GDPR Transparency as a research method

Jef Ausloos

Institute for Information Law (IViR) – University of Amsterdam

Table of Contents

[NOTE TO READERS]

This draft is still very rough and needs further reflection/expansion on certain key issues. Feedback on the scoping of the paper, ethical dimensions, and relation to other research methods would be particularly valuable.

ABSTRACT

Data-driven research is rapidly becoming mainstream across different disciplines in academia and in investigative journalism. One of the key challenges researchers often struggle with is how to obtain good data. Whether one investigates political micro-targeting, discriminatory insurance practices, or physical exercise patterns across demographics, all require high-quality data. Obtaining the necessary data often depends on the goodwill of the entity in control of that data, frequently a private entity carefully guarding its data-assets. Researchers have tried to find ways round this through a range of technical tools (e.g. browser plugins, data scraping), with varying degrees of success.

This paper purports to add an important *legal* tool to the catalogue of data-driven research methods: i.e. GDPR transparency measures, and data subject rights in particular. The reinforced data subject rights in the General Data Protection Regulation (GDPR) have been lauded (and denounced) for their apparent potency. Especially the rights of access, portability and 'explanation' are said to ensure important values such as autonomy, accountability and fairness. To make true on this promise, these rights can (should?) be used as a methodological tool for investigative research. Not just to scrutinize the data practices of those holding the data (e.g. uncovering discrimination), but also to enable broader research objectives (e.g. informing health policies). Indeed, these rights offer a legal hook to effectively crowdsource data-driven research.

Against this backdrop, the paper will describe the merits, challenges, limitations and best practices of using data subject rights as a research method. It will do so by building on past experiences of some first initiatives in the EU and assessing how the growing number of 'download my data' tools are useful/less. After an introductory section setting out some of the issues currently faced in data-driven research, the paper will explain how they might (not) be tackled by some of the key features of using GDPR transparency measures as a research method. The following section (4) is perhaps the core of the paper, describing the legal and practical operationalisation of these novel research methods, as well as the requirements for making them meaningful. The final section then takes a step back, positioning GDPR transparency measures within their broader (research methods and economic-technical) surroundings.

1 *Introduction: The GDPR as a Curse and a Blessing to Data-Driven Research*

Data Protection versus Investigative Research – To many investigative researchers, whether in academia or journalism, data protection rules have traditionally been seen as an important hurdle. They are perceived to impose rigid restrictions on what can be done with (personal) data, e.g. data minimisation, data protection impact assessments, information requirements, etc. Uncertainty and/or anxiety as to data protection compliance can lead to stifle data-driven research initiatives. It has been shown that at pre-GDPR data protection rules have resulted in over-restrictive research governance inside, traditionally risk-averse, academic institutions.[1] Over-restrictive policies may elicit either retreat from data-driven research or simple disregard of rules among researchers. The GDPR does not appear to improve the situation much. While it recognises the importance of scientific research by installing several derogations from the default data protection rules, it also leaves a lot unsaid and outsources crucial interpretative guidance to Member States (cf. Art.89(2)-(3)). As a result, under-resourced/staffed (review boards within) academic institutions are struggling to facilitate data-driven research. In sum, even if 2018, with both the entry into force of the GDPR and the Cambridge Analytica scandal, has put data protection firmly on researchers' agenda, important hurdles remain to be overcome.

Data Protection Complementing Investigative Research – Without wishing to disregard the issues data protection law raises for data-driven research, this paper aims to highlight how it can constitute an important tool *enabling* such research. One type of GDPR provisions in particular shows important promise on doing so: i.e. transparency measures. While Article 6(1)a establishes transparency as a core principle to EU data protection law, the central provisions fleshing it out can be found in Artt.12-15 and 20. These set out the modalities (Art.12) and the content of *ex ante* (Artt.13-14) and *ex post* (Artt.15 and 20) transparency. As will be explained later on, these provisions enable breaking through walled, corporate gardens in order to obtain high-quality data that can feed into investigative research. Of course, GDPR transparency rights are no panacea, and bring their own set of issues and limitations. It is therefore important to clearly delineate their scope and potential.

2 *Issues in Data-Driven (SSH) Research*

Data-driven research in the Social Sciences and Humanities (SSH) Is facing several important challenges. Complexification and privatisation of data eco-systems are two noteworthy ones. The 'big data' mantra is still growing strong, expanding both in breadth and depth. New, improved and cheaper sensors, combined with growing processing capabilities enable an incessant influx of 'data'. Interconnectivity and the 'agile turn' also precipitate highly-dynamic and constantly reinforcing data eco-systems, with a natural tendency of concentration of data in the hands of central nodes.[2] Both these central nodes, as well as the surrounding eco-systems more broadly are predominantly in private hands, rendering the vast majority of data being captured and generated proprietary. It is not within the scope of this paper to map the highly diverging incentive-structures of these private entities

[1] See the extensive work by Erdos: David Erdos, 'Stuck in the Thicket? Social Research under the First Data Protection Principle' (2011) 19 International Journal of Law and Information Technology 133; David Erdos, 'Systematically Handicapped? Social Research in the Data Protection Framework' (2011) 20 Information & Communications Technology Law 83; David Erdos, 'Constructing the Labyrinth: The Impact of Data Protection on the Development of "Ethical" Regulation in Social Science' (2012) 15 Information, Communication & Society 104.

[2] Seda Gürses and Joris van Hoboken, 'Privacy after the Agile Turn' in Jules Polonetsky, Omer Tene and Evan Selinger (eds), *Cambridge Handbook of Consumer Privacy* (CUP 2017) <https://osf.io/preprints/socarxiv/9gy73/> accessed 12 November 2017.

in sharing 'their' data. Suffice to say that these entities will only rarely release data entirely and/or unconditionally, whether for legal, economic or technical reasons. Hence, the 'privatisation of data' has a deep impact on SSH research, as corporate entities become *de facto* gatekeepers of what in effect may be described as data-monopolies. This essentially constrains research to what lies in the private entity's interests (notably profit and PR), rendering it hard to impossible for outside actors (notably civil society and academia) to perform research in parallel. Moreover – as has been amply critiqued already – sealing off data (processing operations) also renders it very hard to scrutinise the practices of these corporate entities themselves.[3]

RESEARCHER RESPONSES TO DATA PRIVATISATION – Researchers have found many different ways of going around the increased sealing-off of data by private entities. In the context of this paper, four categories of obtaining access to data nonetheless can be identified, each with their own benefits and drawbacks. *Firstly*, some researchers/institutions obtain access to privately held data via 'data philanthropy' initiatives[4] and/or through amicable relationships they might entertain with the relevant actors (e.g. Facebook's Social Science One initiative;[5] Consumer Data Research Centre[6]). Beneficial as these may be to the respective researchers, such an approach risks further solidifying existing power dynamics in academia (and the private sector). Moreover, researchers/institutions may have several reasons not wanting to associate with private entities as a precondition for doing research. The last three approaches are therefore more adversarial in nature. *Secondly*, researchers/institutions may find creative ways to re-purpose existing tools (e.g. SDK, API) on offer by the private entities, in order to get access to data. The problem with this approach is that it will often go against the respective Terms of Service, and researchers may therefore risk legal or retaliatory action (e.g. lawsuit or being kicked off platform). Perhaps even more problematic from a privacy and data protection point of view is that this approach does not preclude bad faith (or at least ethically questionable) actors to obtain access to the respective data, the quintessential example being Aleksander Kogan / Cambridge Analytica. *Thirdly*, researchers/institutions may also rely on entirely independent technical/methodological tools to obtain useful data otherwise sealed-off by private entities. Examples include the development of browser plugins to monitor social media (WhoTargetsMe,[7] Algorithms Exposed,[8] FB-Forschung[9]), search engines (e.g. DatenSpende)[10] and/or browsing behaviour more generally (e.g. Robin)[11] or more traditional surveys. Such an approach has the advantage of being more resistant to retaliatory action by the respective entities,[12] and slightly lower risk of bad actors (still requiring active recruitment of – and thus presumably more informed – research subjects). Still, this approach has raised a lot of questions as to the quality of data (notably as to comprehensiveness and representativeness). *Fourthly*, researchers/institutions may put their hopes in regulatory intervention, forcing more transparency. So far, legal instruments primarily focus on transparency of public sector information (i.e. freedom of information acts or the EU's Public Sector

[3] See references in: Muhammad Ali and others, 'Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes' [2019] arXiv:1904.02095 [cs] 5 <http://arxiv.org/abs/1904.02095> accessed 8 April 2019.

