Researchers' access to information from regulated services

Report prepared under section 162 of the Online Safety Act 2023 and submitted to the Secretary of State in accordance with section 162(5) of that Act.

Office of Communications

Researchers' access to information from regulated services

Presented to Parliament pursuant to section 162(5) of the Online Safety Act 2023.

July 2025

This publication is available at www.gov.uk/official-documents.

Any enquiries regarding this publication should be sent to us at Researcheraccess@ofcom.org.uk.

ISBN 978-1-5286-5878-2

E03395491 07/25

Printed on paper containing 40% recycled fibre content minimum

Printed in the UK by HH Associates Ltd. on behalf of the Controller of His Majesty's Stationery Office

Contents

Section

Annex

1. Ofcom's duties and relationships

1.1 Ofcom is the UK's communications regulator, overseeing sectors including telecommunications, post, broadcast TV and radio, and online services. Ofcom was appointed the online safety regulator under the Online Safety Act in October 2023. The Online Safety Act 2023 (the "Online Safety Act") makes providers of certain online services – including social media, search, and pornography services – legally responsible for keeping people, especially children, safe online.

1.2 This report has been produced in accordance with section 162 of the Online Safety Act, which sets out that Ofcom should publish a report that:

a) describes how, and to what extent, persons carrying out independent research into online safety matters are currently able to obtain information from providers of regulated services to inform their research;
b) explores the legal and other issues which currently constrain the sharing of information for such purposes; and
c) assesses the extent to which greater access to information for such purposes might be achieved.

1.3 In parallel to our report, the UK Parliament has enacted legislation that allows the UK Government to create, if it chooses to do so, a new framework to enable researchers to access information regarding online safety matters held by regulated services.[1] It has indicated that Ofcom's report will provide an evidence base to inform the design of any future access framework supporting research into online safety matters.[2]

1.4 Ofcom currently engages with the two parties central to data access for research purposes – researchers and services – which are discussed frequently in this report. Here, we clarify these relationships and disclose how we work with both parties. See Section 3 of the report for our definitions of these parties.

1.5 Ofcom engages with researchers and institutions carrying out research in various ways, providing opportunities to exchange ideas and help us fulfil our role as an evidence-based regulator. We commission work from researchers, procure data from researching institutions, use external research to guide our online safety agenda, and attend research conferences. In April 2024, we published our Online Safety Research Agenda, outlining areas for future research in the online safety space and inviting engagement and collaboration with the wider community to further shared research goals. Ofcom also occasionally hosts seconded academic researchers and provides in-kind support to researchers for grants (such as letters of support). External research forms part of the evidence base that informs Ofcom's regulatory work.

[1] The Data (Use and Access) Act 2025 received Royal Assent on 19 June 2025.
[2] UK Parliament, 2024, Data (Use and Access) Bill Explanatory Notes.

1.6 Ofcom also engages with services in the course of its regulatory duties. We engage through formal methods, such as information requests, and have also established a supervision team to lead targeted oversight of a range of in-scope services. This targeted oversight includes understanding services' measures in detail, assessing how well they protect users, and pushing for timely improvements where necessary.

2. Overview

2.1 Access to high-quality information about the digital ecosystem is vital for empowering people with greater knowledge about online safety matters. This information can provide important insights about how harms manifest (both online and offline), enable users to make informed decisions about their digital habits as well as support researchers and policymakers in assessing the efficacy of measures intended to mitigate those harms.

2.2 Researchers can also act as a powerful mediator between services and users by highlighting online safety issues in the public interest and of societal importance. However, researchers face increasing barriers to accessing information held by services. Recent years have seen restrictions on information access and consequently a growing information asymmetry between researchers and services has emerged. This has led to reduced availability of data access tools including application programming interface (APIs) and public insight tools, increased legal risk for activities like scraping, and a low-trust environment between researchers and services.

2.3 This has reduced the scope for independent research and evidence based on service data. This undermines our collective ability to grasp the dynamics of our digital environment, potentially obscuring risks and allowing harms to go unmitigated. Other sectors have addressed similar challenges through clearer standards, independent oversight, and regulatory mandates. Comparable interventions may be needed in online safety to enable safe, secure, privacy-protecting and meaningful access to information for the purposes of research into online safety matters.

2.4 Different researchers across academia and civil society may have varying information needs. Some work with large-scale historical data, while others seek real-time access to data. Current access models often fail to accommodate this diversity. While some services highlight voluntary transparency efforts, and compliance with emerging legal requirements including the Online Safety Act, researchers remain concerned about the lack of timeliness, consistency, and specificity of information access. Researchers also argue that access to individual-level data is essential to understanding online harms and safety, introducing privacy, data protection, and business risk challenges. Both researchers and services face infrastructure and resource constraints. Smaller organisations, in particular, risk exclusion from access schemes with high barriers to entry.

2.5 We have sought views from a diverse range of stakeholders in preparing this report. We requested evidence of how researchers have overcome information-sharing constraints in sectors beyond online safety, where such cases might illustrate effective data governance or data-sharing mechanisms. This is discussed in more detail in Section 3 ‘About this report’.

2.6 This report outlines three potential policy options that can facilitate greater researcher access to information about online safety matters, which the UK Government may consider as part of the design of any future access framework.

Achieving greater researcher access

1. Clarify existing legal rules

Relevant authorities could provide additional guidance on what is already legally permitted for researcher access on important issues, such as data donations and research-related scraping.

This could help researchers understand what data they can collect and aids services in defining their sharing obligations, which could promote more consistent practices and potentially increase access to individual-level data. However, it places the technical and financial burden of data collection on researchers, potentially disadvantaging those with fewer resources or technical skills and creating risk of misinterpretation. Additionally, since access would remain limited to public and donated data, the diversity and depth of research outcomes may be constrained, and significant expansion of access to non-public data[3] remains unlikely.

2. Create new duties, enforced by a backstop regulator

Services could be required to put in place systems and processes to operationalise data access. This could take the form of a direct access model, which could include standardised service-led procedures for researcher accreditation and data handling, with the backstop regulator responsible for enforcement. Services themselves would be responsible for the tasks of accrediting researchers, providing researchers with data directly or providing the interface through which they can access it and offering appeal and redress mechanisms. A new regulatory body would need to be established, or an existing organisation would need to be given new powers.

A direct access model could reduce security and legal risk by offering a formal mechanism for researchers to access information. In theory, this approach could have limited administrative costs as it would not require an intermediary to perform complex accreditation processes and enable timely access to mandated data types or categories without requiring tailored datasets. However, its effectiveness depends on what data is in scope. The model could require significant technical investment and offer limited research flexibility, as researchers have no input into data selection, timing or structure. The likely limited scope in data could limit research depth and diversity, and the absence of an intermediary could weaken dispute resolution mechanisms which may be relied upon in the absence of intermediary-led accreditation.

3. Enable and manage access via independent intermediary

New legal powers could be granted to a trusted third party which would facilitate and manage researchers' access to data. This could be done through facilitators who vet researchers and provide secure access (for example through 'clean rooms' or application programming interfaces (APIs)), through researchers submitting requests directly to an intermediary which then liaises with services, or through intermediaries which host or manage data access services (either locally or virtually).

There are three ways an independent intermediary could operate – direct access intermediary, notice to service intermediary and repository intermediary models.

[3] In the context of this report, public data refers to information that is readily accessible to the general public and does not require special permissions or authorisation to access. Private data refers to information that is not readily accessible to the public and requires special permissions or authorisation to access. Special category data, often referred to as sensitive data in this report, can be public or private and is subject to data protection laws due to its personal nature.

Direct access intermediary. Researchers could request data with an intermediary facilitating secure access. Services could retain responsibility for hosting and providing data while the intermediary could create an interface by which researchers could request access. This process and the data in scope would be similar to a direct access model but with an intermediary that could set and enforce eligibility criteria and accredit researchers. The intermediary could act as a mediator if disputes arise.

Notice to service intermediary. Researchers could apply for accreditation and request access to specific datasets via intermediaries. Notably, this access request could involve public data, private data and/or special category data (which can be public or private). Services could retain responsibility for hosting and providing data. Proposals could include the scope of research, such as requests for tailored data held by a service, and the preferred data access modality. The intermediary and the relevant service could work together to assess proposals and offer alternatives should original requests be denied. The intermediary could act as a mediator if disputes arise.

Repository intermediary. The intermediary could itself provide direct access to data, including taking responsibility for data governance and providing an interface for data access and/or hosting the data. This could include maintaining an interface for data access and/or hosting the data directly. The intermediary would facilitate access to the relevant and appropriate data, which could include data that would not be accessible in direct access models. This model could take the form of a virtual or local repository. In a virtual repository model, an intermediary would be responsible for an interface that facilitates access to data hosted by services (e.g. via a portal to access APIs). In a local repository model, an intermediary would be responsible for a repository in a centralised, physical location (e.g. a data centre).

Despite variations in design, at a high level, intermediary models share some common strengths and challenges. Intermediary models could help overcome technical and non-technical barriers to data access by providing clear rules and processes, managing accreditation processes, and mediating disputes between researchers and services. They could help foster trust between stakeholders and unlock secure and controlled access to private and sensitive data. Intermediaries could support more diverse research methods and encourage strategic research coordination among researchers, easing the burden on services. However, implementing these models also gives rise to significant cost, expertise, data protection, and other complexity challenges. It is unclear whether any existing authority has the required competencies to perform this intermediary role effectively and/or whether a new body would need to be established.

For further information on these policy options and models, see Section 6.

2.7 Through our analysis, it has become evident that no single model is likely to meet the full range of researcher needs. A layered, flexible approach – combining legal clarity, technical safeguards, and independent oversight – offers the best chance of enabling responsible, timely and useful information access. The policy options (clarifying existing rules, creating new duties enforced by a backstop regulator, and enabling and managing access via an independent intermediary) and models within them (direct access, direct access intermediary, notice to service intermediary and repository models) presented in this report do not need to be considered in isolation and could be regarded as complementary.

2.8 Elements from different models, combined with enabling measures, may present more effective means of facilitating researcher access depending on policy objectives. Where needed, responsibilities for management of a researcher access regime could be shared between technical and governance-focused bodies to manage complexity, support trust-building, and reduce burdens on any single institution. Meaningful data access in support of making people safer will also require a shift in culture – one built on trust, transparency, and a shared commitment to ethical standards.

3. About this report

Evidence-gathering process

3.1 To prepare this report, we gathered evidence through a public call for evidence, private roundtables, working groups, and other bilaterial stakeholder engagements convened by Ofcom, partner organisations, and scientific bodies.

3.2 Between October and November 2024, we held four in-person and two virtual roundtables across the UK. We discussed topics related to researcher access with researchers and researching institutions working on and beyond the field of online safety. In January 2025, we held an additional virtual roundtable to engage with international researchers and researching institutions working specifically on online safety.

3.3 Our Call for Evidence ran from 28 October 2024 to 17 January 2025. We asked for evidence and input on the following:

- how and to what extent independent researchers currently access information from providers of regulated services;
- the challenges that currently constrain information-sharing for these purposes; and
- how greater access to this information might be achieved.

3.4 When we published the Call for Evidence, we reached out directly to the stakeholders named in the Online Safety Act, as well as others whose input we considered particularly relevant, to ensure they were aware of the opportunity to contribute. We received 39 responses.

3.5 Stakeholders consulted during our evidence-gathering process included:

- **Academic researchers and research organisations** representing institutions with expertise in areas such as online harms, functionalities, digital safety, and service governance, including but not limited to the Minderoo Centre for Technology & Democracy, the Knight-Georgetown Institute, and the NYU Center for Social Media and Politics.
- **Civil society organisations with expertise in online harms** representing the interests of children, vulnerable users, and broader human rights, including organisations such as the Ada Lovelace Institute, the Atlantic Council's Digital Forensic Research Lab, and the Institute for Strategic Dialogue.
- **Safety technology organisations** providing expertise in online harms and perspectives on industry-led risk detection and mitigation tools.
- **Research intermediaries and data providers** whose contributions cover access to relevant data and methodological limitations.
- **Services** in scope of the Online Safety Act, reflecting a range of user-to-user and search services, functionalities, risk profiles, and compliance duties, including but not limited to Meta Platforms Inc. and Reddit.

