *HIQ Labs v. LinkedIn.*

[62] Defined as "account held in the name of one or more payment service users which is used for the execution of payment transactions" (Article 4.12 of the PDS2).

[63] Defined as "a service to initiate a payment order at the request of the payment service user with respect to a payment account held at another payment service provider": Articles 4(15) of the PDS2.

[64] Defined as "an online service to provide consolidated information on one or more payment accounts held by the payment service user with either another payment service provider or with more than one payment service provider": articles 4(16) of the PDS2.

> *(a)* ***communicate securely*** *with the account information service providers in accordance with [the common and secure open standards for communication imposed by the Commission]; and*
>
> *(b)* ***treat data requests*** *transmitted through the services of an account information service provider* ***without any discrimination*** *for other than objective reasons.*

This sector-specific legislation complements the B2B portability right of the GDPR as it compels the banks (original controllers) to allow direct transmission of the data subjects' personal banking information to third party providers (payment initiation services or account information services). PSD2 goes further than the GDPR because, on the one hand, it forces the banks to ensure the technical feasibility of this B2B financial account data portability and, on the other hand, it makes this portability continuous as data subjects can request personal data at each transaction, facilitated by APIs.

In order to facilitate and secure such data access and exchange, the Commission adopted regulatory technical standards on the basis of a draft submitted by the European Banking Authority.[65] Those rules impose common and secure open standard for communication between the data giver (the account servicing payment service providers) and the data seekers (the payment initiation service provider or the account information service providers).

The UK went further than the PSD2 with the **Open Banking Programme** which led to a common and open API to access to account information of the customers of the nine biggest banks of the country.[66] This obligation was imposed by the UK antitrust and consumer protection authority, the Competition and Market Authority, in the context on its Retail Banking market investigation in order to increase competition and innovation in the sector.[67]

In practice, the CMA forced those nine largest banks and building societies[68]: to fund and cooperate with an independent new body, Open Banking Implementation Entity (OBIE). The OBIE developed, within a fixed (and short) timeframe, read-only open and common technical and product data standards and read-and-write open and common banking standards for the sharing of transaction data. Those standards ensure that any communication is secure and based on the consent of the customers. Their establishment has been coordinated with the EU standards developed by the EBA and made compulsory by the European Commission.

The main role of the OBIE is to (i) design the specifications for the Application Programme Interfaces (APIs) that banks and building societies use to securely provide Open Banking, (ii) support regulated third party providers and banks and building societies to use the Open Banking standards, (iii) create security and messaging standards, (iv) manage the Open Banking Directory which allows regulated participants like banks, building societies and third party providers to enroll

[65] Commission Delegated Regulation 2018/389 of 27 November 2017 supplementing Directive 2015/2366 of the European Parliament and of the Council with regard to regulatory technical standards for strong customer authentication and common and secure open standards of communication, OJ [2018] L 69/23, arts.28-36.

[66] See https://www.openbanking.org.uk/

[67] See CMA Final Report of 9 August 2016 on the Retail Banking Investigation and CMA, pp. 441-460 and CMA Order of 2 February 2017 on the Retail Banking Investigation, Sect. 10 to 14 and the Associated Explanatory Note, paras.28-39. All documents are available at: https://www.gov.uk/cma-cases/review-of-banking-for-small-and-medium-sized-businesses-smes-in-the-uk

[68] Allied Irish Bank, Bank of Ireland, Barclays, Danske, HSBC, Lloyds Banking Group, Nationwide, RBS Group and Santander.

in Open Banking, (v) produce guidelines for participants in the Open Banking ecosystem and (vi) set out the process for managing disputes and complaints.

As underlined in the Furman Report (2019, p.70), ‘one positive example from Open Banking is the effectiveness of requiring at least a subset of firms to implement and deliver the solution. Without such powers, progress is likely to be slow, disjointed and in some cases non-existent. The issue is not just the complexity of agreeing on unified standards but, potentially importantly, misaligned incentives between the largest platforms and consumers. Another lesson is that just requiring common standards is not sufficient and that an active effort is needed to make this work in practice.

The programme starts to show some success. At the end of 2019, it was used by 70 account providers (data giver) and 134 third party providers of payment initiations or account information services (data seekers) and 77% of SMEs and large corporations were already or were planning on using the Open Banking API.[69]

### *2.5.2 Automotive sector: Access to vehicle diagnostic, repair, and maintenance information*

The Regulation on Motor Vehicles of May 2018 imposes an access to some vehicle data. Article 61 of this Regulation deals with manufacturers' obligations to provide vehicle On-Board Diagnostic (OBD) information and vehicle repair and maintenance information and imposes that:

> *1. Manufacturers shall provide to independent operators* ***unrestricted, standardised and non-discriminatory access to vehicle OBD information****,*[70] *diagnostic and other equipment, tools including the complete references, and available downloads, of the applicable software and* ***vehicle repair and maintenance information****.*[71] *Information shall be presented in an* ***easily accessible manner in the form of machine-readable and electronically processable*** *datasets. Independent operators shall have access to the remote diagnosis services used by manufacturers and authorised dealers and repairers.*
>
> *Manufacturers shall provide a* ***standardised, secure and remote facility*** *to enable independent repairers to complete operations that involve access to the vehicle security system.*
>
> *2. Until the* ***Commission has adopted a relevant standard through the work of the European Committee for Standardisation (CEN) or a comparable standardisation body****, the vehicle OBD information and vehicle repair and maintenance information shall be presented in an easily accessible manner that can be processed with reasonable effort by independent operators.*

[69] https://www.openbanking.org.uk/open-banking-2019-review/

[70] *Vehicle on-board diagnostic (OBD) information* is defined as: ‘the information generated by a system that is on board a vehicle or that is connected to an engine, and that is capable of detecting a malfunction, and, where applicable, is capable of signalling its occurrence by means of an alert system, is capable of identifying the likely area of malfunction by means of information stored in a computer memory, and is capable of communicating that information off-board’: art.3(49)

[71] *Vehicle repair and maintenance information* is defined as: ‘all information, including all subsequent amendments and supplements thereto, that is required for diagnosing, servicing and inspecting a vehicle, preparing it for road worthiness testing, repairing, re-programming or re-initialising of a vehicle, or that is required for the remote diagnostic support of a vehicle or for the fitting on a vehicle of parts and equipment, and that is provided by the manufacturer to his authorised partners, dealers and repairers or is used by the manufacturer for the repair and maintenance purposes’: Regulation 2018/858, art.3(48).

> *The vehicle OBD information and the vehicle repair and maintenance information shall be made* ***available on the websites of manufacturers using a standardised format*** *or, if this is not feasible, due to the nature of the information, in another appropriate format. For independent operators other than repairers, the information shall also be given in a machine-readable format that is capable of being electronically processed with commonly available information technology tools and software and which allows independent operators to carry out the task associated with their business in the aftermarket supply chain.*

Then, Article 63 of the Regulation deals with the fee for the access to that information and provides that:

> *1. The manufacturer may charge* ***reasonable and proportionate fees*** *for access to vehicle repair and maintenance information other than the records referred to in Article 61(10). Those fees shall* ***not discourage access*** *to such information by failing to take into account the extent to which the independent operator uses it. Access to vehicle repair and maintenance information shall be offered* ***free of charge to national authorities, the Commission and technical services****.*
>
> *2. The manufacturer shall make* ***available*** *vehicle repair and maintenance information, including transactional services such as reprogramming or technical assistance,* ***on an hourly, daily, monthly, and yearly basis****, with* ***fees for access to such information varying*** *in accordance with the respective periods of time for which access is granted.*
> *In addition to time-based access, manufacturers may offer transaction-based access for which* ***fees are charged per transaction*** *and not based on the duration for which access is granted.*
> *Where the manufacturer offers both systems of access, independent repairers shall choose systems of access, which may be either time-based or transaction-based.*

Finally, Annex X of the Regulation sets up some common relevant standards for the data access.[72]

Thus, this Regulation on Motor Vehicle complements the GRPR and gives a sector-specific data access right for relevant car data to independent repairs in order to stimulate competition and innovation on this aftermarket.

[72] In particular, point 6 of the Annex X.

### *2.5.3 Energy sector: Access to consumer data*

In order to stimulate competition and innovation among electricity suppliers, the new Electricity Directive of June 2019 imposes the sharing of consumer data, including metering and consumption data as well as data required for customer switching, demand response and other services. Article 23(2) of the Directives provides that

> *Member States shall organise the management of data in order to ensure* ***efficient and secure data access and exchange****, as well as data protection and data security.*
> *Independently of the data management model applied in each Member State, the parties responsible for data management shall provide access to the data of the final customer to any eligible party (…). Eligible parties shall have the requested* ***data at their disposal in a non-discriminatory manner and simultaneously****. Access to data shall be easy and the relevant procedures for obtaining access to data shall be made publicly available.*

Regarding the fee for data access, Article 23(5) of the Directive provides that:

> *No additional costs shall be charged to final customers for access to their data or for a request to make their data available.*
> *Member States shall be responsible for* ***setting the relevant charges for access to data*** *by eligible parties.*
> *Member States or, where a Member State has so provided, the designated competent authorities shall ensure that any* ***charges imposed*** *by regulated entities that provide data services are* ***reasonable and duly justified****.*

Here again, the Electricity Directive complements the GDPR by requiring Member States to set up a specific regime for consumer data sharing and exchange between electricity suppliers.

03

# TECHNICAL ASPECTS OF DATA PORTABILITY AND DATA SHARING

# 3. Technical aspects of data portability and data sharing

We now discuss technical aspects of data portability and data exchange, including a description of the way personal data is stored, the way data export features are currently implemented, standards, and emerging systems and scenarios.

In this technical description, we mostly focus on the data managed by large technology companies, such as those that participate in the Data Transfer Project (namely, Apple, Facebook, Google, Microsoft, Twitter). Smaller companies, and companies from other fields that do manage personal data, may proceed similarly, or may have more ad-hoc process in place to allow users to exercise their right to data portability.

## *3.1 Data Models and Formats*

Personal data comes in a wide variety of forms: structured meta-information provided by the user (account information, contact lists, preferences, etc.), contributed content (social networking posts, multimedia content, etc.), telemetry information (log of user activity), collected sensor data (such as geolocation traces). An attempt at a categorisation of such user data is provided in the ISO/IEC 19944 standard[73] but such a categorisation is necessarily coarse and incomplete. The data models and formats used to represent, store, and exchange data are strongly dependent on the kind of data used.

At a very high level, it is customary to distinguish[74]:

- *structured data*, which follows a rigid tabular format, with data records formed of a pre-defined set of fields;
- *semi-structured data*, which follows a hierarchical shape, mixing structured content and potentially unstructured text, with a range of data attributes that need not be defined ahead of time;
- *unstructured data*, which do not fit in either of the previous categories, and comprises plain text in natural language, multimedia data (images, sounds, videos), and other arbitrary binary data.

Structured data are typically exchanged as CSV[75] files (a simple data format for tabular data), SQL[76] files (a standard for relational database systems), or spreadsheet files, such as OpenDocument[77] or Microsoft's Office Open XML[78]. Note that none of these formats is ideal for data exchange: CSV is not fully standardised, with for instance no specified way of describing the character encoding used; existing implementations of SQL do not strictly follow the SQL standard,

[73] International Organization for Standardization. (2017). *Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use*. ISO/IEC 19944:2017.
[74] *Serge Abiteboul, Peter Buneman, Dan Suciu. (1999). Data on the Web: From Relations to Semistructured Data and XML. Morga Kaufmann*
[75] Internet Engineering Task Force. (2005). *Common Format and MIME Type for Comma-Separated Values (CSV) Files.* RFC 4180.
[76] International Organization for Standardization. (2003) *Database Language SQL.* ANSI/ISO/IEC 9075:2003.
[77] International Organization for Standardization. (2006). *Information technology — Open Document Format for Office Applications (OpenDocument) v1.0.* ISO/IEC 26300:2006.
[78] International Organization for Standardization. (2016). *XLSX Strict (Office Open XML)*. ISO 29500-1:2008-2016.

which makes it difficult to use SQL for data exchange unless a specific database system is targeted; spreadsheet formats are very complex and include many features beyond the scope of simple structured data representation.

Semi-structured data can be represented using a variety of standards. XML[79] was designed as a simple data exchange format for semi-structured information, which comes with a broad ecosystem of tools and associated standards. JSON[80] is originally the way object literals are written in the JavaScript programming language, but has been repurposed as an alternative to XML, with the main advantage of being less verbose. RDF[81] is a framework for representing information in the form of semantic graphs: an RDF statement links a semantic concept (the *subject*) to another semantic concept or data value (the *object*) through a semantic predicate that indicates the relation between subject and object. RDF was introduced in the setting of the Semantic Web, and is adapted to arbitrary semi-structured information; it comes with a variety of serialisation formats, such as Turtle[82] or RDF-XML[83].

In both the structured and semi-structured cases, file formats only specify a syntactic layer on the way information is represented. To make sense of the data, it is necessary to know the *schema* of the data, i.e. what fields and data attributes exist, and what constraints on the data values should be respected. CSV and spreadsheet files do not offer any schema description capabilities, apart from giving names to columns. SQL includes a *data definition language* (DDL) that specifies the types and constraints on fields used in SQL tables. XML documents can similarly include a *document type definition* (DTD), a simple way of describing the respective document's schema, and use more elaborate XML schema languages, such as XML Schema[84], in order to express more complex constraints. An analogous of XML Schema for JSON does exist under the name JSON Schema[85], but is currently not fully standardised and less frequently used. Finally, RDF data can be associated with complex schemas and logical constraints, expressed in the form of the RDF Schema[86] language or OWL[87] ontologies.

Beyond the syntax (given by the file format) and the schema and constraints (given by the schema annotations, when available), data needs to be interpreted with respect to a specific *semantics*, which gives meaning to data fields and attributes. This is sometimes called a data *dialect* or *vocabulary*. In some application areas, these dialects are standardised: for instance, GPX[88] is a de-facto standard for exchanging GPS traces as an XML dialect; jCard[89] is a standard for contact lists as a JSON dialect. In the RDF world, de-facto standards emerge by re-using and combining data

[79] World Wide Web Consortium. (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. W3C Recommendation
[80] Internet Engineering Task Force. (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. RFC 8259.
[81] World Wide Web Consortium. (2014). *RDF 1.1 Concepts and Abstract Syntax.* W3C Recommendation
[82] World Wide Web Consortium. (2014). *RDF 1.1 Turtle: Terse RDF Triple Language.* W3C Recommendation.
[83] World Wide Web Consortium. (2014). *RDF 1.1 XML Syntax*. W3C Recommendation
[84] World Wide Web Consortium. (2012). *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures*. W3C Recommendation.
[85] Internet Engineering Task Force. (2019). *JSON Schema: A Media Type for Describing JSON Documents*. Internet Draft.
[86] World Wide Web Consortium. (2014). *RDF Schema 1.1*. W3C Recommendation.
[87] World Wide Web Consortium. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. W3C Recommendation
[88] Dan Foster. (2004). *GPX: the GPS Exchange Format*. https://www.topografix.com/gpx.asp
[89] Internet Engineering Task Force. (2014). *jCard: The JSON Format for vCard.* RFC 7095

vocabularies: Schema.org[90] is a collaborative effort for proposing data vocabularies of use to search engine companies; Dublin Core[91] is a de-facto standard of a data vocabulary describing digital and physical works. When no prior dialect exists, ad-hoc dialects are created by the data controller, which requires additional documentation. In this situation, when data is exchanged between two data controllers which use different dialects (this is called *schema heterogeneity*), it is necessary to transform it from one schema to another, using *schema mappings*[92] from the source to the destination schema; these schema mappings are most of the time hand-written by data engineers, though there is sometimes the possibility of automatically learning them from examples[93].

Finally, in the case of unstructured data, note that, for some specific applications, there are also standards, such as multi-media file formats, which are out of scope of this report. Arbitrary text and binary data are stored and exchanged in an ad-hoc manner, and also require documentation.

## 3.2 Data Storage and Accessibility

We now discuss how personal data may be stored and accessed by the data controller. Again, we focus on personal data managed by large technology companies.

In almost all cases, it is necessary for the data to be (very) efficiently accessible upon request. This is true whenever data may be used by the data controller in real-time applications, e.g. for display when a web page is accessed. This means any data item of interest needs to be retrievable with a latency of the order of a second or less. This is sometimes made formal: for example, the architecture used by Amazon (such as for retrieving customer and order data) provides service-level agreements (SLAs) for complex aggregation of data of the form: “a response must be given within 300ms for 99.9% of requests”.[94]

There are a few exceptions, when data is stored in ways that preclude its efficient retrieval. In particular, data for which no real-time access is needed (historical transaction data of banking systems, query logs of web search engines when they are not used for other purposes, etc.) are sometimes moved to archiving systems[95] where access may incur a large latency or even involve a manual process. However, these cases are rare in systems of large technology companies, as most personal data is used for producing web content in one form or another.

To support fast access, data is stored in some form of database management system. The most common type of such systems are traditional relational database management systems[96] (commercial systems such as Oracle, IBM DB2, Microsoft SQLServer; or open-source ones such as

[90] Schema.org Community Group. (2020). *Schema.org*. https://schema.org/

[91] Dublin Core Metadata Initiative (2020). *DCMI Metadata Terms*. www.dublincore.org/specifications/dublin-core/dcmi-terms/

[92] *Phokion G. Kolaitis. (2005). Schema mappings, data exchange, and metadata management. PODS 2005: 61-75.*

[93] *Balder ten Cate, Víctor Dalmau, Phokion G. Kolaitis. (2013). Learning schema mappings. ACM Trans. Database Syst. 38(4): 28:1-28:31.*

[94] *Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Werner Vogels. (2007). Dynamo: Amazon's highly available key-value store. SOSP 2007: 205-220*

[95] IBM. (2013). *A roadmap for intelligent archiving*. IBM InfoSphere Optim Archive e-book. www.ibm.com/downloads/cas/PA9YRY1N

[96] Avi Silberschatz, Henry F. Korth, S. Sudarshan. (2010). *Database System Concepts (Sixth Edition)*. McGraw Hill.

MySQL, PostgreSQL, or MariaDB), also called SQL systems as they rely on the SQL standard language. Such systems manage large quantities of structured data (easily scaling to terabytes of content), typically stored on magnetic disk (hard drives), with the possibility of data being distributed[97] on a small cluster of servers, with latencies for typical queries of the order of a dozen of milliseconds.

SQL database systems are not adapted to all storage and access tasks, however: they are ill-equipped to store and query semi-structured data, to store extremely large amounts of data (of the order of a petabyte or more) or to provide extreme performances (e.g. sub-millisecond latencies). As an alternative, a wide variety of systems have been proposed under the collective term of *NoSQL*[98]. This term covers very different kinds of systems. On the one hand, there are systems that propose a different data organisation than the relational, tabular model: XML databases (e.g. BaseX, eXist, MarkLogic server) for handling and querying XML-structured data; triple stores (e.g. Virtuoso, Jena, RDFox) for RDF data; document stores (e.g. CouchDB, MongoDB) for storing semi-structured documents with limited query capabilities. On the other hand, there are systems that focus on performance and storage of extremely large amounts of data: wide column stores (e.g. Google's Bigtable, Apache Hbase, Apache Cassandra) with a tabular structure but flexible schema, and extreme scaling capabilities; key-value stores (Amazon DynamoDB, Voldemort) with extreme performances.

Though traditional SQL systems are still by far the dominant mode of data storage[99], most large technology companies (including Facebook, Google, Amazon, Twitter, LinkedIn, etc.) have made the switch from traditional relational database systems to NoSQL systems focussing on performance, because of their extreme needs in terms of latency, data volume, or query throughput. In some cases, these NoSQL databases are used in combination with traditional SQL databases, depending on applications. There is also a trend to go towards *NewSQL* systems (such as Google's Spanner) that provide SQL support, rich features as with SQL systems, but higher performance by moving data to main memory or eliminating traditional bottlenecks of SQL systems (locking, logging, cache management).