[4] E.g.: https://www.mastercardcenter.org/action/call-action-data-philanthropy/

[5] Facebook, 'Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections' (*Facebook Newsroom*, 9 April 2018) <https://newsroom.fb.com/news/2018/04/new-elections-initiative/> accessed 23 April 2019.

[6] https://www.cdrc.ac.uk/about-cdrc/.

[7] https://whotargets.me/en/

[8] https://algorithms.exposed/

[9] https://fbforschung.de/. This tool combines a data-gathering plugin with occasional surveys with participants, enabling more in-depth information than what can merely be observed.

[10] https://datenspende.algorithmwatch.org

[11] Balázs Bodó and others, 'Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2017) 19 Yale J.L. & Tech. 133.

[12] Though certainly not immune, as illustrated by plugins of ProPublica and WhoTargetsMe slightly being blocked by Facebook changing some of its HTML code. Ariana Tobin Jeremy B. Merrill, 'Facebook Moves to Block Ad Transparency Tools — Including Ours' (*ProPublica*, 28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 17 April 2019; Digital, Culture, Media and Sport Committee, 'Disinformation and "Fake News": Final Report' (House of Commons 2019) 8 64.

Information Directive), but new initiatives are underway to open up privately held data as well.[13] This approach requires patience, legal expertise and persistence and might still fail to enable access to fine-grained information needed by researchers.[14]

In sum, researchers wishing to obtain access to the vast amounts of data collected by private companies – whether to investigate those companies' practices or use the data for other research goals – may face many challenges.[15] Issues range from important information quality concerns, to advantaging privileged researchers/institutions, (technical) arms races and risks of malicious actors obtaining access to information. In the face of these complex challenges and issues, there is a need for a more coherent, robust and sustainable approach to data-driven research. This paper aims to explore how GDPR rights may contribute, even if only slightly, to this broader goal. As will appear, several of the issues mentioned above may be mitigated when using GDPR transparency measures to obtain research data. Having said that, it is important to see these measures as merely additional tools in the data-driven researcher's toolbox. I.e. they do not replace, but *complement* existing methods (e.g. enabling cross-referencing with a diary study). Importantly, they only provide an avenue for *obtaining* high-quality data, not for understanding it.

3 Features of GDPR Transparency Measures as a Research Method

While often presented as a major hurdle, the GDPR – and transparency measures within it – can also be qualified as a boon to data-driven research. Before digging into the specific operational details of these transparency measures in section 4, this section will briefly explore some of their key features.

3.1 Break Through Information Monopolies (but security risks)

As mentioned before, corporate entities have become *de facto* gatekeepers of vast amounts of data about individuals, how they interact with each other and their environment, as well as the digital infrastructures underlying modern society. This exclusionary access precipitates vicious circles of more (advanced) data (research methods) and insights in the hands of a few dominant players. In light of these players primarily being corporate entities, the research agenda will generally be prioritized in terms of what fits within a commercial strategy (and resource constraints), leaving unexplored much research goals. This is not to say that any and all data sealed off by corporate entities necessarily comprise huge, but latent, research potential. Nor that opening up will automatically lead to the data actually being used for (laudable/desirable) research purposes. What *can* be observed though, is the *a priori* power corporate entities have over what data researchers can access, how it is accessed and often even how it is interrogated.[16]

Adversarial approaches to extracting useful data (e.g. scrapers) are often short-lived due to counter-measures (resulting in cat-and-mouse games). Research into the dataflows in smartphone(-app) data ecosystems in particular often faces infrastructural resistance to empirical research.[17]

[13] Cf. The European Commission's mission to 'Build a European Data Economy', involving several legislative initiatives, several of which promoting more transparency of privately-held data (https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy). Notably: https://ec.europa.eu/digital-single-market/en/platforms-to-business-trading-practices; https://ec.europa.eu/digital-single-market/en/guidance-private-sector-data-sharing

[14] Indeed, one of the key differences between freedom of information requests and the GDPRs transparency rights is that the latter enable access to subject-specific information.

[15] See for example: Digital, Culture, Media and Sport Committee (n 13) 64. Recognising that "There is now no practical way for researchers to audit Facebook advertising."

[16] Cf. ibid. Explaining how Beer draws a parallel with Derrida's Archons (i.e. 'lords of the archives').

[17] Michael Dieter and others, 'Store, Interface, Package, Connection. Methods and Propositions for Multi-Situated App Studies' [2018] SFB 1187 Medien der Kooperation - Working Paper Series 1, 12.

Because of their legal backing, GDPR transparency measures, in contrast, are more resistant to counter-measures. As such, they may constitute a valid methodological alternative for exploring trends, actors, data flows, etc. that are (deliberately and/or technically) obfuscated.

In practice, companies will often still resist providing comprehensive data in response to access requests, even when specifically insisted upon. A notable case in point would be the British Uber drivers who were refused access by Uber to location data and the timestamps for when the App was turned on/off.[18] Other examples include Twitter, Facebook and Apple refusing access to personal datasets (e.g. browsing-behaviour, voice-assistant data), claiming it would be too burdensome and/or they do not have a duty to deliver the data in question.[19] Considering the fact that these cases relate to personal data that is known to be collected by the respective companies, one may question the efficacy of using GDPR transparency measures for more exploratory purposes. Indeed, it might be hard to impossible to establish a response to an access request is (deliberately) incomplete.[20] In short, the ability of GDPR transparency measures to pierce through corporate data protectionism may be significantly diminished in practice (essentially requiring legal action to perhaps achieve the desired goal).

Security concerns may further constrain the potential of GDPR transparency measures (and access rights in particular) in prying open corporate black boxes. For example, accommodating subject access requests may result in data breaches, when not well thought-through.[21] Indeed, authentication of data subject requests in particular can be quite complicated in practice. As will be explained later (cf. Section 4.1.), the GDPR does anticipate situations where the identity of the requestor may be hard to verify (Art.12), though remains unclear about practical solutions (rightfully so).[22] In short, transparency request may face closed doors because of legal liability concerns over (supposed) identity verification issues. Other legal defences – intellectual property protection[23] – may also be available depending on the circumstances.

3.2 Relative Ease (but bad actors)

One of the main impetuses behind the overhaul of EU data protection law – and the GDPR in particular – was to bolster and facilitate the exercise of data subject rights.[24] Indeed, even after the

[18] Which would enable them to 'calculate holiday pay and claims for minimum-wage back pay when their earnings dip below that, such as during fallow periods when they have no rides but are logged in to the app and prepared to accept passengers'. 'Uber Drivers Demand Their Data' [2019] *The Economist* <https://www.economist.com/britain/2019/03/20/uber-drivers-demand-their-data> accessed 24 April 2019.

[19] Michael Veale, Reuben Binns and Jef Ausloos, 'When Data Protection by Design and Data Subject Rights Clash' (2018) 8 International Data Privacy Law 4; Jef Ausloos, 'Paul-Olivier Dehaye and the Raiders of the Lost Data' (*CITIP blog*, 10 April 2018) <https://www.law.kuleuven.be/citip/blog/paul-olivier-dehaye-and-the-raiders-of-the-lost-data/> accessed 23 April 2018.

[20] Jef Ausloos and Pierre Dewitte, 'Shattering One-Way Mirrors – Data Subject Access Rights in Practice' (2018) 8 International Data Privacy Law 4, 15.

[21] Chris Norval and others, 'Reclaiming Data: Overcoming App Identification Barriers for Exercising Data Protection Rights' [2018] arXiv:1809.05369 [cs] 4 <http://arxiv.org/abs/1809.05369> accessed 20 September 2018.

[22] It is worth pointing to several researchers that are working on this issue from a more technical background: Norval and others (n 24); Coline Boniface and others, 'Security Analysis of Subject Access Request Procedures How to Authenticate Data Subjects Safely When They Request for Their Data' (2019) 2 <https://hal.inria.fr/hal-02072302/document> accessed 4 April 2019.

[23] Defenses based on intellectual property protection can never result in blocking transparency rights altogether. Cf. Gianclaudio Malgieri, 'Trade Secrets v Personal Data: A Possible Solution for Balancing Rights' (2016) 6 International Data Privacy Law 102.