- **Government departments and domestic and international public sector bodies**[4] including the Department for Science, Innovation and Technology (DSIT), the Information Commissioner's Office (ICO), the Centre for Data Ethics and Innovation (CDEI),[5] UK Research and Innovation (UKRI), and others with statutory or strategic relevance to online safety, including Digital Service Coordinators in the European Union.

3.6 To respect the confidentiality of certain stakeholders, particularly those who contributed in a personal capacity or shared sensitive or unpublished material, this report does not list all consultees. Their input nonetheless informed our analysis and contributed to the development of our evidence base.

3.7 To supplement this evidence, we have further relied on external research outputs produced by organisations such as the Ada Lovelace Institute, the Mozilla Foundation, the Open Data Institute, the European Digital Media Observatory, and the Royal Society. Some these reports have helped outline the case studies that we present in Section 6.

3.8 Our broad engagement and research ensured that the evidence base underpinning this report is comprehensive and balanced, in accordance with our duties outlined under Section 162 of the Online Safety Act.

Definitions as stated in the Online Safety Act 2023

3.9 There are several concepts that we reference throughout this report. This section provides definitions of these primary concepts to ensure consistency and clarity of analysis.

Independent research

3.10 Section 162(2) of the Online Safety Act sets out a definition of "independent research" which we are using for the purposes of this report. It reads as follows: "a person carries out 'independent research' if they carry out research on behalf of a person other than a provider of a regulated service".

Online safety matters

3.11 Section 235(4) of the Online Safety Act defines "online safety matters" as the matters to which Ofcom's online safety functions relate. Under this section, Ofcom's online safety functions are defined as those under the Online Safety Act and a small number of relevant regulatory functions under the Communications Act 2003.

Regulated service

3.12 Section 4(4) of the Online Safety Act defines a "regulated service" as a regulated user-to-user service, a regulated search service, or an internet service (other than a regulated user-

[4] Some of which we were required to consult with under section 162(3) of the Online Safety Act.
[5] The CDEI became the Responsible Technology Adoption Unit (RTA) in February 2024. As of January 2025, the RTA is no longer a standalone unit and its functions are embedded in other relevant teams in the Department for Science, Innovation and Technology (DSIT). We have engaged with stakeholders who have previously worked at the RTA and we consider this to have satisfied the original requirement set out in the Online Safety Act.

to-user service or a regulated search service) that is within section 80(2) of the Online Safety Act (including a service of a kind described in Schedule 2).

Data and information

3.13 Section 162 of the Online Safety Act refers to 'researchers' access to information" in relation to this report. However, most stakeholders tend to use 'access to data' when discussing researcher access. Similarly, the comparable provision in Article 40 of the Digital Services Act 2022 (the "Digital Services Act") refers to "data access and scrutiny."

3.14 We are aware that in knowledge management literature, 'information' and 'data' are considered separate concepts. In the 'informational hierarchy', data is the raw ingredient of information while information is 'data with meaning'.[6]

3.15 Given the use of both terms in the literature to refer to the concept of researchers having access to resources from which they can derive insights and recognising that 'access' can encompass both direct data access and the provision of information derived from data requests (without the underlying data), 'access to data' and 'access to information' should be understood as having the same meaning for the purposes of this report.

Independent research

Research and researchers

3.16 Through our engagement, we have identified a range of concerns regarding the definitions of 'research', 'researcher', and 'independent research'. Stakeholders have highlighted that narrow definitions can inadvertently preclude researchers and organisations from conducting valuable work in the public interest, particularly those operating outside academic or institutional frameworks.

3.17 In preparing this report, we engaged with researchers from a wide range of disciplines and organisations, including academic institutions, civil society organisations, fact-checking organisations, think tanks, and safety tech organisations. While these groups may differ in focus and structure, they share a common goal: to better understand and reduce online harms and improve the safety of digital environments. Despite operating in a variety of different ways, most researchers are committed to evidence-based approaches and public interest outcomes.

Independence

3.18 Independence underpins the credibility, integrity, and value of research. It ensures freedom from influence and allows findings to be guided by evidence rather than agenda. It also ensures that data is accessed ethically and not misused in ways that could threaten services' intellectual property. However, stakeholders have highlighted that overly rigid criteria around independence risk precluding researchers conducting valuable work in the public interest.

[6] Forster, M., 2015. Refining the definition of information literacy: the experience of contextual knowledge creation. Journal of Information Literacy, 9(1). http://dx.doi.org/10.11645/9.1.1981. [accessed 23 June 2025]

3.19 Independence from commercial interests is often emphasised; funding models across the research landscape problematise this criterion. Academic researchers generally depend on research council funding, government grants, and institutional support. Non-academic researchers, such as those affiliated with think tanks and civil society, draw on a wider array of funding sources, including government contracts, foundation grants, academic collaborations, consultancy income, and crowdfunding. Academic and non-academic researchers involved in public interest work also receive funding or engage in partnerships with services or other commercial actors. This reflects the practical realities of financing and delivery. A shrinking funding landscape has further driven the adoption of hybrid funding models as a means of sustaining research.

3.20 A robust research environment supports a plurality of perspectives and safeguards researchers' autonomy. To that end, the concept of independence must similarly be understood as independence from governmental or other institutional influences.

3.21 These nuances illustrate the need for careful consideration of researcher eligibility criteria and raise important questions about how such criteria are interpreted within regulatory regimes. There is a need to ensure that public interest researchers are not inadvertently precluded from being recognised as legitimate contributors to the research ecosystem.

Differing data needs

3.22 While similar in many ways, researchers' needs are not one and the same. Researchers' methods and outputs vary, ranging from large scale quantitative analysis and qualitative interviews to policy evaluation and technical testing. Academic researchers often work on long-term projects and publish in peer-reviewed journals. Non-academic researchers typically produce more immediate materials such as reports, blogs, or media articles, which can be less formal and can provide timely insights and practical recommendations.

3.23 These differing objectives and methods require different types of data. The way data is made available, whether through institutional data access models (such as partnerships or regulatory frameworks) or technical data access modalities (such as APIs or clean rooms), fundamentally dictates what kind of research can be conducted. Ensuring balanced and fit-for-purpose access is not just a matter of fairness – it is essential for enabling a diverse and robust research ecosystem. We have taken this into account in the analysis presented throughout this report.

Report structure

3.24 Section 3 outlined our methodology for undertaking this report and the evidence-gathering process. Section 4 presents the current access methods for independent researchers, followed by Section 5 which highlights the constraints stakeholders face in enabling data access to study online safety matters. Section 6 of the report presents policy options that may enable greater access for researchers, including by helping clarify routes currently available, or by introducing new avenues for access. The Annexes includes additional information and analysis in support of the report.

4. Current access to data on online safety matters

Introduction to access methods available to researchers

4.1 There are a range of different methods and tools that currently enable independent researchers to access and collect data and information from services and about users of services, even in the absence of a regulatory regime for data access. These include but are not limited to APIs, voluntary service-researcher partnerships, ad and content libraries, web scraping, transparency reports, data donations, data purchasing, and social listening tools.

DATA ACCESSED DIRECTLY FROM SERVICES
Application programming interfaces (APIs)
Voluntary research partnerships
Ad libraries
Transparency reports
DATA ACCESSED INDIRECTLY FROM SERIVCES
Data scraping
Avatar research
Purchasing data and access from commercial entities
Research forums and consortia
DATA ACCESSED DIRECTLY FROM USERS
Data donations
Voluntary widgets
Data trusts
Data cooperatives

4.2 Researchers generally characterise the current level of access to information as inadequate, referencing several modalities of access that have been scaled back in recent years. Many services have moved away from offering at-scale, easy, free API access. Some APIs have become more expensive, such as that offered by X (formerly known as Twitter), while others have been replaced by complex data-sharing agreements that create significant barriers and lead to unbalanced access to data. Researchers have shared that these data-sharing agreements are often limited to researchers with academic affiliations or those with

sufficient legal resources to navigate complex contractual agreements with services. Other common access modalities, such as CrowdTangle, have been terminated indefinitely.

4.3 To provide a comprehensive overview of current data access practices, we have grouped tools and methodological approaches[7] into three categories: data accessed directly from services, data accessed indirectly from services, and data provided directly from users. To accurately reflect the evolving landscape, we found it necessary to include indirect access methods alongside those explicitly offered by services. See Section 5 and Annex 2 of this report for further information regarding issues constraining researcher access.

Data accessed directly from services

DATA ACCESSED DIRECTLY FROM SERVICES
Application programming interfaces (APIs)
Voluntary research partnerships
Ad libraries
Transparency reports

Application programming interfaces (APIs)

4.4 Historically, many researchers have accessed data via APIs, which are computing interfaces that allow for automated interaction between different software, including the ability to retrieve data in a scaled and standardised way. This data may include metrics, text, and media (such as image or video and associated metadata). The service establishes and maintains the API infrastructure and takes decisions on granting researchers access. The service is in control of what data is made available for download via the API. Once a researcher has access to an API, they can usually query the database using a programming language without having to submit further requests for API access, allowing them greater flexibility. APIs can enable researchers to quickly test hypotheses and conduct exploratory research. Automated data collection and processing can facilitate scaled research and democratise access to data by making datasets accessible to a broader range of researchers. In some instances, APIs provide access to real-time or near real-time data and can be tailored to specific researcher needs, supporting innovative studies and timely responses to developments on a service.

4.5 It is relevant to note that the levels of access and fees for APIs can vary:

- An **'open' API** is the most accessible, as it is open to anyone and does not charge fees. Open APIs allow researchers to make requests (so-called 'API calls') to a service with an automatic response and immediate ability to download the data. The data made available is usually aggregated and contains large volumes of data points, making this data access modality particularly useful for quantitative studies and natural language

[7] In this report, we focus on researchers' access to information for the specific purposes of conducting research into online safety matters. While some of the tools and methods we share in the report may be relied on in the context of topics or themes that go beyond what may strictly be regarded as 'online safety matters', we consider them relevant to our analysis.

processing (NLP) research. However, many open APIs are designed for purposes other than enabling researcher access, such as supporting software development, meaning they may not contain the kinds of data needed for quality online safety research.

- A **'public' API** is accessible to the public but requires some level of authorisation or approval before it can be accessed. Public APIs may charge for access or they may use a 'freemium' fee structure, where basic functionalities are available for free. Additional features, such as higher volumes of data or higher quotas of API calls, come with a fee.
- A **'partner' API** is only made available to approved parties and is not open to the public. A partner API typically has more safeguards in place to ensure the data protection and security standards are met and may provide access to more sensitive data or data tailored in response to particular research questions or needs. A partner API may also charge for access or use a freemium fee structure.

Voluntary research partnerships

4.6 Voluntary research partnerships are collaborative arrangements where services proactively grant researchers or research institutions privileged access to internal data for the purposes of a pre-defined research project or to jointly establish and deliver data-sharing mechanisms, without being legally mandated to do so.

4.7 Voluntary partnerships can grant researchers access to large volumes of data they would otherwise be unable to access. Depending on the level of collaboration, they may offer access to services' in-house expertise to understand the data. Voluntary research partnerships are typically well-suited for academic-oriented or longer-term research projects where sustained collaboration and mutual interest can be developed over time. These partnerships are typically less commonly used for timely or reactive research requirements where rapid initiation and execution are needed, such as in the context of reacting to unfolding global events. This leaves voluntary partnerships disproportionally suited to a particular methodology of research and type of researcher and/or research proposal.

4.8 Voluntary research partnerships often make use of Privacy Enhancing Technologies (also known as "PETs") to ensure that the security of the service and the privacy of the service's users is respected. These are a range of technologies and procedures that enable access to data will reducing the risk associated with data use.[8] We discuss PETs in more detail in Annex 3 of this report.

4.9 It is worth noting that data-sharing agreements under voluntary partnerships can include conditions which stipulate certain conditions researchers must fulfil, such as being affiliated with an academic institution, and/or being independent from commercial interests. This can leave certain researchers (for example, unaffiliated researchers or researching institutions with hybrid funding models) unable to access data of similar quantity and quality. [9] As access to data facilitated by services has decreased, its availability has become all the more valuable for researchers.

[8] The Royal Society, 2023. From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis. From privacy to partnership | The Royal Society. [accessed 23 June 2025]

[9] The British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, p.11.

Ad libraries

4.10 Ad libraries are collections of online advertisements, which can be supported with APIs[10], with corresponding information on who funded them, how much was spent, and general information on the users that saw them. Ad libraries were designed to increase transparency around online political advertising and became widely used by social media[11] in 2018 in the wake of the 2016 Brexit referendum and the US presidential elections.

