cerre Centre on Regulation in Europe

DATA ACCESS PROVISIONS IN THE DMA

ISSUE PAPER

November 2022

Jan Krämer

As provided for in CERRE's bylaws and procedural rules from its "Transparency & Independence Policy", all CERRE research projects and reports are completed in accordance with the strictest academic independence.

This paper is part of a larger CERRE project entitled 'Effective and Proportionate Implementation of the DMA' which is a collection of nine papers focusing on the trade-offs around the different possible interpretations of the regulation. The project, within the framework of which this report has been prepared, received the support and/or input of the following CERRE member organisations: Apple, Arcep, ARCOM, Booking.com, Comreg, DuckDuckGo, Google, Mediaset, Meta, Microsoft, Ofcom, Qualcomm, Spotify, TikTok, Vodafone. However, they bear no responsibility for the contents of this 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 of any sponsor or of members of CERRE.

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 an Academic Co-Director at CERRE and a Professor at the University of Passau, Germany, where he holds the chair of Internet & Telecommunications Business.

Previously, he headed a research group on telecommunications markets at the Karlsruhe Institute of Technology (KIT), where he also obtained a diploma degree in Business and Economics Engineering with a focus on computer science, telematics and operations research, and a Ph.D. in Economics, both with distinction.

He is editor and author of several interdisciplinary books on the regulation of telecommunications markets and has published numerous articles in the premier scholarly journals in Information Systems, Economics, Management and Marketing research on issues such as net neutrality, data and platform economy, and the design of electronic markets.

Professor Krämer has served as academic consultant for leading firms in the telecommunications and Internet industry, as well as for governmental institutions, such as the German Federal Ministry for Economic Affairs and the European Commission.

His current research focuses on the role of data for competition and innovation in online markets and the regulation of online platforms.

1. INTRODUCTION

Data, especially data on consumer behaviour, is an essential input factor in the digital economy. It facilitates and improves personalisation, product designs, recommendations and predictions, and targeted advertising, among other things. In this vein, data-driven advantages can spur positive feedback-loops (data-driven network effects) which in turn create barriers to entry. The DMA foresees several provisions whereby gatekeepers need to share data which were created by users while using the gatekeeper's core platform service. Such data must be shared with end users (**Article 6(9)**) and business users (**Article 6(10)**) who were involved in creating the data through real-time and continuous data portability. Moreover, data that was created by users while using an online search engine, must be shared with other online search engines (**Article 6(11)**).

The DMA has two main goals: contestability and fairness. However, the data access provisions mentioned above seem to be especially motivated by the goal of contestability. This is expressed, for example in Recital 3 (emphasis added): "Contestability is reduced in particular due to the existence of very high barriers to entry or exit, including high investment costs, which cannot, or not easily, be recuperated in case of exit, *and the absence of, or reduced access to, some key inputs in the digital economy, such as data*." Moreover, Recital 32 states that "The features of core platform services in the digital sector, such as network effects, strong economies of scale, and *benefits from data have limited the contestability of those services and the related ecosystems.*"

This view is strengthened by the specific recitals relating to the above provisions. In relation to data portability by business users and end users, Recital 59 explains that the provision is necessary "to ensure that gatekeepers do not undermine the contestability of core platform services, or the innovation potential of the dynamic digital sector, by restricting switching or multi-homing"**.** Likewise, Recital 61, relating to access to search and query data, highlights that "access by gatekeepers to such ranking, query, click and view data constitutes an important barrier to entry and expansion, which undermines the contestability of online search engines."

However, Recital 34 also makes clear that "contestability and fairness are intertwined". Moreover, Recital 33 states the notion of fairness includes that "business users should have the ability to adequately capture the benefits resulting from their innovative or other efforts". One can argue that the goal of fairness therefore includes that business users receive access to the data that was created through their efforts using a core platform service.

In this issue paper, the provisions of the DMA on data portability and data access to search and query data of search engines are considered in more detail, in particular with regard to open questions concerning their implementation in view of the stated goal of contestability.

2. DATA PORTABILITY FOR END USERS AND BUSINESS USERS

Albeit data portability is already a fundamental right of users under the **General Data Protection Regulation (GDPR) (Article 20), this right is augmented in the context of contestability by the DMA**.

> ***Article 6(9)***
> *The gatekeeper shall provide end users and third parties authorised by an end user, at their request and free of charge, with effective portability of data provided by the end user or generated through the activity of the end user in the context of the use of the relevant core platform service, including by providing, free of charge, tools to facilitate the effective exercise of such data portability, and including by the provision of continuous and real-time access to such data.*

There are three main differences between Article 20 of the GDPR and Article 6(9) of the DMA. First, of course, the GDPR applies horizontally to all data controllers, whereas the **DMA only applies to gatekeepers**. Thus, enhanced data portability is only available at a very limited number of firms. However, by Article 17(4) also emerging gatekeepers, which do not yet meet the thresholds for gatekeepers laid out in Article 3, can be mandated to offer such enhanced data portability (including data portability for business users as discussed below).

Second, and most impotantly, consumers must be provided **data access continuously and in real-time**, likely through APIs. The GDPR only calls for one-off data transfers provided in “a structured, commonly used and machine-readable format”, and data controllers have up to 30 days to provide the data. This often tedious and slow process, where the data may be outdated by the time they are received, has been considered one of the main reasons why data portability under the GDPR is not used widely to date, and thus not an effective instrument to spur competition and innovation (Krämer et al. 2020). At the same time, it should be acknowledged that continuous and real-time data portability is much more complex to implement than a one-off data transfer and raises additional technical feasibility issues that need to be considered in the implementation of this provision.

Third, the gatekeeper must provide “**tools to facilitate the effective exercise of data portability**”. This is worth discussing, and we will return to this below.

In addition, the DMA also extends the idea of data portability to business users as follows:

> ***Article 6(10)***
> *The gatekeeper shall provide business users and third parties authorised by a business user, at their request, free of charge, with effective, high-quality, continuous and real-time access to, and use of, aggregated and non-aggregated data, including personal data, that is provided for or generated in the context of the use of the relevant core platform services or services provided together with, or in support of, the relevant core platform services by those business users and the end users engaging with the products or services provided by those business users. With regard to personal data, the gatekeeper shall provide for such access to, and use of, personal data only where the data are directly connected with the use effectuated by the*

> *end users in respect of the products or services offered by the relevant business user through the relevant core platform service, and when the end users opt in to such sharing by giving their consent.*

2.1 Scope of Data Access

2.1.1 Data portability for end users

The scope of data that is 'provided' by end users, and thus subject to data portability, is already contentious under the GDPR's data portability right (cp. Krämer et al. 2020). One can roughly distinguish between volunteered, observed, and derived data (Crémer et al. 2019). *Volunteered data* is explicitly and intentionally revealed by a user (an email-address or 'likes', for instance). *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 (location data or clickstream data, for example). *Inferred data* is derived through refinement and recombination from volunteered and observed data, such as by use of data analytics such as clustering, filtering or prediction.