[24] DG Justice, 'Summary of Replies to the Public Consultation about the Future Legal Framework for Protecting Personal Data.' (European Commission 2010); Committee on Civil Liberties, Justice and Home Affairs, 'Report on a Comprehensive Approach on Personal Data Protection in the European Union (Voss Report)' (European Parliament 2011) 2011/2025(INI) 8–9; European Commission, 'Communication from the Commission to the European Parliament, the Council, the European

GDPR was promulgated (but before it entered into force), empirical research laid bare significant discord between subject access rights in theory and in practice.[25] Access requests are often received with surprise, confusion and irritation. To the extent access requests were responded to, individuals often obtained part of their personal data and in different formats (e.g. .xml, .doc, .xlsx, plain text). The primary way through which the GDPR purports to improve the situation, is by clearly specifying a number of data subject right modalities that need to be complied with by controllers and processors, under threat of considerable fines. Anecdotal evidence suggests that since the entry into force of the GDPR in May 2018, there is an improvement in how access requests are received and responded to, at least with regard to 'simple'/straightforward requests (e.g. account-details). There still appear to be many bumps in the road regarding more advanced requests (e.g. Facebook's tracking data, Apple's Siri-voice data).

A comprehensive list of information to be provided, combined with specific modalities as to how they should be provided, renders access requests relatively easy, at least in theory. Indeed, there are no *a priori* thresholds to exercising the right[26] and those processing personal data have an explicit duty to facilitate rights (and transparency more broadly). The way in which many online services are dealing with this, is by implementing 'download my data' functionalities. While the data obtained through these functionalities is often not complete, they at least offer some sort of uniformity as to how access rights are responded to, in turn making it easier for potential researchers to analyse a large number of such crowdsourced responses.

Removing friction in the transparency process, also risks rendering it easier for bad (faith) actors to obtain unwarranted access to information. This is exemplified by the tight-roping exercise of authenticating users. In order to prevent data from being shared with impersonators, for example, it will be important to authenticate the request.[27]'' Different methods may be more or less appropriate (and/or burdensome) depending on the circumstances at hand.[28] For instance, it might be considered unnecessary and disproportionate for services that already have a two-factor authentication login, to require individuals sending a copy of their passport. Similarly, requiring a (redacted) passport-copy would not be appropriate coming from organisations who do not have the 'real-life' identity of the individual in the first place (e.g. in which case a device/browser identifier might be more suitable).

Facilitating access also makes it easier to organise and coordinate access requests. When done for legitimate and laudable research aims, this may appear positive. It is not hard to contemplate situations where the merits of easy coordination of access requests may be more questionable. In theory, coordinating access requests could be deployed as a DDoS attack by competitors, activists and/or others. It is also not unconceivable that competitors may start exploiting access requests in order to obtain relevant information on business-practices. It is too soon to tell whether these are purely hypothetical situations, or even if they are 'bad' per se when taking a broader perspective (e.g. extra incentive to smooth out internal compliance procedures; enabling market transparency and competition). For the purposes of this contribution, however, the focus is on how the GDPR's transparency modalities are used for legitimate research purposes.

Economic and Social Committee and the Committee of the Regions Safeguarding Privacy in a Connected World. A European Data Protection Framework for the 21st Century'.

[25] See notably: Ausloos and Dewitte (n 23); René LP Mahieu, Hadi Asghari and Michel van Eeten, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect' (2018) 7 Internet Policy Review <https://policyreview.info/articles/analysis/collectively-exercising-right-access-individual-effort-societal-effect> accessed 16 July 2018.

[26] Though in some circumstances, authentication might raise a *de facto* hurdle that could be quite burdensome (cf. Section 4.1.).

[27] Cf. Article 29 Working Party, 'Guidelines on the Right to Data Portability' (2017) Guidelines WP242 14 <http://ec.europa.eu/justice/article-29/documentation> accessed 17 January 2017.

[28] Boniface and others (n 26) 1.

3.3 Legally Enforceable (but ethical questions)

GDPR transparency measures are not just (theoretically) *easy* to exercise, they are legally enforceable. Indeed, the GDPR now foresees administrative fines of up to 20 million EUR (or 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher).[29] Less dramatically, but nonetheless significant, are the increased powers of data protection authorities to investigate, audit and take a range of injunctive actions.[30] These 'sticks' aim to further ensure transparency measures are proactively complied with by those processing personal data pre-empting the need for individuals to go to a data protection authority or court.

Whereas the law clearly pushes for more transparency with regard to data processing operations, some might question whether using it as a way for obtaining research data, does not align with its intentions. Traditionally, European data protection has had a close connection with the right to privacy and one might argue that it is especially aimed at safeguarding the respective individual's interests. Using data protection transparency measures to gather research data therefore appears to misuse/retrofit a legal device for unintended purposes. As such, its legal enforceability might be called into question. Taking a step back, however, it is important to dissociate data protection law from *privacy* and from being individually-focused. Historically, data protection law has always been about regulating the data infrastructures underlying society (from large, centralised data mainframes to the complex ecosystem today).[31] Indeed, the GDPR in particular can be considered as traffic rules, to regulate data-infrastructures. This includes a legal toolbox that gives some level of control over personal data and/or how it is processed to certain stakeholders (notably individual data subjects, data protection authorities, but also academics, journalists, artists and civil society organisations). Most importantly perhaps; the GDPR's transparency measures are intent-agnostic. This is crucial in understanding why, for example, using the right of access to obtain research data, does not go against the regulation's spirit per se. Of course, this does not preclude certain uses of the GDPR to be qualified as 'abuse of rights' within national civil law.

From the researcher's perspective, ethical issues still emerge (as with any other method of obtaining potentially very sensitive personal data about individuals). Researchers are still required to follow internal procedures (cf. ERB, data management plans, etc.) of research integrity.[32] This is of course no guarantee that problematic practices still pass, case in point being Kogan / Cambridge Analytica. Recognising GDPR transparency measures and data subject rights (of access/portability) in particular as valid methods for obtaining research data, may have unintended/undesirable consequences. For example, one might imagine it could create incentives among researchers (research institutions) to adopt similar coercive practices inducing data-sharing, widespread in the private sector today. The generally higher trust in research institutions (as opposed to the private sector) may also be exploited and offended. It will therefore be critical to maintain and ensure very high ethical standards when deciding to use the GDPR as a method for obtaining research data. *[...]*

[29] Art.83(5)

[30] Art.58.

[31] The distinction between privacy and data protection is a much debated issue that far exceeds the scope and purpose of this paper.
While the issues regulated by data protection law certainly overlap with individuals' right to privacy, it would be myopic to consider the overlap to be complete (and/or one to be comprising the other).

[32] Research integrity, here, is used in its widest meaning, including research ethics. Cf. Serge Horbach and Willem Halffman, 'Promoting Virtue or Punishing Fraud: Mapping Contrasting Discourses on "scientific Integrity""' (Printeger Project 2016) Deliverable D2.2 <https://printeger.eu/wp-content/uploads/2016/10/D2.2.pdf> accessed 28 February 2019; Gloria González Fuster and Serge Gutwirth, 'Promoting Integrity as an Integral Dimension of Excellence in Research. Legal Analysis' (Printeger Project 2016) Deliverable DII.4 <http://printeger.eu>; Heidy Meriste and others, 'Normative Analysis of Research Integrity and Misconduct' (Printeger Project 2016) Deliverable D2.3 <https://printeger.eu/wp-content/uploads/2016/10/D2.3.pdf> accessed 28 February 2019.

One way to ensure research integrity when using the GDPR to obtain research data, is by developing a platform or dashboard through which individuals can actively manage and sanitise their data. For example, they should be given a chance to go through the data received upon an access request and filter out some information (and the platform can already filter out all data not necessary for the specific research project by default).[33] Such a platform/dashboard should of course consider the way in which access requests are complied with in the specific context at hand. Especially regarding large datasets, it might be useful to visualise and offer more aggregate controls as well.

3.4 Data Quality (but 'quality' of data)

The GDPR ordains a detailed list of information to be made available whether proactively or upon request. The right of access – and, to a lesser extent the right to data portability – specifically require a full copy of all personal data being processed as well as a number of other types of information on how the respective individual's data *in particular* was processed (e.g. *specific* recipients and/or sources of personal data at stake). The GDPR further lists a minimum of more generic types of information to be specified on a push (e.g. privacy/data policy) or pull (e.g. some categories of information to be provided upon an access request) basis. Legal enforceability of these provisions, but also PR, constitute significant incentives for the information and data to be of high quality (i.e. truthfully reflecting the *actual* practices and data being processed). Even given the problematic track-record of many responses to access requests *not* containing the entire set of personal data actually being processed, one might hope that the categories of data that *are* provided are in themselves accurate/complete.

There is a noticeable trend, especially among large online services, to standardise the way in which access requests are accommodated (notably through 'download my data' options). Such standardisation is further stimulated by growing number of requests, but also enables the scaling of coordinated requests, e.g. for research purposes. Indeed, standardisation renders it less labour-intensive to collect and interpret data and information, enabling to automate (parts of) the process. Automation – and GDPR transparency obligations in general – also reduce individual subjectivity (deliberate or not) inherent to other research methods (e.g. diary studies) to which they may constitute useful companion pieces.