4.11 Most prominent ad libraries differ in their coverage, level of accessibility, and utility, based on the policies of – and regulatory requirements covering – the respective services hosting the ads. Since 17 February 2024, certain services regulated under the Digital Services Act that feature advertisements have been required to provide repositories containing information about these advertisements.[12] Ad libraries can be useful for accessing clean, programmatic, and well-defined data about online advertising. They are provided and operated by services with minima resource investment and do not present the same legal or ethical challenges as more covert forms of data collection.

Transparency reports

4.12 Some services regularly publish transparency reports to inform the public of their content moderation efforts, data handling practices, and compliance with legal obligations. Transparency reporting has recently become a legal obligation for certain services under the EU's Digital Services Act, which mandates the public disclosure of specific metrics and processes related to user safety, security, and legal and regulatory compliance.[13] In the UK, the Online Safety Act includes similar provisions, requiring categorised regulated services to publish transparency reports.[14] Even before these legal requirements, some services voluntarily released such reports, typically detailing content removals or restrictions, enforcement actions against accounts in violation of policy and other relevant metrics.[15]

4.13 Transparency reports are publicly available and contain information that researchers can use to understand trends in content moderation, government and law enforcement requests for user data, enforcement of terms of service and other policies, and actions taken against harmful behaviour and content, among other topics. This information is usually relatively easy to access and freely available.

---

[10] Medina Serrano, J.C, Papakyriakopoulos, O. and Hegelich, S., 2020. Exploring Political Ad Libraries for Online Advertising Transparency: Lessons from Germany and the 2019 European Elections. https://dl.acm.org/doi/pdf/10.1145/3400806.3400820. [accessed 23 June 2025]

[11] Leerssen, P., Ausloos, J. et al., 2019. Platform ad archives: Promises and pitfalls. Internet Policy Review, 8(4). https://doi.org/10.14763/2019.4.1421. [accessed 23 June 2025]

[12] Official Journal of the European Union L 277/1 of 27 October 2022. Publications Office. [accessed 23 June 2025]

[13] Harling, A-S., Henesy, D. and Simmance, E., 2023. Transparency Reporting: The UK Regulatory *Perspective. Journal of Online Trust and Safety, 1(5).* https://doi.org/10.54501/jots.v1i5.108. [accessed 23 June 2025]

[14] For further information on Transparency reporting under the Online Safety Act, see Ofcom's Consultation on transparency guidance

[15] Harling, A-S., Henesy, D. and Simmance, E., 2023. Transparency Reporting: The UK Regulatory *Perspective. Journal of Online Trust and Safety, 1(5).* https://doi.org/10.54501/jots.v1i5.108. [accessed 23 June 2025]

Data accessed directly from services in other jurisdictions: Digital Services Act 2022 Article 40

The EU's Digital Services Act is the main legislation currently enabling researcher access in the EU and created new access pathways for researchers to online data. This is an EU regulation, enacted as Regulation (EU) 2022/2065, which establishes a framework to increase transparency, safety and accountability in the online environment. There are two main provisions which are intended to enable greater researcher access to online safety information, Article 40(12) and Article 40(4).

Under Article 40, researchers can request access to publicly available data from Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs).[43] VLOPs and VLOSEs are required to provide access to researchers whose request meets the criteria set out in the Digital Services Act.

The European Commission's draft Delegated Regulation further sets out the procedures for data access under Article 40. Researchers submit data access applications to the nationally appointed Digital Services Coordinator ("DSC"), who verifies the details of the applications and then decides whether to pass on the data request to the data provider on the researcher's behalf. If granting access, providers must abide by certain requirements regarding data formation and documentation. This ensures that the data is usable for the researcher and has the necessary security and quality safeguards. The Delegated Regulation also mandates the establishment of the Digital Services Act Data Access Portal, a centralized platform for submitting and managing data access requests, where all "reasoned requests" must be made publicly available by the DSC.[44]

Article 40 of the Digital Services Act combines direct access (option 3A) and notice to service (option 3B) intermediary functions, and relies on accreditation to mitigate security, legal, and ethical concerns to enable greater access. There are two primary means of access to data for researchers who want to conduct research that contributes to the detection, identification and understanding of systemic risks[45] in the EU:

Eligible researchers can access publicly accessible data in online interfaces managed by VLOPs and VLOSEs (Article 40(12)) – See "direct access intermediary" model in Section 6**.**

Researchers captured by Article 40(12) include those affiliated to not-for-profit bodies, organisations and associations, and who meet the following relevant criteria set out in Article 40(8):

- are independent from commercial interests;
- can disclose the research funding;
- are able to comply with data security and confidentiality obligations, and protect personal data; and
- will only access data that is necessary and proportionate to the purpose of their research (being the detection, identification and understanding of systemic risks in the EU).

Access is determined on a case-by-case basis, with ultimate discretion resting with providers of VLOPs and VLOSEs, who consider the safety and security of their service and users in assessing the requests. Article 40(12) provides access without undue delay to data (including, where technically possible, to real-time data), provided that the data is publicly accessible in services' online interface by researchers, including those affiliated with not-for-profit bodies, organisations, and associations.

To assess compliance with Article 40, EU member states' DSCs and the European Commission (EC) monitor the measures taken by providers of VLOPs and VLOSEs to give researchers access to publicly accessible data. The EC can open formal proceedings where a VLOP or a VLOSE is suspected of breaching the Digital Services Act, and take further enforcement action, such as imposing interim measures or making non-compliance decisions.

Vetted researchers can apply for access to non-public data held by VLOPs and VLOSEs by submitting a data access application to the appropriate DSC(s) (Article 40(4)) – See "notice to service intermediary" model in Section 6.

To apply for access to non-public data held by VLOPs and VLOSEs, researchers must fulfil the criteria set out in Article 40(8). In addition to the criteria set out under Direct access), researchers must also show that:

- they are affiliated with a research organisation[46] or are affiliated with other entities dedicated to public interest research;
- the research will be carried out for the detection, identification, and understanding of systemic risks in the EU (Article 34(1)), and/or to assess the adequacy, efficiency, and impacts of the risk mitigation measures outlined in Article 35; and
- the research results will be made publicly available and free of charge.

Researchers can submit their application for access to the DSC of the EU member state where their affiliated research organisation is based or where the provider of the VLOP or VLOSE is based. The DSC assesses whether researchers fulfil the given criteria and therefore obtain vetted status.

The DSCs have a mandate to facilitate research where possible but must also consider the interests of providers and users with regards to the protection of personal data and confidential information (such as trade secrets), and the security of the service. The decision to request data access from providers rests with the DSC of where the provider is based.

Data accessed indirectly from services

4.14 To complement direct access modalities, mitigate the limitations of certain modalities, or access data if direct access is not available to them, some researchers employ indirect methods to access data. These methods include (but are not limited to) data scraping, passive observation, or the use of third-party tools and intermediaries.

DATA ACCESSED INDIRECTLY FROM SERIVCES
Data scraping
Avatar research
Purchasing data and access from commercial entities
Research forums and consortia

Data scraping

4.15 Data scraping refers to the automated collection of data from digital sources where information is not readily available in a structured or downloadable format. This may include crawling, parsing, and/or screen capture. Researchers collect this data autonomously, meaning they are not reliant on services to share it. In the face of decreasing access to service-operated APIs (see Section 4) some researchers have said that scraping is one of the only remaining options for systematic gathering of information about online content.[16] They also view data scraping as one of the only ways to verify that data provided via a service is complete and unmanipulated (for example, to audit APIs or other service-provided datasets).[17]

Avatar research

4.16 Avatar research involves the creation of an account that represents a specific persona on a service. The researcher either manually engages with content as that persona or uses automated tools (such as bots) to simulate interactions, while documenting what is encountered as a result. This avatar is typically based on certain criteria (such as demographic or behavioural criteria) that the researcher wishes to explore. Avatar research differs from passive observation – which can also involve the creation of research accounts for reasons related to operational security – in that passive observation involves observing online behaviour without interacting or otherwise influencing an environment.

4.17 Avatar research provides a way to examine whether and how users, particularly vulnerable groups such as children, are likely to encounter harmful content.[18] It can highlight potential risk factors for encountering harmful content, since it allows researchers to compare the experiences of avatars with different behavioural or demographic traits, noting correlations between these factors and the nature of the content encountered. Due to the limited scalability of this methodology, the data generated is typically used in qualitative research to mimic and study digital experiences of a limited number of persona types. Many services' terms of service prohibit the creation of accounts used for avatar research, sometimes referred to as 'sock puppet accounts'. This can pose legal and ethical challenges for researchers using this methodology.

[16] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.4.

[17] For further discussion, see AI Forensics' report on researchers' experiences using TikTok's API: TikTok's Research API: Problems without explanations. The authors audited the quality of the data available through the TikTok API using data donations and scraping, and concluded that TikTok's API does not provide up-to-date, complete and consistent information when evaluated against the data collected through alternative methods. In response, a spokesperson for TikTok expressed concern that researchers misunderstand the design of their tools.

[18] Ofcom, 2023. Research pilots for understanding children's online experiences. Research pilots for understanding children's online experiences - Ofcom. [accessed 23 June 2025]

Purchasing data and access from commercial entities

4.18 Some researchers purchase data about services for research purposes. This data can be provided as a dataset or be accessed via a third-party analytics platform. Purchasing data from commercial data brokers can offer a relatively easy, albeit potentially costly, way for researchers to acquire data that may otherwise be inaccessible to them. The data is typically provided in a cleaned, organised, and labelled format, thus reducing both researchers' workloads and the need for technical expertise at the data collection and processing stages.

4.19 Researchers can also purchase access to social listening tools. These are software applications that monitor and analyse public online content and engagement across services. In many cases, these tools use a combination of official APIs, licensed data feeds, partnerships with services, and scraping to collect data for onward analysis. While many of these tools were initially created for businesses to track public sentiment and brand reputation,[19] online safety researchers also use these tools to help them understand and gather insights from services. Many social listening tools provide APIs that allow researchers to access and retrieve data programmatically, using custom queries and filters, such as specific keywords or n-grams,[20] hashtags, URLs, or mentions, and focusing on specific topics, timeframes, and geographic locations. This data can be exported in various formats for further analysis using statistical software or custom scripts. Some social listening tools also provide dashboards with the option to aggregate data from multiple sources and services and offer real-time or near real-time monitoring. These tools can facilitate cross-service research and provide access to information about services quickly and without resource-intensive researcher accreditation or proposal vetting processes, which is particularly suited to exploratory research.

Research forums and consortia

4.20 Research forums and consortia are collaborative structures where individual researchers, research organisations, data providers, and others can collaborate and share insights, resources, and data to pursue research objectives.

4.21 Research forums can be beneficial for the pooling of resources and expertise. They can contribute to raising the quality of the research landscape as well as supporting a more egalitarian distribution of resources and opportunities among researchers. For data providers, collaborating with researchers in the context of research forums can offer more reassurance about researchers' capability, resources, and reliability. This may motivate data providers to share more data and information than usual. However, while such research forums and consortia can foster collaboration, they are usually not data-centric organisations, which means their primary aim is not necessarily to enable or increase researchers' access to data.

---

[19] Rosenblatt, M., Curran, T., and Treiber, J., 2018. Building brands through social listening. Building Brands through Social Listening. [accessed 23 June 2025]

[20] "An n-gram is a collection of n successive items in a text document that may include words, numbers, symbols, and punctuation. N-gram models are useful in many text analytics applications where sequences of words are relevant, such as in sentiment analysis, text classification, and text generation." Source: MathWorks. What Is an N-Gram? What Is an N-Gram? - MATLAB. [accessed 23 June 2025]

Data accessed directly from users

4.22 To complement direct and indirect access modalities, mitigate the limitations of certain modalities, or access data if direct and indirect access is not available to them, some researchers employ research methodologies to access data directly from service users.

DATA ACCESSED DIRECTLY FROM USERS
Data donations
Voluntary widgets
Data trusts
Data cooperatives

Data donations

4.23 Data donations refer to the willing and purposeful sharing of data with researchers by data subjects. Under the UK's data protection regime, individuals have the right to access their personal data held by data controllers, such as services. By exercising these rights, in the form of a 'subject access request', individuals may obtain a copy of this data and choose to share it with researchers. In certain circumstances, individuals may also exercise their right to data portability under UK data protection legislation. This entitles an individual to receive personal data provided to a data controller in a structured, commonly used, and machine-readable format. Individuals can then share this information with researchers.