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. However, in its "Guidelines on the right to data portability" the European Data Protection Board (EDPB) defines provided data as those data "provided by the data subject by virtue of the use of the service or device". The EDPB also suggests that this can include data that was explicitly provided by the data subject ("volunteered data") as well as data that was implicitly provided by the data subject ("observed data").[1] But the question remains how far reaching one can interpret the scope of 'observed data'.

Accordingly, as the DMA is using the same language as the GDPR and, in Recital 59, clearly labels its data portability provision as a complement to that under the GDPR, the same ambiguity on the precise scope of data portability now extends to the DMA. Given that Recital 59 also refers to the "innovation potential" and, as laid out above, seeks to promote contestability in the context of data portability, a wider scope of observed data could be assumed. This could, for instance, also include clickstream data of consumers then. In reverse, the DMA must also be interpreted in light of the proportionality of its obligations and in view of IP rights, data security, and privacy-related innovation. While we have already pointed to the ambiguity of the scope of observed data to be included in a portability request under GDPR, the **similar language in the DMA suggests that the same scope as under GDPR should also apply to the DMA. That would exclude derived data from portability requests under the DMA**.

While it is clear that only raw data can be ported, under a generous interpretation of 'provided' data (especially with regard to 'observed' data), it is **reasonable to ask how much context to the ported data needs to be provided** so that data subjects (and third parties to which the data is ported) can truly assess the information content of that data. For example, in the context of clickstream data, such

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

data is meaningless unless the content that was consumed or which ads were clicked on is also shared. But the contextual information (for instance the content or the ad) was of course not ‘provided’ by the user, although the user engaged with that content. So, should contextual information be in scope of data portability? What are the 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? Given that data needs to be shared in real-time and continuously, potentially vast amounts of data are subject to data portability if such a wider scope is assumed. When exactly is technically infeasibility admissible as a defence for a data rich firm in the digital economy? Generally, Article 34 of the GDPR requires a data controller to put in place appropriate security measures for all personal data in its possession, including that which it is provided further to the rights of access and portability. The DMA needs to be consistent with those requirements.

Finally, one may ask if the answer to these questions should factor in at all what the intended use of the data is, or where consumers are porting the data to. If the goal of this provision is just to facilitate switching and multihoming, then it may be **justified to limit data portability to the personal data required to enable switching and multihoming**, but no more. For example, which ads the consumer clicked on may well be personal data (and as such eligible for portability under GDPR), but not be material for switching to another service provider offering a similar service, and thus must not be provided under the end user’s DMA portability right (in a real-time and continuous manner). If such a logic were admissible, then there would be difficult case-by-case decisions to make on which data are material for switching. For example, even though the alternative provider may not require the clickstream data on ads for improving its service offering directly, that information may nevertheless improve the service quality indirectly, as it helps to increase ad revenues, which can be re-invested in service quality. Further, viable alternative providers may not yet exist at the time where the Commission needs to specify the scope of the data portability provision for a given gatekeeper, and then it would have to anticipate which data may be useful for switching to a future (innovative) service provider. This seems challenging.

In conclusion, a **wider scope of the data to be ported, independent of the use and the destination of the data, seems preferable**.

2.1.2 Data portability for business users

Generally, the same issues on scope also arise in the context of data portability for business users, especially relating to personal data of end users engaging with the product or service of the business user.

In addition, it is noteworthy that, in relation to Article 6(10), Recital 60 seems to allow for some ‘*adversarial portability*’ by business users. That is “a gatekeeper should not use any contractual or other restrictions to prevent business users from accessing relevant data”. This could be understood as a free pass for business users to use web scraping or other tools not directly provided or authorised by the gatekeeper in order to access the information that the gatekeeper may provide to those business users via its own website or other interfaces. This would mean **that whatever data the gatekeeper chooses to make available to the business user via its own interfaces (such as**

performance metrics, aggregate user data, and so on) in any form, could be in scope for data portability by business users. Although the data must be 'relevant', this does not seem to put a strong restriction. If the data were not 'relevant' to the business user, then it probably would not have been provided to them via some interface in the first place. Moreover, the judgment on whether data is 'relevant' to a business users can only be made on a case-by-case basis and must also anticipate future business users.

Another issue of the scope of data portability relates to the definition of a "business users' product or service". Business users are only entitled to port data that was created in relation to their product or service. *But what if a business user is just one of many on a platform that offer the identical product or service?* Say a business user offers a product on an online marketplace and is listed as one of many sellers of that very product. The online marketplace shows the product only once (for instance in a search result or on a dedicated product page), and from there it links to several business users from which the product can be bought. Which of the many business users is entitled to the data that was created through the engagement of an end user with that product in the online marketplace? Only the business user which was listed as the "default" buy option? All business users, albeit some may have offered the product at a price that would not have led to an engagement by the end users in the first place? This issue may not arise in contexts where the product or service offered through the core platform service is uniquely linked to a specific business user, but as the example highlights, in the case of some core platform services that linkage between product and business user is not unique.

KEY QUESTIONS

- What is the **scope of observed user data that needs to be provided** for end users pursuant to a portability request? In particular, does contextual information on the data need to be provided to make data portability effective?

- Can the gatekeeper bring forward a **technical feasibility constraint**, given that data needs to be ported in real time.

- Is '**adversarial portability**' allowed for business users under the DMA?

- Are there **limits** to data portability of a given business users **when the business user's products** or services are offered by several business users?

2.2 Consumer Consent

2.2.1 Regardingdata portability by end users

In the context of real-time and continuous data portability, consumers should be able to give their **consent on a fine-granular level** regarding which data is to be potentially transferred. All-or-nothing transfers are often not necessary, and would create more transaction costs, both technically (network load or space requirements, for example) as well as economically (larger privacy concerns). The granularity of consent should be part of the regulator's specification procedure in relation to the data portability obligations. Of course, gatekeepers shall not influence consent or dissent by offering commercial incentives or disincentives.

A particular concern in the context of data portability is a potential conflict with the *rights of other data subjects*. Consumers may want to port data that has been co-created (for instance, chat protocols) or is shared by others via a core platform service (for example, pictures of (several) data subjects), or is otherwise linked to others (such as address books, or pictures where other people are tagged). In this case, the platform service may be required to ask for consent to include such data in a data portability request. Otherwise, the gatekeeper may just exclude such data from data portability diminishing its value, especially with regard to the intended goal of facilitating switching and multi-homing for end users. Thus, next to the end user's consent to initiate a data portability request, **end users may also need to consent (possibly in advance) to other users' data portability requests**.