It is worth re-emphasising that even if much of the GDPR transparency measures are individually-focused (i.e. data subject rights of access or portability), the research goals do not necessarily need to be. Indeed, one of the key benefits of crowdsourcing/coordinating the gathering of research data through data access requests, for example, is that it enables to explore broader trends (e.g. discriminatory ad-practices, news dissemination/propagation), with more granularity than other legal transparency measures allow (e.g. FOIA / PSI).

Importantly, the quality of research data/information obtained through GDPR transparency measures should not be overstated. Some data/information may be omitted (deliberately or not), and the level of transparency provided may not account for broader contextual elements necessary to understand. Nor does the use of GDPR transparency measures as a method for obtaining research data solve the issue of representativeness. This last issue may pose particular issues when the demographics of the entire data-set are unclear (e.g. what is a representative sample of platform users, if it is unclear what the demographics of *all* users on the platform are?). Researchers should therefore remain critical of the information they obtain in this manner and devise well-thought-out strategies (recognising limitations) for doing so.

[33] Cf. Article 29 Working Party, 'Guidelines on the Right to Data Portability' (n 30) 12.

4 Operationalising GDPR Transparency Measures as a Research Method

Whereas the previous section aimed to outline some key features of using GDPR transparency measures as a research method, this section will take a closer look at the actual operationalisation of doing so. This will be done in two parts, taking a theoretical and a more practically-oriented perspective respectively. The first subsection describes the actual GDPR provisions, including their requirements and modalities. The second subsection is aimed at exploring some concrete examples of how the GDPR can – in practice – be used to obtain data for research purposes. Building on these theoretical and practical perspectives, the third sub-section will identify key requirements for making sure GDPR transparency measures can *actually* be used as a valid and useful method for obtaining research data.

4.1 In Theory

Fundamental Right – Transparency has been part of data protection discussions and rule-making from the very start. It is seen as a crucial element in countering information and power asymmetries resulting from growingly powerful computing capabilities. Unsurprisingly, transparency was also included as an important component of the right to data protection in the EU Charter of Fundamental Rights. Indeed, not only should data be processed *fairly*,[34] Article 8's second paragraph proclaims that everyone 'has the right of access to data which has been collected concerning him or her.' More recently, ensuring transparency of automated processing, and profiling in particular, is also a big concern for the Council of Europe, as illustrated in the (recently modernised) Convention 108[35] and earlier recommendations.[36]

Qualifying Transparency Measures:

- Overarching v concrete
- Stand-alone v enabling
- Target audience
- General v specific
- Push v pull
- Ex ante v ex post

Qualifying Transparency in the GDPR – Transparency provisions come in different shapes and tastes in the GDPR. It is an overarching principle that informs the interpretation and application of all GDPR provisions (cf. Art.5(1)a),[37] but also takes shape in certain concrete rights and obligations (cf. Artt.13-15). Among these concrete rights/obligations, transparency can either be an end in itself (e.g. Art.13-14) or be a critical component of safeguarding others (e.g. Art. 4(11) and 7 (valid consent); Art.22 (automated decision-making); Art.33 (data breach notification); Art.42 (certification mechanisms)). Transparency provisions can have different target audiences, from an individual data subject (e.g. Art.15), to the public at large (e.g. Art.13-14), or supervisory authorities (e.g. Art.30(4)). It is also possible to distinguish between general versus specific transparency requirements, the former comprising more abstract information about data practices (e.g. privacy policy) and the latter being tailored to the recipient (e.g. access request). Transparency provisions can also be differentiated on the basis of the time they kick in, *before* or *after* data is first being collected/processed. A final distinction that can be made is that between push and pull transparency provisions, i.e. who needs to take the initiative for transparency: the controller or the target audience (largely corresponding to transparency *obligations* versus transparency *rights*). All of these ways of categorising transparency provisions in the GDPR somewhat overlap, but are helpful nonetheless in qualifying the relevant forms

[34] Cf. Damian Clifford and Jef Ausloos, 'Data Protection and the Role of Fairness' (2018) 37 Yearbook of European Law 130.

[35] In particular Artt.8-9.

[36] E.g. Council of Europe, 'Recommendation on the Protection of Individuals with Regard to Automatic Processing of Personal Data in the Context of Profiling CM/Rec(2010)13'.

[37] Clifford and Ausloos (n 38).

of transparency later on. Indeed, they help better situate the twofold goal of transparency measures in the GDPR. On the one hand they have a protective dimension, ensuring demonstrable accountability. On the other hand, some of these measures also have an important empowerment dimension, putting controls in individuals' hands to be more informed. Both dimensions can be considered to contribute to a common goal of redistributing power stemming from information/data asymmetries.

CORE TRANSPARENCY MEASURES – The epicentre of transparency measures in the GDPR can be found in Articles 13-15. The first two articles list the information that controllers (i.e. those responsible for the data processing) need to provide proactively, i.e. at their own initiative and *before* they start processing personal data. In substance, Article 13 (focused on situations where personal data was obtained from individuals *directly*) and Article 14 (personal data was obtained *indirectly*) differ very little (cf. Table 1). These provisions can first and foremost be qualified as *protective* ('push' transparency) measures, forcing controllers to give proper thought to, and be upfront about, their processing operations and enabling to hold them to account later on. As such, they also serve as a useful compliance-testing tool for data protection authorities and/or other interest-groups. Having said that, Articles 13-14 also have an empowering facet to them. After all, they make data subjects (i.e. those to whom the personal data being processed relates) aware of processing taking place and as such can be seen as a *sine qua non* for empowering individuals to invoke one or more of their rights (e.g. object, erasure, portability). The most important components of *ex ante* transparency relate to the scope and purposes of processing, the risks involved, the retention period and how to exercise data subject rights.

Information Requirement	**Art. 13**	**Art. 14**
Identity and **contact details of the controller** and, where applicable, its representative	1(a)	1(a)
Contact details of the data protection officer, where applicable	1(b)	1(b)
Purposes of the processing for which the personal data are intended and **legal basis**	1(c)	1(c)
Categories of personal data concerned	-	1(d)
Where the processing is based on point (f) of Article 6(1), **the legitimate interests pursued** by the controller or by a third party	1(d)	2(b)
Recipients or categories of recipients of the personal data, if any	1(e)	1(e)
Details on potential **data transfers to third countries**	1(f)	1(f)
Retention period, or if that is not possible, the criteria used to determine that period	2(a)	2(a)
Existence of the data subject rights to access, rectification, erasure, restriction of processing, to object, and to data portability	2(b)	2(c)
Where the processing is based on consent, the existence of the **right to withdraw consent** at any time	2(c)	2(d)
Right to lodge a complaint with a supervisory authority	2(d)	2(e)
Whether the provision of personal data is a statutory or contractual **requirement**, or a requirement necessary to enter into a contract, as well as **whether the data subject is obliged to provide the personal data** and of the possible **consequences of failure** to provide such data	2(e)	-
Source from which the personal data originate, and if applicable, whether it came from publicly accessible sources[38]	-	2(f)
Existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, **meaningful information about the logic involved**, as well as the **significance and the envisaged consequences** of such processing for the data subject.	2(f)	2(g)

Table 1 - Ex Ante Transparency

Ex Post Transparency – Article 15 complements the information obligations in Artt.13-14, by granting data subjects an explicit *right* to obtain additional information (Table 2). Article 15 can be qualified as an *ex post* empowerment measure and essentially gives individuals the ability to force more transparency, specifically with regard to their own specific situation.[39] Indeed, *ex ante* transparency generally relates to information that is relevant to all processing operations and individuals (general information). Article 15 is focused on making sure transparency is also effective at the individual level. Importantly, this includes the provision of details on 'any personal data used for profiling, including the categories of data used to construct a profile.'[40] Controllers should be transparent about the full set of personal data they use to create the profile, where they got each piece of information from, what the exact purposes are of the profiling, and who it has been (or might

[38] Recital 61 does acknowledge that '[w]here the origin of the personal data cannot be provided to the data subject because various sources have been used, general information should be provided.'

[39] Ausloos and Dewitte (n 24).

[40] Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (Article 29 Working Party 2018) Guidelines WP 251 17 <http://ec.europa.eu/justice/article-29/documentation>.

be) shared with. Moreover, they will also have to specify the retention period(s) and the availability of data subject rights, including the ability to file a complaint with the DPA.