4.24 Data donations are noted by several stakeholders as potential mitigations for ethical challenges such as the consent of the data subject(s).[21] [22] Stakeholders have emphasised that services do not currently make it easy for users to donate their data, even though individuals are able to exercise data portability and access rights under the UK GDPR.[23]

4.25 The various forms of data donations are discussed further below.

Voluntary widgets

4.26 Participants can voluntarily download vetted tracking software onto their device and donate the data captured to researchers. Examples of the use of this method include the work of Who Targets Me[24] and New York University (NYU).[25] Both researching institutions have developed browser extensions that capture online advertisements users encounter to build their public ad repositories. One of the main benefits of this methodology is the fact that data subjects voluntarily provide their data to researchers with full and informed consent. This can offer a mitigation to many of the ethical and legal issues that may arise when data is collected and used in ways that do not guarantee data subjects' consent.

[21] [✂]
[22] British Academy response to October 2024 Call for Evidence, p.2.
[23] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN), response to October 2024 Call for Evidence, p.18.
[24] Who Targets Me, Who Targets Me – Making digital political ads more transparent and accountable. [accessed 23 June 2025]
[25] Ad Observer, Ad Observer. [accessed 23 June 2025]

Data trusts

4.27 Instead of managing financial or physical assets, data trusts oversee data on behalf of the data provider, and in the interests of a set of defined beneficiaries.[26] They provide a framework for sharing data based on independent oversight and transparent rules for access and use.[27] Data trusts also provide mechanisms to enable individuals to share their data in a privacy-protecting manner, including to support research. The beneficiaries declare how and for what purpose they want their data to be used, and the trustee exercises the beneficiaries' data rights on their behalf. It is also possible to delegate responsibility to the trustee to decide on what data processing is in their beneficiaries' interests. This allows a trustee to make processing decisions on the beneficiaries' behalf according to a pre-agreed set of rules or principles.[28]

Data cooperatives

4.28 Data cooperatives are a mechanism for individuals to collectively pool their data to pursue a certain interest or objective.[29] The cooperative is owned and operated by its participating members. As the data is managed collectively, each data donor is able to influence its use and the safeguards in place. Data cooperatives can be used as a mechanism for members to share data to assist research.

[26] Delacroix, S. and Lawrence, N.D., 2019. Bottom-up data Trusts: disturbing the "one size fits all" approach to data governance. International Data Privacy Law, 9(4). https://academic.oup.com/idpl/article/9/4/236/5579842. [accessed 23 June 2025]

[27] Open Data Institute response to October 2024 Call for Evidence, p.8.

[28] For example, OpenCorporates, the largest open database of companies in the world, founded OpenCorporates Trust in 2017 shifting to a trust structure based on their "Legal-Entity Data Principles". For more information, see: Principles : Open Corporates, and Trust – OpenCorporates. [accessed 23 June 2025]

[29] For example, the Drivers Cooperative collects data from on-demand drivers, and aggregates and analyses this data to deliver insights that help users understand mobility logistics. Insights are also shared with city planners to drive better infrastructure and transportation planning. For more information, see: The Drivers Cooperative. [accessed 23 June 2025]

5. Barriers to sharing information[30]

Introduction

5.1 We identified a range of factors which constrain both services and researchers in sharing data. These factors include a variety of legal, ethical, security, technical, information quality, financial and operational challenges. These are the primary criteria that we have used for assessing the three policy options for improving researcher access to data, which are discussed in Section 6. We discuss each of these challenges in more depth in the following sections.

Legal and ethical constraints

5.2 Services and researchers must navigate a complex legal landscape when sharing data for research purposes. This includes managing legal risk, complying with data protection laws across jurisdictions, and interpreting legal ambiguity around data collection methods, such as scraping. Researchers share that these challenges are often compounded by uncertainty around the legality of sharing, storing and analysis of data, often resulting in a risk-averse approach to sharing data.

5.3 UK data protection law sets out the requirements for processing personal data. But this is only one part of the picture. In many cases, data-sharing agreements are used to facilitate lawful research access, but we also heard that these and services' terms of services can introduce additional complexity and can be difficult to navigate. Many researchers face ongoing uncertainty about best practices, particularly when navigating overlapping or competing legal, contractual and policy issues related to gathering and handling data.

Data scraping, data purchase and data transfer practices

5.4 Restrictions on data scraping have intensified in recent years, in part as a response to the rise of large language models (LLMs). Large language models are often trained on data scraped from the internet. This has drawn heightened attention to the practice.[31] In the US in recent years, there have been several instances of services pursuing lawsuits against those scraping their data, such as *Meta v. Bright Data*,[32] and *X Corp v. Center for Countering Digital*

[30] The views expressed in this chapter regarding data protection law are based on the responses to our Call for Evidence. They do not necessarily represent an accurate or comprehensive reflection of current data protection law.

[31] Megan, A. B., Gruen, A. et al., 2024. Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. [2410.23432] Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. [accessed 23 June 2025]

[32] Meta Platforms Inc., Plaintiffs, v. Bright Data LTD., Defendants, Order Denying Meta's Motion For Partial Summary Judgment: And Granting Bright Data's Motion for Summary Judgement, 01/23/24, Meta v. Bright Data. [accessed 23 June 2025]

Hate.[33] Additionally, the ICO and eleven other data protection and privacy authorities have also published a joint statement calling for the protection of personal data from unlawful data scraping on social media sites.[34] In October 2024, the ICO, along with other signatories, published a Concluding Statement that builds on this Joint Statement.[35] The co-signatories (including the ICO) acknowledged the importance of socially beneficial research but reminded social media companies and other organisations that host publicly accessible personal data that, when allowing large-scale access or collection, they must ensure that they are complying with applicable data protection laws, including by ensuring that there is a lawful basis for granting access or permitting collection of personal data.[36] The Concluding Statement also recognises that using APIs can serve as a potential safeguard against unlawful scraping.[37] The ICO has also recently clarified its position on the lawful basis for web scraping for the purposes of training generative AI models.[38]

5.5 Our evidence suggests that the tightening of restrictions by services has created significant challenges for researchers. They report lacking clear legal guidance on when scraping is permitted, forcing them to assess their individual and institutional risk tolerance. Even when scraping (or data processing more broadly) is permissible, such as for 'scientific or historical' research purposes, the precise legal rules are complicated and the definition of qualifying research, while broad, may be challenging to apply in practice (See Annex 1). For example, research into online safety matters often involves 'special category data' and, if that is the case, then to rely on the research provisions any data processing must be 'necessary for the performance of a task carried out in the public interest'. The sense from researchers of the lack of clarity around the subject and the threat of entering expensive legal proceedings with services can have a chilling effect which risks deterring researchers.[39] Although we note that some recent scraping-focused lawsuits against researchers have ultimately been dismissed.[40]

5.6 Some researchers also purchase data from data brokers as an alternative source. This method can also present legal and ethical challenges. Historically, data brokers have sometimes utilised poor practices for data collection, such as sharing data without the data

[33] X CORP., Plaintiff, v. Center for Countering Digital Hate, Inc., et al., Defendants, Order Granting CCDH Motion to Dismiss and Strike, 03/25/2025 gov.uscourts.cand.416212.75.0.pdf. [accessed 23 June 2025]
[34] Information Commissioner's Office (ICO), Joint statement on data scraping and the protection of privacy. [accessed 23 June 2025]
[35] Office of the Privacy Commissioner of Canada, Concluding joint statement on data scraping and the protection of privacy - Office of the Privacy Commissioner of Canada. [accessed 23 June 2025]
[36] See Clause 25: Concluding joint statement on data scraping and the protection of privacy - Office of the Privacy Commissioner of Canada. [accessed 23 June 2025]
[37] See Clause 26: Concluding joint statement on data scraping and the protection of privacy - Office of the Privacy Commissioner of Canada. [accessed 23 June 2025]
[38] Information Commissioner's Office (ICO), The lawful basis for web scraping to train generative AI models. [accessed 23 June 2025]
[39] Open Data Institute, 2024. Exploring global challenges of regulating researcher access to platform data. Exploring global challenges of regulating researcher access to platform data | The ODI. [accessed 23 June 2025]
[40] Keller, J. R., Moriniere, S. and Tinsman, C., 2024. What is 'public data'? And who should be allowed to collect and use it? What is 'public data'? And who should be allowed to collect and use it? | by Jared Robert Keller | Canvas | Medium. [accessed 23 June 2025]

subjects' knowledge or consent,[41] and exploitation of regulatory loopholes.[42] This may dissuade researchers from utilising broker-collected data due to ethical risks and concerns around data acquisition and validity, as it is often unclear and difficult to establish how data was collected. [43]

5.7 There are also uncertainties around how to transfer data safely and securely. Stakeholders have shared that both researchers' and services' legal teams often advise erring on the side of caution and taking a conservative approach to receiving and storing service data.[44] This advice stems, in part, from a perceived lack of standardised processes for transferring data and limited practical guidance on how data controllers should facilitate data transfers.[45] Applying the principle of data minimisation and managing individuals' rights to request deletion of personal information[46] are also legal requirements that some researchers do not feel comfortable taking on.

Perceived risk around public versus private data

5.8 Online safety researchers often say they need access to individual-level personal data[47] [48], thereby necessitating compliance with data protection law. Researchers may inadvertently access personal information even after applying standard data cleaning procedures, and individual users can sometimes be identified despite anonymisation efforts.[49] For example, datasets may contain large amounts of largely unstructured text, images or other unlabelled or mislabelled data which could either directly contain or provide proxies for identifying personal information.[50] Additionally, data can become more 'sensitive'[51] depending on how

---

[41] Fair, L. What Goes on in the Shadows: FTC Action Against Data Broker Sheds Light on Unfair and Deceptive Sale of Consumer Location Data. [accessed 23 June 2025]

[42] Mishra, S. 2021. The dark industry of data brokers: need for regulation? International Journal of Law and Information Technology, 29(4), https://doi.org/10.1093/ijlit/eaab012. [accessed 23 June 2025]

[43] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.5.

[44] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) response to October 2024 Call for Evidence, p.9.

[45] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) response to October 2024 Call for Evidence, p.8.

[46] Under Article 17 of the UK GDPR, people have the right to have their personal data erased. However, there is a built-in exception for research where the research exemptions apply. Article 17(3)(d) states that, if you are processing data for research-related purposes, the right to erasure does not apply in so far as complying with the right is likely to render impossible or seriously impair the achievement of research-related purposes.

[47] "At its core, scientific research requires individual-level data to understand and explain social phenomena (…) Individual-level data can help to explain the conditions under which someone is likely to suffer mental health difficulties as a result of their online experiences and which, if any, forms of online intervention are most likely to help…In order to allow researchers to conduct much-needed causal inquiry, as well as other vital research that supports innovations and insights in the public interest, scientific researchers will need to have access to personal data. Quite simply, the ability of multi-disciplinary researchers around the world to conduct pioneering and socially important research is tied to the availability of these data." Source: European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, pp.4-5. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]

[48] [✂]

[49] [✂]

[50] Garfinkel, S.L., 2015. De-Identification of Personal Information. National Institute of Standards and Technology, U.S. Department of Commerce. De-Identification of Personal Information. [accessed 23 June 2025]

[51] Forthcoming Ada Lovelace Institute publication, p.7.

it is used and the inferences made from it. These risks expose researchers to both legal and ethical consequences.

5.9 An important challenge facing researchers and services is the unclear boundary between 'public' and 'private' data[52]. Some researchers told us that this ambiguity makes it difficult for them to understand what information they can reasonably request for either category in data access applications.[53] [54] [55] The Knight-Georgetown Institute is developing a cross-industry framework to clarify access to public service data, aiming to help researchers understand the relationships between services and individuals, communities, and societies.[56]

5.10 The UK data protection framework regulates all personal data, whether public or private, with extra protections for 'special category' data[57] and provisions for lawful research purposes (see Annex 1 for more information). However, researchers told us that understanding and navigating these legal requirements can be difficult, particularly those without necessary legal and technical support.