Neither the DMA nor GDPR currently provide guidance on this issue, that is, who is responsible for obtaining consent from others, or whether a data portability request can be limited to only that portion of the data which is not affected by rights of others.

It should also be understood that obtaining consent doesn't shield the parties involved from compliance with the GDPR, which includes compliance with the principles of lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality, and accountability. The GDPR imposes obligations on all actors of the data supply chain and the data ecosystem – the gatekeeper, the recipient of the data, and other business partners.

2.2.2 Regarding data portability by business users

In the context of data portability by business users an additional complication (in comparison to data portability by end users) arises from the fact that non-personal (aggregate) data and personal data relating to some end user must be differentiated. For example, an end user may have provided his or her gender and age via the core platform service to the business user. While the business user may be entitled to port aggregate information about its users' demographics, the information about a specific user is only accessible if the user has opted into data portability for that very business user. That is, **each business user must ask each of its users whether they agree to data portability**. There are at least two implementation issues associated with this:

First, **when is the consumer being asked for consent?** To exemplify this issue take the following case. Say a user browses through an online marketplace, looking at various products. The age and gender of that user is known to the online marketplace, and it can thus be linked to the clickstream data that the user is generating while browsing. The user enters a keyword and is presented a list of products. After looking at some products in that listing, the user eventually buys a product. When should a user be prompted for consent, and with regard to which personal data? Should the user consent already when clicking on the product page, possibly to each business user who provides that product individually? That would clearly be valuable for business users, as they may determine from the clickstreams why a consumer has not bought. However, this would also clearly not be practical and quickly render the online marketplace unusable from an end user's perspective. Should the user then only be asked for consent if he or she has bought a product? Article 6(10) only seems to require 'engagement', but does not further specify whether there is a threshold to what qualifies as 'engagement'. And which data should that consent then cover? The whole customer journey leading up to that sale? Only the customer journey on the product page of the product that was eventually bought? And does that data only include observed data (say which reviews the customer has read on the product page), or also personal data provided by the user to the core platform service in advance to the customer journey (age and gender in this example), but which has not been provided in the customer journey leading up to that sale or as part of the sales process?

Second, **how is consent being obtained?** Through the business user (via channels outside of the core platform service) or via the core platform service directly? Recital 60 notes that the gatekeeper "should enable business users to obtain consent of their end users". This seems to suggest that gatekeepers need to implement tools directly through the core platform service that enable business users to obtain consent. Say, in the online marketplace example, a consumer can tick a box before concluding a purchase as to whether his or her data can be ported to the business user. That would require at least one additional click for users – in a context where every click leading up to a final purchase can be one click too many. Are gatekeepers then allowed to have consumers opt in to data portability for all business user interactions that they encounter using the core platform service at once? This seems overly broad and may not be in line with the notion of informed consent under GDPR that is required also under the DMA (as noted in Recital 60 when referencing Regulation (EU) 2016/679 and Directive 2002/58/EC). Relatedly, can a business user ask for consumer consent to portability for all the core platform services that it may operate on, or does this need to be obtained per core platform service?

KEY QUESTIONS

- How fine **granular must the consent** for data portability obtained be, and should it include consent to data portability requests of others?

- What is the **threshold for 'engagement'** in order for a business user to be entitled to data portability?

- Does end **user consent** to portability need to be obtained for each business user, or for each core platform service separately?

2.3 Tools to Facilitate Data Portability

Interestingly, Article 6(9) relating to data portability by end users requires gatekeepers to provide end users with "tools to facilitate the effective exercise of such data portability" free of charge. This is worth discussing for several reasons.

First, it is noteworthy that this obligation to provide 'tools' does not exist for data portability by business users, probably since business users are expected to have their own 'tools' for that purpose – or because third-party tools are being developed and offered to business users. By contrast, as mentioned above, business users are provided with an implicit authorisation to make use of 'adversarial data portability', which is not the case in relation to portability by end users.

One may **wonder why end users seemingly do not have the right to use 'adversarial data portability', but instead need to rely on the tools provided by the gatekeeper**. This matters, because users may have access to more data through their user interface when using the core platform service than what may fall under the scope of personal data portability (see above for a discussion on the scope of data portability requests). One may also expect that third-party tools that pull data from the user interface are being in case this was legitimised by the DMA. In fact, those tools often already exist, but have been shut off by some of the possible gatekeepers.[2] One important objection against such adversarial data portability may be that some of the data that is being shown to users via their user interface is in fact not their personal data, at least not theirs alone, as it also touches on rights of others (cf. discussion above on consent). For example, they may see personal pictures of others, or chat protocols that have been co-created with others. Although a user may see this personal data of others, and others may have shared it with the whole world, this does not mean that the data can be legally ported. There has already been controversy about this issue in context to data portability under the GDPR (Graef 2020; Krämer et al. 2020) and the debate is still not resolved. In this regard, the use of tools provided by the gatekeeper to exercise data portability, instead of adversarial data portability, can also be understood as a means to be able to control the lawfulness of data portability better. A gatekeeper can ask consumers for their consent to the porting of their personal data that they shared with others (such as personal photos that they publicly uploaded) in order to ensure that only such data are being ported (through the tool).

Second, there may also be an unintended consequence in relation to the tools provided for data portability. Data portability, especially continuous and real-time data portability, can spur the emergence of Personal Information Management Systems (PIMS) (cf. Krämer, Senellart & de Streel 2020), which are believed to be an important pillar for consumer empowerment in the digital economy (EDPS 2020). PIMS are envisaged to facilitate consumers' data control and the exercise of their rights to data portability, including accessing, storing, visualisation, and possibly monetisation of their data. PIMS may also allow for (automated) consent management across services. While the technical details and economics of PIMS are beyond the scope of this report, it is important to note that the **DMA's provision to offer tools for data portability to consumers may lead to a crowding out of independent PIMS**. Such PIMS may then be provided by gatekeepers instead of independent third parties.

Ultimately, the tools provided by gatekeepers may not only facilitate data portability from their core platform service to another provider, but also data portability requests between various services more generally. In fact, in 2018 some of the major tech firms, under the lead of Google, already started an

[2] An example is the tool 'Ad Observer' that has been developed by researchers from New York University to study misinformation on Facebook. See https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

open source project, called the **Data Transfer Project**[3], with exactly this purpose. While the project has not progressed significantly since then, the DMA may spur its development. This should be scrutinised closely, as otherwise this may unintendingly equip gatekeepers with even more sources to consumer data.

KEY QUESTIONS

- Must consumers rely on the **tools provided by gatekeepers** to exercise their right to continuous and real-time data portability under the DMA?

- Can **third-party** tools also access the data with the same performance and quality?