In addition to the core system used to store data, an additional *caching* layer is also often used, to provide faster response to common queries by storing responses in main memory. Redis is a popular open-source such *in-memory* data store.

Whatever the technology used, it is important to note once more that, as long as some data is needed in a real-time applications, it will need to be accessible efficiently, with typical sub-second latencies.

### 3.3 Application Programming Interfaces (APIs)

Before describing technical ways that can be used to implement the right to data portability, it is necessary to introduce the basics of *web services*, more commonly called (Web) *Application*

[97] M. Tamer Özsu, Patrick Valduriez. (2020). *Principles of Distributed Database Systems (Fourth Edition)*. Springer.
[98] Christof Strauch. (2011). *NoSQL Databases*. White paper, Stuttgart Media University. www.christof-strauch.de/nosqldbs.pdf
[99] DB-Engines. (2020). *DB-Engines Ranking*. https://db-engines.com/en/ranking

*Programming Interfaces*[100], or *APIs* for short. They will form a key component in the implementation of various data export and transfer scenarios.

It is common for a user to be able to access his or her personal data through a web-based interface (a website, or possibly a smartphone application). This web-based interface is usually only meant for human users. It is technically possible to program a system that can access the same personal data as the users, by using the login credentials of the user and mimicking the interaction of a human with this web-based interface. However, this is usually discouraged by the developer of the interface, and often even disallowed in the terms and conditions of use of the interface.

A web service, or API, is a technical interface to access the data that is meant to be used by programs, in particular by third-party software that introduces novel applications of the data. Although there is no requirement to offer such APIs, they are commonplace as they allow the data controller to specify what kind of access third-party software have, with even the possibility of introducing monetisation of richer forms of data access. APIs usually allow access to a subset of the features accessible to human users (for instance, they may disallow modifying the data), and in particular to a subset of the corresponding data. However, they may also provide access to more data than what is available through the web interface for users, because such data is not meant to be processed by humans. In any case, it is common to impose *rate limitations* on APIs, in terms of the maximum number of calls to the API authorised within a given time period.

Although web services were traditionally implemented in SOAP[101], a technology allowing complex sequences of structured messages, modern APIs use a much simpler *RESTful* approach, this name meaning that they follow a Representational State Transfer (REST) architecture. A specificity of REST architectures is that the use of the API is assumed to be *stateless*: no information is kept from one call to the API to the next, all relevant information needs to be given in the parameters to the API call. This ensures scalability of APIs, allowing for many concurrent requests.

In order to be used to access personal, potentially private, data, the use of APIs need to be combined with an *access delegation* protocol, that verifies that the call to the API has been authorised by the user whose data is being accessed. The most commonly used such protocol is OAuth 2.0[102]. OAuth involves three parties: the data controller (which controls the API), the data subject (whose data is being accessed), and the *client,* which is the application that accesses the API. The typical workflow for authorisation is as follows: the client refers the data subject to a web page of the data controller, which indicates to the data subject the set of permissions that are requested; if the data subject approves the request, the data controller will issue the client an opaque *access token,* which the data controller associates with a set of permissions and a duration validity. This token then needs to be used by the client every time the API is called, to demonstrate that it has been authorised to make the request by the data subject.

[100] Leonard Richardson, Sam Ruby, Mike Amundsen. (2013). RESTful Web APIs: Services for a Changing World. O'Reilly Media.
[101]James Snell, Doug Tidwell, Pavel Kulchenko. (2009). Programming Web Services with SOAP: Building Distributed Applications. O'Reilly Media.
[102] Internet Engineering Task Force. (2012). *The OAuth 2.0 Authorization Framework*. RFC 6749.

Now we have introduced the important notions of API, rate limitation, access delegation, and access token, we are ready to describe data export and transfer scenarios.

### *3.4 Data Export Modes*

We now review different possible technical modes for implementing the right to data portability, and discuss the current state of implementations. We first distinguish between *data export* and *data transfer*: data export is when the data subject receives the data from the data controller, as per Article 20(1) of the GDPR. *Data transfer* is when the data is exchanged directly from one data controller to another, without the need for the data subject to be involved, as per Article 20(2) of the GDPR. Let us first consider the simpler case of data export.

We distinguish three different modes of implementation of the data facilities:

- *Asynchronous data export.* The data subject requests the data at a given point of time, which the data controller produces after some possibly lengthy period of time and makes available for download.

- *Pull-based data export.* The data subject requests the data at a given point of time, which the data controller provides as a response, within usual interactive response times (of the order of one second or less).

- *Push-based data export.* The data subject registers with the data controller his or her intent to retrieve the data; as soon as new data is produced or collected, the data controller sends it to the data subject.

In all three modes, there is also the question of the scope of the request of the data subject: the data subject may request the entirety of its personal data, or part of it, specified by a query: this may be, for instance, the part produced or managed by a specific application of the data controller, or all data obtained since a given time (which we call time-based queries). Using the push-based mode, or a combination of the pull-based mode and time-based queries, the data subject is able to obtain a *continuous* view of its personal data (in push-based mode, nearly in real time; in pull-based mode, every time the data subject issues a request, e.g. every hour or every day). This is not directly possible in the absence of time-based queries or in the asynchronous mode, since the former requires a possible complex analysis of what has changed between two snapshots of a user's personal data, and the latter does not provide near-instantaneous responses.

We now review the current practice of a sample of large technology companies (Facebook, Google, Microsoft, and Twitter) with respect to the access provided to personal data (under the right to data portability and beyond).

- All propose an asynchronous data export mode[103], with usually very limited query capabilities, mostly allowing to select specific applications or categories for which to export the data. No guarantee is given about the delay needed to produce the data, which is for

[103] See the following instructions: https://www.facebook.com/help/1701730696756992 for Facebook, https://support.google.com/accounts/answer/3024190 for Google, https://docs.microsoft.com/en-us/power-automate/gdpr-dsr-export-msa for Microsoft, https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive for Twitter.

practical purposes of the order of the minute or hour (Google for instance indicates "This process can take a long time (possibly hours or days) to complete.", while Twitter states that the download link will be available within 24 hours). In addition to this delay, additional restrictions may apply: Twitter for instance states that only one request may be issued every thirty days. Finally, this data export process is not automatable: it requires the user to manually select the corresponding option in her or his profile.

- All propose a set of specific APIs[104] that allow for pull-based retrieval of *some* part of the personal data, with usually complex query capabilities, including time-based queries. Contrarily to the asynchronous data export mode, access to these APIs is meant to be automated. As previously mentioned APIs require access authorisation before providing any private data, but do not typically allow access to all private data. For example, the Facebook API does not provide access to the full timeline of a user; Google discontinued in 2013 the Latitude API, which allowed access to a user's geolocation history[105] and has not provided any alternative. In addition, APIs almost always come with drastic rate limitations.[106]
- Very few push-based data export facilities are available. An example is the real-time Twitter streaming API[107] that can be used to export all *public* data of a given user.

The outcome of the asynchronous data export mode is a compressed ZIP archive of various files, of diverse formats and schemas; most common formats are JSON for semi-structured information and multimedia file formats for images, audio, etc., but Google, notably, includes in its data export a wide variety of formats, including XML dialects and CSV.

The output of APIs is usually JSON files, again in a wide variety of dialects, with little to no standardisation from one company to the other.

If most personal data is efficiently accessible (see Section 0), why may it take hours or days to provide asynchronous data? A reason is the potentially very large amount of data involved. Indeed, Google's data export includes all photos uploaded by a user to Google Photos, or all documents shared on Google Drive. This may represent gigabytes of data or more, which takes a non-negligible amount of time (indeed up to several hours) to copy, package, and compress within a ZIP archive, even when directly accessible. Note that this is mostly specific to unstructured data formed of multimedia or arbitrary binary files, as the structured and semi-structured personal data of a user tend to represent a more reasonable volume (for instance, geolocation traces, which are one of the most frequently updated form of personal data, amount to only a few kilobytes per day, i.e. no more than a few dozen megabytes over the entire life of a user).

It would be very much possible for technology companies to provide pull- or push-based data exports of all personal data, as long as it is not expected that a single request results in the download of the entire dataset as in the asynchronous mode; handling multimedia content, for example, could mean issuing one request to get the list of all files, and accessing each file one by

[104] See the documentation: https://developers.facebook.com/docs/ for Facebook, https://developers.google.com/apis-explorer for Google, https://developer.microsoft.com/en-us/web/apis/ for Microsoft, https://developer.twitter.com/en/docs/api-reference-index for Twitter.

[105] https://web.archive.org/web/20150814192105/https://support.google.com/gmm/answer/3001634?p=maps_android_latitude&rd=1

[106] See, for instance, https://developer.twitter.com/en/docs/basics/rate-limiting for Twitter.

[107] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter

one. This would be compatible with the way data is stored and accessed, and would allow a data subject to obtain a continuous view of his or her personal data.

### *3.5 Data Transfer Modes and Emerging Systems*

We now go beyond data export and consider data transfer scenarios[108], where the personal data of a data subject is transferred from one data controller (the *source*) to another (the *destination*). To illustrate, this can mean transferring a user's photos from Google Photos to Apple's iCloud, a user's documents from Microsoft Office 365 to Google Docs, a user's text message from Whatsapp to Skype, or a user's event calendar on Facebook to Apple Calendar.

The different modes we discussed in the previous section (asynchronous, pull-based, and push-based), as well as the notion of continuous data access and that of scope of a query (what data items to consider), are still relevant for data transfer. But additional challenges arise:

- As previously discussed, because of *schema heterogeneity*, the data formats used to model and store personal data can vary considerably from one platform to the other. In order to transfer data between these platforms, appropriate schema mappings need to be applied to transform the source data representation into the destination one.
- Should the data transfer be *one-way* (from one platform to the other) or *two-way* (from one platform to the other and back)? In the case of two-way transfer, how should *synchronisation* and conflict resolution be handled (i.e. what to do if the same data item has been updated on both platforms since the last time data was transferred)?

These are real challenges, but to simplify we will assume that schema heterogeneity is dealt with hand-crafted schema mapping (as is most often done in practice), and we will only consider one-way transfers.

As mentioned in the previous section, the asynchronous mode of data export requires a human intervention and is not meant to be automated. For this reason, in the case of data transfer, systems use pull-based (or occasionally push-based) APIs.

There are two main ways to implement data transfer:

- **Direct exchange:** Data is transferred directly from the source data controller to the destination data controller. Either the source or destination data controller may initiate the transfer, though it is most commonly done by the destination, as read access to data is more commonly provided by APIs than write access. Whatever the case, the data controller that initiates the data transfer now acts as a client to the API of the other data controller, using an access delegation protocol to obtain an access token to prove that it has been authorised by the data subject. Note that, before being added to the destination data controller, data needs to be mapped from the source schema to the target schema by the data controller that initiated the data transfer.
- **Exchange through a third-party:** Data is transferred from the source data controller to a third-party system (such as a PIMS, see below), with the third-party acting as a client of the

[108] Fing. (2018). *Data-responsible Enterprises: User Experience and Technical Specifications.* http://mesinfos.fing.org/wp-content/uploads/2018/03/PrezDataccess_EN_V1.21.pdf

API of the source data controller. Then the third-party system uses an API of a target data controller providing write access to transfer the data to that third-party. The third party is responsible for mapping the source data to the target schema.

Note that the first scenario requires the destination (respectively, source) data controller to implement themselves the data transfer capability and the mapping from the source schema (respectively, to the target schema), but only a single API is needed. In the second scenario, the data transfer and schema mappings are handled by the third-party system, but there need to exists both a read-access API on the source data controller and a write-access API on the target controller.

In the first scenario, the data subject needs to grant permission to one data controller to access the API of the other. In the second scenario, the data subject needs to grant permission to the third-party system to access the APIs of both data controllers. In both cases, there is a trust issue since the permission associated to access tokens are rarely fine-grained enough that the data subject can be confident that the only use that will be made of them is for the data transfer task. For instance, their validity in time may extend beyond the time needed for the data transfer.

We now discuss two emerging types of systems that implement data transfer: personal information management systems, or PIMSs, and the Data Transfer Project.

### *3.5.1 Personal Information Management Systems (PIMSs)*

A Personal Information Management System[109] (or PIMS for short) is a system that allows a user to build an integrated view of her own personal data, e.g. emails and other kinds of messages, calendar, contacts, web search, social network, travel information, work projects, etc. Such information is commonly spread across different services. The goal is to handle all the personal data of a user in a system that the user controls and trusts; this may mean repatriating all the data on the PIMS, or keeping the data distributed but using the data integration[110] methodology to interface with the different data sources.

The former approach is sometimes called data *warehousing* (as all retrieved data is stored in a local data warehouse) whereas the latter is called the *mediator* approach (a mediator system is charged of rewriting all user queries to queries over the original data sources). The mediator approach is more scalable, as it does not require as much local storage and computation resources as the warehousing approach, but heavily relies on the availability of powerful enough APIs to express translated user queries. Both the data warehousing and mediator approach also require the careful, often manual, design of schema mappings to overcome the heterogeneity between the information schema used by the PIMS and that of the various data sources.

On the basis of such PIMS, novel applications can be built that exploit the fact that the entire personal information is (possibly virtually) available. For instance, a user may formulate queries

[109] Serge Abiteboul, Benjamin André, Daniel Kaplan. (2015). *Managing your digital life*. Commun. ACM 58(5): 32-35. Also European Commission Services, *An emerging offer of Personal Information Management Services: Current state of service offers and challenges*, November 2016, available at: https://ec.europa.eu/digital-single-market/en/news/emerging-offer-personal-information-management-services-current-state-service-offers-and

[110] *Maurizio Lenzerini. (2002). Data Integration: A Theoretical Perspective. PODS 2002: 233-246.*

such as "What kind of interaction did I have recently with Alice B.?", "Where were my last ten business trips, and who helped me plan them?"; the system has then to orchestrate queries to the various services (which means knowing the existence of these services, and how to interact with them), and integrate information from them (which means having data models for this information and its representation in the services), e.g. align a GPS location of the user to a business address or place mentioned in an email, or an event in a calendar to some event in a web search.

With respect to data transfer, PIMSs act as a separate data controller, with direct exchanges from external data controllers to the PIMSs. Though this is not their main function, PIMSs may also offer the possibility of pushing the data to other data controllers, acting in this case as a third party between the source and destination data controllers. By nature of the role of a PIMSs, continuous data exchanges are favoured: at any point in time, a PIMS aims at having a complete and up-to-date view of the personal data of a user.

The PIMS initiates API calls, controls access tokens, and implements schema mappings. It is therefore crucial that it is fully trusted by the user. Several models allow this: Cozy Cloud[111] is a company that offers private cloud services (with each user hosted on a private virtual machine) on which PIMSs can be deployed; Digi.me[112] offers a PIMS with fine-grained control on what private data is sent to which data controller.

### *3.5.2 Solid*

An alternative but similar approach is proposed by the Solid[113] project led by Tim Berners-Lee. While the project is still in its infancy, the goal is to build a fully decentralised space for personal information, with data distributed over multiple *personal online data stores (pods)* located on different hosts, with a mechanism allowing a user to grant third-party applications fine-grained access to specific data items on specific pods. Each pod can thus be seen as a data controller, with Solid acting as a specification of how pods interface with the world, and as an overall infrastructure on top of these pods.

At the time being, Solid is still under active development, and the question of importing data from legacy data controllers has not been solved. It is currently a software project of a somewhat moderate size. An imperfect software metric often used to estimate the rough complexity of a system is the number of source lines of code; for the Solid server, this is roughly 32k at the time being. For comparison, the number of lines of code of the Nextcloud[114] open-source project, a widely deployed platform for self-hosting of cloud services with some limited PIMS capabilities, is around 730k. Other common metrics (such as the number of times the project was *forked* on the Github platform, which is in the hundreds for Solid) consistently point to a project of moderate size and impact – roughly a tenth of that of a widely deployed system such as Nextcloud (two thousand forks).

[111] https://cozy.io/fr/
[112] https://digi.me/
[113] https://solid.mit.edu/
[114] https://github.com/nextcloud/server

### *3.5.3 The Data Transfer Project (DTP)*

The Data Transfer Project[115] is a technical initiative that was launched in 2018 by large technology companies. Specifically, Apple, Facebook, Google, Microsoft, Twitter are associated with the project. The main outcome of this initiative is the development of a specification and of an open-source platform for data transfer. Though these five companies are nominally involved, the project inherits from Google's former *Data Liberation Front*, and Google is by far the main contributor to the DTP platform (with Facebook also contributing, but at a similar level as independent contributors), as is shown by the following analysis of the source code of the system (analysis of the Git repository as of March 17, 2020):

| Institution | Proportion of changes (commits) | Proportion of changes (source code lines) |
|---|---:|---:|
| **Google** | 83,21 % | 80,96 % |
| **Facebook** | 10,05 % | 3,26 % |
| **Others** | 6,74 % | 15,78 % |

The DTP aims at supporting both direct exchange and exchange through a third-party, though direct exchange is the mode that is favoured. The DTP envisions the role of a *hosting entity*, which is the entity in charge of initiating the data transfer. It is usually meant to be either the source or destination data controller, but it can also technically be a third-party system (indeed, the provided demonstration server acts as a third-party hosting entity). As previously discussed, the hosting entity is the one that controls the access tokens, and therefore needs to be trusted.

At the time being, the DTP does not consider continuous data transfer; the main application scenario is a bulk transfer at the initiative of a data subject who decides to move from one platform to another.

Note that the Data Transfer Project is under development at the moment and cannot be considered as a stable product. A number of import or export connectors have been implemented to interface with various platforms, but there are very few public-facing sites that do use the DTP (the most prominent being a specific use case on Facebook: users have been able very recently to use it to transfer their photos to Google Photos[116]). However, it is possible to test the infrastructure by using the provided demonstration server. Software metrics for the DTP are of the same order of magnitude as that of Solid: 44k of lines of code, hundreds of forks; well under those of a project such as Nextcloud, or of other projects that a company like Google or Facebook dedicate effort into, such as their machine learning computation frameworks, respectively Tensorflow[117] (2.5 million lines of code, 80 thousand forks) and PyTorch[118] (1 million lines of code, 10 thousand forks).

[115] https://datatransferproject.dev/
[116] https://engineering.fb.com/security/data-transfer-project/
[117] https://github.com/tensorflow/tensorflow
[118] https://github.com/pytorch/pytorch

### 3.6 Summary: How to Make Data Portability Work?

As we have seen, the current way data portability rights can be technically exercised is minimal and far from ideal:

- it is only possible to perform asynchronous data exports;
- there is no guarantee on the delay between the request and the availability of the data;
- data provided in the export does not follow any specific standard, and includes a wide variety of data models, file formats, schemas, dialects, in a way that is not compatible between one data controller to the other;
- data exchange facilities are not implemented by data controllers.

We claim that this is not inevitable, and that there are no strong technical challenges in providing continuous pull- or push-based data exports, with limited delay (as long as specific solutions are implemented for large unstructured data). Indeed, the fact that large data controllers provide similar (though not complete) features through APIs means there is no particular burden in implementing them, and that such functionalities would not be a cause for performance issues beyond specific concerns about large files, which can be addressed separately. In order to better exploit exported data, data controllers should aim for more standardisation in data models (e.g., using common RDF dialects).

Data exchange capabilities are currently impeded by the problem of schema heterogeneity, but assuming this problem is resolved (either by a standardisation of the data export models, or by manually compiled schema mappings), they do not pose any particular technical challenge. Data exchange through a trusted third-party (as in PIMSs, or as in the DTP where the hosting entity is on a trusted external host) has the advantage that access tokens need not be provided to the original data controllers.