5.11 Services also told us that they face a complicated legal and regulatory environment. Some services told us that their ability to share information for research purposes often depends on obtaining consent from relevant data subjects.[58] Meeting these obligations presents both operational and legal challenges, which means that services may choose to restrict data access rather than risk compromising user rights.[59]

Perceived ambiguity in data-sharing agreements and services' terms of service

5.12 When researchers receive data from services, they typically sign data-sharing agreements, which set out contractual terms and conditions. While these agreements help define roles, responsibilities and liabilities, they create significant challenges for researchers. Despite having formal agreements, researchers report uncertainties about what activities are actually permissible. Due to this uncertainty, and researchers' perception that the terms of data sharing agreements can be difficult to understand and navigate, researchers remain concerned about the legal liabilities they may face if their conduct is deemed to breach

[52] In the context of this report, public data refers to information that is readily accessible to the general public and does not require special permissions or authorisation to access. Private data refers to information that is not readily accessible to the public and requires special permissions or authorisation to access. Special category data, often referred to as sensitive data in this report, can be public or private and is subject to data protection laws due to its personal nature.

[53] It should be noted in this context that data protection law applies to both public and private personal data.

[54] Open Data Institute response to October 2024 Call for Evidence, pp.3, 4, 5, 7.

[55] Knight-Georgetown Institute response to October 2024 Call for Evidence, p.9.

[56] Knight-Georgetown Institute, Publicly Available Platform Data Expert Working Group – Knight-Georgetown Institute. [accessed 23 June 2025]

[57] Specific categories of personal data that are likely to be more sensitive.

[58] It is important to note here that, as the ICO's guidance states, consent to participate in a research study is distinct from consent as a UK GDPR lawful basis to process personal data. Even where there is a separate ethical or legal obligation to get consent from people participating in research, this is distinct from the lawful bases under data protection law that services can rely on (e.g. for UK GDPR purposes, consent is one of the lawful bases that can be relied upon, but there are others – see Annex 1). See ICO guidance on principles and grounds for processing.

[59] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence, p.7.

those terms.[60] [61] Services are also aware of the risk of reputational damage from research findings when data is improperly collected.[62] Even when negative findings do not directly breach the terms of the data-sharing agreement, researchers may nonetheless face legal risks if services claim the data has been misinterpreted.

5.13 Researchers find these agreements overly complex,[63] especially where they concern personal data, details of licensing terms, liabilities, and intellectual property rights.[64] Navigating such agreements is particularly challenging for those with limited resources.[65] This complexity, combined with services' risk concerns, makes it difficult to achieve balanced distribution of legal rights and responsibilities.

5.14 If data is accessed outside of a formal data-sharing agreement, researchers' conduct on a given service is still governed by the services' terms of service. Researchers have also flagged that services' terms of service or community guidelines often constrain the collection of data by prohibiting practices such as scraping and avatar research, as well as the use of third-party technologies such as browser extensions.[66]

5.15 While some researchers view many of these restrictions as overly cautious, services must balance empowering users and protecting their rights (as well as their own legal position). Services target certain tools or methods because they perceive them as sources of legal and operational risk. Without a formal framework for research data access rights, services may take cautious approaches to data collection tools, partly because they cannot guarantee these tools are always used for legitimate research purposes.

Challenges associated with sharing data across jurisdictions

5.16 Many services are multinational, operating across different legal jurisdictions. This creates challenges for both researchers and services who need to navigate data protection legislation spanning multiple nations.[67] This can become particularly relevant where the data holder, data provider, researcher or data subjects are based in different jurisdictions. There is some consensus among those in the research community that they would benefit from greater clarity on whether data from one jurisdiction may be stored and processed in another and how this intersects with international data rights.[68] [69] These ambiguities lead to some reluctance from both researchers and services when it comes to collaborating and exchanging data across jurisdictions.

5.17 Services highlight the complexity and burden associated with having to unpick and comply with overlapping legal requirements relevant to providing researchers with access to data in relation to privacy law and other areas.[70] In some cases, this leads to increased caution

[60] OpenMined response to October 2024 Call for Evidence, p.7.
[61] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, pp.6, 7, 18.
[62] Reddit response to October 2024 Call for Evidence, p.5.
[63] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.18.
[64] Smart Data Research UK response to October 2024 Call for Evidence, p.6.
[65] Smart Data Research UK response to October 2024 Call for Evidence, p.2.
[66] Institute for Strategic Dialogue response to October 2024 Call for Evidence, pp.11, 15.
[67] For further information, see ICO guidance on international transfers of personal data.
[68] British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, p.12.
[69] [✂]
[70] Google response to October 2024 Call for Evidence, p.8.

around entering into data-sharing initiatives. The issue is compounded by the emergence of several varied regulatory frameworks in similar areas of online safety; an example cited in the context of trust and safety was transparency reporting.[71] To some extent these regulatory burdens are inescapable and come with services that are provided globally across different legal jurisdictions. It nonetheless brings into focus how lawmakers and regulators may be able to facilitate better outcomes by ensuring some degree of alignment with similar regimes that may be operating in other countries.

Security constraints

5.18 Services cite security risks as a major constraint to offering data access to researchers. These concerns fall into two categories: data security issues that can occur during the access process, such as inappropriate or outdated encryption, and systems security risks where data disclosure could enable unauthorised use of the service.[72]

5.19 Services worry that publicising available research data types through channels such as public inventories may provide useful information to malicious actors and leave services vulnerable to security breaches.[73] [74] These risks could potentially affect national security if exploited by bad actors or foreign adversaries.[75] Security threats may also harm users, who could be exposed to scams, harmful profiling, or other forms of harms.[76] Publishing data inventories and information on how data is structured also carries risks to services' businesses, including potential disclosure of proprietary algorithms and confidential business strategies, or third-party data.[77]

5.20 While these concerns may be justified, in principle, it is not the case that data structures or inventories *per se* give rise to all these risks. The actual level of risk depends on specific proposals and implementation details. Services were open to exploring ways in which types of information held might be made available in ways that are secure and privacy preserving.[78] Projects such as the Online Safety Data Initiative, commissioned by the Department of Digital, Culture, Media & Sport (DCMS), have also called out the importance of addressing security challenges in data sharing. The project calls out a range of potential mitigations to security challenges, such as the use of PETs and data protection frameworks such as Zero Trust principles.[79]

[71] Pinterest response to October 2024 Call for Evidence, p.2.
[72] Leerssen, 2023. Call for evidence on the Delegated Regulation on data access provided for in the Digital Services Act – Summary & Analysis. European Commission. pp. 19, 20.
[73] Vinted response to October 2024 Call for Evidence, p.5.
[74] Snap Inc. response to October 2024 Call for Evidence, p.2.
[75] Google response to October 2024 Call for Evidence, p.10.
[76] [✂]
[77] Vinted response to October 2024 Call for Evidence, p.5.
[78] Snap Inc. response to October 2024 Call for Evidence, p.2.
[79] Online Safety Data Initiative, 2021. Putting data security at the heart of the Online Safety Data Initiative – Online Safety Data Initiative. [accessed 23 June 2025]

Information security

5.21 Other technical constraints relate to handling sensitive user data, which is challenging due to the lack of standardised security protocols.[80] Services are also constrained by the inability to share data without compromising user privacy, which includes the removal of personally identifiable information and other data that could lead to the potential reidentification of users.[81] [82] Securely storing data in a privacy-preserving way post-data transfer can be particularly challenging.[83] Services have expressed concern about the risks of reidentification of individuals if data is not properly anonymised, as well as the potential for the misuse of information for harmful profiling.[84] Privacy-compliant data anonymisation is also challenging from a technical perspective, especially when considering the need for data to stay useful, accurate, and representative.[85] While PETs hold a lot of promise to facilitate the sharing and analysis of sensitive data, at-scale implementation of PETs can be expensive and technically demanding.[86] Crucially, while PETs can offer a mitigation to security and reidentification risks, the trade-off is often reduced value in the data. [87] We discuss PETs in more detail at Annex 3 of this report.

Technical constraints

5.22 Issues relating to standardisation, interoperability, tracking how data is used and analysed, as well as the complex infrastructure required to facilitate data sharing are some of the most prominent technical constraints which limit services' ability to share data with researchers.

5.23 Our evidence suggests that there are challenges related to the size, formatting and interoperability of data to be shared with researchers, [88] [89] as well as the technical challenges the arise from presenting large volumes of data in near real-time.[90]

5.24 There is a lack of technology or tools that enables the tracking and monitoring of data once it is shared or made available to download.[91] This would help services understand how the data is used, manipulated, and shared onwards. [92] Services are understandably concerned about the potential reputational risk of sharing data for research purposes without knowing exactly how the data is to be represented or analysed. Misinterpretation or misuse of shared data could lead to misleading conclusions, harm to the services' credibility or negative public perception.[93] [94]

---

[80] NYU's Center for Social Media and Politics response to October 2024 Call for Evidence, p.6.
[81] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence, p.8.
[82] Vinted response to October 2024 Call for Evidence, p.3.
[83] Reset.Tech response to October 2024 Call for Evidence, p.7.
[84] [✂]
[85] Vinted response to October 2024 Call for Evidence, p.4.
[86] Open Data Institute response to October 2024 Call for Evidence, p.6.
[87] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.17.
[88] Reset.Tech response to October 2024 Call for Evidence, p.7.
[89] Vinted response to October 2024 Call for Evidence, pp.4, 6.
[90] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence, p.8.
[91] Google response to October 2024 Call for Evidence, p.9.
[92] Vinted response to October 2024 Call for Evidence, p.5.
[93] Vinted response to October 2024 Call for Evidence, p.4.
[94] Reddit response to October 2024 Call for Evidence, p.5.

5.25 A primary benefit of API access is that the service operating the API has greater control over the process of data-sharing as compared to many other data access modality. The service establishes and maintains the data collection infrastructure and grants access to the requested data to the researcher, who usually stores it themselves.

5.26 Accessing data through APIs may not always be straightforward however, as it may require specialist technical skillsets in coding and data analysis.[95] [96] Data provided by services through APIs may have inconsistencies, errors or anomalies, resulting in so-called 'noisy' data. Such data may require substantial preprocessing and augmentation before it is ready for analysis, which could include both manual content analysis and compute-intensive tasks like training models.[97] The analysis and storage of certain content formats, such as audio or audiovisual content, can also be technically challenging for researchers.[98] [99] [100] If the data received is provided in a proprietary data format, it can be difficult to share data in an interoperable manner across different research tools and institutions.[101] Receiving large volumes of data may overwhelm researchers' ability to analyse it, particularly if the dataset is both large and inadequately labelled or formatted.[102] Pre-defined API endpoints can also present a challenge when research questions arise requiring data not anticipated when an API was created. Some researchers have recommended that APIs be flexible to enable researchers to specify how they want to query data.[103] At the same time, changes to API designs within services can pose constraints and disrupt ongoing research.[104] For example, if the data available via an API changes, this limits the ability to conduct longitudinal analysis, forces researchers to make methodological adjustments, and compromises data consistency.[105] Given that service-provided APIs are service-specific, some researchers have called for standardisation in API access both between and within services to ease access and analysis[106] [107] and enable comparative research studies.[108]

5.27 However, APIs remain a convenient solution for those researchers who have the technical skillset to analyse such data. APIs are commonly accompanied by documentation regarding how they should be used, and many are standardised for readability among those who know a coding language such as Python or R. The flexibility that many APIs have historically offered should also be accounted for, with researchers being able to query for keywords, set date ranges and language, and access other features of the data they wish to retrieve.

---

[95] NYU's Center for Social Media and Politics response to October 2024 Call for Evidence p.8.

[96] Name Withheld 1 response to October 2024 Call for Evidence, p.2.

[97] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.22.

[98] [✂]

[99] Institute for Strategic Dialogue response to October 2024 Call for Evidence, pp.11, 13.

[100] Working Group on Gaming and Regulation at the NYU Stern Center for Business and Human Rights response to October 2024 Call for Evidence, p.13.

[101] British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, p.12.

[102] Reset.Tech response to October 2024 Call for Evidence, p.7.

[103] OpenMined response to October 2024 Call for Evidence, p.10.

[104] [✂]

[105] [✂]

[106] British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, pp.17, 20.

[107] [✂]

[108] Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to October 2024 Call for Evidence, p.4.