2.4 Effective and High Quality Data Access

Both Article 6(9) and 6(10) require the implementation of data portability to be *effective*. Beyond the points already mentioned, what could effective data portability mean? Recital 59 suggests that effectivity could relate to the *format* in which the data are accessible (through the interfaces or 'tools' provided by the gatekeeper), as it notes that such data shall be "effectively accessed and used by the end user". Moreover, 'effective' could also entail that data transfers are safe to use for end users. That is, 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;

Effectiveness of data portability can also relate to aspects of transparency and adherence to common standards. Thus, **where possible, data portability should make use of *open standards* and protocols, which are free to use and transparent for developers** (see, for instance, Furman et al, 2019, pp.71-74). To this end, Article 48 of the DMA allows the regulator to request European standardisation bodies to develop standards for portability. This is also elaborated on in Recital 96: "The implementation of some of the gatekeepers' obligations, such as those related to data access, data portability or interoperability could be facilitated by the use of technical standards. In this respect, it should be possible for the Commission, where appropriate and necessary, to request European standardisation bodies to develop them." Examples may be drawn from the Australian Consumer Data Right (CDR) initiative,[4] which has also relied on a standardisation body.

However, the development of standards may require much time, and regulators will have to rely on effective implementations by the gatekeepers in the meantime. In any case, albeit challenging and time consuming, it will be important to **harmonise data formats and interfaces for data portability across the various gatekeepers** subjected to the DMA's data portability provisions. Harmonisation and standardisation are important, because it allows third party tools, such as PIMS, to better

[3] See: https://datatransferproject.dev
[4] See https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0

integrate with the largest possible set of firms and thereby to facilitate switching and multihoming (cf. Krämer et al 2020). In other words, instead of having one tool per gatekeeper, it would be better to have one tool that is able to connect to all gatekeepers for the purposes of data portability.

As mentioned above, effective data portability also means that **consumer consent for data portability can be provided effectively**. This relates to the granularity of consent (see above), but may also include the possibility to give automated (rule-based) consent, for instance, through tools such as PIMS. It will have to be clarified, however, to what extent under what conditions such automated consent (that is, consent automatically derived from the users' explicitly provided privacy preferences) would qualify as "informed consent" under Article 7 GDPR.

Effectivity should also relate to *availability and performance* of the interface used for data portability. In the context of the Revised Payment Services Directive (PSD2 Directive), **performance and reliability was measured against the data provider's other consumer-oriented interfaces**.[5] This seems to be a reasonable approach also in the context of the DMA's data portability provisions.

Finally, it is also worth highlighting that **'high quality' of data portability** is explicitly mentioned, next to 'effectivity', only in Article 6(10), but not in Article 6(9). Recital 59 clarifies what may be meant by 'high quality' in relation to data portability: "Gatekeepers should also ensure, by means of appropriate and high quality technical measures, such as application programming interfaces, that end users or third parties authorised by end users can freely port the data continuously and in real time. This should apply also to any other data at different levels of aggregation necessary to effectively enable such portability." Does the mentioning of 'high quality' only in Article 6(10) mean the performance standards are possibly lower in relation to data portability by end users, relative to data portability by business users? This would not seem reasonable in light of the goals of the provision, that is to facilitate switching and multihoming. If this distinction in the 'quality' of data portability is indeed intended by the regulator, where should we then draw the line between 'high quality' and 'effectiveness' of data portability?

KEY QUESTIONS

- To what extent does data portability need to fulfil **security, availability, performance, and standardisation** criteria in order to be 'effective'?

- Are the **criteria different** for end user data portability vs. business user data portability?

[5] See Article 32 as well as Recitals 23-25 in the Delegated Regulation (EU) 2018/389, amending the PSD2 Directive (EU) 2015/2366.

3. DATA ACCESS FOR SEARCH ENGINES

Online search engines are a particularly important service in the digital economy as they are usually the entry point of an online session. Scholars have long argued that the market for (general) online search engines may lack contestability due to strong data-driven network effects, and that data sharing may be an appropriate remedy (Argenton & Prüfer 2012). The DMA now includes a provision exactly to that effect.

> ***Article 6(11)***
> *The gatekeeper shall provide to any third-party undertaking providing online search engines, at its request, with access on fair, reasonable and non-discriminatory terms to ranking, query, click and view data in relation to free and paid search generated by end users on its online search engines. Any such query, click and view data that constitutes personal data shall be anonymised.*

3.1 Tension Between Privacy and Contestability

The provision in Article 6(11) requires gatekeepers to provide ranking, query, click, and view data to third-party search engines with the explicit goals of fostering competition and contestability in relation to that gatekeeper. However, at the same time the provided data cannot contain personal data, which shall therefore be sufficiently anonymised. Recital 61 explains that the gatekeeper should "ensure the protection of the personal data of end users, including against possible re-identification risks, by appropriate means, such as anonymisation of such personal data, without substantially degrading the quality or usefulness of the data."

However, there is **a tension between strong anonymisation and maintaining enough level of detail in the data that it is valuable** for third-parties such that they can derive better algorithms and predictions in order to contest the data provider. While it is clear that anonymisation needs to be effective, there is regularly a dispute over the boundary at which data is effectively anonymised, especially as technology and computing power progresses. In the extreme, data could be so strongly aggregated (for example, an average age for all users) that it is not useful anymore for deriving insights from the data. When implementing this provision, navigating this tension between aggregation and detail of the data will probably be the most difficult task for the regulator.

Some observers seem to question that it will ever be possible to balance this tension, as methods and tools for de-anonymisation are continuously being improved and even relatively little detail may already reveal a person's identity (see, for instance, Rocher, Hendrickx, & de Montjoye 2019). It may likely require technical *and* institutional means to achieve this in the best possible way. Some possible approaches are discussed below, which may also be used in combination. However, no matter how the data sharing is implemented, anonymised user data will never have the same 'depth' as the original data set due to this trade-off. This also calls into question whether such data sharing is sufficient for a third-party to truly contest the incumbent who had provided this data. We will return to this below.

The risk of de-anonymisation in a particular data set depends crucially on the uniqueness of the attributes associated with different individuals. It is therefore generally not enough to just remove a personal identifier (for instance, the combination of full name, birthday, and place of birth) and to replace it with a pseudo-identifier (a unique combination of numbers and letters, for example). Although it might not be immediately obvious anymore who is associated with a given data record, the values of the remaining attributes in the data set (such as the combination of blood type, zip code, and age) may still uniquely identify an individual. This is the more likely the more unique the individual values are (a very rare blood type or a very high age, for instance). '**Anonymity' is therefore not a discrete zero-one concept but rather a statistical concept that relates to a particular probability that an individual may be re-identified**.