04

# THE IMPACT OF DATA PORTABILITY ON COMPETITION AND INNOVATION IN DIGITAL MARKETS

# 4. The impact of data portability on competition and innovation in digital markets

## 4.1 *Preliminaries: the economics of data*

Understanding the economics of data portability requires foremost to acknowledge the economics of data. Here we focus at first on the differentiation between data, information and knowledge, only from the latter of which ultimately value can be derived. Moreover, we take a closer look at the asserted non-rivalry of data.

### *4.1.1 The value of data, information and knowledge*

Data per se does not have any economic value as it is merely the (digital) representation of signals that have been received or perceived using some syntax. For example, the receipt of a light signal can be transformed into data by recording the time (e.g., using the syntax HH:MM:SS) and recording the "on" and "off" states (e.g., using the syntax "1" for "on" and "0" for "off"). Such data is transformed into information only if it is combined with semantics. For example, a corresponding semantic would be that the data is on received light as part of a communication effort using morse code.

This gives the data a meaning (here: a message that is communicated), hence transforming it into information. Such information can then be transformed further into actionable knowledge with the additional input of and in combination with other pieces of information. For example, the received message may have been "HELP" and combined with the information that a friend is hiking all by himself in the mountains, in about the location from where the light beams have been received, leads to the actionable knowledge that he is in danger and that a rescue operation should be started.

The same holds true, of course, for clicks (data) on an e-commerce site, which represent which products a shopper considered for purchasing (information) and which can then be used to infer which products the shopper might be interested in (knowledge). Ultimately only such actionable knowledge that is generated from data potentially has economic value and can increase welfare.

Nevertheless, it is customary to refer to all three concepts – data, information and knowledge – simply as "data" in policy circles. Instead, 'raw data' is often differentiated from 'derived' and 'inferred data' (see, e.g., Section 2.1.1). We will therefore follow this practice as well in the following. In order to do so, we need to introduce a related, but distinct terminology in the next subsection.

### *4.1.2 Volunteered, observed and inferred data*

With respect to the economics of data it is also important to note a differentiation in how data was acquired about an individual consumer. As is customary, we distinguish between volunteered, observed and inferred data.

*Volunteered data* is explicitly and intentionally revealed by a user, such as a name and birthday entered into a registration form, a post, tweet or rating submitted, or an image or video uploaded. Consumers are usually aware of the volunteered data that they revealed and often this is the only type of data that consumers think they have revealed when using an online service. In practice, it is often also the only data that is fully made available by data controllers in response to a data portability request according to Article 20 GDPR.

*Observed data* is obtained from the usage of a device, website or service and the user may or may not be aware that such data is collected. This ranges from clicks on products and purchase histories over geo-locations gathered by GPS sensors in smart phones to recording every single interaction of the consumer with the service–potentially even when the consumer does not even know that she is currently interacting, such as in the context of voice assistants that are constantly recording. As detailed in Section 2.1 there is some uncertainty regarding the degree and scope to which observed data should be fully made available according to Article 20 GDPR. Furthermore, as detailed in Section 1, this can potentially comprise large data sets, but technically it would be feasible (at least for large online service providers) to provide such data sets.

*Inferred data* is derived through refinement and recombination from volunteered and observed data, e.g. by use of data analytics such as clustering, filtering or prediction. The result can be a complex preference profile of a consumer or a recommendation. Inferred data can actually already be – using the definition introduced in Section 4.1.1. – knowledge that can provide actionable insights. Thus, inferred data is ultimately the basis for competition between data-intensive firms, whereas volunteered data and observed data are the 'raw data' inputs.

The distinction between volunteered, observed and inferred data, albeit not being a legal definition, is also important in the context of data portability. Article 20 GDPR clearly includes volunteered data, while it is commonly understood that inferred data is not included. With respect to observed data, it is currently not completely clear how far the users' right to port data goes and whether and to what extent it is covered by the right to data portability (see Section 2.1.1). However, in its "Guidelines on the right to data portability" the European Data Protection Board suggests that observed data are to be included in the right to data portability.[119]

### *4.1.3 Non-rivalry of data, and its limits*

Moreover, data is *non-rival*, which means that the same data can, in principle, be used by different entities at the same time. Moreover, the same data could also be shared and collected by different entities without depleting the source of data for others. For example, many observers could have collected the data on the light signals sent at the same time without interfering with the ability of others to do the same. Likewise, the data on the light signals could have been shared without having to give it up. But data is also *excludable*, which means that the data controller can impose technical or legal constraints to prevent sharing of data. Non-rivalry and excludability are distinct concepts and should not be seen as the two sides of the same medal. Although the consumption of data is non-rival, there may be economic impacts of data that are non-rival. This is sometimes overlooked in the policy debate. In the following, we therefore want to put non-rivalry of data (as opposed to the excludability of data) in perspective in order to emphasise benefits and a risks of data sharing and data portability.

[119] See https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf

#### 4.1.3.1 Rivalry in the collection of data

First, *specific data* (e.g., on products liked on a particular e-commerce site, or links clicked on a particular search engine*) cannot just be collected by anyone interested*. Just like a focused ray of light, e.g. emitted by a laser, may just be received in a particular location and not by a random observer. In this sense, although data consumption of data is non-rival, *the collection of data may be rival and is therefore inherently concentrated.*

In this context it must be noted, however, that there is a lively debate with respect to the degree to which the collection of data is rival. On the hand, some scholars claim that data is ubiquitous, as consumers are willing to share their data over and over again with different services, frequently multi-home similar services, and that specialised data brokers make data available to everyone who wants to buy it (see, for example, Lambrecht and Tucker 2015[120], and Tucker 2019[121]).

This is contrasted by the empirical findings that – despite the multitude and variety of websites and online services available – consumers' attention is highly concentrated on a few sites and even fewer firms. In other words, only those firms are in the right 'location' to actually collect consumer data at a large scale. For example, the European Commission found in the context of the Google AdSense case that Google had a market share of generally over 90% in 2016 the market for general search in all Member States.[122] Similarly, in its investigation of Facebook, the German Federal Cartel Office (Bundeskartellamt) found that Facebook had market share in the market for social networks of over 95% (with respect to daily active users) in Germany in December 2018.[123]

In similar vein, even fewer firms are currently able to collect tracking data across multiple sites. For example, Englehardt and Narayanan (2016) [124] measured which third-party web trackers were deployed at the top 1 million websites. They find that Alphabet/Google (with trackers deployed at about 70% of all sites), followed by Facebook (trackers deployed at about 30% of all sites), are also in a unique position to track users' activity across various (third-party) websites. Very similar results are obtained by Ghostery, a browser extension that blocks third party trackers.[125] The situation is likely to become even more pronounced as Google has recently announced to disallow third-party cookies in Google's Chrome browser, which many view as a step that bolsters Google's and Facebook's dominance in web tracking (Financial Times, 2020[126]), because these companies have alternative means to track users across the web, e.g., through services such as 'Google Analytics' or 'Login with Facebook'.

[120] Lambrecht, A. and Tucker, C.. Can Big Data Protect a Firm from Competition? (Dec. 18, 2015). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2705530

[121] Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. Review of Industrial Organization, 54(4), 683-694.

[122] See https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770

[123] See https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html?nn=3591568

[124] Englehardt, S., & Narayanan, A. (2016, October). Online tracking: A 1-million-site measurement and analysis. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 1388-1401).

[125] See https://www.ghostery.com/study/ and Macbeth, S. (2017). Tracking the Trackers: Analyzing the Global Tracking Landscape with GhostRank. Available at: https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf

[126] Financial Times (2020). 'Cookie apocalypse' forces profound changes in online advertising. Available at: https://www.ft.com/content/169079b2-3ba1-11ea-b84f-a62c46f39bc2

Taken together, this already points to the conclusion that the collection of observed user data (as opposed to volunteered user data) is indeed often rival, because for key services (such as search, or social networking) the market is highly concentrated and only a few firms are able to track user activity across the web. Thus, observed data is not ubiquitously available, and it is also usually not feasible nor socially desirable to duplicate the collection of the same observed data. This would mean that users would have to conduct the same search, the same post or the same purchase on several platforms; and it would mean that even more web trackers are being built into the websites that we visit. Thus, rivalry in the collection of data is not necessarily a problem, but it does provide a strong rational for sharing data.

#### *4.1.3.2 Rivalry in deriving value from data*

Second, the economic value of data likely depends on how many others have access to the same data, or put more precisely, can derive the same insights from data. For example, both Ishihashi (2019)[127] and Gu (2018)[128] highlight by means of a game-theoretic model that the value of data collected from consumers may drop significantly (in their theoretical models to zero) if more than one firm possesses it.

Once the data is created (e.g. generated by using the service of a firm, which has 'paid' for the data by offering a free service), consumers will give it up to a second firm, even at a 'price' close to zero, because each additional sharing of data does not bear opportunity costs. This is a direct consequence of the non-rivalry of data. This means that if data sharing is frictionless and bears zero transaction costs for consumers, firms eventually possess identical sets of data. A potential buyer of this data is only interested in acquiring such data once because each data set is a perfect substitute for the other. This means that firms engage in fierce price competition selling the data – known as Bertrand competition. Eventually they compete each other down to marginal costs, which means that they sell the data for a price close to zero. If, however, only one firm would have possessed the data, it could have demanded a non-zero price. In this sense, although the consumption of data is non-rival, the *economic value that can be derived from data is rival*.

If taken literally, this provides a strong rational for *not sharing* data, as this would destroy any incentive to collect data in the first place. However, three important clarifications are in order.

First, and foremost, the above argument does not differentiate between 'data' and 'knowledge', because it essentially only considers data intermediaries, which collect and sell raw data. Even though two firms may have access to the same raw data set (in terms of volunteered and observed data), they may derive different insights from it ('inferred data' or 'knowledge' in our terminology), which is ultimately the basis for competition.

Second, and relatedly, the above argumentation has abstracted from cases where the data is not sold to third parties on some data market, but rather used internally (e.g. for marketing purposes or for improving the service quality)—or where data is combined with other data available to the firm and the enriched data set can be sold as a unique data set, overcoming the competition in the data market.

[127] Ichihashi, S. (2019). Non-Competing Data Intermediaries (Jun. 29, 2019). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3310410

[128] Gu, Y., L. Madio, and C. Reggiani (2019). Data brokers co-opetition (Feb. 13, 2019). Available at SSRN http://dx.doi.org/10.2139/ssrn.3308384.

Third, the above argumentation has abstracted from transaction costs, such as additional privacy concerns of sharing the data set with another firm, or the effort in selling additional data in return for only a low additional benefit. If these transaction costs are non-negligible, this will reduce the non-rival nature of data sharing, which leads to less data sharing and eventually decreases the competition in the data market.

Taken together, this means that more prevalent sharing of 'raw' user data, will likely render the market for data intermediaries, which simply acquire and sell raw data, but do not offer advanced analytics on such 'raw' data, more competitive and possibly unlucrative. However, this does not destroy the incentives to compete on the basis of insights derived from data. Rather, as raw data becomes more prevalent, the focus of competition is likely to move more from collection to analytics, which likely stimulates innovation rather than stifling it (see Section 4.3.2). Indeed, as data collection is inherently concentrated (see Section 4.1.3.1) and the services through which (observed) data is collected usually exhibit strong network effects (see Section 4.2.2), a stronger competition at the data analytics level (based on knowledge) seems much more feasible and desirable than competition at the data collection level.

### *4.1.4 The quality of data, and its relationship to volunteered and observed data*

Volunteered, observed and inferred data are also useful concepts for discussing different qualities of data. Generally, the *quality of data* can be measured along the dimensions of

- fitness for use (is the data suited to derive the desired insights?)
- accuracy (does the data represent the facts?)
- completeness (how many data points are missing?)
- timeliness (how fast can data be collected and how quickly is it outdated?)

Volunteered data is derived from direct human input. That is, this data may be inaccurate, e.g., because wrong information (e.g. a wrong email address, fake name or fake review) have been submitted intentionally or unintentionally. But often the accuracy of the volunteered data is also essential for the quality of the service, which provides consumers with an incentive to provide accurate data (e.g., a correct liking of songs in a music streaming service will trigger a better recommendation for new songs).

Moreover, volunteered data is prone to being incomplete, and it may outdate relatively fast, because it is not automatically updated after it has been provided. However, volunteered data is usually structured, because it has been collected in a structured way, such as through forms, 'like' buttons, or on a rating scale. Thus, it can immediately be used as input to generate inferred data.

By contrast, observed data is less prone to deliberate manipulation, because it is derived from actual behaviour and sensors. Moreover, observed data tends to be more complete and timelier, because it is recorded automatically. The accuracy and fitness for use is often very context dependent. For example, click data from an e-commerce session can be very noisy and sparse, because the user might just be browsing through random products and in each product category only very few products are explored. In another session, the similar click data can be very accurate and dense, as a consumer explores several similar products and puts some of them in the shopping basket, but finally only buys one. Similarly, data from sensors (e.g., GPS sensors) can be highly accurate at times and inaccurate at other times, depending, for example, on geography and environmental conditions.

Quality of data with respect to fitness for use also depends highly on the context. Highly accurate GPS data, for example, may be necessary to identify which products a consumer was interested in when visiting a department store, whereas coarser data may still be acceptable to identify which stores a consumer has visited in a mall. In any case, observed data is often less structured and must be cleaned and structured in a way that allows to derive actionable knowledge.

Finally, the quality of inferred data depends not only on the quality of the observed and volunteered data, but also on the amount of observed and volunteered data. With respect to the analysis of data, empirical studies suggest that in many (big) data analytics applications, (i) there is a minimum required scale, (ii) there are benefits from larger data sets, and (iii) these benefits are marginally decreasing as data sets become very large. More precisely, Junqué de Fortuny et al. (2013)[129] and Martens et al. (2016)[130] demonstrate that prediction accuracy increases for larger data sets of fine-grained user behavior data (observed data). Whereas benefits decrease marginally as prediction accuracy approaches the theoretical benchmark (cf. Li, Ling, Wang, 2016[131]), the studies show this convergence is not yet reached in many popular application settings. Furthermore, for the online advertising industry, Lewis and Rao (2015) [132] find that only very large amounts of data allow firms to measure whether advertising campaigns are indeed successful. Thus, empirical studies and general indications point to the presence of scale economies from data collection and data analysis.

Consequently, having access to more data (e.g., not only volunteered but also observed data) will, in many applications, yield a better quality of the inferred data, i.e. the actionable knowledge, and thus offer higher profit opportunities for firms. Therefore, the application scope of data portability, i.e. whether it is restricted to volunteered data or also encompasses observed data, is also crucial from an economic perspective.

## 4.2 *Data portability and competition*

The previous subsection highlighted that

- particularly observed data is a valuable raw input for data-intensive business models in the digital economy
- the collection of observed data in the digital economy is inherently concentrated and only a few digital firms are in a unique position to collect it.

In this context, the question arises how newcomers and start-ups may get access to the required observed data, in order to be able to compete on the basis of inferred data, i.e. knowledge and insights generated from these raw inputs. More generally, this raises the question how and if data portability indeed increases the competitiveness of digital markets.

[129] Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: is bigger really better? Big Data, 1(4), 215-226.

[130] Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. MIS Quarterly, 40(4), 869-888.

[131] Li, X., Ling, C. X., & Wang, H. (2016). The convergence behavior of naive Bayes on large sparse datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(1), 1-24.

[132] Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics, 130(4), 1941-1973.

First, we take the perspective of a consumer, and highlight that switching to a new service may impose two types of ‘costs’ that can result in consumer lock-in. The first type of cost is a transaction cost from switching. The second type of cost is related to network effects. We describe both in turn. Then, we discuss in more detail why firms may not want to import data from other providers, thus rendering consumers’ right to export data mute for the purpose of switching providers.

### *4.2.1 Data portability and data-induced switching costs*

It is often argued that consumers do not switch from one digital service to another because they shy away from the transaction costs to give away their (volunteered) data again at the new service. This seems especially problematic in cases where large amounts of data have been volunteered over a long time in which the current service was used. For example, in the case where thousands of songs have been liked while using an online streaming service, liking the same songs again at a new service seems an unreasonable burden. This transaction cost is a classic switching cost, i.e. a fixed cost for setting up a service that occurs only once. When a consumer evaluates two services—the one that she is currently using, and the new one—the difference in expected utility must at least exceed the switching cost, in order for the consumers to switch.

The classic literature on switching costs (see, e.g., Klemperer 1987a[133]) finds that switching costs can constitute a significant barrier to entry, shielding incumbents from competition. In digital markets, switching cost may vary substantially depending on the context. However, the classic literature also finds that when established services compete for customers in the presence of switching costs, then competition is fierce for ‘new’ customers, whereas ‘old’ customers tend to be exploited (see, e.g., Klemperer 1987b[134]; Farrell and Shapiro 1988[135]). However, in the long run, markets tend to be less competitive in the presence of switching costs (see, e.g., Beggs and Klemperer, 1992[136]).

Generally, services whose quality depend to a high degree on customisation and personalisation (e.g., services in which recommendations play a significant role) are more prone to be subject to switching costs. However, often it may not just be volunteered data that constitutes a switching cost, but also the observed data. For example, the current music streaming service may also have recorded which songs where actually listened to, how often each song was played, for how long, and at what time of the day. Like the volunteered data, this observed data can be a very useful input for the next music streaming service.

The right to data portability can lower these switching costs by making the volunteered data and observed data readily available in a “structured, commonly used and machine-readable format” to the consumer, who can then pass it on to the new provider. Thus, in light of the classic switching cost literature, the right to data portability can make digital markets more competitive in the long run and lower entry barriers for new service providers. This is commonly viewed as beneficial to

[133] Klemperer, P. (1987a). Markets with consumer switching costs. The Quarterly Journal of Economics, 102(2), 375-394.

[134] Klemperer, P. (1987b). The competitiveness of markets with switching costs. The RAND Journal of Economics, 138-150.

[135] Farrell, J., & Shapiro, C. (1988). Dynamic competition with switching costs. The RAND Journal of Economics, 123-137.

[136] Beggs, A., & Klemperer, P. (1992). Multi-period competition with switching costs. Econometrica: Journal of the Econometric Society, *60*(3),651-666.

consumer welfare and one of the strongest economic arguments for the right to data portability as it is currently implemented in the GDPR.

However, more recently a new strand of economic literature has re-investigated the classic results and specifically considered the welfare implications of the right to data portability. In a game-theoretic model, Wohlfarth (2019)[137] highlights that the right to data portability can have an effect on the amount of data this is collected by data-intensive firms. Without the right to data portability, market entrants are forced to design services that economise on the use of data in order to be able to attract consumers. However, as data can be easily ported to the entrant, the new provider has less incentives to economise on data use and increases the amount of data collected. In this sense, the GDPR's right to data portability (Article 20) runs contrary to the GDPR's principle of data minimisation (Article 5.1c); not only from a legal point of view, as pointed out in Section 2.1.3, but also with respect to the economic incentives of data collection. Wohlfarth shows that this economic trade-off can eventually lead to a reduction in consumer surplus.

In a similar vein, Krämer and Stüdlein (2019) also analyse the economic effects of data portability on market entry in a game-theoretic model. They focus on the firms' incentives to disclose user data, e.g., in the context of targeted advertising, with and without the right to data portability. They show that the right to data portability is likely to benefit the 'old' customers of the incumbent, especially those that do switch to the new provider, as switching costs are reduced and competition is increases. However, the 'new' customers of the entrant, i.e. those that were not previously customers of the incumbent, are likely to be worse off, because the entrant's competitive position is strengthened under the new right to data portability. Without data portability, the entrant would have competed more fiercely for these new customers. In reverse, this means that its customers are worse off than without data portability. Again, this highlights that not all consumers need to benefit from a right to data portability – although this right unambiguously lowers switching costs.