Finally, we note that the technical challenges related to using APIs are not unique to independent researchers accessing data but are also reflective of the fact that researchers have different technical and methodological skillsets and preferences. Options to mitigate this challenge could include accessible technical support and upskilling opportunities for researchers, as well as ensuring the availability of less technically demanding data access modalities for researchers with limited technical capability. Several research consortia and researching institutions have formed to propose possible avenues to address these challenges, among others.[109] [110]

Information quality and availability constraints

Data quality concerns

5.28 Research data from services suffers from significant quality issues that undermine its utility for meaningful research. Data from APIs often contains inconsistencies due to services' filtering, content moderation, or technical errors. The lack of transparency around data collection and modification processes makes research reproducibility challenging.[111] [112]

5.29 Researchers frequently receive data without essential context, such as service policies, user demographics, or significant events during the data collection period, which can have an impact on how the data is interpreted.[113] [114] [115] [116] Generally, data obtained directly from services through APIs, voluntary partnerships, ad libraries or transparency reports can be difficult to validate, and often lacks the necessary metadata. Data from ad libraries and transparency reports, in particular, may not be detailed enough to conduct meaningful research.[117] [118] [119] [120] [121]

[109] For work currently undertaken in this space, see Columbia-Hertie Working Group on Building Capacity for Data Access, Analysis, and Accountability.
[110] For work currently undertaken in this space, see UK Research and Innovation's Social Platforms Data Access Taskforce.
[111] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.18.
[112] Name Withheld 1 response to October 2024 Call for Evidence, p.5.
[113] Open Data Institute response to October 2024 Call for Evidence, p.6.
[114] British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, pp.13, 15.
[115] Smart Data Research UK response to October 2024 Call for Evidence, p.7.
[116] National Society for the Prevention of Cruelty to Children (NSPCC) response to October 2024 Call for Evidence, p.8.
[117] Global Witness response to October 2024 Call for Evidence, p.4.
[118] Reset.Tech response to October 2024 Call for Evidence, p.5.
[119] Open Data Institute, 2024. Exploring global challenges of regulating researcher access to platform data. Exploring global challenges of regulating researcher access to platform data. [accessed 23 June 2025]
[120] British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, p.3.
[121] Center for Democracy & Technology. Transparency Reports. Transparency Reports - CDT. [accessed 23 June 2025]

5.30 Commercial data brokers present additional quality challenges because their data is typically collected for commercial purposes, making it unsuitable for research. Their collection methods are often unclear or difficult to verify, raising further questions about data acquisition and validity.[122] Data scraping, while potentially useful for validating API data, requires significant technical skills and introduces legal and ethical uncertainty that many researchers cannot navigate.

5.31 Services face their own constraints in producing high-quality and accurate data for independent research purposes. Privacy regulations in various jurisdictions limit data availability beyond specific timeframes. It has also been suggested[123] that data should be structured and classified to maximise its utility for researchers, as the provision of unnecessary or unstructured data may overwhelm researchers and ultimately hinder their research.

5.32 Some services noted that it would be preferable and least burdensome if they were only expected to supply data that they already collect, which is often for commercial or product-driven purposes. However, researchers argue that such data is insufficient[124] for research into online safety matters.[125] [126] Commercial or product-driven data requires extensive cleaning and preparation for meaningful analysis and may lack metrics or information relevant to online safety research. Where such information is captured, it may not be of sufficient quality or quantity.

5.33 Researchers typically require comprehensive metadata (data about the data) – for example, how it has been selected – to understand how the data they are working with was collected. This allows them to establish its validity, reliability, and generalisability. Metadata is often incomplete or entirely unavailable, which can affect researchers' ability to quality-check the data and can therefore have an impact on the quality of the research.

5.34 The lack of standardisation across services creates additional challenges. Different services use varying data formats, collection methods, and metric definitions. This can lead to discrepancies, missing data, and an inability to conduct cross-service and longitudinal research.[127] While each services' unique context and design may justify different data standards, these particularities complicate cross-service analysis. A minimal standardisation approach through a gateway schema could potentially improve multi-service analysis while reducing the burden on both services and researchers, maintaining the useful specificity of individual datasets.

5.35 One example from other sectors of facilitating greater data access and use in the context of many different datasets, researcher needs, and data sources is the Ocean Data and Information System (ODIS) and the technology that lives on top of that system, the Ocean InfoHub (OIH).

---

[122] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, pp.3, 9.

[123] Reset.Tech, response to October 2024 Call for Evidence, p.7.

[124] Open Data Institute response to October 2024 Call for Evidence, p.6.

[125] National Society for the Prevention of Cruelty to Children (NSPCC) response to October 2024 Call for Evidence, p.8.

[126] Center for Countering Digital Hate response to October 2024 Call for Evidence, p.8.

[127] Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to October 2024 Call for Evidence, p.4.

Ocean InfoHub and the Ocean Data Information System

Ocean stakeholders, including researchers, state actors, civil society and industry, could all stand to benefit from deeper insight into various ocean matters. One of the challenges is that while many of these stakeholders hold data that could contribute to further understanding, each group only has partial information, shaped by their own objectives, standards and technologies.

Recognising that most parties would benefit if there were means of sharing and combining these insights and sources, the ODIS was set up to coordinate around oceanic data and enable greater sharing and access for all parties.

ODIS is designed to provide a "foundation stable enough to accommodate many stakeholders, experiences, indices, and models". It is a "federation of systems that uses common conventions to share and exchange their (meta)data. ODIS is not a new portal or centralised system under the control of a single authority, but a partnership of distributed, independent systems voluntarily sharing (meta)data and information along co-developed and clear conventions in the pursuit of common goals".

The overarching long-term goal of ODIS is to provide a sustainable and responsive digital ecosystem where users can discover data, data products, data services, information, information products, and services.[128]

Constraints on data availability and accessibility

5.36 Researchers are currently limited in their ability to conduct high-quality research on a diverse range of topics. Examples include not having access to data for longitudinal research[129] or research which uses video data and audiovisual (combined video and audio) data.[130] This is particularly impactful given the audiovisual content across social media services. [131] [132]

5.37 The evidence we gathered suggests that end-to-end encrypted technologies present significant constraints for research data collection. Multiple researchers noted that encrypted data is unavailable for systematic data collection because only the sender and recipient can share their own messages.[133] [134] [135] Services that provide end-to-end encryption as a functionality are primarily messaging services, such as WhatsApp, Telegram and Signal. Encryption remains a vital tool for user privacy and secure communications, making it a

[128] National Academies of Sciences, Engineering, and Medicine. 2024. Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. https://doi.org/10.17226/27520. [accessed 23 June 2025]

[129] [✂]

[130] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.11.

[131] NYU Center for Social Media and Politics response to October 2024 Call for Evidence, p.3.

[132] Working Group on Gaming and Regulation at the NYU Stern Center for Business and Human Rights response to October 2024 Call for Evidence, pp.4, 8, 15.

[133] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) response to October 2024 Call for Evidence, p.9.

[134] Center for Countering Digital Hate response to October 2024 Call for Evidence, p.2.

[135] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.14.

widely adapted feature of many services. In most end-to-end encrypted implementations, services cannot access encrypted content. However, services collect varying levels of metadata about messages sent on encrypted services, which could become available to online safety research.

5.38 Decentralised services, such as BlueSky and Mastodon, present similar challenges due to the fragmentation of their networked services, which reduces opportunities for systematic data access.[136] These architectural challenges are illustrative of broader concerns within the research community about accessing relevant material in a timely manner.[137] [138] [139] [140] [141]

Concerns about data integrity and the decision-making processes related to data sharing and analysis

5.39 Online safety research often takes place in a low-trust environment. Services have voiced concerns about the credibility and motivations of researchers requesting access to their data and the secondary risk of malicious actors targeting researchers to access sensitive information and deliberately cause harm. This may motivate some services to set restrictive conditions for researcher eligibility. Services may also refuse applications out of caution. At present, it can be difficult for researchers to understand why an access request has been refused, or to appeal a refusal. As a result, we have understood that many researchers are of the opinion that services exert disproportionate control and unjustifiably limit which researchers are granted access, what data they can access, what research questions they can pursue, and what findings they can publish. [142] [143] Researchers and services have shared concerns about the integrity of these decision-making processes in the absence of a researcher access framework which clearly defines issues of scope and eligibility and has robust transparency and dispute mechanisms in place.

5.40 Even in cases where data is provided, the low-trust environment may still lead to concerns around the process of provision and the integrity and completeness of that data. While it may be necessary for services to withhold data in some cases, researchers have expressed concerns that services are unnecessarily withholding due to an abundance of caution or concerns around reputational damage. [144] [145] [146] If the data received is of poor quality, this will limit researchers' ability to conduct valuable research. Erosion of confidence in data provision and data use has a knock-on effect; it can draw into question the validity of research findings and reduce participation in voluntary research partnerships between

[136] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.13.
[137] Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE) response to October 2024 Call for Evidence, p.4.
[138] National Society for the Prevention of Cruelty to Children (NSPCC) response to October 2024 Call for Evidence, p.2.
[139] Global Witness response to October 2024 Call for Evidence, p.6.
[140] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.11.
[141] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online response to October 2024 Call for Evidence, p.8.
[142] Knight-Georgetown Institute response to October 2024 Call for Evidence, pp.2, 6.
[143] Casas, A., Dagher, G., and O'Loughlin, B., 2025. Academic Access to Social Media Data for the Study of Political Online Safety. https://doi.org/10.31235/osf.io/7pcjd. [accessed 23 June 2025]
[144] British Academy response to October 2024 Call for Evidence, p.6.
[145] Name Withheld 1 response to October 2024 Call for Evidence, p.4.
[146] Open Data Institute response to October 2024 Call for Evidence, p.4.

services and researchers as well as users and researchers (such as data donations). When a service voluntarily provides data with stipulations, such as conditions related to data misrepresentation or reputational harm, it could affect researchers' independence and shape their perception of the service's credibility.

5.41 Our view is that trust in the ecosystem will not be restored as a result of a single intervention. Greater transparency around the decision-making process for making data available, services' grounds for rejecting applications, and better tracking of data provenance and subsequent analysis is required. Establishing an appropriate data access framework is needed to enable such transparency and harness its benefits.

Financial constraints

Cost of APIs

5.42 One of the most prominent constraints researchers face are the often-substantial fees required to use services' APIs to retrieve the data needed for their research. This is especially true for large-scale data collection or long-term projects and longitudinal studies. Long-term studies can become costly due to fixed monthly rates for a certain level of access or usage, and large-scale data collection can be expensive when services have a pay-per-use or pay-per-call charge based on the number of API requests sent. Providing only one, fee-paying access model can be viewed as a barrier to access for different types of research. Researchers operate on limited budgets and the cost of API access fees can be particularly prohibitive for small academic institutions, civil society organisations, early career researchers, and individual researchers without affiliation. [147] [148]

Cost of storing, managing, providing, and analysing data

5.43 Managing and storing large volumes of data requires robust infrastructure, which can be expensive.[149] This includes costs for servers, cloud storage, and data management tools.[150] The cost of processing and analysing data for research purposes can also be high, especially when advanced techniques like machine learning are involved. The data management and analysis of data from services may involve purchasing software licenses and computational resources, in addition to potentially hiring specialised personnel. This is particularly the case where researchers are working with data collected for commercial- or product-related purposes (which is often the only type of data available from services), rather than research purposes. Such data can be additionally resource intensive to augment and clean to make it suitable for use in research.[151] [152] Services highlight the development and maintenance costs associated with making large volumes of data available for researchers, including costs for computing power, high-bandwidth networking, servers, storage and cloud services.[153]

---

[147] Name Withheld 1 response to October 2024 Call for Evidence, p.6.

[148] [✂]

[149] Xiao, L.Y. response to October 2024 Call for Evidence, p.4.

[150] Steinhart, G. and Collister, L. 2024. The Cost and Price of Public Access to Research Data: A Synthesis. The Cost and Price of Public Access to Research Data: A Synthesis. [accessed 23 June 2025]

[151] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.19.

[152] Casas, A., Dagher, G., and O'Loughlin, B., 2025. Academic Access to Social Media Data for the Study of Political Online Safety. https://doi.org/10.31235/osf.io/7pcjd. [accessed 23 June 2025]

[153] Vinted response to October 2024 Call for Evidence, p.6.

5.44 As with API fees, we note that such challenges around financial accessibility may be more pronounced for certain researchers, such as junior researchers and those working in regions of the world where funding is less accessible. This can contribute to reinforcing existing inequalities within researcher communities. The same may be said for smaller services, who may find it particularly challenging to meet the financial requirements of providing researchers with high-quality data at scale.