Information Requirement	**Art. 15**
Confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the **personal data**	1
Purposes of the processing	1(a)
Categories of personal data concerned	1(b)
Recipients or categories of recipient to whom the personal data **have been or will be disclosed**, in particular recipients in third countries or international organisations	1(c)
Retention period, or if that is not possible, the criteria used to determine that period	1(d)
Existence of the data subject rights to rectification, erasure, restriction of processing, and to object	1(e)
Right to lodge a complaint with a supervisory authority	1(f)
Where personal data are not collected from the data subject, any information on the **source**	1(g)
Existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful **information about the logic involved, as well as the significance** and the envisaged consequences of such processing for the data subject	1(h)
In case of transfer to third country, information about the appropriate safeguards	2

Table 2 - Ex Post Transparency

DATA PORTABILITY – The new right to data portability offers some promise for use in order to obtain research data. Put very briefly, Art.20 grants data subjects the right to receive their personal data, 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.[41] Moreover, data subjects can request their personal data to be directly transferred from one controller to another where technically feasible. It is not hard to see how this provision may make the process of sharing personal data with researchers a lot smoother.[42] Indeed, in contrast to the right of access, the right to data portability actively recognises the value and ability for data subjects to move their personal data from one entity to another. Having said that, three important constraints limit the practical scope of this right. *Firstly*, it only applies to personal data that the data subject has provided to the controller, excluding for

[41] The format should be interoperable (i.e. 'the ability of disparate and diverse organisations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organisations, through the business processes they support, by means of the exchange of data between their respective ICT systems.'). Machine-readable has been defined in EU law as: 'a file format structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure. Data encoded in files that are structured in a machine-readable format are machine-readable data. Machine-readable formats can be open or proprietary; they can be formal standards or not. Documents encoded in a file format that limits automatic processing, because the data cannot, or cannot easily, be extracted from them, should not be considered to be in a machine-readable format. Member States should where appropriate encourage the use of open, machine-readable formats.' Recital 21 of Directive 2013/37/EU, cited in: Article 29 Working Party, 'Guidelines on the Right to Data Portability' (n 30) 16–18. The Working Party further specifies that '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, CSV,...) along with useful metadata at the best possible level of granularity'

[42] Even the Working Party 29 (uniting all EU data protection authorities) explained how the right might be useful to learn more about music consumption by using the right with streaming services or assessing carbon footprint by using the right with loyalty cards. ibid 4–5.

example 'inferred' and 'derived' data.[43] *Secondly*, it only applies where processing is based on 'consent' or 'necessity for the performance of a contract' as a lawful ground, effectively exempting four other grounds (including the important 'legitimate interests' ground). *Thirdly*, and probably the least restrictive constraint, the right to data portability only applies in situations where the respective personal data is processed 'by automated means'. Based on Art.20(4) controllers might (partially) fend off portability requests when they can establish that accommodating them would 'adversely affect the rights and freedoms of others'. One could argue that this is the case when the respective personal data includes data about other individuals (e.g. transactional information). Though research purposes benefit from a more lenient approach in the GDPR, proactive steps may still be necessary to minimise (risks to) third party personal data in the research project (cf. Art.89).[44]

TRANSPARENCY MODALITIES – The GDPR also lists a number of modalities, so as to ensure transparency is effective. The key provision for this is Article 12, but some specific modalities can also be found within the respective provisions discussed above. Importantly, individuals cannot be charged a fee for claiming transparency[45] and there are strict timing requirements as well as broader conditions for the way in which transparency is provided (Table 3). The European Data Protection Board (EDPB) has further specified that controllers should actively consider the audience's 'likely level of understanding' when accommodating transparency (e.g. appropriate level of detail, prioritising information, format, etc).[46] This means the controller will need to consider the context of data processing, the product/service experience, device used, nature of interactions, etc.[47] As a result, the information obligation may also differ throughout time.[48] In order to render transparency more meaningful, non-written material (e.g. interactive AV material) may complement the information provision.[49] As explained in Recital 60, information 'may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner, a meaningful overview of the intended processing. Where the icons are presented electronically, they should be machine-readable.' A teleological reading of the GDPR also requires controllers to share personal data in a machine-readable format, unless requiring disproportionate effort (unlikely in the context of online services). At the very least, it is an explicit requirement for responding to data portability requests. Obtaining the respective data in machine-readable format is of course crucial in the context of research purposes. Finally, it is worth keeping in mind that controllers have a duty to facilitate – by 'implementing appropriate technical and organisational measures' – the exercise of data subject rights (Artt.12(2) and 25) and only work with processors who can guarantee doing the same (Art.28), e.g. through standard-setting and/or API's.[50]

TRANSPARENCY AND COMPLEX DATA PROCESSING – As observed elsewhere, we can increasingly locate examples where data subjects' personal data is being processed without the accompanying data

[43] Working Party 29 does however advocate for a broad interpretation, encompassing both 'data actively and knowlingly provided by the data subject' as well as 'observed data provided by the data subject by virtue of the use of the service or the device'. Data such as search histories, browsing/location behaviour, 'raw data' collected through 'mhealth devices' therefore fall within the scope of the right to data portability. ibid 9–11.

[44] For example, this could done by pseudonymisation/anonymisation and the dashboard for selecting data prior to transmission, mentioned above in section 3.3. See also: ibid 12.

[45] This also means that a controller may not require you to be a paying customer as a condition to accommodate your rights. Article 29 Working Party, 'Guidelines on Transparency under Regulation 2016/679' (2018) WP260 13 <http://ec.europa.eu/justice/article-29/documentation>. Previous empirical work has demonstrated that certain controllers effectively only enable access requests filed by people who have an account with the service and/or have bought something with the service before. See: Ausloos and Dewitte (n 24) 12–13.
It is also worth pointing to the fact that 'download my data' functionalities are generally only available to registered users.

[46] See also Recital 60 Article 29 Working Party, 'Guidelines on Transparency under Regulation 2016/679' (n 49) 11.

[47] This may require running (and documenting) trials before 'going live'. ibid 14.

[48] Cf. ibid 16–17.

[49] ibid 12. The EDPB further lists the following examples: cartoons, infographics or flowcharts. Where transparency information is directed at children specifically, controllers should consider what types of measures may be particularly accessible to children (e.g. these might be comics/ cartoons, pictograms, animations, etc. amongst other measures).'

[50] Which, admittedly, is easier said than done. Cf. Norval and others (n 25) 4.

subject rights effectively being enabled.[51] Apple, Twitter and Facebook, for instance, have so far refused to accommodate access requests to certain categories of personal data, claiming that doing so is 'impossible' or 'extremely impractical' because of how they designed their system infrastructure.[52] This worrying trend – especially considering the size and resources available to these companies – is currently undergoing some pushback by investigating data protection authorities across the EU.[53] Indeed, even if the increasing complexity of data processing ecosystems may render it hard to accommodate the core transparency requirements,[54] it does not exonerate controllers. To the contrary, Recital 58 highlights transparency is even more important in complex situations involving many actors.[55] When the controller processes a large quantity of personal data, Recital 63 does permit the controller to request the data subject to specify the information or processing activities to which the request relates. One way to enable effective transparency regarding large data-sets, is to provide remote direct access.[56]

[51] Veale, Binns and Ausloos (n 23).

[52] Concretely, personal requests by the author and colleagues have shown that (a) Apple refuses to give access to Siri voice-data because they consider those recordings not to constitute personal data in the sense of the GDPR; (b) Facebook argues it does not have the 'infrastructure capacity' to enable individuals to access all of the browsing behavior collected by the company; and (c) Twitter claims it would involve disproportionate effort to grant users access to a record of all the hyperlinks they clicked on the platform. See also: Ausloos (n 23).

A partner at the international law firm Fieldfisher recently also claimed that facilitating an access request has cost their clients up to a quarter of a million GBP. Jeroen Terstegge, 'Hazel Grant at #GPS19: "Dealing with a Single Data Subject Access Request May Cost an Organization as Much as a Quarter of a Million British Pounds" #GDPR #DSAR #privacypic.Twitter.Com/E8rxBnhMSN' (*@PrivaSense*, 1 May 2019) <https://twitter.com/PrivaSense/status/1123655137971646467> accessed 2 May 2019.

[53] Irish Data Protection Commission, 'Annual Report 25 May - 31 December 2018' (DPC 2019) <https://www.dataprotection.ie/sites/default/files/uploads/2019-02/DPC%20Annual%20Report%2025%20May%20-%2031%20December%202018.pdf> accessed 28 February 2019.