Despite these nuances, if data portability indeed lowers switching costs, this is likely to increase the competitiveness of markets. However, not the least, this will also depend on whether consumers actually make use of data portability, and whether data is actually imported by other services.

### *4.2.2 Data portability and network effects*

Network effects arise whenever a consumer's value of a good or service depends on how many other consumers are using the same good or service. Network effects are ubiquitous in digital markets, and often services are explicitly designed to incorporate network effects. For example, in social networks, network effects arise, because participation in the network is more valuable the more other people are also using the same social network. This is a direct network effect. But more than often, indirect network effects are present. In this case the value of the service increases because of the presence of more complementors to the service. For example, an operating system is valuable mostly due to the availability of software complements that run on this operating system. Likewise, an e-commerce website may be valuable to a consumer due to the number of product reviews on that website, which depend only indirectly on the number of users. Indirect network effects are also at the core of platform markets (a.k.a. two-sided markets), which bring

[137] Wohlfarth, M. (2019). Data Portability on the Internet. Business & Information Systems Engineering, 61(5), 551-574.

together two distinct user groups (such as buyers and sellers). At least one of the two groups values the presence of the other group on the platform, thereby creating an indirect network effect. A prototypical example is an app store, where consumers value the presence of many app developers, and, in reverse, app developers value the presence of many consumers.

Network effects are important in the context of data portability and the competitiveness of markets for two main reasons.

#### *4.2.2.1 Data portability and user-side network effects*

First, network effects create a coordination problem. Because the value of the service depends directly or indirectly on how many others are using it, consumers want to be where everybody else is. This also creates a lock-in situation, distinct from that of simple switching costs, because switching a provider seems only reasonable if everyone switches at the same time. It is important to note that, contrary to the case of data-induced switching costs, data portability does not alleviate this type of lock-in. This would require some (protocol) interoperability (see Crémer, de Montjoye and Schweitzer 2019) of the services, whereby services interoperate to a degree where ultimately users can interact seamlessly albeit being on different networks – like users of different telecom networks can communicate with each other. Then users can switch to a new provider without losing access to the network effect exerted by users who remain with the old provider. Consider a social network for example. Even if a user would be able to take its data to a new social network, it would still not be able to interact with the users that remained on the old network. Indeed, in this context, it has been argued that "identity portability" (Gans 2018[138]) or "social graph portability" (Zingales and Rolnik, 2017[139])—both a form of protocol *inter*operability—would be desirable to overcome user-side network effects Identity portability means that a person can switch to a new network and take her identity with her, so that all messages related to that person are forwarded to the new network, and vice versa. The idea of identity portability is thus comparable to interconnection in conjunction with number portability on telecom networks.

However, demanding (protocol) interoperability over and beyond (data) portability also has some caveats, especially with respect to the need for regulatory oversight (like in telecoms networks), and the ensuing risk of barriers to innovation due to the necessity to remain within the standard for interoperability. As others have noted (see, e.g., Crémer, de Montjoye and Schweitzer 2019), this seems warranted only in specific applications such as text messaging services and social networks, where the benefits of interoperability (through increasing the network effect and competition *in* the market) are likely to outweigh the risk (of reduced innovation and competition *for* the market).

Finally, there is a noteworthy interaction between network effects and switching costs, laid out in Suleymanova & Wey (2011). Markets with strong network effects tend to monopolise, because consumers tend to gravitate to the service or platform that already exhibits the largest network effects. In other words, once a critical mass of users has been reached, markets tip towards the largest player. Switching costs can dampen this process, because they create an economic friction

[138] Gans, J. (2018). Enhancing Competition with Data and Identity Portability. Brookings Institute - The Hamilton Project, Policy Proposal 2018-10. Available at: https://www.brookings.edu/wp-content/uploads/2018/06/ES_THP_20180611_Gans.pdf

[139] Zingales, L., & Rolnik, G. (2017). A way to own your social-media data. The New York Times. Available at: https://www.nytimes.com/2017/06/30/opinion/social-data-google-facebook-europe.html

(transaction cost) that prevents customers from switching to the service with higher network effects as easily. In this vein, switching costs may allow two networks to co-exist at the same time. However, this is usually not an efficient situation in the presence of network effects. Moreover, the argument rests on the assumption that there are two services, albeit with different market shares, which both have a viable and stable user base. In practice, many digital markets with strong network effects have already tipped and new entrants do not have a viable and stable user base so that switching costs (or non-portability) would protect them from churn. Thus, we argue that in many relevant scenarios the interaction of data portability and network effects is not anti-competitive. But as laid out above, it is also not pro-competitive in the sense that data portability affects user-side network effects per se. Rather, data portability may impact analytics-based network effects, which may then have a pro-competitive effect. We describe this in the following.

#### *4.2.2.2 Data portability and analytics-based network effects*

Second, indirect network effects can also arise with respect to data analytics capabilities. Here network effects yield a positive feedback loop for algorithmic learning that can constitute an effective entry barrier (see Lerner 2014[140] for a thorough discussion): The more consumers are using a service, the more (volunteered and observed) data is created on which analytics can be performed and algorithms can be trained, which in turn results in an improvement of the service (e.g., better recommendations, better search results), which in turn leads to more consumers. For example, a dominant search engine is likely to provide better results simply because it records more search queries (volunteered data) and records more clicks on search results (observed data), which can then be used to derive better results lists for future searches.

This means, in this case barriers to entry are not created by switching costs in the narrow sense (indeed switching a search engine hardly entails any switching costs due to setting up the service), nor are they due to a lack of access to the network of users (on the same or the other market side). Here it is rather the lack of access to the data that is created by fellow users – a type of indirect network effect – that creates a barrier to entry. This lack of data limits the ability of a new service provider to compete on the basis of algorithmic insights and data analytics, i.e. on the basis of inferred data or knowledge. This argument is explored more formally, for example, in Hagiu and Wright (2020)[141], who show that this competitive advantage of the incumbent prevails under various assumptions about the shape of the learning curve from data. Moreover, Schaefer, Sapi and Lorincz (2018) [142] provide empirical evidence that such network effects in algorithmic learning exist in the context of search engines.

Thus, if enough users would consent to a transfer of their raw data, and if it were possible to continuously transfer data through a standardised interface (API), then data portability could potentially promote entry and competition. It is important to highlight that the provision of data to competitors would be initiated by the consumers and, in each case, only entail the data of that consumers. This is very different to an access request entailing (anonymised) input data across a

[140] Lerner, A. V. (2014), The Role of 'Big Data' in Online Platform Competition (August 26, 2014). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2482780

[141] Hagiu, A. and Wright, J. (2020). Data-enabled learning, network effects and competitive advantage. Mimeo. Available at: http://andreihagiu.com/wp-content/uploads/2020/05/Data-enabled-learning-20200426-web.pdf

[142] Schaefer, M., G. Sapi and Lorincz, S. (2018). The effect of big data on recommendation quality: The example of internet search, Düsseldorf Institute for Competition Economics Discussion Paper No 284. Available at: http://www.dice.hhu.de/fileadmin/redaktion/Fakultaeten/Wirtschaftswissenschaftliche_Fakultaet/DICE/Discussion_Paper/284_Schaefer_Sapi_Lorincz.pdf

large number of users, initiated by another firm, e.g., by a competitor under the essential facilities doctrine. Although some commentators highlight that such access to input data may be a possibility to restore market contestability (e.g., by Argenton and Prüfer, 2012[143], Krämer and Wohlfarth, 2018[144] and Schweitzer et al., 2018[145]), the focus of the present report is on user-initiated data portability. The advantage of data portability is that also personally identifiable data can be transferred, and thus there is no trade-off between competition and privacy, which is inherent to access requests that are not user-initiated. However, at the same time, it is unlikely that all users initiate a transfer of their data. Thus, the data set that is ported under data portability is likely to be more detailed on specific data subjects, but less representative for the user base as a whole. Whether or not such a data set is useful for a competing or complementing firm, is context specific and depends on the degree to which consumers make use of data portability, of course.

Finally, it is noteworthy to mention in this context that data portability may also be viewed with caution, because this can lead to situations in which ultimately consumers and competitors are worse off. In particular, Lam and Liu (2020)[146] argue by means of a game-theoretic model that the right to data portability encourages consumers to reveal more data to the incumbent, because consumer are less concerned about data-induced switching costs that may arise later when considering to switch to a new market entrant (see Section 4.2.1). However, as consumers reveal more data, they also create a higher data analytics network effect at the incumbent, which indeed strengthens the competitive position of the incumbent vis-à-vis a new market entrant, and raises entry barriers. While data portability facilitates switching (which lowers entry barriers and raises consumers' surplus), this effect can be completely offset by the increase in the data analytics network effect (which raises entry barriers and may prevent efficient entry). In summary, the authors therefore conclude that data portability can have an adverse effect on entry and long-run efficiency, although (or indeed because) data portability lowers switching costs. Note that this arguments rests strongly on the assumption that data portability leads to a different data revelation behaviour of consumers at the incumbent.

## 4.3 *Data portability and innovation incentives*

The previous subsection has focused on the impact of data portability on competition and contestability of markets, i.e. adopted a more static efficiency perspective. We now turn to a dynamic efficiency perspective and consider the impact of data portability, and more generally of data access on innovation incentives.

There has been a lively scholarly and policy debate about data access and innovation (see, e.g., Crémer, de Montjoye and Schweitzer, 2019[147]; Furman et al. 2018[148]), which we do not intend to

[143] Argenton, C., & Prüfer, J. (2012). Search engine competition with network externalities. Journal of Competition Law and Economics, 8(1), 73-105.

[144] Krämer, J., & Wohlfarth, M. (2018). Market power, regulatory convergence, and the role of data in digital markets. Telecommunications Policy, 42(2), 154-171.

[145] Schweitzer, H., Haucap, J., Kerber, W., & Welker, R. (2018). Modernisierung der Missbrauchsaufsicht für marktmächtige Unternehmen (Vol. 297). Nomos Verlag. Executive Summary in English available at: https://pdfs.semanticscholar.org/ba99/aa34216249bcd6e036d8efe1f99bcb1798cd.pdf

[146] Lam, W. M. W., & Liu, X. (2020). Does data portability facilitate entry?. International Journal of Industrial Organization, 69, 102564. Available at: https://doi.org/10.1016/j.ijindorg.2019.102564

[147] Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). Competition policy for the digital era—final report. Available at: https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf

repeat here. However, we wish to highlight the main trade-offs involved in order to lay the groundwork to derive appropriate policy recommendations in the context of data portability and data interoperability.

With regard to innovation, it is important to differentiate between the innovation incentives and capabilities of the firm that provides access to data and the firms that receive access to data. Moreover, it is important to differentiate whether such data is used to compete with the data provider or whether it is used for other purposes, such as offering complementary or completely new services. We consider these scenarios in turn.

### *4.3.1 Innovation by the incumbent: Conventional wisdom and kill zones*

Although the consumption of data is non-rival (although there may be rivalry in the collection and monetisation of data, see in Section 4.1.3), data is excludable, which – in an economic sense – means that a firm can exert exclusion rights on data assets. Without mandated access to data, data-intensive firms can utilise their economic control over data in order to make economic profits – be it by selling access to data or by using the data to improve their product or service in order to gain a competitive advantage. It is, by now, evident that data-rich firms can be highly profitable and this creates an economic incentive to invest in data collection and analysis. This spurs innovation, ranging from innovative services (that allow for a collection of data) to innovative data storage and data analytics. In this view, losing control of those data would lead to what economists call a "hold-up problem". That is, the lack of sufficient appropriability on data renders the economic benefits of data uncertain and leads to a reduction in investment and innovation. This is, of course, conventional wisdom among economists, the very reason why intellectual property rights exist (i.e. a legal instrument for data *excludability*), and an argument that is not specific to data. In this sense, innovation incentives in the context of data are particularly strong when data can be used exclusively, and if in consequence a market can be monopolised.

In a similar vein, it is conventional wisdom in economics that there is a (non-linear) relationship between innovation incentives and competition, although there is continued research on the topic. Innovation is a means to provide a better service or product and to differentiate from competitors. This tends to increase profits and provides innovation incentives. In line with an Arrowian view, in a monopolistic environment, where high entry barriers already exist (be it by network effects or switching costs, or something else), innovation incentives tend to be low, because there is no competitive advantage to be gained from innovation. But in line with a Schumpeterian view, in markets with very high degrees of competition, innovation incentives also tend to be low as well, because innovation rents are quickly competed away and firms are often lacking sufficient scale for innovation activities.

Taking both arguments together, and in line with ample empirical evidence, innovation incentives tend to be the highest in oligopolies with only a few firms (see, e.g., Aghion et al. 2005)[149]. In this sense, if data portability indeed induces more competition in digital markets with high data-induced

[148] Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. Report prepared for the Government of the United Kingdom, March. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

[149] Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. The Quarterly Journal of Economics, 120(2), 701-728.

entry barriers, then this would likely increase incentives to innovate. In particular, in the context of digital markets, innovation incentives are particularly high if a market has not yet tipped and there is still competition *for* the market; or, possibly even more importantly, if digital markets were indeed contestable. This would mean that, despite a de-facto monopoly, entry barriers remain low and the incumbent needs to constantly defend its incumbency through innovation.

There is some doubt, however, as to whether the market inhabited by some big tech firms are indeed contestable and whether data portability would indeed lead to more competition in established markets. On the one hand, we have already detailed that data portability cannot overcome user-induced network effects per se (see 4.2.2), such that important barriers to entry remain, irrespective of the degree of data portability, if a new service were to compete head-to-head. On the other hand, there is growing empirical evidence that some firms may have established 'kill-zones' around their core business model (see, e.g., Kamepalli et al. 2020[150] and Scott Morton et al. 2019[151] for a thorough discussion, but also related news reports[152]). This means that innovative start-ups, which may become competitors to a big tech firm's data-centric business model, may either be bought by the big tech firm, or it is quick to incorporate the innovation into its own service. In the latter case the incumbent has a comparative advantage relative to start-ups or smaller firms due to its deep financial pockets, and existing economies of scale as well as network effects (e.g., in data analytics). In this way the incumbent can successfully prevent customer churn and, at the same time, deny start-ups a viable and stable customer base. Such 'kill zones' also seem to have an effect on the venture capital market, where start-ups that complement the incumbent's business model are more likely to receive venture capital than start-ups that challenge the incumbent (for a discussion see, e.g., Smith, 2018[153], Rinehardt, 2018[154] and Kamepalli et al. 2020[155]). For the same reasons, there is also a growing consensus that data-intensive mergers should be reviewed more carefully and with adapted tools by competition authorities, (see, e.g., Bourreau and de Streel, 2020[156]; Crémer, de Montjoye and Schweitzer 2019; Motta and Peitz 2020[157]; Scott-Morton et al. 2018). In this context, recall also our discussion in Section 2.4 on the (high) legal barriers with respect to access to data under the essential facilities doctrine.

[150] Kamepalli, S. K., Rajan, R. G., & Zingales, L. (2020). Kill Zone. CEPR Discussion Paper No. DP14709. Available at: https://ssrn.com/abstract=3594344

[151] Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Melamed, D. & Morgenstern, J. (2019). Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report. Draft. Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business.

[152] See, for example, The Economist (2018). Into the danger zone
American tech giants are making life tough for startups. Available at: https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups; Financial Post (2018). Inside the kill zone: Big Tech makes life miserable for some startups, but others embrace its power. Available at: https://business.financialpost.com/technology/inside-the-kill-zone-big-tech-makes-life-miserable-for-some-startups-but-others-embrace-its-power

[153] Smith, N. (2018). Big Tech Sets Up a 'Kill Zone' for Industry Upstarts. Available at https://www.bloomberg.com/opinion/articles/2018-11-07/big-tech-sets-up-a-kill-zone-for-industry-upstarts;

[154] Rinehardt, W. (2018). Is there a kill zone in tech? Available at: https://techliberation.com/2018/11/07/is-there-a-kill-zone-in-tech/

[155] Kamepalli, S. K., Rajan, R. G., & Zingales, L. (2020). Kill Zone. CEPR Discussion Paper No. DP14709. Available at: https://ssrn.com/abstract=3594344

[156] M. Bourreau and A. de Streel, Big Tech Acquisitions: Competition & Innovation Effects and EU Merger Control, CERRE Issue Paper, February 2020 available at https://www.cerre.eu/publications/big-tech-acquisitions-competition-and-innovation-effects-eu-merger-control

157 Motta, M. and M. Peitz (2020). "Big Tech Mergers" CEPR Discussion Paper 14353, available at https://cepr.org/content/free-dp-download-31-january-2020-competitive-effects-big-tech-mergers-and-implications

In summary, this means that, irrespective of the degree of data portability, we conclude that data portability would lead to more or less competition and innovation in established digital markets per se. It may, however, spur innovation in complementary and new digital markets, which we argue next.

### *4.3.2 Innovation at the service level vs. innovation at the analytics level*

In Section 4.2.2.2 we have already discussed the positive feedback loop that provides an incumbent digital service provider with a competitive advantage in terms of data analytics capabilities. We now return to this issue from an innovation perspective. Data (volunteered and observed) is often accumulated as the results of *innovation at the service or product level*, which led consumers to use and thereby to contribute personal data. By contrast, inferred data is the result of *innovation at the data analytics level*. An important observation in this context is that, given the raw data, it does not necessarily require an innovation at the service level per se to achieve an innovation at the data analytics level.

However, as discussed previously, innovations with respect to inferred data (i.e. data analytics innovations) rest upon the input of raw data (observed and volunteered data), which typically can only be amassed if the firm also runs a successful service at the service or product level. This creates a virtuous innovation cycle for incumbents. Innovations at the analytics level facilitate innovation at the service level, which again spur innovations at the analytics level. While there are certainly inherent efficiencies in this virtuous cycle, it may be viewed as problematic that innovation can to a large degree only occur 'in house', whereas truly innovative ideas often come from outsiders, often (business) users (see, e.g., van Hippel, 2005[158]).

Indeed, innovation at the data analytics level may spur innovation at the service level in a completely different domain.[159] For example, Google Flu Trends[160] exemplified that search data cannot just be used to improve the search engine's results, but also to predict the spread of the flu. But it has also been demonstrated that there was significant scope for improvement over Google's algorithm (see, e.g., Lampos et al. 2015[161]).

Similarly, an innovation at the service level may not get off the ground, if it is not fed with sufficient raw data to begin with. For example, collaborative-filtering based recommender systems suffer from a well-known 'cold-start problem' (see, e.g., Bobadilla et al. 2012[162]). That is, in order to provide good results, the recommender system needs to be fed with sufficient user data (observed and volunteered data) in order to be able to find similarities between users from which recommendations can then be derived. For example, suppose it were an innovation at the service level to offer customers personalised recommendations for clothing and styling. If the idea is found to be intriguing enough by potential customers, it would – at least at the beginning – not be required to be very innovative at the analytics level, because collaborative-filtering algorithms for

[158] See, e.g. von Hippel, Eric (2005), Democratizing Innovation, MIT Press. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=712763

[159] See also Prufer J. and C. Schottmüller (2017), Competing with Big Data, *TILEC Discussion Paper 2017-006.*

[160] https://www.google.org/flutrends/about/

[161] Lampos, V., Miller, A. C., Crossan, S., & Stefansen, C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. Scientific reports, 5, 12760. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4522652/

[162] Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-based systems, 26, 225-238.

such a purpose would be readily available. The main challenge would be to overcome the cold-start problem, however, so that if new customers try the service for the first time, it would offer useful recommendations already.

Thus, there is reason to believe that innovation activities would be significantly increased, if it were possible that innovation at service level and innovation at the analytics level could occur independently, i.e. in different organisations. Thanks to the non-rivalry of data, this would not mean that the current data controller loses access to the data, and thus, can continue to be innovative both at the service *and* the analytics level, taking advantage of the virtuous feedback cycle.