3.1.1 Technical means

K-anonimity

In computer science, two concepts are frequently used to describe the degree of anonymity in a given data set. The first concept is k-anonymity: **A data set is said to have k-anonymity if the information for each person contained in the data set cannot be distinguished from at least k-1 other persons who are also contained in the same data set**. Consequently, the larger k, the larger is the degree of anonymisation of a data set. K-anonymity can generally be achieved by suppression of attributes (for example, deleting name, dates, or address) or by generalisation of attributes (such as transforming names to initials, dates to years, and addresses to zip codes). K-anonymity usually does not involve any randomization of attributes and it can be seen that in large data sets, especially 'deep' data sets with many attributes, anonymity may nevertheless be compromised.

Differential privacy

A second, more recent and more sophisticated concept is differential privacy. Roughly speaking **differential privacy is not a discrete concept (as k-anonymity), but a probabilistic concept and requires randomization of attribute values (adding some random noise to GPS data, for instance).** The goal is to create a data set for which it is not possible (with some statistical guarantees) to know whether an individual's data is contained in the data set. This is important because de-anonymisation attacks typically match data from different data sets from which it is known that they contain a given individual. This can be achieved, for example, by running several similar queries to a data base, with the goal to obtain (anonymised) data sets that differ only by the entry of one person. While data sets with a k-anonymity property are susceptible to such attacks, data sets with differential privacy are not, due to randomization. There are several algorithms to achieve differential privacy, and this is subject to ongoing research in cryptography. In practice it may be difficult and computationally burdensome to achieve differential privacy, especially if data is shared in a continuous manner. A more practical approach is therefore **not to store accurate data about individuals at all, but to add some noise already when data is collected**. This is a technique that is already applied by Apple and Google for select applications in iOS/macOS and Chrome,[6] but it is not known whether it also applies in relation to online search engines. This also highlights that differential privacy is not just a theoretical option,

[6] Green, M. (2016). What is differential privacy? https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/

but can indeed be applied in the context of large-scale data collection as is typical for prominent digital services. This may also mean, however, that regulated firms may not only be mandated to share their data, but also mandated to collect (or rather *not* collect) their data in a certain way, in order to enable privacy-preserving sharing of that data later.

Synthetic data

The anonymisation of search logs while preserving useful information is a relatively recent and emerging field of research (Hong et al. 2009), but also a domain of computer science in which progress is being made quickly. **Promising developments seem to be the creation of 'synthetic search logs'** which contain plausible search sequences, but are created from a machine learning model and do not relate to an actual person (see, for instance, Krishnan et al. 2020).

3.1.2 Institutional means: Data trusts and data sandboxing (in-situ access)

Next to such technical means, there are also institutional means to protect privacy, which can also be combined. A common institutional proposal is to establish a trusted data intermediary (*data trust*). To ensure this, the trust needs to be independent from the regulated entity, of course. The main idea is that **user data (from the various entities that are mandated to share data) is collected by a data trust in its original raw and detailed form (see, for example, Graef & Prüfer 2021). The trust could then combine the data and anonymise it properly**. Such anonymisation of the joint data set directly would be preferred over anonymisation of separate data sets at the source, because it would reduce the risk of de-anonymisation through re-matching of the different data sets, each of which may have different attributes omitted or generalised.

Moreover, the data trust may not need to reveal any raw data directly but **could act as a data sandbox** instead. This means that third-parties would need to submit their algorithm for analysing the data to the trust, who would then run it on their behalf on the detailed raw data. The third-party would receive back the trained algorithm, but never see the raw data itself. Data sandboxing could also be applied at the original data source directly (Graef & Prüfer 2021), which is then referred to as *in-situ access* (Martens et al 2021).

It is however **not clear whether access of the data only through data trusts and data sandboxing is sufficient** to allow access seekers to combine the provided data with their own in order to really be able to obtain a novel dataset – which, as argued above, would probably be necessary in order to be truly able to contest the incumbent. Otherwise, the shared data will just be an (inferior) subsample of the data that is available to the incumbent.

In addition, there are several practical issues with data trusts and data sandboxes, especially when applied to vast amounts of data, as those collected by online search engines. For all practical purposes a **data trust would require an enormous infrastructure** to be able to store, aggregate and anonymise the data (continuously) in any meaningful way. For example, Google Search alone processes over 80,000 search queries every second on average, which translates to almost 7 billion searches per day.[7]

[7] Internet Live Stats, 2020, https://www.internetlivestats.com/one-second/#google-band

It seems one would have to duplicate much of the gatekeeper's data centre infrastructure to achieve this. Who would then finance and operate this, and be liable in case of failure or data breaches? Would this be in line with the Commission's sustainability goals?

Likewise, data sandboxing is an intriguing theoretical idea, but it would require an even larger infrastructure to have sufficient computing power required for running probably complex algorithms on the data. Since these would operate on the detailed raw data, it would also require enormous effort and expertise to make sure that the algorithms do not compromise privacy. It seems to be a formidable task for the regulator to police this – that is, to ensure that the data are neither too highly aggregated (so they don't prove useful), nor too little aggregated (so they may compromise privacy).

If algorithms are run directly on the infrastructure and raw data of the original data controller (in-situ access), then this would also **put a significant computational burden and cost on the regulated firm**. In turn, this would warrant some financial compensation (discussed in more detail below in relation to FRAND access provisions). It also needs to be feared that the original data controller would be able to acquire business sensitive information about the third-parties through the algorithms that are run on its infrastructure.

Nevertheless, regulators may want to entertain the idea that data trusts or data sandboxing (at a data trust or via an in-situ access) **may be feasible if confined to subsets of the data, particularly with a focus on recency**. In the short term only in-situ access seems practical, as the gatekeeper already has an appropriately sized infrastructure in place. In this case scrutiny must be placed on the confidentiality of the competitors' algorithms. In addition, in-situ access does not alleviate privacy risks completely, as data requests may be so specific that they can be matched to individuals and anonymisation techniques (cp. to the concept of 'differential privacy' above) must be applied nevertheless.

3.1.3 Data portability as a complement to anonymisation

Finally, it is worth highlighting that there may be a fruitful interaction between the data portability provisions and anonymisation in this context.

Gatekeepers operating an online search engine and being subjected to Article 6(11) are also subjected to Article 6(9) and 6(10). As the number of business users are defined by the commercial websites listed by an online search engine (see Appendix to the DMA), this potentially gives a large number of business users portability rights vis-à-vis a search engine. **Depending on the scope of access under this portability right (see above), this could also include query and click data, and – if the end user has consented to it – personal data**.

Likewise, the user could directly transfer (continuously and in real-time!) personal data (which should include clickstream data, as discussed above) under his or her portability right to a third party, such as a search engine. Of course, only a subset of the end users will ever make use of this. Thus, the ported data are not representative by any means, but provide more 'depth' to the data than what can be shared through Article 6(11), and thus data portability can act as a complement (but not as a substitute) to it.