[54] Rene Mahieu, Joris van Hoboken and Hadi Asghari, 'Responsibility for Data Protection in a Networked World – On the Question of the Controller, "Effective and Complete Protection" and Its Application to Data Access Rights in Europe' (Social Science Research Network 2019) SSRN Scholarly Paper ID 3256743 <https://papers.ssrn.com/abstract=3256743> accessed 4 April 2019.

[55] Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (n 44) 25.

[56] Recital 63 and ibid 17.

Modality	Content	Provision
Fee	Free of charge**, but controllers may charge reasonable fee when:** ⇨ Requests are manifestly unfounded or excessive ⇨ Any further copies of personal data are requested.	**Rec.59; Art. 12(5)** Art. 12(5)a Art. 15(3)
Time limit	**For ex ante transparency, at the moment of obtaining** the personal data, when collected from the data subject **directly**,	Art. 13(1)
	when collected **indirectly**: (a) within **a reasonable period** after obtaining, but **within one month**, considering processing-context; (b) if used for communication with the data subject, at the latest at the time of the first communication; or (c) if a disclosure to another recipient is envisaged, at the latest when the personal data are first disclosed.	Art.14(3)
	For ex post transparency, without undue delay and, in any event **within one month** of receipt of the request (whether the controller intends to take action or not). Period may be **extended** by two further months where necessary, taking into account the complexity and the number of the requests.	Rec.59; Art.12(3)-(4)
Form for exercising	Controllers should provide means for requests to be made **electronically**, especially where personal data are processed by electronic means.	Rec.59
Form for answering	Information shall be provided **in writing**, or by other means, including, where appropriate, by **electronic means**. Where possible, direct remote access to a secure system should be made available. When requested by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means.	Rec.63; Art.12(1)
Intelligibility	In a **concise, transparent, intelligible and easily accessible form**, using **clear and plain language**, in particular for information addressed specifically to a child.	Rec.58; Art.12(1)
Verification of identity	Controllers may request **additional information necessary to confirm the identity** of the data subject, and should use all reasonable measures to do so, in particular in the context of online services and online identifiers.	Rec.64; Art.12(6)
Limitations	Union or Member State **law may restrict** by way of a legislative measure the scope of the obligations and rights provided for in Articles 12 to 22 when such a restriction respects the essence of the fundamental rights and freedoms and is a necessary and proportionate measure in a democratic society to safeguard: see list Art. 23(1)a-j. These measures must contain specific provisions at least, where relevant, as to: see list of Art. 23(2)a-h.	Rec.73; Art.23(1)-(2):
	When **processing is carried out for journalistic purposes or the purpose of academic artistic or literary expression**, Member States shall provide for exemptions or derogations from Chapter III (rights of the data subject) if they are necessary to reconcile the right to the protection of personal data with the freedom of expression and information.	Recital 153; Art.85(2)
	Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in Art. 89(1) in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.	Rec. 156; Art. 89(2)

Table 3 - Transparency Modalities

HURDLES – The GDPR contains several provisions that might effectively raise hurdles to (full) transparency. Article 11(1) explains that controllers do not have to retain personal data *only* for the ability potentially accommodate data subject rights at a later stage. Put differently, the requirement to accommodate data subject rights does not prevent them from anonymising their datasets. Be that as it may, data subjects still have the possibility to provide the controller with additional information so as to (re-)identify their data in anonymised data-sets (Art.11(2)).[57] Article 15(4) further specifies that providing copies of personal data to the data subject should not adversely affect the rights and freedoms of others (e.g. when personal data cannot be dissociated from personal data of another person). Recital 63 also hints at the fact that controllers may oppose accommodating data subject rights when this would impact their trade secrets or intellectual property. Still, there will be a significant burden of proof on the controller's shoulders and this defence cannot result in flat-out denying data subject rights altogether. Finally, it should be said that Article 23 grants Member States the ability to install specific exemptions to the rights of access/portability in their national laws. While most of the situations in which such exemptions can be prescribed relate to specific contexts (Art.23(1))[58] and are subject to conditions (Art.23(2)), there is one catch-all included that makes it hard to anticipate Member State level derogations to access/portability rights. Especially because this catch-all – enabling Member States to restrict data subject rights in order to safeguard 'the rights and freedoms of others' – may be deployed in any kind of national legislations (so not just the GDPR implementation laws). In any case, such exemptions/derogations ought to be formulated and interpreted restrictively/narrowly. Hence, it is fair to say that the default positions should be that data subject rights *are* applicable, *unless* the controller can clearly establish the applicability of a (national) exemption/derogation.[59]

4.2 In Practice

Having gone through the legal details of GDPR transparency measures, it is worth briefly reflecting on how their use as a research method might operate in practice. Several consecutive steps can be identified in preparing a research project:

1. **Aim**. What is your research goal? What purpose are you gathering data for?
2. **Data**. What specific data do you need to achieve said purpose?
3. **Legal Measure**. What GDPR transparency measure is appropriate for obtaining said data (if any)?
4. **Scope**. What does your (ideal) research sample look like?
5. **Recruitment Strategy**. Based on the scope, how to identify and recruit research participants accordingly?
6. **Interaction Strategy**. How will you (and your participants) interact with the respective controllers?
7. **Data Collection Strategy**. How will you actually gather the data you need?
8. **Ethical Review**. Follow standard procedure for having research project pass the ethical review board.

AIM – As with any other research project, it is crucial that researchers first define and scope their actual research goals (i.e. research questions, etc). This will naturally enable answering the question of *what* data is *actually* needed to answer the main research questions. In an SSH context, examples include investigating the functioning and effects of content-personalisation[60] and recommender

[57] In practice, this may lead to a frustrating back-and-forth between data subject and controller, as further detailed in: Veale, Binns and Ausloos (n 23); Ausloos (n 23).

[58] E.g. national security; defence; public security; prevention, investigation, detection or prosecution of criminal offences.

[59] While certainly an interesting and much needed exercise, mapping the different implementations of Article 23 across EU Member States, even when only focusing on GDPR implementation laws, far reaches beyond the scope of this paper.

[60] 'Get Ready for a New Era of Personalized Entertainment' (*TechCrunch*) <http://social.techcrunch.com/2019/04/13/get-ready-for-a-new-era-of-personalized-entertainment/> accessed 15 April 2019.

systems;[61] uncovering the online advertisement ecosystem;[62] identifying (the scale of) political microtargeting;[63] scrutinising the operation of platform economy operators;[64] mapping data flows in mobile app ecosystems;[65] exposing 'dark patterns'[66] or problematic business-models;[67] etc. Rather than focus on phenomena, researchers might also decide to focus on how specific values, norms or rules are affected (by the data economy). For example, by looking at media pluralism and diversity on user-generated-content platforms;[68] (price, gender, race, age, political) discrimination;[69] impact of ride-hailing apps on the environment, city congestion or labour rights;[70] or how online porn, dating apps and social media affect relationships and sex life/health.[71]

DATA – After having decided on the research aim, researchers need to identify what specific data they need. For example, when researching working conditions of Uber or Deliveroo drivers, it might be absolutely necessary to obtain very detailed data on times of idling, driving with/for a client and location.[72] This will allow identifying which controller(s) to approach,[73] what GDPR transparency measures are in fact available and whether there might be other tools that could be helpful. Indeed, oftentimes GDPR measures will not be enough, not in the least because the research questions also require data/information that can only be obtained from participants directly (e.g. subjective opinions or demographic data). Generally, there will be multiple options available and researchers may wish to do a brief SWOT analysis of the different methods, based on their expertise, resources, time, etc. For example, researchers wishing to do a longitudinal study of how (political) ads and content are curated on Facebook, may wish to (also) use GDPR measures in the face of the platform often changing API or its HTML code, exactly in order to block independent research.[74] It might also be useful to use GDPR

[61] World Wide Web Foundation, 'The Invisible Curation of Content: Facebook's News Feed and Our Information Diets' (2018) <http://webfoundation.org/docs/2018/04/WF_InvisibleCurationContent_Screen_AW.pdf> accessed 6 March 2019.

[62] Johnny Ryan/MV's complaints + Privacy International, 'Submission to the Information Commissioner - Request for an Assessment Notice / Complaint of AdTech Data Brokers. Criterio, Quantcast and Tapad (the 'AdTech Data Brokers')' <https://privacyinternational.org/sites/default/files/2018-11/08.11.2018%20Final%20Complaint%20AdTech%20Criteo%2C%20Quantcast%20and%20Tapad.pdf> accessed 8 April 2019; 'Our Complaints against Acxiom, Criteo, Equifax, Experian, Oracle, Quantcast, Tapad' (*Privacy International*) <http://privacyinternational.org/advocacy-briefing/2426/our-complaints-against-acxiom-criteo-equifax-experian-oracle-quantcast-tapad> accessed 8 April 2019; Johnny Ryan, 'Regulatory Complaint Concerning Massive, Web-Wide Data Breach by Google and Other "Ad Tech" Companies under Europe's GDPR' (*Brave Browser*, 12 September 2018) <https://www.brave.com/blog/adtech-data-breach-complaint/> accessed 1 May 2019.