#### *4.3.3 Lack of empirical studies on data portability and innovation*

While it is without doubt that we have seen an unprecedented wave of innovations in digital markets, the above arguments provide some reasoning that the level of innovation could be even higher, if data portability were more prevalent. To be clear, we are not aware of conclusive empirical evidence that has tested this hypothesis. In fact, while there is a substantial legal literature on data portability (see Section 2), and some theoretical work (see Section 4.2), we are not aware of any empirical studies on how data portability specifically has altered competition or innovation incentives in digital markets. It is probably also difficult to establish a conclusive cause-and-effect relationship at all, as data portability usually comes in package with other privacy rights, and because in the dynamic environment of digital markets it is very difficult to establish the counterfactual for innovation.

There is some tentative evidence, however, in the case of Open Banking, which is probably one of the most important natural experiments in this context. Although there was competition between banks, the emergence of new financial services (fin techs) has spurred following the availability of API-based common interfaces that made continuous data portability possible[163]. This seems to suggest that data portability has indeed facilitated innovation activities in this sector.

[163] See https://www.openbanking.org.uk/wp-content/uploads/2019-Highlights.pdf

05

# THE ECONOMICS OF PERSONAL MANAGEMENT INFORMATION SYSTEMS

# 5. The economics of Personal Management Information Systems

Personal Information Management Systems (PIMS) have been introduced mainly from a technical perspective in Section 1. Given their possibly central role in the context of data portability, we now discuss PIMS from an economic perspective. Particularly, we will focus on the questions whether and under which conditions PIMS may indeed be economically sustainable.

## 5.1 *Key functionalities of PIMS*

As highlighted in Section 0, PIMS come in a variety of shapes, but their central premise is to empower users to regain control over their personal data a variety of otherwise decentralised services in one central place. The core vision is to provide users with a central dashboard, where they can grant and revoke their consent for data processing — at a fine-grained level – with any given data controller, and exercise their rights, especially the right to data portability (Art. 20 GDPR) and the right to erasure (Art. 17 GDPR). Hence, in policy and technical circles, PIMS are often regarded as the silver bullet, which is the missing building block for a fair and transparent data economy. Accordingly, associations like the MyData movement[164], which originated in Finnish policy circles in 2014, are currently gaining increasing attention. DG Connect published a report on PIMS already in 2016[165] and the idea is still prominently discussed in the European Commissions' recently adopted Data Strategy[166]. The idea of PIMS is much older, however, and dates back to the mid 90s when Laudon (1996)[167] envisioned the creation of a national information market, where data subjects can deposit their information in bank-like institutions and are compensated for the use of their data.

More specifically, according to a review of proto-typical PIMS by MyData[168], the key functionalities of PIMS are

- Identity management: Authentication at various services
- Permission management: Overview of data transactions and connections, including management of legal rights and consent
- Service management: Linking various data sources
- Value exchange: Accounting and capturing the value of data, including remuneration (personal data broker)
- Data model management: Managing semantic conversions (schemas) from one data model to another
- Personal data transfers: Implementing interfaces (APIs) for standardised and secure data exchange between various data sources and data recipients

[164] See https://mydata.org
[165] See https://ec.europa.eu/digital-single-market/en/news/emerging-offer-personal-information-management-services-current-state-service-offers-and
[166] See https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
[167] Laudon, K. C. (1996). Markets and privacy. Communications of the ACM, 39(9), 92-104. Available at: https://dl.acm.org/doi/pdf/10.1145/234215.234476?casa_token=gULl7E4j-L8AAAAA:AqclO47nFQjPs62Gz1bj4ROeHOl47OoceFOKFn8u7eE01TUIVVJtddSqBxxkEpnz1Xgpilpa2VNXbg
[168] See https://mydata.org/wp-content/uploads/sites/5/2020/04/Understanding-Mydata-Operators-pages.pdf

- Personal data storage: Storing data from various sources, including data generated directly at the PIMS.
- Governance support: Ensuring compliance with legal frameworks
- Logging and accountability: Keeping historic logs of any data access and exchange facilitated by the PIMS

However, it is not always useful for PIMS to offer all of these functionalities. For example, as we will highlight below, whether or not PIMS should engage in value exchange (which we denote as personal data brokers) is debatable. It is also noteworthy that in the context of the digital economy, some of the key functionalities of PIMS are currently offered by large digital platforms directly. For example, the Data Transfer Project (see 0), which is backed by some of the largest digital platform provider, is a PIMS focused on personal data transfers and data model management. But possibly more importantly, large online platforms also offer online identity management solutions, i.e. registration and authentication of a user at various online services. For example, this is currently offered like Google, Facebook, Amazon, Microsoft, LinkedIn or Twitter. Thereof "Sign in with Google" and "Sign in with Facebook" are probably the most well-known.[169]

This begs the question whether, in the context of the digital economy, PIMS stand a chance to operate independently of the big tech firms, as neutral stand-alone brokers that can truly empower users to exercise control over their personal data. We explore this issue from an economic perspective in more detail below. This means that users see added value in the adoption of a PIMS, and that the PIMS can find a sustainable business model. We see problems on both accounts.

## 5.2 *Lack of (de-facto) standards and APIs*

The central premise of PIMS for users is that they offer a centralised dashboard that seamlessly integrates with the various services that they are using, offering key functionalities such as identity management, permission management and data transfers. This requires a common set of de-fact standards and high-performance APIs (see Section 0) through which a PIMS would be able to access the various services and users' data.

However, to date, such common standards are lacking. Instead, data integration is rather done through individual solutions, customised for each service, either using existing APIs (with rate and other access limitations) or through web scraping. Some view this as a central role of PIMS, because in this way PIMS enable some limited portability in an otherwise incompatible and non-standardised data ecosystem.

However, we view this with some scepticisms, because this approach is not scalable to a large set of services, and access hinges on the goodwill of the data provider, who may at any time make changes to the data access or undermine it. At the same time, without a significant user base, any given PIMS does not have sufficient leverage to set a standard on its own.

By contrast, large digital platforms, such as Google or Facebook, have successfully leveraged their vast user base to induce many independent service providers to implement their standards, such as their single-sign-on solutions.

[169] See, for example, https://www.avg.com/en/signal/is-it-safe-to-log-in-with-facebook-or-google

As explained in Section 0, in theory, widespread availability of APIs or a common export standard would alleviate this problem, because then network effects do not matter anymore. As long as all entities (PIMS and data controllers) can communicate thanks to common standards and interfaces, even a PIMS with only a few customers would be able to offer its consumers a comprehensive service. Several PIMS could even co-exist and thus PIMS could even compete for customers, as switching would be easy, thanks to the common standards, APIs, and the right to port volunteered data.[170] In this context, one must differentiate, however, at least between the

- standards for managing consent
- standards for authenticating the user and
- standards for actually transmitting and possibly storing personal data

in order to enable key functionalities of PIMS. While OAuth (2.0) seems to be the de-facto standard for authentication, which is also used by "Login with Google" or "Login with Facebook" as well as in the Data Transfer Project, there are many implementation details that yet need to be considered (see Section 1), and even with a common standard, centralised controlled could be retained, e.g., through the centralised control of crucial resources (such as token management in the context of OAuth) and rate management of APIs.  Yet, in the other two domains even more isolated implementation approaches exist (see Section 0), and there is currently an ongoing development and debate how to design such standards. Recently, solutions based on blockchain designs have surged (see, e.g., Zyskind and Nathan 2015[171] as well as several industry initiatives[172]), because these promise a decentralised framework that could do without a centralised control and oversight. It yet remains to be seen, however, whether these solutions are practical and scalable.

## 5.3 *Lack of sustainable business models*

Second, there seems to be a lack of a sustainable business models for PIMS that is not build on selling users' data. Indeed, if we look beyond the need for standards and API access to connect a user's various data sources in a centralised PIMS, the question arises how the business model of a privately-financed 'neutral' data broker can ever be sustainable. In principle, there are three potential sources of revenues for a purely privately-financed PIMS, data markets, data controllers and consumers. If all of these turn out to be not sustainable, there may also be a role for public subsidies. We discuss each of these possibilities in turn.

### *5.3.1 Generating revenue from data-driven services or on data markets*

By data markets we mean any market where (access to) data or insights derived from data can be monetised. In particular, this can be advertising markets, the market for customer analytics services, and the market for data intermediaries (selling access to raw data). PIMS would then generate revenues in much the same way as the original data controllers (such as Google or Facebook) from which the data was transferred to the PIMS. Why would consumers then want to transfer their data to the PIMS at all? We see three reasons:

[170] Clearly, consent notifications given to a PIMS would qualify as volunteered data.

[171] Zyskind, G., & Nathan, O. (2015, May). Decentralizing privacy: Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops (pp. 180-184). IEEE.

[172] For example, by Microsoft (https://qz.com/989761/microsoft-msft-thinks-blockchain-tech-could-solve-one-of-the-internets-toughest-problems-digital-identities/ ) and Orbiter (http://www.orbiter.de/english/)

First, in this way consumers could exert some competitive pressure on data-rich platforms. In theory, PIMS could even have better data on its customers than any given data controller, precisely because PIMS have the possibility to aggregate data from various data controllers. That is, data sets might have greater 'depth' (i.e. more variables/columns per observation/row). In practice this is not very likely, however. At least not compared to large online platforms which have the ability to track consumers' activity across multiple websites and services. In reverse, PIMS can only sell data from consumers that use the PIMS, and thus, data sets have less 'breadth' (i.e. less observations/rows). Even if the PIMS would have the same ability to generate insights from data and to offer data-intensive services, the extent to which PIMS can indeed exert competitive pressure and be a successful actor on the data markets is not clear. In this context, it is important to recall our discussion on the potentially fierce competition that could come along with selling identical data sets (see Section 4.1.3.2).

Second, users would have more control over where and which data is sold (data trust). This could be an incentive to transfer data to the PIMS in its own right. However, this additional control and transparency relates only with respect to the additional data sales by the PIMS, and not those done by the data controller from which the data was transferred. Thus, if privacy is of concern to users, they first create an additional problem (selling more of their data) which they can then partially fix. This does not seem to be a very convincing incentive for consumers to transfer data. This may change, however, if, like in the California Consumer Protection Act (CCPA), consumers would additionally have the right to opt out of the sale of their personal data at the original data controller. Precisely, CCPA (Cal. Civ. Code § 1798.135(a)(1))[173] states that a business that falls under the CCPA[174] shall

> *"Provide a clear and conspicuous link on the business's Internet homepage, titled "Do Not Sell My Personal Information," to an Internet Web page that enables a consumer, or a person authorized by the consumer, to opt-out of the sale of the consumer's personal information. A business shall not require a consumer to create an account in order to direct the business not to sell the consumer's personal information."*

If consumer had the same right under European law, this would mean that a consumer could deny the original data controller to sell its data, and transfer it to a PIMS, who would then sell the data respecting the user's fine-granular control and consent options. This would indeed offer consumers more control over which data and to whom data is sold. PIMS could even compete among each other on the basis of finer control rights for the sale of data.

However, this would likely induce the original data controller to also offer consumers finer control rights with respect to how their data is sold – instead of just the full opt-out mandated by law. This, in turn, would give consumers less incentives to port their data to the PIMS in the first place. Consequently, from this view – and only if CCPA-like regulation would be adopted in Europe as well – PIMS could induce large online platforms to give users more control rights over how their data is used, because the market would become more contestable; but under this view, PIMS would

[173]See https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.135

[174] The CCPA applies to any business, including any for-profit entity that collects consumers' personal data, which does business in California, and satisfies at least one of the following thresholds: i) Has annual gross revenues in excess of USD 25 million; ii) Buys or sells the personal information of 50,000 or more consumers or households; or iii) Earns more than half of its annual revenue from selling consumers' personal information.

probably never actually have a significant amount of customers, and would eventually only serve as a competitive threat to achieve market contestability according to the contestable markets theory (Baumol, 1985)[175]. It is questionable whether this business model is sustainable, especially if setting up a PIMS involves significant fixed costs or venture capital, because PIMS would constantly be in a potential 'kill zone'. Clearly, everything else being equal, consumers would find it easier to control their data directly at the original platform than to port it to a PIMS first. This gives the original platform a competitive advantage over a PIMS that would allow it to foreclose the PIMS from entry. Nevertheless, the threat of entry by a PIMS remains, depending on shadow costs of entry, and disciplines the incumbent accordingly.

Third, and probably the most important incentive for consumers to transfer data to a PIMS under this revenue-generation scheme, is that the PIMS could pay consumers for their data. In other words, the PIMS would become a Personal Data Broker (PDB), who sells personal data on behalf of the users, and offers users financial rewards in return (also called value exchange above). Consequently, PDBs are not just promising users more control over who they sell the data to, but foremost that users can financially participate from the commercialisation of their data. This is also the vision that was expressed already by Laudon (1996) and later by Larnier (2014)[176], who also coined the term "data as labor" (Arrieta-Ibarra et al., 2018)[177]. Indeed, such PDB business models are currently being pursued in practice, such as the joint venture between digi.me and UBDI (which stands for "Universal Basic Data Income")[178]. However, similar previous PDBs, such as Datacoup[179], have already failed and paid consumers only minimal rewards. According to Wikipedia[180] in the trial phase, Datacoup offered each user up to USD 5 per month, and in the beta phase up to USD 8 per month in return for access to user accounts of various social networks such as Facebook and LinkedIn, as well as to debit and credit card transactions. However, in November 2019 Datacoup announced its users that it is closing down, and had actually never sold any of their data up to this point. Instead, all payments had been made from the Datacoup treasury account. Other examples of PDBs are people.io (who seem to face similar issues[181] as Datacoup), Datum[182] (where data can be sold in return for cryptocurrency), ItsMyData[183] (which plans to pay consumers in the future, but does not do so yet[184]), and Wibson[185] (where users can earn tokens that can be redeemed in a marketplace; the market place has not been launched yet, however[186]). Even the large telecom operator Telefonica has announced a PIMS with PDB[187] in 2017, which they call 'Aura' (in fact a partnership with people.io), but this project has never taken off the ground.

Thus, while there is an emerging offer of PIMS that promise consumers to redeem them for their data (in the future), none of them currently seem to have a sustainable business model. Rewards

[175] Baumol, W. J. (1986). Contestable markets: an uprising in the theory of industry structure. Microtheory: applications and origins, 40-54.
[176] Lanier, J. (2014). Who owns the future?. Simon and Schuster.
[177] Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Weyl, E. G. (2018, May). Should We Treat Data as Labor? Moving beyond" Free". In AEA Papers and Proceedings (Vol. 108, pp. 38-42).
[178] See https://www.marketplace.org/shows/marketplace-tech/an-app-that-pays-you-for-your-data-yes-actually/
[179] See https://www.datacoup.com
[180] See https://en.wikipedia.org/wiki/Datacoup
[181] See https://uk.trustpilot.com/review/people.io
[182] See https://www.datum.org
[183] See https://itsmydata.de/?lang=en
[184] See https://www.faz.net/aktuell/wirtschaft/digitec/start-up-it-s-my-data-moechte-die-demokratisierung-der-daten-16328619.html
[185] See https://wibson.org
[186] See https://medium.com/wibson/wibson-update-01-03-2020-352e9a422438
[187] See https://www.ft.com/content/3278e6dc-67af-11e7-9a66-93fb352ba1fe

are either very low or not being paid out yet. This is also in line with the game-theoretical model by Haberer, Krämer and Schnurr (2019)[188] who show that the incumbent platform will strategically react to the emergence of PDBs by adapting the quality of its online service. In cases where the PDB is a relatively weak competitor on the data market (i.e. the PDB is not very successful in monetising user data on the data market), the PDB is either foreclosed by the incumbent, or will only be able to pay out a minimal reward. Overall consumer welfare will decrease in this range, because the incumbent platform reduces its quality in order to deter the PDB. Consumer benefit only if the PDB is a relatively strong competitor (i.e. is very successful in monetising user data). In this case, the PDB pays users a positive and significant reward. However, in this case the platform will also start to charge users for access and not offer its service for 'free' anymore. In this way, the platform can appropriate some of the additional consumer surplus that was created by the PDB. This highlights that PDBs may well change the business model of incumbent platforms from a free (e.g., advertising based) to a paid (e.g., subscription based) business model.

Moreover, paying users for their data also gives rise to an ethical issue. Such practice would quickly reveal that the data of some users is more valuable than the data of others. Even worse, the 'valuable users' are likely to be the most economically advantaged anyway. One interesting feature of the current zero-price (ad funded) business models in the digital economy is that everyone can access the same services, irrespective of how valuable their own data actually is. PDBs could change that and indeed, some low value users might find they have to start paying for services that were previously 'free', whilst high value users get paid to use them.

Relatedly, Bergemann, Bonatti and Gan (2020)[189] as well as Acemoglu et al (2019)[190] highlight the 'social dimension' of data, which reduces the value and monetary compensation for individual data points. Their argument is that data revealed by one individual also reveals information about other, similar individuals. This creates a data externality. When similar users have already revealed data to a data intermediary (a platform, or a PIMS), then the value of additional data by similar users is lower. This leads to an unravelling, whereby consumers with the lowest privacy preferences sell their data first, so that the data intermediary can acquire (statistical) information about users at relatively little costs. This social externality of data fundamentally undermines the idea that 'data ownership' of one sort or another actually empowers consumers to receive a 'fair' and significant remuneration for their personal data.

### *5.3.2 Generating revenue from data controllers*

An alternative way to generate revenues for PIMS is to offer online service providers a convenient tool by which they can be compliant to the seemingly complex and evolving legal frameworks that have been established by GDPR, CCPA, and others yet to come. In this case, the PIMS serves as compliance service, which is to the benefit of the user (who can exercise his or her rights conveniently) and of the online service provider (who does not have to worry about compliance).

[188] Haberer, B., Krämer, J., & Schnurr, D. (2018, August). Standing on the Shoulders of Web Giants: The Economic Effects of Personal Data Markets. Working Paper. Available at: https://pdfs.semanticscholar.org/1067/8e8c90d6fbf319eacc91cba9ab691845b1c2.pdf

[189] Bergemann, D., Bonatti, A., & Gan, T. (2020). The economics of social data. Cowles Foundation Discussion Paper No. 2203R. Available at: https://ssrn.com/abstract=3548336

[190] Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). Too much data: Prices and inefficiencies in data markets (No. w26296). National Bureau of Economic Research. Available at: https://economics.harvard.edu/files/economics/files/acemoglu_spring_2020.pdf

Such a business model is pursued, e.g., by Datawallet.[191] Interestingly, Datawallet initially started out with the idea of a PDB in the sense discussed above. However, the company recently shifted focus and now clearly advertises itself as a compliance tool for service providers. The revenue model rests exclusively on charging service providers, but not on charging consumers. Nor do they seek to make money by selling user data on their own.

It is unlikely, however, that this business model will attract the current data rich firms as customers. Large online platforms have sufficient scale to handle compliance with GDPR and CCPA on their own. Thus, the business model is clearly targeted at small and medium sized services and in this sense a welcomed addition to the data ecosystem. However, PIMS pursuing this business model will have little impact on the data ecosystem for personal data, because they do not exert competitive pressure on large data rich firms. This also means that this business model may well be sustainable, because it is unlikely that such PIMS are entering the 'kill zone'.