However, in contrast to the data portability provisions discussed above, the search engine data to be shared under Article 6(11) does not need to be shared in real-time and continuously. This would also not be possible while at the same time, anonymisation techniques shall be applied. However, regulators should still put an emphasis on recency of the data to be shared, including an appropriate frequency at which the data is to be updated.

KEY QUESTIONS

- What is the appropriate **institutional means** (data trusts, data sandboxing/in-situ access, APIs) through which search engine data shall be shared in order to balance contestability goals and privacy?

- How far reaching are the powers of the EC to impose a certain means and/or to impose a **certain data collection procedure to facilitate sufficiently anonymised** yet useful data sharing?

3.2 Which data could/should be provided?

In general search, the data bottleneck lies in the search queries, associated context information, and behavioural data on how users interacted with the search results (Krämer et al. 2021). The data bottleneck is not, however, the web index (that is, the register of all websites), as this data can be more easily duplicated by (potential) competitors.

Any shared data must therefore at least contain information about the search queries that users have presented to the search engine provider. This already brings about one central difficulty. **Search queries are inherently personal** and can reveal significant information about an individual. They may also reveal a person's identity relatively easily. A famous example is the case of Ms. Arnold, who was identified from a list containing 20 million web search queries conducted by a total of 657.000 Americans over the period of just three months. Although the data set was released by AOL in a pseudo-anonymised way (evidently not respecting k-anonymity or differential privacy), she was re-identified based on her search queries alone (Barbaro & Zeller 2006).

Since web queries are based on text, it is not as straightforward to add some noise to the search terms without rendering them useless (cf. 'differential privacy'). Moreover, as has been pointed out by the CMA's study on 'Online platforms and digital advertising market' (CMA 2020, p. 12)[8] as well as recent research (Klein et al. 2022), it is precisely the rare search terms that are particularly valuable for training a prediction model improving ranking quality.

[8] Also specifically Appendix I of that study. Available at https://assets.publishing.service.gov.uk/media/5fe4957c8fa8f56aeff87c12/Appendix_I_-_search_quality_v.3_WEB_.pdf

The risk of re-identification is less pronounced if one would not associate specific search queries with a unique user identifier, which allows to associate different searches of an individual over a period of time. However, without such a user identifier, the data set loses traceability, for instance, on the search process of an individual, which is an important for improving for the quality of search engines.

A second major challenge is to define the scope of the contextual information relating to the search results page properly. Aggregate, or even individual search query data is only one part of the relevant information that users reveal to a search engine. The other part is how they have interacted with the search results page, for example, which links were clicked subsequent to a given search and in which order. But sometines it may be even more informative which links consumers did *not* click and thus did not find relevant after a given search. For a proper assessment of clicks, it would also be necessary to know which other elements were shown on the search results page in addition to the organic search results. For a long time now, Google's search results page does not only contain "10 blue links" anymore, but in addition, and depending on the search query, sponsored search results and other 'boxed' elements such as a news carousel, flight search, a shopping comparison, or an immediate answer to the search query are displayed. In fact, an increasing percentage of search sessions end with the search results page, and consumers never follow up and click on a search result. It is estimated that so called Zero-Click Searches amounted to about 50% of all searches on Google.com in June 2019 (Fishkin 2019). Likewise, research has shown that clicks on organic search results are heavily influenced by whether and how sponsored search results and 'boxed' results are presented (Edelman and Lai, 2016).

Evidently, the search results page (ranking) is already inferred data of the search engine, and forcing the release of detailed information about the search results page pertaining for every query would probably go too far with regards to the proportionality principle under EU Law and Art. 8.(1) DMA, as it would undermine past and future innovation efforts. Based on this data, third parties may be able to reverse-engineer the gatekeepers ranking algorithm. However, it may be justified to release such information for samples of queries, or to limit the details of the data relating to the search results page. One could release, for example, only the first clicked result.

In order to advance the discussion on the appropriate scope of shared search logs, Krämer et al. (2021) suggested to think in three main categories: i) data on the query itself, ii) data on the search results page, and iii) data on the user. The below table exemplifies which pieces of information can belong to each category.

Table 1: Categories and scope of search data to be considered for sharing

DATA ON QUERY	DATA ON THE SEARCH RESULT PAGE (SERP)	DATA ON THE USER
Keywords (such as raw search string, synthetic search string)	Clicked URLs (first clicked result, last clicked result, all clicked results)	Unique identifier
Timestamp (week, day, hour, seconds)	Zero-Click search (yes/no)	Device metadata (for instance, mobile/ desktop, browser metadata)
Connected queries in the same session	Results ranking (top 3, top 5, top 10)	Location data (IP-address, GPS)
	Layout of the SERP (sponsored results, one-boxes)	Other available user attributes (age and gender from account data, for example)

This is certainly not a complete list, but it invites policy makers to think how different data, each at various level of granularity (listed in parentheses), can be mixed and matched from the different categories, and this would result in significantly different data sets that may be shared.

Another issue relates to the **impact that the legal geographical reach of the DMA may have on the scope of data that can be provided**. Can the DMA only ever demand relevant data to be shared that was collected in the EU, or provided by citizens of the EU? Or can it demand from gatekeepers operating in the EU that all of their relevant search and query data, no matter where and from whom it is collected, and no matter where it is stored, fall under the scope of the DMA's sharing obligation? Given the fact that data-driven learning and improvements of the search algorithm provided to EU citizens is also (at least potentially) determined by data that was provided outside of the EU, the latter may be justified.

KEY QUESTIONS

- What is the **precise scope** of data to be shared with respect the detail on the query, the search results page, and the user?

- What is the **scale of data to be shared** (for instance, full or random samples, only data from within the EU)?

- What is the appropriate **timeliness of the data** (frequency of updates and recency of the data)?

3.3 Who can receive access to search data?

The answer to this question may seem obvious at first, as Article 6(11) limits access to other 'search engines'. However, this calls into question what exactly the purpose of this obligation is, how narrowly defined the 'search market' is, and whether one should evaluate the seriousness of the access seeker to contest the access provider.