Operational constraints

Processes for obtaining data seen as lengthy or complex

5.45 While thorough vetting and application processes for data access are necessary to ensure researcher legitimacy, project quality and data security, many researchers find current processes excessively arduous and lengthy. Services typically request detailed information about the researchers, including their research institution and area of expertise, but researchers report that these applications are overly burdensome, with lengthy delays and slow responses that hamper time-sensitive online safety research efforts.[154] [155]

5.46 The requirements for accessing data can be particularly challenging for non-academic researchers to meet, as processes are largely designed in a way that favour academic cycles and structures. Article 40 provisions of the Digital Services Act,[156] [157] requires detailed information on project scope, funding, specific data needs, and ethics board approval. Non-academic research projects often operate differently and cannot provide such information, creating barriers to data access. We also note that the European Commission has initiated investigations against several services for alleged non-compliance with Article 40.12,[158] with stakeholders reporting frustration over excessive portal requirements and access denials without clear justification.[159] [160]

5.47 Voluntary research partnerships present additional equity concerns. Some stakeholders told us that services tend to favour senior, well-known researchers from prestigious, well-funded institutions in the Global North. These partnerships can reinforce existing inequalities by deprioritising more junior researchers, as well as those from less prestigious or well-funded institutions.[161]

---

[154] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) response to October 2024 Call for Evidence response to October 2024 Call for Evidence, p.6.
[155] OpenMined response to October 2024 Call for Evidence, p.3.
[156] NYU's Center for Social Media and Politics response to October 2024 Call for Evidence response to October 2024, p.4.
[157] [✂]
[158] European Commission, 2025. Supervision of the designated very large online platforms and search engines under DSA. Supervision of the designated very large online platforms and search engines under DSA | Shaping Europe's digital future. [accessed 23 June 2025]
[159] NYU Center for Social Media and Politics response to October 2024 Call for Evidence, p.4.
[160] National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) response to October 2024 Call for Evidence, p.6.
[161] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, p.3.

Resource constraints

5.48 Both researchers and services face significant resource demands from current data-sharing arrangements. Researchers require personnel to process requests and build technical infrastructure for data transfers,[162] while services dedicate substantial staff resources to support existing initiatives.[163] Resource-intensive methods like data donations and avatar research cannot be effectively scaled under current arrangements.

5.49 These resource requirements are expected to increase with the rollout of the Digital Services Act in the EU and the online safety regime in the UK. Services have expressed concerns about being overwhelmed by unreasonable or disproportionate requests. [164] [165] [166] [167]

5.50 Services overwhelmingly believe that calculating request proportionality should be handled by an independent third party or by the body responsible for establishing any potential researcher access regime, rather than by researchers. They suggest offering researchers data they already hold that is reasonably accessible and legally shareable.

5.51 Several approaches have been proposed for managing proportionality. Some suggest that the data access should be proportional to whether it helps further legislative aims.[168] Another approach would be to only grant access if the research aims to address an existing knowledge gap, with an intermediary body providing proportionality guidance.[169] An independent intermediary could consider pre-processing requests, consulting services on scope, access and deadlines before deciding whether the request should be substantiated in order to reduce the burden.[170] See Section 6 for further discussion of intermediaries.

5.52 A categorisation framework has been suggested whereby only certain services would be subject to data-sharing requirements, considering factors such as service structure and data collection practices. This would account for risk levels to avoid having a disproportionate impact on lower-risk services. An explicit categorisation framework would allow services to establish data-sharing infrastructure in advance of requests.[171] [172]

5.53 Services highlight operational challenges of collecting additional data in anticipation of data access requests they may or may not receive,[173] or needing to gather new data for requests covering information they do not presently hold.[174] [175] For context, one service shared that

---

[162] Reset.Tech response to October 2024 Call for Evidence, p.8.
[163] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence response to October 2024 Call for Evidence, p.8.
[164] Reddit response to October 2024 Call for Evidence, p.6.
[165] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence response to October 2024 Call for Evidence, p.8.
[166] Snap Inc. response to October 2024 Call for Evidence, p.1.
[167] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence response to October 2024 Call for Evidence, p.9.
[168] [✂]
[169] [✂]
[170] Snap Inc. response to October 2024 Call for Evidence, p.3.
[171] Pinterest response to October 2024 Call for Evidence pp.2-3.
[172] Reddit response to October 2024 Call for Evidence, p.6.
[173] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence, p.7.
[174] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence, p.8.
[175] Google response to October 2024 Call for Evidence p.9.

creating new logging or tracking systems could take six months to a year.[176] Further operational challenges may be present for services, researchers, and any potential involved parties, depending on the scope and specific requirements of any data-sharing framework.

[176] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence response to October 2024 Call for Evidence, page 8,

6. Achieving greater data access[177]

Introduction

6.1 The challenges of safe, rights-preserving researcher access affect all parties looking to gain insight into online safety, including researchers, services, users, policy makers, and the public. These challenges are not unique to the online safety landscape. Similar access challenges have already occurred in health, environmental, and commercial sectors and fields, such as astronomy and oceanography. These sectors took a long time to develop secure, rights-preserving data access solutions, with incentives, data and challenges specific to each field.

6.2 In this section we explore the different possible ways to facilitate greater researcher access to online safety information. We build on the themes discussed in the previous section, proposing different approaches to address the range of challenges raised by the stakeholders we engaged with. We consider that to meaningfully achieve greater access it would be necessary to clarify existing rules and/or create new duties for services. In this section we discuss these approaches, including options for independent intermediaries to oversee any new duties and important implementation considerations.

6.3 We outline a range of ways that greater researcher access could be achieved in a safe and rights preserving manner. These are presented as three broad policy options, and three models within the third policy option with regards to how a potential data access intermediary could operate.

6.4 Our analysis focuses primarily on the institutional relationships between parties, rather than specific technologies or research methods, as different policy options and institutional models could support various technological and methodological solutions. The models reviewed apply to both public and private data,[178] recognising that public data can potentially be misused depending on the data accessed and the access modalities. For example, public data properly assembled could be used to generate inferences and insights that allow for the reconstruction of private, sensitive or other potential harmful data outside the scope of the research query. Certain forms of nominally private data could be accessed in a way that does not present a risk of harm. For example, the total number of users in a given age range. How each model interacts with different data types will vary based on the nature of the data, its potential sensitivity and the potential use or misuse of it.

[177] The views expressed in this chapter regarding data protection law are based on the responses to our Call for Evidence. They do not necessarily represent an accurate or comprehensive reflection of current data protection law.

[178] In the context of this report, public data refers to information that is readily accessible to the general public and does not require special permissions or authorisation to access. Private data refers to information that is not readily accessible to the public and requires special permissions or authorisation to access. Special category data, often referred to as sensitive data in this report, can be public or private and is subject to data protection laws due to its personal nature.

6.5 The following analysis is not exhaustive nor definitive: it may be possible to achieve positive outcomes by using some of the options discussed in combination or by relying on other approaches entirely. We note that all models are agnostic to private and public data, and most of the policy options and models could conceivably include real-time monitoring capabilities[179] that are important to gain insight into current and emerging phenomena.[180] See the Annex 6 for further discussion of real-time access.

6.6 We evaluate how potential policy options and models within them can address existing challenges faced by both services and researchers, and how effectively they support high-quality, reliable online safety research. Each policy option and model are assessed against a set of six criteria, which reflect the key concerns raised by stakeholders regarding the viability of any future researcher access regime outlined in Section 5.

6.7 The six criteria are:

i) **Legal and ethical considerations:** Complies with relevant laws and regulations, including privacy and data protection requirements.
ii) **Security considerations:** Supports technical data protection safeguards and protections against security breaches.
iii) **Technical feasibility:** Uses existing technologies and infrastructures to support efficiency.
iv) **Cost considerations:** is financially sustainable, does not impose excessive costs for researchers, services, or any other parties.
v) **Operational effectiveness:** promotes easy-to-use data access processes, avoiding any disproportionately onerous measures that implicitly hinder researcher access.
vi) **Strategic effectiveness:** supports online safety objectives and provision of quality information for research purposes.

6.8 The three policy options we present are:

i) Clarify existing rules
ii) Create new duties for services
iii) Establish and manage access via independent intermediary

Clarify existing rules

6.9 The first policy option maintains the current regulatory framework without introducing new regulatory measures or obligations. This approach builds on existing practice, complementing existing regulatory products,[181] recognising that services, data donors and researchers must already comply with relevant data protection and privacy laws. The UK GDPR governs data processing for research purposes, and like its EU counterpart, encourages a broad approach to the concept of 'research'.[182] The ICO has produced guidance on interpreting the scope of scientific and historical research according to these broad parameters.[183]

[179] With exceptions being 'scraping' and data donations, both of which are currently constrained due to legal and ethical uncertainty.
[180] Institute for Strategic Dialogue response to October 2024 Call for Evidence, p.15.
[181] For example, the ICO's guidance on the data protection research provisions.
[182] See Recital 159 of the UK GDPR.
[183] For further information, see ICO guidance on research-related processing.

6.10 While these existing frameworks, depending on the jurisdiction,[184] can facilitate data access for research purposes, some stakeholders perceive a lack of legal clarity around exercising rights and understanding liabilities, as described in Section 5.[185]

6.11 Clarification could open additional avenues of research in addition to clarifying existing methods. We briefly explore these points with reference to data donations and data scraping.

6.12 Data donations do not require an amendment to data protection law (see Annex 1 for more information about the current data protection framework in the UK). In our Call for Evidence, respondents indicated that a number of research projects have successfully used data donations as a data collection mechanism.[186] Researchers must maintain secure data hosting infrastructure and limit access to the research teams only, with data subjects having reasonable expectations of privacy for the use of their donated data. Since data is donated with the consent of the data subject, this approach reduces ethical concerns associated with direct data access from services.

6.13 However, data donations face significant limitations. Subject access requests are time-consuming for services and subjects[187] and must be made personally by the data donors[188], potentially deterring participation and limiting the pool of available data.[189] Data portability requests often involve long processing times[190] and do not always return data in a practically useable format.[191] Monitoring software can potentially violate services' terms of service and users' rights, requiring frequent updates for any service design changes. Given the sensitive and controversial nature of many online safety-related matters, users may be reluctant to share their data on such matters with researchers.[192] Most critically, data donations can carry a risk of biased and unrepresentative samples, and are difficult to scale for large-scale, robust online safety research. See Annex 2 for additional discussion of data donations.

6.14 The UK GDPR provides certain exceptions for research purposes, making data scraping potentially permissible when necessary for the performance of a task carried out in the public interest (see Annex 1 for more information about the current data protection framework in the UK). However, data protection frameworks in other jurisdictions may not provide a public task research exception, and even when it does exist, there may be

---

[184] Office of Privacy Commissioner Canada, 2024. Clause 25, Concluding joint statement on data scraping and the protection of privacy. Concluding joint statement on data scraping and the protection of privacy - Office of the Privacy Commissioner of Canada. [accessed 23 June 2025]

[185] Open Data Institute, 2024. Exploring global challenges of regulating researcher access to platform data. Exploring global challenges of regulating researcher access to platform data | The ODI. [accessed 23 June 2025]

[186] Dommett, K., Orben, A., and Zendle, D. response to October 2024 Call for Evidence, pp.8, 15.

[187] There are exemptions from the right of access if a controller is processing for research-related purposes. For further information, see ICO guidance on the research provisions.

[188] Subject access requests can also be made by a third party with the relevant permission from the data subject.

[189] NYU Center for Social Media and Politics response to October 2024 Call for Evidence, p.11.

[190] For more information on time limits for compliance with data portability requests, see ICO guidance on the right to data portability.

[191] NYU Center for Social Media and Politics response to October 2024 Call for Evidence, p.11.