### *5.3.3 Generating revenues from users*

Some observers have noted (e.g., Section 4.3.3 of the Opinion of the German Data Ethics Commission (2020)[192]) that any business model that depends on generating revenues from profit maximising data controllers is problematic per se. PIMS should act in the best interest of consumers, and not in the best interest of those that handle or monetise consumers' data. Business model, which collect a flat subscription fee from users, which does not rely on the type or amount of data handled by the PIMS, are therefore preferred. Evidently, the question is how sustainable this business model is. Especially, if PIMS rely on a common set of standards, and therefore entry costs are relatively low, competition between PIMS that rely only on a flat subscription fee from users is likely to be fierce. At the same time, PIMS should offer a secure and reliable architecture for controlling personal data, and should not see cost-cutting as their primary concern to stay in business. This tension may only be resolved by effectively limiting the number of PIMS available, e.g., through licensing.

### *5.3.4 No revenue generation*

The preceding discussion highlighted that privately funded PIMS, which rely on revenue generation from users, from the data controllers or on the data markets, may either not be sustainable or not have a significant impact on the data ecosystem. This may give rise for governmental intervention or PIMS which are not financed privately. If PIMS are indeed seen as a central element to empowering users, state subsidised or even state-run PIMS may in fact be the only option to address this market failure.

However, two potential caveats of state-run PIMS are worth mentioning here. First, the state is often a bad investor and innovator compared to private firms. This seems especially problematic in a highly dynamic and complex environment like the data economy. Second, it is not clear – from the perspective of the users – that the state is the better controller of personal data. In some jurisdictions, consumers may have larger distrust in the government handling their data than a

[191] See https://www.datawallet.com

[192] German Data Ethics Commission (2020). Opinion of the Data Ethics Commission. Available at: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3

private firm. Although there may be technical solutions to ensure that data indeed remains private, and cannot be intercepted by the state (e.g., through cryptographic means such as blockchain solutions), it is not clear whether this is indeed a convincing argument for non-experts. Moreover, in some jurisdictions, such as the US, consumers have heightened privacy rights vis-à-vis the state compared to their privacy rights vis-à-vis private firms. In the European Union this does not apply, however.

A final option may be to rely on open-source, not-for-profit solutions for PIMS. It is not unlikely that such solutions may emerge, particularly when there are agreed-on standards on which such solutions can be built. Ongoing projects, such as the Data Transfer Project or Solid are indeed examples for such open-source not-for-profit solutions. But also in these case, policymakers may take a more active role in facilitating the emergence and use of such PIMS, for example by setting common standards or by reducing information asymmetries though audits. We will return to this point in our policy recommendations.

# 06

# INCREASING THE EFFECTIVENESS OF DATA PORTABILITY IN THE DIGITAL ECONOMY

# 6. Increasing the effectiveness of data portability in the digital economy

Having laid out the complex legal, technical and economic considerations that arise in the context of data portability, within and possibly beyond the current EU legal framework, we now collect the gathered insights and derive concrete policy recommendations for increasing the effectiveness of data portability in the digital economy.

## 6.1 *The issues*

The right to data portability has been in effect for just about two years, but to date empirical and theoretical research on its economic consequences is scant. In the previous Section, we have identified several issues but also possible solutions under the given legal framework, which give rise to a number of recommendations how data portability can be made more effective in the context of the digital economy.

First, we highlighted that the **collection of personal data is highly concentrated** in the digital economy. The issue arises primarily with respect to observed data (tracking data, clickstream data, behavioural data) and to a lesser extent with volunteered data. Volunteered data also tends to be more static, whereas observed data has a more dynamic character, i.e. it is generated at a much higher rate. It is therefore primarily the access to observed data, which is seen problematic under the current legal regime. While we have made clear that observed data should be included in data portability requests, the static and infrequent nature of a data portability request often diminishes the usefulness of observed data for other applications. Here, a more dynamic and continuous data portability would be desirable to overcome this issue.

Second, we have also argued that **widespread data portability, including both volunteered and observed data, is likely to render digital markets more competitive and innovative**. While there is a lack of empirical studies to back or refute this claim, we have argued that freeing personal data from organisational silos would enable more decentralised innovation, which could also occur more independently at the service and the analytics level. We have also argued that, due to inherent concentration in the collection of observed data, it is desirable to have competition rather at the level of inferred data and analytics, but not in the collection of data. Taken together, this provides a strong rationale to facilitate data portability of `raw’ user input data (i.e. both volunteered and observed), but not derived and inferred data, as much as possible. This will also likely require to educate consumers on their rights, to make the data available to them transparent, and to derive technical solutions (through PIMS or other means) so that data portability is just a click away.

Third, there are **numerous technical difficulties that arise from different standards and data formats that may be used following a data portability request**. In particular, the sending provider must not adhere to a certain standard and can change it at any given point in time. We have described the various technical complexities in Section 3. These uncertainties regarding standards and their perseverance can make it very costly for the new provider to offer an interface to import data. In return, this means that more stringent and common standards for data portability are a key to ensuring that data is more widely imported and used. The provisions in GDPR, which merely call for a “structured, commonly used and machine-readable format” are not enough. If the same type of data (e.g., photos, videos, search logs) would be made available in the same format, irrespective of the provider, then it would be more feasible to develop and provide respective import adapters.  A more widespread availability of such adaptors and re-usability of ported data would also raise the awareness among users and encourage them to port their data. The transfer could further be facilitated by PIMS, who could perform schema mappings between various services.

Fourth, given the novelty of the right to data portability, firms also raise **legal concerns that might arise when including data in data portability requests and when accepting data from other providers**. This includes potential conflicts of rights, especially regarding the porting of data provided by the data subject on other data subjects (e.g., address books, or pictures in which other people are tagged). But legal concerns also arise with respect to liability issues, such as who is responsible if data is lost or modified in the transfer process. The White Paper on Data Portability by Facebook (2019)[193] summarises these legal concerns well. As explained in Section 2, some of those concerns can be addressed with the current legal rules. However, in order to encourage that more is included under the scope of data portability and that firms are more willing to import data, especially in the context of the digital economy, more legal certainty and guidance would be welcomed. Moreover, there may be a role for regulatory testbed, where innovative start-ups accepting ported data, could work more closely together with the privacy-regulator in order to develop legally sound and economically viable solutions.

Fifth, we have highlighted that, from a technical perspective, **PIMS are an important and welcomed addition to the data ecosystem**. However, the existing offers are still in its infancy and we have also raised doubts that, from an economic perspective, PIMS may find a sustainable business model, especially if they are indeed acting as a neutral data broker. A minimum requirement to make PIMS feasible is to develop common standards and APIs through which PIMS can interact with the various services in a standardised and immediate way.

Sixth, **to date there is limited evidence that data portability is widely used**. Rather, we think that the root of the problem lies in the evident **chicken-and-egg problem**. Not at least for the reasons given above, currently very few providers to indeed accept ported data from users. If data is imported, it is often not done via the data set that a user has exported following a data portability request, but rather through existing APIs or other workarounds. In reverse, this means there is a lack of use cases for consumers to exercise their right to data portability. We believe that more continuous and standardised data portability is key to overcoming this chicken-and-egg problem.

Moreover, the experience from telecom markets (number portability) shows that portability became widely adopted when the consumer merely needs to give consent, but the (technical) details of exchange are deliberated by the sending and receiving data controllers directly according to some standardised process. The experience from other industries, foremost the Open Banking Order in the UK, highlights that third-parties often do see a value in importing data, and that data importing becomes more likely when standards are in place that allow for a continuous importing of data. In the case of Open Banking, after a slow start, there has been a continuous increase in both the number of third-parties accessing the available APIs as well as in the number of API calls being made.[194]

[193] Facebook (2019). Charting a Way Forward: Data Portability and Privacy (September 2019). Available at: https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf

[194] See https://www.openbanking.org.uk/providers/account-providers/api-performance/

Finally, it important to note that, **even if data portability would be frictionless for consumers and providers, there may are also be good economic reasons not to import data**, e.g., because they have not created this data themselves and therefore have difficulties to assess the quality of data. For example, it is often argued that it would be valuable to import reputation and product review data from another platform. Even if such data can be ported, it is questionable to what degree, say a positive review of a user on an e-commerce platform has economic value for an accommodation platform. Similarly, it will be very context specific, whether or not data in the provided granularity is useful. For example, even within the firms contributing to the Data Transfer Project, which already offers a relatively frictionless environment for data portability, as of March 2020, Facebook (including Instagram) and Twitter only allow to export data (video and photos), but provide no adapter to import data.

Taken together, we therefore see scope for improvement in three areas: (i) effective enforcement of the current legal framework, (ii) a new right for continuous data portability, tailored for the digital economy, and (iii) enabling PIMS through standards. We discuss each in turn.

## 6.2 *Effective enforcement and clear scope of data portability under GDPR and DCD*

A first set of recommendations entails effective enforcement and legal certainty on existing legal frameworks for data portability, particularly Article 20 General Data Protection Regulation (GDPR). The objective of the GDPR and Article 16 of the Digital Content Directive (DCD) is to facilitate a one-time switching of consumers between services. As explained in Section 2.2, both legal provisions are complementary and allow consumers to comprehensively port their personal *and* non-personal data when switching from one digital services provider to another.

However, as we have highlighted throughout the report, in the context of fast-moving digital markets, the provisions of the GDPR and DCD may not be enough to actually achieve this purpose (see Section 4). Data porting under DCD requires to terminate the contract with the previous provider, and therefore does not allow for multi-homing in order to trying out new services. Since digital services are often experience goods, whose value is known to users after they have been used, consumers may be reluctant to invoke Article 16 DCD too often. Likewise, the scope of Article 20 GDPR regarding portability of observed data is not very clear, and will be clarified in future case proceedings. Moreover, there are several tensions arising from Article 20 with other provisions in the GDPR such as data minimisation (see Section 2.1.4), which would require clarification. This is especially the case in the context of the digital economy, where the collection of personal data is ubiquitous, and often occurs in the form of observed rather than volunteered data. This is particularly evident in the tracking of consumers behaviour across several websites (see Section 4.1.3.1). Taken together, this creates legal uncertainty not only for providers but also for consumers.

### *6.2.1 Legal certainty on the scope and trade-offs of data portability right*

Thus, a first priority for policymakers is to increase the legal certainty with regard to **the scope and the limits of data portability under Article 20 GDPR.** In the context of the digital economy, where data is always processed by automated means and every click is potentially recorded, the tensions between purpose limitation, data minimisation and data portability are particularly immanent. More guidance is needed on issues like:

- To what extent exactly is **observed data** to be included in a data portability request? As laid out under Section 4.3, a wider scope of data portability, including both volunteered and observed data, is desirable to stimulate data-driven innovation outside the current silos and is covered by the GDPR according to the EDBP.
- In particular, does observed data include **location, tracking and clickstream data** (before being analysed or refined)? If so, how much context to such clickstream data should and needs to be made available so that data subjects can truly assess the information content of that data (e.g., exactly which content was consumed, exactly which ads were clicked on)? What are objective legal, economic or technical reasons not to make location, tracking and clickstream data available? For example, are concerns about data security and about a possible loss of reputation due to data leakage or misuse at the end of the receiving data controller admissible? When exactly is technically infeasibility admissible as a defence for data rich firm in the digital economy?
- Is there an obligation for data controllers to install measures and tools so that every data subject must make an explicit decision on whether **they *consent or dissent* in case another data subject** asks to port data that affects their data rights (e.g., if a photo is to be ported on which the data subject is tagged)? What about data subjects who do not have a contract with the data controller (but, e.g., a photo with their name tagged nevertheless exists with that data controller and is to be ported)?
- If some portable data **affects data *rights of other data subjects*** (and some of those data subjects have dissented to porting), does this mean that no data can be ported, or must the data controller offer to port at least the portion of the data that does not affect data rights of other subjects?

Some of these issues relate directly to the questions that data rich firms have raised (e.g., Facebook in their White Paper on "Data Portability and Privacy"[195]) when trying to move forward in the context of data portability. Legal clarity which is in line with the realities of the digital economies is needed so that Art. 20 GDPR, will be effective.

We realise that at some point these questions can become so complex that a **case-by-case analysis** is necessary. In this case, it should be clear what are the main interests of the trade-offs and where firms and consumers can find legal guidance on the balancing of those trade-offs in a timely manner. In particular, in these cases, providers willing to facilitate data portability for consumers should be able to receive specific guidance by the privacy regulator in a cooperative approach. In this context, it is also worthwhile to discuss the use of **sandbox-regulation**, as it is

[195] See Egan, E. (2019). Egan, E. (2019). Charting a way forward: Data portability and privacy [White Paper]. Facebook. https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf

the case in Open Banking, in order to provide a safe-harbour under which data portability can be developed further.

### *6.2.2 User-friendly transparency on data*

A second priority is that there should be **more transparency about the categories and extent of personal data that firms in the digital economy hold about a certain data subject**. This information should be readily available to users already before a formal access request (Article 15(3) GDPR) or data portability request (Article 20 GDPR) is initiated. Data subjects already have these rights under Articles 12 and 15 GDPR, but currently there still seems to be, in some cases, a lack of transparency concerning the actual extent of data collection pertaining to each data subject (e.g., on the extent of tracking data).

In our view, this information can be made more transparent and accessible to data subjects in the context of digital service providers, e.g., through the use an appropriate dashboard in the respective user's privacy settings. To be clear, several large online platforms, including Google and Facebook, already provide comprehensive dashboards.[196] Such dashboards could then also be used to consent to data portability requests of other data subjects for individual data categories.

### *6.2.3 Effective monitoring and enforcement*

A third priority is that there should be an **effective monitoring and enforcement of the existing provisions on data portability** under GDPR. This requires first that the scope and the limits of these provisions are clear in the context of the digital economy (see first priority) and that users are well aware about the data that is available about them and can be ported (see second priority). Then, there should be an effective monitoring and enforcement of the:

- timeliness in pursuing data portability requests relating to Article 12(3) GDPR,
- completeness of data (volunteered and observed data) in data sets created for portability,
- admissibility of technical feasibility constraints,
- admissibility of fees for data portability requests, particularly in the context of repeated requests relating to Article 12(5) GDPR.

## *6.3 Continuous Data Portability*

Data portability under the scope of Article 20 GDPR, when clarified and enforced effectively as recommended in Section 6.2, is a welcome and necessary step to empower consumers to exercise their privacy rights. In combination with Article 16 DCD, it should also facilitate switching from one digital service provider to another. However, we believe that one-off data portability according to Article 20 GDPR may not be sufficient to truly empower users in digital markets to foster competition and innovation. Often consumers want to try out a new service provider immediately, and that provider may be in need to cold start with the users' data in order to offer an immediately appealing service. But the GDPR does not give the consumers the right to immediate and very frequent access to their data. Consumers may have to wait up to a month or longer to receive the portable data from their current provider, and may face constraints regarding the frequency of these requests. Moreover, often consumers do not want to immediately switch to a new provider

[196] See, e.g. https://www.facebook.com/your_information/ and https://www.facebook.com/off_facebook_activity/

completely, but multi-home between providers first.[197] In this case consumers may not switch if they have to terminate their contract with the other provider in order to exercise their right to data portability (as under Article 16 DCD), and also a much more frequent porting of data that was is provided by Article 20 GDPR would be desirable.

### *6.3.1 Objectives and legal tools*

We believe that a more diverse data ecosystem is necessary and fruitful, where collection, analytics and monetisation of data are balanced across a number of stakeholders, each of whom is incentivised through a healthy state of competition in the respective area that it operates in and where innovation can spur, because (personal) data is freed from corporate silos if consumers consent to it (see Section 4).

As we have highlighted in Section 4.1.3, some of the personal data generated by data rich firms in the digital economy will not be easy to replicate. The real advantage of many firms rich in personal data is that they can combine personal data from many different sources seamlessly and in real time in order to create detailed user profiles. A one-off data transfer with a delay of up to a month is not consistent with this reality.

Moreover, there is currently a lack of standards and tools to enable data portability at the click of a button. However, as we have laid out in Section 3, in the context of the digital economy the technical barriers to establish such de-facto standards – albeit challenging – can be overcome and some tools, like the Data Transfer Project, are already demonstrating workable examples.

Article 20 GDPR (as well as its complementing Article 16 DCD) was not intended for continuous (i.e. very frequent) data transfers which would empower users to port their data (particularly observed data) from one service to the various other services that they may be using in (near) real-time. Following the deliberations in this report, we believe that this would be necessary to empower users to stimulate competition and innovation.

Moreover, many commentators agree that ex-post competition law is not the right instrument to address the data access and portability issues (see Section 2.4), especially if the purpose is to promote innovation (e.g., at the data analytics level) in general, and not to contest a specific market. The legal barriers to obtain data access under competition law are typically very high (for good reasons), interventions take a long time and, most importantly, competition law is not well-suited to develop timely and effective remedies in the complex environments of digital markets (see inter alia, e.g., Crémer et al. 2019, Furman et al. 2019[198], Feasey and Krämer 2019[199]). The advantage of data portability is that it can offer personal identifiable consumer level data. But its disadvantage is that only a relatively small fraction of consumers will ever port data, such that the data is not representative. Thus, data access requests under competition law (or some other legal framework) will continue to play a role in the future.

[197] In this sense also Crémer et al., 2019, p.82.

[198] Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. Report prepared for the Government of the United Kingdom, March.

[199] Feasey, R., Krämer, J. (2019). Implementing effective remedies for anti-competitive intermediation bias on vertically integrated platforms. Centre on Regulation in Europe (CERRE) Policy Report, 10/2019. Available at: https://www.cerre.eu/sites/cerre/files/cerre_intermediationbiasremedies_report.pdf

In summary, we therefore argue to investigate the need and feasibility of an new proportionate rule **enabling consumers to transfer their personal data (as under Article 20 GDPR) and their non-personal data (as under Article 16 DCD) in a timely and frequent manner from their existing digital service provider to another provider, at any given point in time.** This is what we refer to as '**continuous data portability**'.

This is not an entirely new policy proposal. It relates immediately to the "Smart Data" initiative in the UK which, is initially focussed on regulated industries, beginning with the Open Banking, but also seeks to include digital markets in the future.[200] Similar steps are being taking under the new Consumer Data Right (CDR) in Australia. The policy proposal also relates to the recently adopted European data strategy, who recognises that the "absence of technical tools and standards" makes the exercise of data portability burdensome.[201] Indeed, even several of the largest tech firms recently expressed their efforts to give users more control over their data and privacy.[202] Facebook CEO Mark Zuckerberg explicitly urged governments for more regulation, and identified data portability as one of four areas where such action should be taken.[203] The envisioned regulation on continuous data portability would be a step in this direction.

As there is a possibility that such regulation **amplifies the legal and economic risks and trade-offs** inherent to data portability, it is key that the legal uncertainties raised in Section 6.2 are thoroughly addressed first. Moreover, in accordance with the proportionality principle, the obligation to implement and enable continuous data portability should only be applicable when its benefits are likely to exceed its costs.

### *6.3.2 Guidelines for implementation*

First lessons for the implementation of such an extended right to data portability can be drawn from the Free Flow of Data Regulation (FFDR) and also from Open Banking. Like in the FFDR, as a first step, we propose a **participatory, adaptative**[204] and soft approach in the first phase. Thus, the regulation could require the establishment of **codes of conduct and agreements on common standards and APIs**, including performance criteria for the availability of these. We suggest that the following points should be included in such guidelines in any case:

- Consumers must be able give their *consent on a fine-granular level* regarding which data is to be transferred. All-or-nothing transfers are often not necessary, and would create more transaction costs, both technically (e.g., network load, space requirements) as well as economically (larger privacy concerns). They would also run counter the legal requirements of data minimisation under GDPR; firms shall not influence consent or disconsent by offering commercial incentives or disincentives.

[200] See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/808272/Smart-Data-Consultation.pdf

[201] Communication from the Commission of 19 February 2020, A European strategy for data, COM(2020) 66, p.10.