[63] Frederik J Zuiderveen Borgesius and others, 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14 Utrecht Law Review 82.

[64] 'Uber Drivers Demand Their Data' (n 22).

[65] Reuben Binns and others, 'Third Party Tracking in the Mobile Ecosystem' <https://arxiv.org/abs/1804.03603> accessed 24 October 2018; Dieter and others (n 4).

[66] Noam Scheiber, 'How Uber Uses Psychological Tricks to Push Its Drivers' Buttons' *The New York Times* (2 April 2017) <https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html> accessed 25 October 2017; Forbruker Rådet, 'Deceived By Design. How Tech Companies Use Dark Patterns to Discourage Us from Exercising Our Rights to Privacy' (2018) <https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf> accessed 28 June 2018.da

[67] Shoshana Zuboff, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilization' (2015) 30 Journal of Information Technology 75.

[68] Judith Möller and others, 'Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity' [2018] Information, Communication & Society 1; Balázs Bodó and others, 'Interested in Diversity' (2018) 0 Digital Journalism 1.

[69] Joanna Mazur, 'Right to Access Information as a Collective-Based Approach to the GDPR's Right to Explanation in European Law' [2019] Erasmus Law Review 178; and the many examples cited in: Ali and others (n 6).

[70] E.g. by requesting access to detailed temporal data in connection to location data.

[71] E.g. by investigating social media / porn / dating app behaviour of HIV-infected people.

[72] Things which Uber is currently refusing to give in the UK, cf. 'Uber Drivers Demand Their Data' (n 22).

[73] This may actually require some preliminary tests in complex environments, where it might not be straightforward who is controller for what data exactly. Cf.Gürses and van Hoboken (n 3); Dieter and others (n 20) 11.

[74] Jeremy B. Merrill (n 15).

transparency measures to complement and/or verify data gathered via other means.[75] This is also the stage where researchers may wish to think about ethical and legal principles such as 'data minimisation'. For example, in the context of urban planning, researchers may only need historic and/or aggregated location data. The decisions made in this stage may also inform the default settings in a potential dashboard through which research participants interface with their personal data.

LEGAL MEASURE – Once it is more or less clear *what* data researchers need, it is possible to identify the relevant GDPR transparency measures. Generic information may be obtained by relying on Articles 13-14 GDPR, effectively coming down to analysing privacy/data policies (e.g. through automated textual analysis tools),[76] alongside other publicly available documents. However, the real added value of GDPR transparency measures will often be how it enables access to very granular, subject-specific information (from browsing and location behaviour to reading-patterns, energy-consumption, etc), the core provisions being Artt.15 and 20. Given its emphasis on providing the data in a 'structured, commonly used and machine-readable format', the right to data portability (Art.20) will generally be the most interesting for both qualitative and quantitative researchers wishing to run automated analysis tools. Yet, because of its narrow scope (cf. Section 4.1.), the right of access (Art.15) might be the second-best option. Moreover, the right of access does not just grant access to (any and all) personal data itself, but also to important metadata which might be of equal research-interest (e.g. retention periods, source and recipients, etc.).

SCOPE – The next step will be to identify the research sample. Identifying a representative research sample is a requirement in any data-driven research project and will thus not be further discussed here. What can be said though, is that depending on the research aim and required data, it may not be necessary to rely on research participants at all. For example, testing GDPR compliance with data subject rights, can be conducted by just one person (which of course, may fail to account for potential differences in accommodating such rights depending on personal characteristics of the respective data subject).

RECRUITMENT STRATEGY – Assuming multiple participants are required, there are many different ways in which they can be recruited. Research aim and required data will determine the representativeness criteria. The desired scale, both in depth (level of detail) and breadth (number), of the data will also play a significant role in researchers' recruitment strategy. Resources and/or technical know-how – e.g. to automate (part of the) requests or develop a dashboard through which participants can select/remove information beforehand – will further influence the scalability. If the representativeness does not depend on the personal characteristics of research participants, it is possible to work with students[77] or – more informally – colleagues, friends and family.[78] If personal attributes *are* relevant (e.g. uncovering discriminatory practices), conventional sampling rules apply. Regardless, researchers will have to consider participants' skill-set when crowd-sourcing data-gathering using the GDPR. This can be dealt with to some extent, by providing explanations and/or personal/technical assistance and tools.[79] The eventual sample of research participants (and their anticipated skillsets) will also inform the data collection strategy.

INTERACTION STRATEGY – Using GDPR transparency measures for obtaining research data necessarily implies data/information is obtained from a third party, often a company. As mentioned before, the interaction with those companies can take many different forms along a spectrum of collaborative to

[75] E.g. Tobias Urban and others, 'The Unwanted Sharing Economy: An Analysis of Cookie Syncing and User Transparency under GDPR' [2018] arXiv:1811.08660 [cs] <http://arxiv.org/abs/1811.08660> accessed 10 April 2019.

[76] See for example: Marco Lippi and others, 'CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service' [2018] arXiv:1805.01217 [cs] <http://arxiv.org/abs/1805.01217> accessed 29 August 2018.

[77] Ausloos and Dewitte (n 24).

[78] *Mahieu et al.?*

[79] The author has done several relevant projects with students, interacting with them in different ways. Unsurprisingly, the project with a limited number of students (3), with bi-weekly follow-up calls, was the most successful. Cf. Ausloos and Dewitte (n 24).

adversarial. While the former might often be most productive, it raises a whole number of other issues (privileged access, biases, etc.). Adversarial approaches may be more appropriate in some cases, but might require a lot more active involvement and/or even taking legal action (e.g. complaint with the DPA[80] or legal action[81]). In between these two extremes, there are a plethora of possible options. On the more collaborative side of the spectrum, one could imagine company and researchers agreeing to include a specific tag in participants' access requests so that they are prioritised and/or responded to in a predefined format. Researchers may also simply rely on available 'download my data' functionalities already offered by many online services. On the more adversarial side of the spectrum, finally, researchers may try different strategies in pressuring companies to share data in specific formats and/or challenge 'inability to identify' defences.[82]

DATA COLLECTION STRATEGY – After all of the previous steps, it is possible to determine the most adequate data collection strategy. This will most importantly be determined by what data is in fact needed, but also by the legal measure relied on, participants' capabilities, interaction strategy and the available resources. In some cases, researchers may wish to gain access to the 'raw data' set that is returned to participants. It is fair to assume that researchers may wish to receive this data in a 'structured, commonly used and machine-readable format' (cf. right to data portability, Art.20), enabling automated analysis.[83] Sometimes, additional information may be needed, such as attitudes towards responses or demographical data not previously known. Such information may be collected through surveys or interviews. It is worth referring once more to the added value of having a research participants' dashboard of some sort. This would facilitate following up, explaining and empowering participants more comprehensively, as well as make the process more scalable overall. In any case, it will generally be advisable for researchers to already test the waters – especially in more adversarial contexts – on what data can realistically be expected back in return to an access/portability request. Indeed, such preliminary tests, may in fact bring to light previously unknown data flows and/or controllers, which in turn may result in having to readjust the interaction strategy.[84] And finally, as with any form of transparency – particularly in algorithmic environments - researchers should remain diligent and aware of interpretative limitations of whatever data/information is gathered through GDPR measures.[85]

ETHICAL REVIEW – The last step before actually moving on to *actually* conducting the research, is to obtain approval by the research institution's (ethical) review board. Notably in the context at hand, researchers will need to pursue the data minimisation principle and only gather the (personal) data they *actually* need for their stated research purposes. Having said that, this review process is no different than other research projects and will therefore not be elaborated upon here.

4.3 Requirements

Drawing from the above, it is possible to make some general conclusions as to key requirements for effectively using GDPR transparency measures as a valid method for gathering research data. Three (categories of) interacting requirements can be identified: (a) norms and ethics; (b) technical constraints; and (c) expertise.

[80] E.g. Ryan (n 66); 'Our Complaints against Acxiom, Criteo, Equifax, Experian, Oracle, Quantcast, Tapad' (n 66).

[81] E.g. 'Uber Drivers Demand Their Data' (n 22).

[82] Cf. work Paul-Olivier Dehaye.