First, does the access seeker have to be in the area of general search, or can it be also in more narrow search markets, such as product search or location search? **The definition of an 'online search engine' in the DMA suggests that only general search engines are subject to the access obligation**,[9] but does this also extend to the access seeking search engine? This interpretation is very important. If the scope of third parties that can seek access is rather small and limited to other search engines, then the goal of this provision would clearly be to enable other general search engines to contest the gatekeeper. In this case, what if an undertaking just cooks up a general search engine in order to receive data 'as a search engine', but use that data for a totally different business really? Should there be a vetting procedure determining the seriousness to contest the incumbent by an access seeker. This seems problematic. And if not, is this worrisome with respect to contestability? As search engine data may well be repurposed to innovate and pursue different types of services, this would be welcomed from an innovation point of view, and this could also provide a stepping stone for entry of new digital firms (not necessarily in the search market), which may ultimately be able to become a sizable competitor and thus to increase contestability in digital markets. If, however, access is indeed limited to other search engines in the narrow sense, then this innovation and competition potential coming from search data is forfeited. In reverse, if the scope is meant to be rather wide, why has it been limited to 'search engines' in Article 6(11)?

Relatedly, is it realistic that a third-party search engine takes the main search engine head on and tries to improve its search algorithm with the ambition to take over the leading position in the general

[9] The definition of an 'online search engine' is borrowed from the Platform-to-Business Regulation (EU 2019/1150) , Article 2(5), where it is noted that « 'online search engine' means a digital service that allows users to input queries in order to perform searches of, in principle, **all websites**, or all websites in a particular language, on the basis of a query **on any subject** in the form of a keyword, voice request, phrase or other input, and returns results in any format in which information related to the requested content can be found; » [emphasis added].

search market? Since Article 6(10) limits access to search engines, the DMA seems to rather assume this narrow view of contestability *á la Baumol* (1986). However, there is reason to **doubt that such frontal contestability is feasible**. As detailed above, the access provider will only be able to make a fraction of its data trove ever available to third parties, due to technical limitations and due to privacy reasons. Moreover, significant investments in (duplication of) infrastructure are necessary to lead this market. Competing search engines deliberately try to differentiate themselves, for instance, in the dimension of privacy or sustainability. However, many competing search engines also rely on syndication agreements with other search engines (especially with Microsoft's Bing) for search results, which also limits their potential to differentiate themselves.

Suppose contestability in the narrow sense is desired and realistic, then the main search engine is supposed to be challenged by a not yet leading search engine. However, the thresholds for the number of users and business users at which a search engine qualifies as a core platform services are relatively low. This is especially so because business users do not necessarily have to actively sign up with the search engine, but only have to appear in the search index and be listed as a search result. In consequence, also not yet leading search engines may well have to provide access to its search and query data to third parties under the DMA, including to the main engine, as there is no restriction in Article 6(11) or anywhere else in the DMA that would prevent gatekeepers to access data by other gatekeepers made available under the DMA.[10] If contestability of the main search engine is indeed the short term goal of Article 6(11), which seems to be the most likely interpretation, such reciprocal data access may rather weaken than strengthen the position of the most likely challenger.

3.4 Fair, reasonable, and non-discriminatory access

Gatekeepers do not have to provide access for free, as in the data portability provisions, but have to provide access on FRAND terms. It is worth considering the two main parts of FRAND access separately, that is , "non-discriminatory" and "fair and reasonable". The former relates to who shall receive access and whether each data seekers receives the same data. The latter relates to an appropriate price for such access. Both are very challenging problems on their own and discussed in turn.

3.4.1 Providing non-discriminatory data access

One interpretation of 'non-discriminatory' could be that access has to be non-discriminatory in the sense that every third-party search engine should receive the same type of data. Another interpretation is that the gatekeeper must offer a menu of data access possibilities, and the third-party search engines can choose their appropriate access service from that menu. Both interpretations bear issues, as discussed below.

Different search engines have different needs for data, depending on their specific business model and value proposition. If a new search engine were to take on Google, then it would first need to pick at least one area (say product search, or people search) where it can potentially offer a true improvement over the incumbent's offering. But to that end, the access seeker would need to be able

[10] This is remarkably different to the draft proposal of the Data Act, where gatekeepers are supposed to be denied from accessing data made available under the Act.

to specify what kinds of data are most useful to it, or at least to be able to pick from a comprehensive menu of option.

In this regard, the access provision under Article 6(11) bears some specific challenges typically not present in other cases where FRAND access is required. That is, providing more detail in one dimension, almost certainly requires –through the need for anonymisation and technical limitations – to reduce detail in another dimension of the data.

- If the interpretation of FRAND is that every search engine shall receive the same data, then such tailoring of data access would not be possible. The access provider must then offer a one-size-fits-all access, which ultimately may not be truly helpful for anyone to be able to challenge the incumbent.
- If the interpretation is that access seekers can pick from a menu of options, then there may likely be issues with identifying the right specification of those options, and there certainly are additional challenges for anonymisation. Who is then to determine the kind of data access options? Can access seekers specify what kind of data they would want or can the gatekeeper determine what options are appropriate?

In the latter case, the gatekeeper likely has distorted incentives and in consequence is not likely to provide the most useful data to potential competitors. In the former case, the preferred data choice of access seekers must still meet anonymisation requirements and is thus limited. More fundamentally, if a menu of data sets is offered, each of which by itself would satisfy anonymisation requirements, then the data set that can be derived from combining these individuals data sets is much less likely to still satisfy the same anonymisation requirements. This would mean that access seekers could be limited to picking only one of the data access options offered (assuming that is sufficient to prevent recombination). Or, what would be worse for 'contestability', each individual data set must be more strongly anonymised (leaving less detail to the data), so that even if the menu of data sets is combined, sufficient anonymisation can be preserved.

3.4.2 What is the appropriate price for access under FRAND terms ?

Finally, the determination of a 'fair and reasonable' access price is a central issue of the FRAND access terms. In general, the **determination of a 'fair and reasonable' price is going to be heavily influenced by the implementation questions discussed above**, such as whether data is provided through a data trust, via data sandboxing (in-situ), what the scope of the data is, and how many data access options are to be provided. Thus, it seems advisable to achieve clarity on the other implementation questions first, before setting a price for access.

Next, a common understanding must be reached on the meaning of 'fair and reasonable' in this context. The price should be fair and reasonable not only to the access seekers, but also to the access provider. To this end, it can be useful to think in two broad price components:

1) the direct costs of providing access, which may be determined through accounting measures or (as is the case in telecom markets) through an engineering cost model
2) a mark-up, which is determined by economic considerations, such as a risk and innovation premium, or opportunity costs.

Such a 'cost-plus' breakdown of the access pricing is roughly also the approach that is used in other regulated network industries where some mandated access regimes with price regulation are in place, such as energy markets and telecoms. However, the lessons from these industries are that it is a formidable task to determine an appropriate price, and it took many years to achieve this.[11] **Much depends on the precise goal of the access regime**. The 'efficient' access price is going to be different depending on whether one wants to preserve innovation incentives of the incumbent (in this case the mark up and price should be higher), or whether one wants to promote entry (in which case the price should be lower). A notable difference to those regulated network industries is also that regulators were looking for an 'efficient' price, and not a 'fair and reasonable' price. In a FRAND regime, the price is not supposed to be unilaterally set by the regulator, but to be negotiated between access provider and access seeker.