[202] See, e.g. https://eandt.theiet.org/content/articles/2020/01/google-ceo-backs-gdpr-says-privacy-should-not-be-a-luxury/

[203] See https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

[204] Ctrl-Shift (2018), *Data Mobility: The personal data portability growth opportunity for the UK economy*, Report for the UK Department for Digital, Culture, Media & Sport; P. Alexiadis and A. de Streel (2020), *Designing an EU Intervention Standard for Digital Platforms*, EUI Working Paper-RSCAS 2020/14

- Data should be able to be shared directly between firms, when the consumers have consented to this. Data portability should be possible without any additional infrastructure at the consumer end. However, this does not preclude the possibility that users employ PIMS to store data or to facilitate this process.
- Relatedly, the *nature and scope of the data ported should be very clearly communicated* to consumers, in plain language; generally the scope of portable data should be the same as under Art. 20 GDPR.
- The data transfer needs to be *secure*, minimising risks for data leakage to parties not involved in the transfer, data modification or loss of data.
- Where possible *open standards* and protocols should be used, which are free to use and transparent for developers (Furman Report, 2019, pp.71-74).
- APIs need to be available with a *high reliability and performance*. They should have the same performance and reliability as the interfaces that consumers otherwise use to interact with the digital service provider (as in the PSD2).

Several options are possible for the **policy process by which these guidelines are transformed into technical solutions**, ranging from industry-led self- and co-regulation to a standardisation body with legal powers.

We believe that the development of standards and codes of conduct should be **industry-led through multi-stakeholder groups**, as in the case of the FFDR. All parties involved should negotiate in good faith to achieve the best possible outcome in the interest of the consumer. Given the international nature of digital services and data standards, possibly, the development can be facilitated by independent international standard setting committees, such as the W3C. However, as such committees typically require unanimous decisions, it needs to be taken care that developments are not vetoed by single parties to protect their market power.[205]

Ideally, the development of standards and technical solutions can be built on existing projects such as the Data Transfer Project or Solid. Of course, the devil is in the detail and implementing this involves challenges, as the implementation of PSD2/Open Banking or cloud-based services like IaaS and SaaS have shown. Given the demonstration project of DTP and Solid, there does not seem to be a compelling technical reason why this is not feasible in a wider context (see Section 3). It is also to be expected that, once standards are defined and APIs are available, there will be a significant effort from the open-source community to provide import and export adapters between various services. There should be a timely deadline after which the progress and implementation status is evaluated by the Commission.

If no sufficient progress has been made by means in establishing standards and operational interfaces within a specified period of time, there **may be a need for stronger governmental intervention** or guidelines to ensure progress is made and neutrality of interests are warranted. For example, in Open Banking the major banks were required to constitute an independent trustee to develop standards. In the case of PSD2, relatively detailed technical provisions were adopted by the Commission on the basis of the participatory work done at the European Banking Authority.

[205] This has occurred, for example, recently where Google blocked a vote to give the W3C's privacy group more powers. See https://www.bloomberg.com/news/articles/2019-09-24/google-blocks-privacy-push-at-the-group-that-sets-web-standards

Similar case-by-case provisions are also done in Australian Consumer Data Right (CDR) initiative.[206]

The ultimate option is to enact a public standards organisation to achieve this end. For example, the Australian government has given a legal mandate the Data Standards Body to develop standards for data access and portability.[207] It works in close collaboration with the competition authority and the data protection authority.

### 6.4 *Enabling and Governing Personal Information Management Systems*

With a larger and continuous flow of personal data, facilitated by a right to continuously port data from large digital service providers, the role of **Personal Information Management Systems** (PIMS) is likely to become very important in practice. They could provide a centralised management of user's privacy settings and consented data flows; ideally aggregating relevant information across the various digital services that the consumer is using, and being able to change settings across several services as needed. In this sense, it would provide a dashboard of dashboards for user's privacy settings.

We think that this point should not be underestimated, because it is crucial that consumers are aware of their given consents and exercise their rights, particularly if this is the basis on which data is being shared between firms. In order to facilitate this, a centralised consent management is seen as crucial, as otherwise recent empirical studies suggest that this may lead to a vertical integration of PIMS with large platforms[208], which would run contrary to the intention of the PIMS being a neutral broker.

In order to enable a centralised consent management, first **additional standards for consent management need to be agreed over and beyond those needed for data transfers**.[209] Here, the same guidelines and recommendations as for the standards development for data transfer (Section 6.3.2) should apply. In order to facilitate this process at an early stage, additional funding for research and development on secure, decentralised and scalable solutions for consent management (e.g., based on blockchain technology) could be made available.

Second, as we have pointed out in Section 5, even if standards for data portability and consent management are developed and the policy recommendations under 6.2 and 6.3 are being pursued, PIMS may struggle to find a **profitable and sustainable business model**. Indeed, it is crucial that PIMS become and remain a neutral intermediary acting purely on the behalf of consumers. This also why it has been suggested, e.g., by the German Data Ethics Commission, that there should be regulatory guidelines on acceptable business models for PIMS, preferring, e.g., business models based on flat monthly fees for consumers. Again, we pointed out in Section 5 that we are doubtful that such a business model would be sustainable without further safeguards. Moreover, it raises the question whether PIMS should not be available free of charge to consumers, because

[206] See https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0

[207] See https://consumerdatastandards.org.au

[208] See Marthews, A. and C. Tucker (2019), Privacy policy and competition, https://www.brookings.edu/wp-content/uploads/2019/12/ES-12.04.19-Marthews-Tucker.pdf

[209] European Commission Services, *An emerging offer of Personal Information Management Services: Current state of service offers and challenges*, November 2016, p.16.

otherwise, there would be a monthly price tag on consumers' privacy management, which may not be in line with European privacy values.

However, we also expect that if such standards are in place, there will be considerable development in open source communities, providing decentralised non-profit solutions. Given the potentially sensitive nature of the data being handled through PIMS, there may still be a need for public oversight, such as through **privacy seals and certification**.[210]

To achieve critical mass for PIMS, a fruitful avenue may also be to build a user base on top of existing or developing identity management solutions. In particular, the European Commission is currently pushing national governments to offer an interoperable European identity management based on public national electronic identification (eIDs).[211] Moreover, during the European Council Meeting held on 10th March 2020, Heads of States and governments agreed to launch an initiative entitled European Digital Identity, "with the aim of developing an EU-wide digital identity that allows for a simple, trusted and secure public system for citizens to identify themselves in the digital space by 2027". This could also be a starting point to **couple identity management with consent management**, and to link the eIDAS regulation to the Digital Services Act, which is expected in about the same time frame.

[210] European Commission Services, *An emerging offer of Personal Information Management Services: Current state of service offers and challenges*, November 2016, p.12.

[211] See https://ec.europa.eu/digital-single-market/en/trust-services-and-eid

07

# CONCLUSIONS

# 7. Conclusions

Personal data portability, in the form of GDPR Article 20, was an important first step in empowering users to take their data, volunteered or observed, to wherever and whoever they wished. This has also led to a wider movement among policy makers to support data mobility, which also includes portability of non-personal-data, more generally. Examples of this are the Digital Content Directive and the Free Flow of Data Directive.

The specific impact of data portability, as currently implemented in the legal framework, is not clear as yet, since there is a lack of empirical studies on the topic. Nevertheless, what is already clear is that many data services do not yet offer import possibilities for ported data, and consequently, data portability in the legal sense is not used widely by consumers in the digital economy. Many commentators have argued that this is not actually a consequence of an inability to use data outside of the context where it was created. Rather, it is the result of a lack of common standards, and of APIs to access the data in a convenient and timely manner. In addition, it has raised legal issues over liability and protection of the rights of others.

In this report, we argue that all of these issues can be overcome, albeit potentially requiring considerable effort. Regarding the legal issues, specific guidance should be offered on how data portability can be facilitated and which data is subject to data portability in the digital economy without violating privacy rights. In particular, we advocate that observed data should clearly be included under the scope of data portability, with a wide interpretation of observed data adopted, including clickstream and tracking data, if available. In order to make the inherent trade-offs salient and to be able to resolve them, an open and constructive dialogue between data-intensive firms in the digital economy and regulators is needed. Such a dialogue could evolve around prototypical use cases for data types to be transferred, such as posts, videos, photos, search logs, clickstreams, geo-location or ad views. Eventually, data portability is also likely to require the collection of explicit consent from consumers on requests initiated by others that include their personal data and - in some cases – greater transparency over the personal data available for porting.

In addition, we also argue that continuous data portability, facilitated by common data standards and APIs, is technically feasible, albeit challenging. In fact, many providers already have such APIs in place, either privately or publicly accessible, often with considerable technical or commercial constraints. Moreover, demonstration projects, such as the Data Transfer Project, have highlighted that continuous data portability is technically feasible. We therefore consider it essential that, within the scope of digital markets, the obligation to offer standardised APIs to enable consumers to continuously port their data should be much more widespread. This approach echoes ongoing policies in the UK and Australia, and we believe that the European Commission, in its Digital Strategy, should follow suit. We believe that standardised APIs that enable continuous data portability are a prerequisite for encouraging more organisations to import personal data, and for encouraging more consumers to initiate such transfers. Ultimately, this is likely to spur innovation and competition in digital markets, although it is unlikely to disrupt existing market structures. Indeed, any such obligation must always bear in mind a proportionality principle; it should not be overly burdensome for small and emerging digital service providers.

Finally, we believe that Personal Management Information Systems (PIMSs) will have a crucial role to play in a digital economy where data portability is widely adopted. In particular, a PIMS should facilitate the complex consent management and offer users a centralised dashboard for monitoring and controlling the flow of their personal data. Data portability must be as straightforward as possible in order not to overwhelm consumers with choices that may ultimately prevent them from exercising their rights. This may well require educating and informing users on their rights through information campaigns running alongside the policy measures proposed here.

Moreover, we are sceptical that PIMSs can be economically self-sustaining and can find a business model where they can be a truly neutral agent that acts purely on behalf of the consumer. We suggest two avenues where a PIMS could be developed. First, standards for consent management could be established that will enable PIMSs to access and control the privacy settings in various services. Combined with appropriate certification, this could drive open-source solutions. Second, the EU could take a more active role by coupling development of its consent standards more closely to its ongoing efforts for a joint European identity management solution.

The proposed measures are likely to be a tedious but essential step in making data portability more effective and thus harnessing its true potential for competition, innovation and empowering users. We believe that the time to act is now. There should be no second guessing on how to make data portability more effective. All the measures proposed are aimed at empowering users, not competitors, and providing them with the tools required to share their data with whoever they wish.

# REFERENCES

# References

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web: From relations to semistructured data and XML*. Morgan Kaufmann.

Abiteboul, S., André, B., & Kaplan, D. (2015). Managing your digital life. *Communications of the ACM*, *58*(5), 32–35. https://doi.org/10.1145/2670528

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2019). *Too much data: Prices and inefficiencies in data markets* (NBER Working Paper No. 26296). National Bureau of Economic Research. https://economics.harvard.edu/files/economics/files/acemoglu_spring_2020.pdf

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. *The Quarterly Journal of Economics*, *120*(2), 701-728. https://doi.org/10.1093/qje/120.2.701

Alexiadis, P., & de Streel, A. (2020). *Designing an EU intervention standard for digital platforms* (Robert Schuman Centre for Advanced Studies Research Paper No. 2020/14). https://dx.doi.org/10.2139/ssrn.3544694

Argenton, C., & Prüfer, J. (2012). Search engine competition with network externalities. *Journal of Competition Law and Economics*, *8*(1), 73-105. https://doi.org/10.1093/joclec/nhr018

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Weyl, E. G. (2018). Should we treat data as labor? Moving beyond "free". *AEA Papers and Proceedings*, *108*, 38-42. https://www.doi.org/10.1257/pandp.20181003

Barker, A. (2020, February 26). 'Cookie apocalypse' forces profound changes in online advertising. *Financial Times*. https://www.ft.com/content/169079b2-3ba1-11ea-b84f-a62c46f39bc2

Baumol, W. J. (1986). Contestable markets: An uprising in the theory of industry structure. In *Microtheory: Applications and origins.* MIT Press.

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., & Carothers, G. (2014). *RDF 1.1 Turtle: Terse RDF Triple Language.* W3C Recommendation. https://www.w3.org/TR/turtle/

Beckett, D., Gandon, F., & Schreiber, G. (2014). *RDF 1.1 XML Syntax*. W3C Recommendation. https://www.w3.org/TR/rdf-syntax-grammar/

Beggs, A., & Klemperer, P. (1992). Multi-period competition with switching costs. *Econometrica: Journal of the Econometric Society*, *60*(3), 651-666.

Bergemann, D., Bonatti, A., & Gan, T. (2020). *The economics of social data* (Cowles Foundation Discussion Paper No. 2203R). https://dx.doi.org/10.2139/ssrn.3548336

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems*, *26*, 225-238. https://doi.org/10.1016/j.knosys.2011.07.021

Bourreau, M., & de Streel, A. (2020). *Big tech acquisitions: Competition & innovation effects and EU merger control*. Centre on Regulation in Europe (CERRE).

https://www.cerre.eu/publications/big-tech-acquisitions-competition-and-innovation-effects-eu-merger-control

Bray, T. (2017). *The JavaScript Object Notation (JSON) data interchange format* (RFC No. 8259). Internet Engineering Task Force. https://tools.ietf.org/html/rfc8259

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2008). *Extensible Markup Language (XML) 1.0 (5th ed.)*. W3C Recommendation. https://www.w3.org/TR/xml/

Brickley, D., Guha, R. V., & McBride, B. (2014). *RDF Schema 1.1*. W3C Recommendation. https://www.w3.org/TR/rdf-schema/

Cate, B. T., Dalmau, V., & Kolaitis, P. G. (2013). Learning schema mappings. *ACM Transactions on Database Systems*, *38*(4), Article 28. https://doi.org/10.1145/2539032.2539035

Competition and Markets Authority. (2016). *Retail banking market investigation*. https://www.gov.uk/cma-cases/review-of-banking-for-small-and-medium-sized-businesses-smes-in-the-uk

Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). *Competition policy for the digital era*. European Commission. https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf

Ctrl-Shift. (2018). *Data mobility: The personal data portability growth opportunity for the UK economy*. UK Department for Digital, Culture, Media & Sport. https://www.ctrl-shift.co.uk/reports/DCMS_Ctrl-Shift_Data_mobility_report_full.pdf

Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J. J., & McBride, B. (2014). *RDF 1.1 Concepts and Abstract Syntax.* W3C Recommendation. https://www.w3.org/TR/rdf11-concepts/

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Voss, P., & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. *ACM SIGOPS Operating Systems Review*, *41*(6), 205-220.

DB-Engines. (2020). *DB-Engines ranking*. https://db-engines.com/en/ranking

Drexl, J. (2017). Designing competitive markets for industrial data. *Journal of Intellectual Property, Information Technology & Electronic Commmerce Law*, *8*(4), 257-293.

Dublin Core Metadata Initiative. (2020). *DCMI metadata terms*. www.dublincore.org/specifications/dublin-core/dcmi-terms/

The Economist. (2018, June 2). *American tech giants are making life tough for startups*. https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups

Egan, E. (2019). *Charting a way forward: Data portability and privacy* [White Paper]. Facebook. https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf

Englehardt, S., & Narayanan, A. (2016). Online tracking: A 1-million-site measurement and analysis. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1388–1401. https://doi.org/10.1145/2976749.2978313

European Commission. (2009). Guidance on the Commission's enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings. *Official Journal of the European Union*, *C45/7*, 7-20.

European Commission. (2015). Directive (2015/2366/EU) on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC. *Official Journal of the European Union*, *L337/35*, 35-127.

European Commission. (2016a). *An emerging offer of Personal Information Management Services: Current state of service offers and challenges*. https://ec.europa.eu/digital-single-market/en/news/emerging-offer-personal-information-management-services-current-state-service-offers-and

European Commission. (2016b). Regulation (2016/679/EU) on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, *L119/1*, 1-88.

European Commission. (2017a). *Guidelines of Article 29 Data Protection Working Party on the right to data portability* (WP 242 rev.01). https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611233

European Commission. (2017b). *The new European interoperability framework*. https://ec.europa.eu/isa2/eif_en

European Commission. (2017c). Commission Delegated Regulation (2018/389/EU) supplementing Directive (EU) 2015/2366 of the European Parliament and of the Council with regard to regulatory technical standards for strong customer authentication and common and secure open standards of communication. *Official Journal of the European Union*, *L69/23*, 23-43.

European Commission. (2018a). Regulation (2018/1807/EU) on a framework for the free flow of non-personal data in the European Union. *Official Journal of the European Union*, *L303/59*, 59-68.

European Commission. (2018b). Regulation (2018/858/EU) on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, amending Regulations (EC) No 715/2007 and (EC) No 595/2009 and repealing Directive 2007/46/EC. *Official Journal of the European Union*, *L151/1*, 1-218.

European Commission. (2019a). Directive (2019/770/EU) on certain aspects concerning contracts for the supply of digital content and digital services. *Official Journal of the European Union*, *L136/1*, 1-27.

European Commission. (2019b). Directive (2019/944/EU) of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU. *Official Journal of the European Union*, *L158/125,* 125-199.

European Commission. (2019c). Directive (2019/1024/EU) of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. *Official Journal of the European Union*, *L172/56*, 56-83.

European Commission. (2019d, May 29). Guidance on the regulation on a framework for the free flow of non-personal data in the European Union. *COM (2019) 250*. https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-250-F1-EN-MAIN-PART-1.PDF

European Commission. (2019e, December 9). *Presentation of codes of conduct on cloud switching and data portability*. https://ec.europa.eu/digital-single-market/en/news/presentation-codes-conduct-cloud-switching-and-data-portability

European Commission. (2020, February 19). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data. *COM (2020) 66*. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

Facebook. (2020a). *Accessing & downloading your information.* https://www.facebook.com/help/1701730696756992

Facebook. (2020b). *Documentation*. https://developers.facebook.com/docs/

Farrell, J., & Shapiro, C. (1988). Dynamic competition with switching costs. *The RAND Journal of Economics*, *19*(1), 123-137.

Feasey, R., Krämer, J. (2019). *Implementing effective remedies for anti-competitive intermediation bias on vertically integrated platforms*. Centre on Regulation in Europe (CERRE). https://www.cerre.eu/sites/cerre/files/cerre_intermediationbiasremedies_report.pdf

Fing. (2018). *Data-responsible enterprises: User experience and technical specifications*. http://mesinfos.fing.org/wp-content/uploads/2018/03/PrezDataccess_EN_V1.21.pdf

Foster, D. (2004). *GPX: the GPS Exchange Format*. https://www.topografix.com/gpx.asp

Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). *Unlocking digital competition: Report of the digital competition expert panel*. Government of the United Kingdom. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. *New York University Law Review*, *94*(4), 737-771.

Gans, J. (2016). *The disruption dilemma*. MIT Press.

Gans, J. (2018). *Enhancing competition with data and identity portability.* The Hamilton Project. https://www.brookings.edu/wp-content/uploads/2018/06/ES_THP_20180611_Gans.pdf

German Data Ethics Commission. (2020). *Opinion of the Data Ethics Commission*. https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3

Google. (2020a). *Download your data*. https://support.google.com/accounts/answer/3024190

Google. (2020b). *Google APIs Explorer*. https://developers.google.com/apis-explorer

Graef, I., Tombal, T., & De Streel, A. (2019). *Limits and enablers of data sharing: An analytical framework for EU competition, data protection and consumer Law* (TILEC Discussion Paper No. 2019-024). https://dx.doi.org/10.2139/ssrn.3494212

Grill, A. (2013). *PeopleBrowsr and Twitter settle Firehose dispute.* PeopleBrowsr. https://www.peoplebrowsr.com/blog/2013/04/peoplebrowsr-and-twitter-settle-firehose-dispute

Grove, A. S. (1996). *Only the paranoid survive: How to exploit the crisis points that challenge every company and career*. Currency Doubleday.