[83] Faced with a wide diversity in scope and format of responses to access requests, researchers have taken different approaches in interpreting them. Mahieu, Asghari and Eeten (n 27) have analysed responses themselves, whereas Ausloos and Dewitte (n 22) had developed surveys with all relevant parameters and had regular interactions with participants in order to ensure proper understanding and completion of the detailed surveys.

[84] Especially in an increasingly modular software environment, the relevant controller may not necessarily be the one thought of instinctively (e.g. front-end service provider). Cf. Gürses and van Hoboken (n 5); Dieter and others (n 4).

[85] Cf. Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973.

NORMS AND ETHICS – Given the relative novelty of this approach, many ethical and legal questions still remain to be answered. These will depend to a large extent on the actual context of the respective research project, which may often involve different disciplines and practices each with their own ethics and normative frameworks. Researchers need to reflect on how these different ethics and frameworks (might) interact and align them accordingly.[86] To the extent the information/data collected includes *personal* data (e.g. when direct access is given to access request responses), the GDPR will of course be a critical legal framework also *regulating* the research itself. The practice of working with artificial personas may solve some issues stemming from working with real-life research participants, but would raise significant legal questions in the present context (notably because the GDPR only applies to natural persons). When working with actual research participants, researchers will also have to be conscious of the implications the research may have on the relationship between controller and participants. Concerns may arise when these relationships precede the research and/or the research takes an adversarial approach.

TECHNOLOGICAL CONSTRAINTS – The effectiveness of using GDPR transparency measures as tools for gathering research data depends on technical constraints. While this may seem obvious in the abstract, it does require close attention during the research design phase. Is it technically/organisationally possible to obtain the desired data from the respective controller in the first place? The extent of GDPR transparency measures might be constrained by a proportionality assessment. The right to have personal data transmitted directly from one controller to another (i.e. right to data portability, Art.20) may be refused when not *technically feasible*. The provision on data protection by design/default (Art.25) also recognises that controllers' technical and organisational measures for accommodating data subject rights may depend on 'the state of the art, the cost of implementation and the nature, scope, context and purposes of processing, as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing'. In practice the standard of what can be considered 'technically feasible' will effectively be set by the respective controllers. Apart from very obvious cases, it will be hard for researchers/data subjects to establish accommodating data subject rights *is* in fact technically feasible. In the same vein, controllers may further weaken the potential of GDPR transparency measures as research methods, by frequently altering relevant parameters (e.g. format, content, location/URL, API, etc).[87] Such practices may equally result in a back-and-forth on whether the data/information is in fact 'concise, transparent, intelligible and easily accessible form, using clear and plain language [...] provided in writing, or by other means, including, where appropriate, by electronic means' (Art.12(1)). In sum, even if theoretically the desired data plainly falls within the scope of data subject rights (access/portability), controllers may raise significant practical/legal obstacles that erode the value in using those rights as research methods in the first place. Finally, researchers should also be aware that controllers may delay response-time with an additional two months (so a maximum of three months between request and answer in total) if they can establish more time is needed to comply in light of the complexity of the request (Art.12(3)). In conclusion, and as mentioned before, good practice dictates researchers to already experiment and explore with GDPR transparency measures in order to anticipate their technical-legal constraints in practice.

TRIANGLE OF EXPERTISE – In order to devise an adequate research design adopting GDPR transparency measures, three types of expertise are needed: domain, technical and legal expertise. The exact mix of expertise required will strong vary depending on the specifics of each situation (just like the available expertise may influence the research design). Put very briefly, *domain expertise* refers to the actual core discipline(s) in which the research aims can be situated (e.g. political communication; environmental law; psychology; etc.). Secondly, and as apparent from the previous paragraph, *technical expertise* might be more or less critical both in devising the research set-up, as well as

[86] ibid 14.

[87] Not a hypothetical concern, as recently illustrated by Facebook's actions against independent research of its ad practices. Jeremy B. Merrill (n 15); Digital, Culture, Media and Sport Committee (n 15) 64.

interpreting the data/information as it is being gathered (potentially necessitating small adjustments to be made in the research design along the way). Thirdly, a minimum level of *legal expertise* in the scope, requirements and operationalisation of GDPR transparency measures will evidently be necessary as well in order to adopt them in a research design. When considering this research method, it will be critical to map the available expertise along this triangle to identify potential knowledge gaps and ensure feasibility of the project overall.

5 Positioning Data Subject Rights as a Research Method / Discussion

The purpose of this section is to take a step back and make some broader reflections on using GDPR transparency measures as research methods and how they can be capitalised on in the most productive manner.

It is also worth mentioning that the potential value of using data subject rights in an academic context extends beyond their ability to gather useful research data. They also have didactic value, as suggested by the positive experiences of several teachers across Belgian and Dutch universities who have done class exercises with these rights (in different disciplines). Moreover, there is also some potential for stimulating and enabling better compliance among the targeted actors. Anecdotal evidence in this author's projects, but also that of colleagues, suggests that certain companies do in fact change certain practices following access requests. This is not to overstate the (potential) value of data subject rights.

[...]

5.1 Complement, not replace

It is important to nuance the value of GDPR transparency measures against the backdrop of a growing field of more or less useful (interdisciplinary) methods for gathering research data. The 'black box society', often characterised by 'infrastructural resistance that might circumvent or side-track empirical research,'[88] necessitates multiple different methods for achieving one's research goals. Not just as a redundancy measure, but also to compare, verify and explore different perspectives. For example, access/portability rights could be used in order to enrich a diary study.

In most cases, GDPR transparency measures will generally best serve the research goal as a *complementary* research method. Especially because they only recently appear to become more widely appreciated as a valid research method. This will allow to gradually iron out some of the misunderstandings and uncertainties currently still inherent to these measures.

Of course, GDPR transparency measures are not immune to existing concerns over research data and transparency more broadly (e.g. intentional obfuscation,[89] perpetuate power/resourcefulness-dynamics,[90] seeing does not imply understanding,[91], etc.).

[88] Dieter and others (n 4) 12.

[89] Ananny and Crawford (n 89).

[90] Barbara Prainsack, 'Logged out: Ownership, Exclusion and Public Value in the Digital Data and Information Commons' (2019) 6 Big Data & Society 2053951719829773.

[91] Ananny and Crawford (n 89); J Joost Beuving, 'Ethnography's Future in the Big Data Era' (2019) 0 Information, Communication & Society 1, 11.

5.2 Capitalising on centralising forces in the data economy

How do the GDPR transparency measures relate to the centralisation of information and power, apparently inherent to (online) data ecosystems? Power asymmetries instinctively seem to thwart meaningful empowerment of the weaker party. Indeed, data protection law in particular has long been criticised for being incapable of pushing back on data-driven power asymmetries.[92] Having said that, when approaching GDPR transparency measures as a research method, centralisation may actually be a valuable feature rather than a bug. It facilitates the process by reducing the number of actors to be contacted and might even enable gaining insights into things that would otherwise be impossible to uncover, at least in theory. For example, the fact that a large portion of internet browsing behaviour is monitored by Google and Facebook – problematic as that may be in itself – does enable researchers to simply target one or two entities to get detailed insights into surfing behaviour, combined with detailed demographic data, media consumption and social interactions. Similarly, a lot can be learned about specific sectors or communities by simply focusing on the corresponding dominant platforms (e.g. hotel/flight booking websites, to second-hand marketplaces, UGC content, etc.).

In fact, powerful actors sitting on central nodes may even be considered to have a responsibility to ensure or at least facilitate (data)transparency of all actors operating on their platform. The clearest example perhaps being mobile app-stores (Apple and Google Play in particular).[93] These mobile platforms may also be crucial in making or breaking access/portability requests to specific app developers directly (i.e. when these actors cannot verify the identity of the data subject without receiving confirmation of specific device identifiers).[94]

One could also think of researchers 'centralising' know-how acquired through their respective projects. For example, by creating an open access repository that lists information relevant to GDPR transparency measures, in a structured (and potentially multi-language) manner. An illustration of how this might operate in practice is wikidata.org, where the page on Uber lists all of the confirmed personal data they have, where to send access requests to, etc.). Such a platform facilitates automating (parts of) the research design, as well as anticipate potential hurdles along the road.

5.3 Best Practices

Clear/concise list with recommendations... . Perhaps as an annex? Flowchart?

[92] Indeed, this was one of the main drivers behind the GDPR. European Commission (n 27).

[93] Ronan Fahy, Joris van Hoboken and Nico van Eijk, 'Data Privacy, Transparency and the Data-Driven Transformation of Games to Services' 11.

[94] Norval and others (n 25) 4–5.