Both parts, the determination of the direct costs, as well as the determination of the mark-up have their own challenges. With respect to direct costs, investments in IT infrastructure are typically lumpy and there exist large economies of scale and scope, which make it difficult to attribute costs to any specific activity. Moreover, marginal costs may be zero. In telecoms the concept of 'incremental costs' was therefore introduced, which measured the difference in cost at a supra-marginal level. Here the gatekeeper would have to demonstrate the costs that actually arise from providing access, and whether these costs occur only once (for setting up the access regime), repeatedly (which each new data set made available), or continuously (as access is being sought). Certainly gatekeepers have an incentive to inflate costs and to pursue clever accounting to prove it. These reasons have led regulators in the energy and telecom domains to use non-accounting/non-self-reporting methods of determining the costs over the years.

With respect to the appropriate **mark-up, the challenge is to balance innovation and sabotage incentives of the gatekeeper with those of the access seekers**. If the mark-up is low, the access provider may invest less in innovation of the service through which it was able to collect superior data, or it may innovate less in its ability to collect data. However, these innovation risks seem to be rather low in the context of gatekeepers under the DMA, especially a search engine, as these firms are financially very strong and have an inherently strong incentive to collect data. This may justify a lower mark-up. What seems to be more relevant are risks of sabotage, that is, efforts of the access provider to impede the quality of the data access (for instance, through providing lower quality data, or reducing the performance of the interface). A higher mark-up would reduce those incentives. Ultimately, according to the famous Efficient Component Pricing Rule (ECPR), the access provider

[11] For example, in telecom markets instead of looking at historic cost values, a complex "forward looking long-run incremental cost approach" (FLIRC) was devised, which built on inputs from an engineering model that would determine the cost that a hypothetically efficient incumbent would have using the currently available efficient technology. In energy markts, efficient costs are often determined through benchmarking – a procedure that is only available if there are several comparable firms in the market.

would have no incentives to sabotage anymore if the mark-up compensates exactly for the opportunity costs of providing access. However, as the goal is to make the market more contestable, a mark-up according to the theoretical concept of the ECPR is certainly too high – but provides an upper bound. In reverse, if contestability and entry is to be achieved, then a lower mark-up is justified, including a mark-up of zero.[12]

Given all these issues with the determination of the cost-based component, the vast infrastructure that potential gatekeepers like Google or Microsoft have available, (which can be used to provide access); and arguing that the mark-up should be zero anyway in order to promote competition and contestability, one may ask whether the 'fair and reasonable' should not be zero after all. Instead of keeping innovation incentives intact through a price, one may also seek to be mindful in this regard with respect to which data ought to be shared. For example, as argued above, forcing the gatekeeper to reveal the combination of full search term and the full search results page (ranking) seems problematic from an innovation perspective. At the same time, the data provided needs to be 'effective' to facilitate contestability, which requires more than just the search terms (see above on which data should be provided).

In conclusion, given the complexity of issues that arise in this context, it seems unlikely that the access provider and access seekers can ever succeed in negotiating a FRAND access between themselves. Thus, the **Commission needs to provide some guidance on the key trade-offs** mentioned, and yet, it is likely that ultimately the issues need to be resolved in courts. This means, access may not be provided for years to come, unless the Commission is willing to impose interim measures. However, also in this case, a specification has to be made and trade-offs have to be addressed.

KEY QUESTIONS

- What is the **process by which data access options** are determined? Who can pick data to be provided and how many different access options must be made available?

- Can a price of zero be '**fair and reasonable**'?

[12] In telecoms, this is known as Pure LRIC, where access providers receive only a compensation for the direct costs of access, but not a mark-up.

REFERENCES

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

Attoresi, M., Moraes, T. & Zerdick, T. (2020). EDPS TechDispatch on Personal Information Management Systems. Available at: https://edps.europa.eu/sites/default/files/publication/21-01-06_techdispatch-pims_en_0.pdf

Barbaro, M. and Zeller, T. (2006). A Face is Exposed for AOL Searcher No. 4417749. New York Times. Available at: https://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0

Baumol, W. J. (1986). Contestable markets: an uprising in the theory of industry structure. Microtheory: applications and origins, 40-54.

Competition and Markets Authority [CMA] (2020). Online platforms and digital advertising market study. Available at: https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study

Crémer, J., de Montjoye, Y. A., & Schweitzer, H. (2019). Competition for the digital era. Report for the European Commission.

Edelman, B., & Lai, Z. (2016). Design of search engine services: Channel interdependence in search engine results. Journal of Marketing Research, 53(6), 881-900.

Fishkin, R. (2019). Less than Half of Google Searches Now Result in a Click. SparkToro. Available at: https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/

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/un locking_digital_competition_furman_review_web.pdf

Graef, I. (2020). The Opportunities and Limits of Data Portability for Stimulating Competition and Innovation- Competition Policy International, Antitrust Chronicle November 2020 (II), Available at SSRN: https://ssrn.com/abstract=3740185

Graef, I., & Prüfer, J. (2021). Governance of data sharing: A law & economics proposal. Research Policy, 50(9), 104330.

Hong, Y., He, X., Vaidya, J., Adam, N., & Atluri, V. (2009, November). Effective anonymisation of query logs. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1465-1468).

Klein, T. J., Kurmangaliyeva, M., Prüfer, J., & Prüfer, P. (2022). How important are user-generated data for search result quality? Experimental evidence. Tilburg University Working Paper.

Krämer, J., Senellart, P., & de Streel, A. (2020). Making data portability more effective for the digital economy: Economic implications and regulatory challenges. CERRE Policy Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3866495

Krämer, J., Schnurr, D., & Micova, S. B. (2020). The role of data for digital markets contestability: case studies and data access remedies. CERRE Policy Report.

Krishnan, U., Moffat, A., Zobel, J., & Billerbeck, B. (2020, April). Generation of Synthetic Query Auto Completion Logs. In European Conference on Information Retrieval (pp. 621-635). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-45439-5_41

Martens, B., Parker, G., Petropoulos, G., & Van Alstyne, M. W. (2021). Towards efficient information sharing in network markets. TILEC Discussion Paper No. DP2021-014, Available at SSRN: https://ssrn.com/abstract=3956256 or http://dx.doi.org/10.2139/ssrn.3956256

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9. Available at: https://www.nature.com/articles/s41467-019-10933-3

cerre Centre on Regulation in Europe

Avenue Louise 475 (box 10)
1050 Brussels, Belgium
+32 2 230 83 60
info@cerre.eu
www.cerre.eu
@CERRE_ThinkTank
Centre on Regulation in Europe (CERRE)
CERRE Think Tank