Gu, Y., Madio, L., & Reggiani, C. (2019). *Data brokers co-opetition*. SSRN. https://dx.doi.org/10.2139/ssrn.3308384

Haberer, B., Krämer, J., & Schnurr, D. (2018). *Standing on the shoulders of web giants: The economic effects of personal data markets* (Working Paper). https://pdfs.semanticscholar.org/1067/8e8c90d6fbf319eacc91cba9ab691845b1c2.pdf

Hardt, D. (2012). *The OAuth 2.0 Authorization Framework* (RFC No. 6749). Internet Engineering Task Force. https://tools.ietf.org/html/rfc6749

Hagiu, A., & Wright, J. (2020). *Data-enabled learning, network effects and competitive advantage* (Working Paper). http://andreihagiu.com/wp-content/uploads/2020/05/Data-enabled-learning-20200426-web.pdf

Hippel, E. (2005). *Democratizing innovation*. MIT Press. https://ssrn.com/abstract=712763

IBM. (2013). *A roadmap for intelligent archiving*. IBM InfoSphere Optim Archive e-book. www.ibm.com/downloads/cas/PA9YRY1N

Ichihashi, S. (2019). *Non-competing data intermediaries*. SSRN. https://dx.doi.org/10.2139/ssrn.3310410

International Organization for Standardization. (2003). *Database language SQL.* (ANSI/ISO/IEC Standard No. 9075:2003). https://www.iso.org/standard/34132.html

International Organization for Standardization. (2006). *Information technology — Open Document format for office applications (OpenDocument) v1.0.* (ISO/IEC Standard No. 26300:2006). https://www.iso.org/standard/43485.html

International Organization for Standardization. (2016). *XLSX Strict (Office Open XML)*. (ISO Standard No. 29500-1:2008-2016). https://www.iso.org/standard/71691.html

International Organization for Standardization. (2017). *Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use*. (ISO/IEC Standard No. 19944:2017). https://www.iso.org/standard/66674.html

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better?. *Big Data*, *1*(4), 215–226. https://doi.org/10.1089/big.2013.0037

Kamepalli, S. K., Rajan, R. G., & Zingales, L. (2020). *Kill Zone* (CEPR Discussion Paper No. DP14709). https://ssrn.com/abstract=3594344

Kewisch, P. (2014). *jCard: The JSON Format for vCard* (RFC No. 7095). Internet Engineering Task Force. https://tools.ietf.org/html/rfc7095

Klemperer, P. (1987a). Markets with consumer switching costs. *The Quarterly Journal of Economics*, *102*(2), 375-394. https://doi.org/10.2307/1885068

Klemperer, P. (1987b). The competitiveness of markets with switching costs. *The RAND Journal of Economics*, *18*(1), 138-150.

Kolaitis, P. G. (2005). Schema mappings, data exchange, and metadata management. *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems – PODS '05,* 61-75. https://doi.org/10.1145/1065167.1065176

Krämer, J., & Wohlfarth, M. (2018). Market power, regulatory convergence, and the role of data in digital markets. *Telecommunications Policy*, *42*(2), 154-171. https://doi.org/10.1016/j.telpol.2017.10.004

Lam, W. M. W., & Liu, X. (2020). Does data portability facilitate entry?. *International Journal of Industrial Organization*, *69*, Article 102564. https://doi.org/10.1016/j.ijindorg.2019.102564

Lanier, J. (2014). *Who owns the future?*. Simon and Schuster.

Lambrecht, A., & Tucker, C. E. (2015). *Can big data protect a firm from competition?*. SSRN. https://dx.doi.org/10.2139/ssrn.2705530

Lampos, V., Miller, A. C., Crossan, S., & Stefansen, C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports*, *5*, Article 12760. https://doi.org/10.1038/srep12760

Laudon, K. C. (1996). Markets and privacy. *Communications of the ACM*, *39*(9), 92-104. https://doi.org/10.1145/234215.234476

Ledger, M., & Tombal, T. (2018). Le droit à la portabilité dans les textes européens: droits distincts ou mécanisme multi-facettes?. *Revue du Droit des Technologies de l'information*, (72), 25-44. http://www.crid.be/pdf/public/8456.pdf

Lenzerini, M. (2002). Data integration: A theoretical perspective. *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '02*, 233-246. https://doi.org/10.1145/543613.543644

Lerner, A. V. (2014). *The role of 'Big Data' in online platform competition*. SSRN. https://dx.doi.org/10.2139/ssrn.2482780

Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, *130*(4), 1941–1973. https://doi.org/10.1093/qje/qjv023

Li, X., Ling, C. X., & Wang, H. (2016). The convergence behavior of naive bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data*, *11*(1), 1–24. https://doi.org/10.1145/2948068

Macbeth, S. (2017). *Tracking the trackers: Analysing the global tracking landscape with GhostRank*. https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf

Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly*, *40*(4), 869-888. https://doi.org/10.25300/MISQ/2016/40.4.04

Marthews, A., & Tucker, C. (2019), *Privacy policy and competition*. Brookings. https://www.brookings.edu/wp-content/uploads/2019/12/ES-12.04.19-Marthews-Tucker.pdf

McLeod, J. (2020, February 7). Inside the kill zone: Big tech makes life miserable for some startups but others embrace its power. *Financial Post*. https://business.financialpost.com/technology/inside-the-kill-zone-big-tech-makes-life-miserable-for-some-startups-but-others-embrace-its-power

Microsoft. (2018, May 25). *Responding to GDPR data subject export requests for power automate.* https://docs.microsoft.com/en-us/power-automate/gdpr-dsr-export-msa

Microsoft. (2020). *Web APIs and embeddable controls*. https://developer.microsoft.com/en-us/web/apis/

Motta, M., & Peitz, M. (2020). *Big tech mergers* (CEPR Discussion Paper No. 14353). Centre for Economic Policy Research (CEPR). https://cepr.org/content/free-dp-download-31-january-2020-competitive-effects-big-tech-mergers-and-implications

OECD. (2019). *Enhancing access to and sharing of data: Reconciling risks and benefits for data re-use across societies*. OECD Publishing. https://doi.org/10.1787/276aaca8-en.

Özsu, M. T., & Valduriez, P. (2020). *Principles of distributed database systems (4th ed.)*. Springer.

PeopleBrowsr. (2012, November 28). *PeopleBrowsr wins temporary restraining order compelling Twitter to provide Firehose access.* https://www.peoplebrowsr.com/blog/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access

Prüfer, J. & Schottmüller, C. (2019). *Competing with Big Data* (TILEC Discussion Paper No. 2017-006). https://ssrn.com/abstract=2918726

Richardson, L., Amundsen, M., Amundsen, M., & Ruby, S. (2013). *RESTful Web APIs: Services for a changing world*. O'Reilly Media, Inc.

Rinehart, W. (2018, November 7). *Is there a kill zone in tech?.* Techliberation. https://techliberation.com/2018/11/07/is-there-a-kill-zone-in-tech/

Schaefer, M., Sapi, G., & Lorincz, S. (2018). *The effect of big data on recommendation quality: The example of internet search* (DIW Berlin Discussion Paper No. 1730). http://www.dice.hhu.de/fileadmin/redaktion/Fakultaeten/Wirtschaftswissenschaftliche Fakultaet/DICE/Discussi on Paper/284 Schaefer Sapi Lorincz.pdf

Schweitzer, H., Haucap, J., Kerber, W., & Welker, R. (2018). Modernisierung der Missbrauchsaufsicht für marktmächtige Unternehmen. Nomos Verlag. https://doi.org/10.5771/9783845296449

Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Melamed, A. D., & Morgenstern, J. (2019). *Stigler committee on digital platforms: Market structure and antitrust subcommittee report*. University of Chicago Booth School of Business. https://research.chicagobooth.edu/-/media/research/stigler/pdfs/digital-platforms---committee-report---stigler-center.pdf?la=en&hash=2D23583FF8BCC560B7FEF7A81E1F95C1DDC5225E

Shafranovich, Y. (2005). *Common format and MIME type for comma-separated values (CSV) files* (RFC No. 4180). Internet Engineering Task Force. https://www.hjp.at/doc/rfc/rfc4180.html

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2010). *Database system concepts (6th ed.)*. McGraw-Hill.

Smith, N. (2018, November 7). *Big tech sets up a 'kill zone' for industry upstarts*. Bloomberg. https://www.bloomberg.com/opinion/articles/2018-11-07/big-tech-sets-up-a-kill-zone-for-industry-upstarts

Snell, J., Tidwell, D., & Kulchenko, P. (2001). *Programming web services with SOAP: Building distributed applications*. O'Reilly Media, Inc.

Strauch, C. (2011). *NoSQL databases* [White Paper]. Stuttgart Media University. www.christof-strauch.de/nosqldbs.pdf

Thompson, H. S., Gao, S., Sperberg-McQueen, C. M., Mendelsohn, N., Beech, D., & Maloney, M. (2012). *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures*. W3C Recommendation https://www.w3.org/TR/xmlschema11-1/

Tombal, T. (2018). Les droits de la personne concernée dans le RGPD. In *Le règlement général sur la protection des données (RGPD/GDPR): analyse approfondie*, (44), 407-557. Larcier. http://www.crid.be/pdf/public/8347.pdf

Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, *54*(4), 683–694. https://doi.org/10.1007/s11151-019-09693-7

Twitter. (2020a). *How to download your Twitter archive*. https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive

Twitter. (2020b). *API reference index*. https://developer.twitter.com/en/docs/api-reference-index

Web Archive. (2020). *Latitude has been retired*. https://web.archive.org/web/20150814192105/https://support.google.com/gmm/answer/3001634?p=maps_android_latitude&rd=1

Wohlfarth, M. (2019). Data portability on the internet: An economic analysis. *Business & Information Systems Engineering*, *61*(5), 551–574. https://doi.org/10.1007/s12599-019-00580-9

Wright, A., Andrews, H., & Hutton, B. (2019). *JSON Schema: A media type for describing JSON documents* (Internet Draft). Internet Engineering Task Force. https://tools.ietf.org/html/draft-handrews-json-schema-02

W3C OWL Working Group. (2012). *OWL 2 web ontology language document overview (2nd ed.)*. W3C Recommendation. https://www.w3.org/TR/owl2-overview/

Zingales, L., & Rolnik, G. (2017, June 30). A way to own your social-media data. *The New York Times.* https://www.nytimes.com/2017/06/30/opinion/social-data-google-facebook-europe.html

Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. *2015 IEEE Security and Privacy Workshops*, 180-184. https://doi.org/10.1109/SPW.2015.27

# Annex I: Draft Codes of Conduct for Data Portability and Cloud Service Switching

## *Infrastructure as a Service (IaaS)*

The **Draft Code of Conduct for Data Portability and Cloud Service Switching for Infrastructure as a Service (IaaS)** provides that:[212]

- *The cloud service shall be capable of importing and exporting Cloud Service Customer (CSC) Infrastructure Artefacts, in an easy and secure way, supporting the following scenarios: CSC to cloud service, cloud service to cloud service and cloud service to CSC. The Infrastructure Cloud Provider (Infra. CSP) shall provide the support to enable the transfer of Infrastructure Artefacts using structured, commonly used, machine readable format.*

- *When exporting CSC Infrastructure Artefacts from a CSC to a cloud service, or between cloud services, the Infra. CSP should provide support to facilitate the interoperability between the CSC's capabilities including the user function, administrator function and business function5 related to the cloud service.*

- *The Infra. CSP should provide Application Programming Interfaces related to the cloud service and, if provided, they shall be fully documented. These APIs should enable the transfer of Infrastructure Artefacts between participating parties. If there are any associated code libraries or dependencies they should be documented and made available.*

- *The cloud service is not required under this Code to transform the CSC Infrastructure Artefacts where the destination environment requires the Infrastructure Artefacts to be in different formats than that offered by the source environment. Parties may agree otherwise in the CSA.*

- *Transfer of CSC Infrastructure Artefacts to and from the cloud service should use open standards and open protocols for Infrastructure Artefacts movement.*

- *Where the CSC data involves Infrastructure Artefacts that rely on a feature or capability of the cloud service, the Infra. CSP shall provide an appropriate description of the environment for their execution and how the service dependencies can be satisfied.*

- *The Infra. CSP should provide a self-service interface that enables the CSC to carry out periodic retrieval of the CSC's data. This functionality can be subject to contract and may include additional costs.*

- *The Infra. CSP shall take reasonable steps to enable a CSC to maintain their service continuity while transferring data between providers, where technically feasible.*

The same draft Code of Conduct Switching and Portability of data related to Software as a Service proposes the following requirement for data export:[213]

> *3.2.1. The source CSP shall have and specify an explicit and structured process for data export. The source CSP should include data management considerations (e.g. snapshots and incremental approaches, records management policies and procedures, bandwidth assessment) and any relevant timescales, notice requirements, customer contact procedures (contact points, escalation etc) and impact on service continuity. This should include relevant SLO and SQO from*

[212] Draft Code of Conduct of 22 November 2019 of WIPO for Data Portability and Cloud Service Switching for Infrastructure as a Service (IaaS) Cloud services, v. 2.9, art. 5.2..

[213] Draft Code of Conduct of November 2019 of WIPO on Switching and Portability of data related to Software as a Service (SaaS), v. 1.5, art. 3.2.

*the SLA. The process and documentation shall cover technical, contractual and licensing matters such that they are sufficient to enable porting and switching.*

*3.2.2. The source CSP shall specify any CSP imposed or enforced obligations on CSCs before exporting data can commence. (i.e. any action required of the CSC to implement the source CSP's processes for data portability as specified in 3.2.1, shall be part of the CSP transparency statement).*

*3.2.3. The source CSP shall specify any known post contractual license fees or other liabilities, for example patent and licensing fees covering use of derived data or data formats or claims and cases that are ongoing.*

*3.2.4. The source CSP shall specify any tools and services incurring additional fees for data export that are required by the source CSP processes for data portability as specified in 3.2.1.*

*3.2.5. The source CSP shall specify any source CSP provided tools or services (including for example addressing integration or interoperability support) that are available to assist the export process and any fees associated with those tools. The source CSP may specify any 3rd party tools or services.*

*3.2.6. The source CSP shall specify whether or not the source CSP's processes for data portability as specified in 3.2.1. allow a CSC to be completely autonomous in exporting data i.e. when the CSC does not need human interaction with the CSP.*

*3.2.7. The source CSP shall specify which data, including derived data (e.g. computed field values, graphics, visualizations) can be exported from the service prior to the effective export date.*

*3.2.8. The source CSP shall specify what, if any, security audit related data (e.g. access logs) is available for export (e.g. logs of user interactions with the cloud service that could be needed for security analysis and for supervisory request).*

*3.2.9. The source CSP shall specify which data standards, formats and/or file types are recommended, used or available for data exporting (e.g. binary, MIME, CSV, SQL, JSON, XML, Avro) for each and every data set available for export including any unstructured data.*

*3.2.10. The source CSP shall provide documentation on the format and structure of the exported data including where it can be sourced and under what terms if from a 3rd party source (including open or industry standard formats or exchanges (e.g. Open Financial Exchange format). As per 3.2.1 above this must be sufficient to enable porting and switching.*

*3.2.11. The source CSP shall specify what cryptographic processes and services it provides, if any, during data export (including unencrypted options) and how encryption keys are managed. The process shall allow the CSC to decrypt the exported Data.*

*3.2.12. The source CSP shall specify any security controls (e.g. access controls) available during data export.*

*3.2.13. The source CSP shall specify any access to, retention period and deletion processes (including notification of deletion) of data, including differing categories of data (including derived data and management data) after the expiration of contract.*

*3.2.14. The source CSP shall specify the costs structure for data export and related procedures.*

*3.2.15. The source CSP shall specify any processes that it supports to maintain data integrity, service continuity and prevention of data loss specific to data exporting (e.g. pre and post transfer data back-up and verification, freeze periods and secure transmission and roll back functionality).*

> *3.2.16. The source CSP shall specify the available mechanisms, protocols and interfaces that can be used to perform data export (e.g. VPN LAN to LAN, Data Power, SFTP, HTTPS, API, physical media…)*
>
> *3.2.17. The Source CSP shall specify any dependencies between the data available for export and other data connected to another cloud service that are created unilaterally by the source CSP and that are not under control of the CSC.*
>
> *3.2.18. The source CSP shall specify any processes, as part of the precontractual transparency document, to disclose use of subcontractors during data portability activity.*

### Software as a Service (SaaS)

The **draft Code of Conduct Switching and Portability of data related to Software as a Service (SaaS)** proposes the following requirement for data import:[214]

> *3.3.1. The destination CSP shall have and specify an explicit and structured process for data import. The destination CSP should include data management considerations (e.g. snapshots and incremental approaches, records management policies and procedures, bandwidth assessment) and any relevant timescales, notice requirements and customer contact procedures (contact points, escalation etc) and impact on service continuity. The process and documentation shall cover technical, contractual and licensing matters such that they are sufficient to enable porting and switching.*
>
> *3.3.2. The destination CSP shall specify any CSP-imposed or enforced obligations on customers before importing data. (i.e. any action required of the CSC to implement the destination CSP processes for data portability as specified in 3.3.1 shall be part of the CSP transparency statement).*
>
> *3.3.3. The destination CSP shall specify any tools incurring additional fees for data import that are required by the destination CSP processes for data portability as specified in 3.2.1.*
>
> *3.3.4. The destination CSP shall specify any CSP provided tools or services (including for example addressing integration or interoperability support) that are available to assist the import process and any fees that are associated with those tools or services. The CSP may specify any 3rd party tools or services.*
>
> *3.3.5. The destination CSP shall specify whether or not the customer can be completely autonomous in importing data i.e. when the CSC does not need human interaction with the CSP.*
>
> *3.3.6. The destination CSP shall specify which data, including any derived data from a source exporting service (e.g. computed field values, graphics, visualizations) can be imported into the service.*
>
> *3.3.7. The destination CSP shall specify what, if any, security audit related data can be imported (e.g. logs of user interactions with the cloud service that could be needed for security analysis and for supervisory request).*
>
> *3.3.8. The destination CSP shall specify which data standards, formats and/or file types are recommended, used or available for data importing (e.g. binary, MIME, CSV, SQL, JSON, XML, Avro) for each and every data set available for import including any unstructured data.*
>
> *3.3.9. The destination CSP shall specify the format/structure required of imported data and where definitions are available and under what terms (including open or industry standard formats or exchanges (e.g. Open Financial Exchange format). The CSP should specify any*

[214] *Ibidem*, art. 3.3.

*available validators and if so what type (e.g. structure, format, storage type, volume, links), from where and under what terms. As per 3.3.1 above this must be sufficient to enable porting and switching.*

*3.3.10. The destination CSP shall specify what encryption processes are used during data import (including unencrypted options) and how encryption keys are managed*

*3.3.11. The destination CSP shall specify any security controls (e.g. access controls) used during data import.*

*3.3.12. The destination CSP shall specify the costs structure for data import and related procedures (e.g. volume restrictions).*

*3.3.13. The destination CSP shall specify any processes that it supports to maintain data integrity, service continuity and prevention of data loss specific to data importing (e.g. pre and post transfer data back-up and verification, freeze periods and secure transmission and roll back functionality).*

*3.3.14. The destination CSP shall specify the available mechanisms, protocols and interfaces that can be used to perform data import (e.g. VPN LAN to LAN, Data Power, SFTP, HTTPS, API, physical media …)*

*3.3.15. The destination CSP shall specify any processes, as part of the precontractual transparency document, to disclose use of subcontractors during data portability activity.*

cerre

Centre on Regulation in Europe

Avenue Louise, 475 (box 10)
1050 Brussels, Belgium

+32 2 230 83 60

info@cerre.eu

cerre.eu

@CERRE_ThinkTank