[192] "When given options for how and why public and private organisations could share people's personal data, the preferences of the public are primarily shaped by the actors involved in the data sharing… people are less inclined to select scenarios involving big technology companies in either the sharing or recipient role." Source: Department of Science and Technology (DSIT), 2024, Public attitudes to data and AI: Tracker survey (Wave 4), 2024. Public attitudes to data and AI: Tracker survey (Wave 4) report - GOV.UK. [accessed 23 June 2025]

limitations on the scope of its application (as is the case under the UK GDPR).[193] See also Section 5 for further discussion about the perceived challenges related to data scraping, data purchase and data transfer practices. Stakeholders report uncertainty about what qualifies as 'research-related purposes', which can create challenges for the fair and reliable use of scraping and discourage researchers due to legal action risks.[194]

6.15 Existing regulators could provide further guidance and clarity on criteria for research-related scraping. Initiatives like the Knight-Georgetown Institute's Gold Standard for Publicly Available Platform Data project[195] and the European Digital Media Observatory's (EDMO) report on Platform-to-Researcher Data Access explore these issues.[196] The EDMO report details the GDPR's permissive regime for scientific research processing, outlining the legal bases including consent, legitimate interests, and public task.[197] Enhanced guidance specific to online safety could provide greater long-term certainty for all stakeholders regarding research plans without requiring legislative or regulatory changes.

6.16 There are examples in other sectors of regulatory authorities providing clarification and guidance on the use of data where it could produce a benefit. The ICO recently clarified its position on the lawful basis for web scraping for the purposes of training generative AI after a consultation on the matter and used it refine its position based on the responses.[198] There are also examples of multiple regulatory bodies working together to clarify existing rules as they apply to specific circumstances,[199] including via coordinating mechanisms such as the Digital Regulator Cooperation Forum (DRCF).[200] [201]

---

[193] Office of Privacy Commissioner Canada, 2024. Clause 25, Concluding joint statement on data scraping and the protection of privacy. Concluding joint statement on data scraping and the protection of privacy - Office of the Privacy Commissioner of Canada. For further information on the public task research exemption, see ICO guidance on public task.

[194] Megan, A. B., Gruen, A. et al., 2024. Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. [2410.23432] Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. [accessed 23 June 2025]

[195] Knight-Georgetown Institute, Publicly Available Platform Data Expert Working Group – Knight-Georgetown Institute. [accessed 23 June 2025]

[196] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p5. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]

[197] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p29. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]

[198] Information Commissioner's Office (ICO), The lawful basis for web scraping to train generative AI models

[199] Joint statement from the FCA, ICO and TPR for retail investment firms and pension providers, Information Commisioner's Office, The Pension Regulator's and the Financial Conduct Authority's, 2024. Joint statement from the FCA, ICO and TPR for retail investment firms and pension providers | ICO.

[200] Joint letter from the ICO and FCA to UK Finance and Building Societies Association, Information Commissioner's Office and Financial Conduct Authority, 2023, Joint letter from the ICO and FCA to UK Finance and Building Societies Association | ICO. [accessed 23 June 2025]

[201] Digital Regulation Cooperation Forum (DRCF), 2025. Tech, Trust and Teamwork: How the FCA & ICO are Helping Innovation Take Off. Tech, Trust and Teamwork: How the FCA & ICO are Helping Innovation Take Off | DRCF. [accessed 23 June 2025]

6.17 It should be noted that the following evaluation of this policy option against our criteria is not exhaustive, but functions as an overview of the most likely impacts of adopting this option. There may be additional, further-reaching implications of clarifying existing rules that emerge in the future. We first provide a summary of the evaluation, followed by more detailed analysis against each criterion.

Summary of considerations for clarifying existing rules

A benefit of this policy option is that it provides all parties with a clearer understanding of what data access is permissible. Researchers would gain increased clarity on the types of data they can and cannot collect, while services would better understand their obligations regarding data sharing and what can be reasonably withheld. This mutual clarity could reduce ambiguity, foster more consistent data sharing practices and potentially facilitate greater access to individual-level data (where that is consistent with the data protection principles, e.g. data minimisation).

However, since data collection under this policy option would be largely researcher-led, the technical and financial burden of collecting, processing, and analysing data would fall primarily on researchers. While these demands are not fundamentally different from existing practices, they may risk disadvantaging less well-resourced or technically skilled researchers.

A considerable drawback is that researcher-led data collection would largely remain limited to public and donated data, potentially limiting the diversity of research outcomes and failing to overcome any limitations associated with these forms of research.

Evaluation against stated criteria

Legal and ethical

6.18 Clarifying existing rules would help to address some of the uncertainties highlighted in Section 5 in relation to current data access practices. Researchers would have increased clarity around what data they can and cannot access, and services would have increased clarity around what data they must allow researchers to collect, and what data can be reasonably withheld.

6.19 Clarifying existing rules would likely lead to an increase in researcher-led forms of data collection, such as data donations and research-related scraping, as opposed to service-led data provision. If the scope of data donations increases, this may mitigate ethical risks around participant consent. If the scope of research-related scraping across services increases, services and researchers will need to comply with the data protection transparency principle,[202] where required. While such outreach may mitigate the issue of users' data being scraped without their knowledge or consent, there may be outstanding questions around whether users sufficiently understand how their data may be used by researchers.

[202] The transparency principle requires organisations to inform data subjects about how their personal data is collected, used and shared. For further information, see ICO guidance on transparency.

6.20 By clarifying existing rules, it would be possible to address legal and ethical concerns around either researchers or services not fulfilling their obligations, or breaching rules defined in the existing laws and regulations.

Security

6.21 Involved parties, such as data providers, data holders, data controllers and data processors, would all have increased clarity around their obligations regarding data protection and security and how to fulfil those obligations. It should be noted that this policy option would not introduce new data protection and security obligations on any party. However, if the scope of data donations and scraping increases as a result of clarifying existing rules, researchers will need to comply with data protection requirements, including security, in relation to larger volumes of data than currently.[203] In the case of data donations, data donors would also need to ensure they meet security and data protection obligations when receiving their data and sharing it with researchers, which may be challenging for those with limited resources and/or technical literacy.[204]

Technical feasibility

6.22 Because data collection under this policy option would be largely researcher-led, the largest technical burden would be placed on researchers during the collection, processing and analysing of the data. It should be noted that these burdens are unlikely to be substantially different from technical requirements researchers currently face, although their scale may increase in case of an uptick in data donations and scraping. This policy option could prove more advantageous to more technically capable and/or well-funded researchers who have the capacity for technically demanding data collection methods such as crawling and scraping and handling large volumes of data. This could create a landscape in which only the most technically capable researchers could access large volumes of data for research.

Cost

6.23 Under this policy option, the financial burden sits largely with researchers. This would include costs associated with developing tools, engaging in crawling and scraping activities, and receiving, cleaning, hosting and analysing data in a safe and compliant manner. In the case of data donations, researchers may need to factor in increased costs related to financial incentives for research participation. These costs are unlikely to be substantially different from costs researchers must already account for in their work, although they may increase should other avenues of data access become further restricted. This policy option could particularly advantage well-funded researchers who can afford to increase their existing research activities as a result of increased clarity. At the same time, increased clarity around the legality of various data collection methods may enable researchers to apply for funding to develop these capacities and conduct research using these methods.

Strategic effectiveness

6.24 The clarification of existing rules would not increase the provision of data directly from services, but it could improve other means of access that do not directly involve the services, such as data donations and public interest scraping. Researchers would be able to safely,

---

[203] For further information, see the ICO's guide to data security.

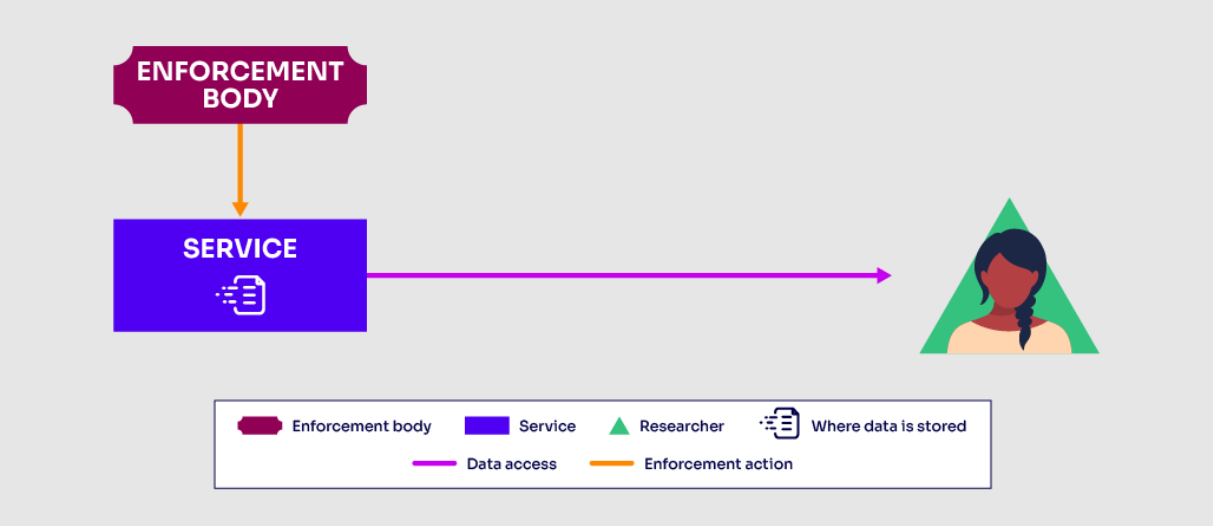
[204] In this regard, Article 32 UK GDPR requires a data controller to implement measures to ensure a level of security appropriate to risk, taking into account the state of the art, the costs of implementation, the nature of the processing and risk to rights and freedoms of natural persons.

legally and reliably access data they deem useful, as far as their resources allow. With better clarity around the legality of various methods, researchers would be better able to use innovative new methodologies, in addition to scaling existing methods. Some methods, such as scraping for public interest research, could allow near-real time analysis, supporting the monitoring of emerging harms, including around specific events such as elections or political crises.

6.25 This policy option is unlikely to substantially increase researchers' access to non-public data. However, it is possible that data which, due to caution in the face of perceived ambiguity,[205] is currently deemed sensitive may become more accessible and deemed in-scope, particularly for public facing data. Some stakeholders suggested that the sensitivity of data can depend on the context of it or the inferences made.[206]

Create new duties for services: Direct access

Figure 1 – A visual representation of the direct access model



6.26 The second policy option involves setting new regulatory requirements for data access by establishing obligations on services to permit access under certain conditions. For the purposes of this report, we refer to this approach as a 'direct access', given that the exchange of data does not directly involve an intermediary body. A regulator could be empowered to enforce these requirements. This approach would be similar to that set out in Article 40.12 of the Digital Services Act, in which very large platforms must provide data access to vetted researchers for systemic risk research.[207]

[205] Rupp. V, Von Grafenstein. M, 2024, Clarifying "personal data" and the role of anonymisation in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection, Computer Law & Security Review, 52. https://doi.org/10.1016/j.clsr.2023.105932 [accessed 23 June 2025]

[206] Forthcoming Ada Lovelace Institute publication, pp.6-7.

[207] Very large online platforms or of very large online search engines shall give access without undue delay to data, including, where technically possible, to real-time data, provided that the data is publicly accessible in their online interface by researchers, including those affiliated to not for profit bodies, organisations and associations, who comply with the conditions set out [defining a vetted researcher], and who use the data solely for performing research that contributes to the detection, identification and understanding of systemic risks in the Union pursuant to Article 34(1).

6.27 A direct access model typically requires services to set up and maintain a public portal for access to data. Basic approaches allow unrestricted access to public data without accreditation, removing the need for individual requests or third-party facilitation. Notably, direct access modalities that do not require researcher accreditation are limited to only providing public data. Examples include ad libraries, providing searchable databases, or APIs enabling bulk data analysis.

6.28 One approach would be to mandate that certain services make specific categories of data available for research purposes. For this to have meaningful effect, it would be sensible to target data that is presently largely inaccessible.

6.29 However, requiring services to allow direct access to all online safety relevant data could lead to unexpected negative outcomes for service users, such as reidentification[208] and associated risks, and services, such as potentially compromising services' security. A more targeted approach focusing on services with certain characteristics (such as user base, business model, or risk factors), may be more appropriate, though care is needed to ensure meaningful research diversity and scope is maintained. A direct access model would maximise its utility if it ensured access to datasets that are large and varied.

Summary of considerations for direct access

Making public data available through a direct access model can help reduce uncertainty and legal risk by providing a clear and formal mechanism for researchers to access information. Because the data flow is unidirectional (from the service to the public) and is intended for broad use, the security requirements are typically lower than those for systems handling private or sensitive data. This model also removes the need for complex accreditation processes, reducing the administrative burden on both researchers and services. Because a mandate could require that specific categories of data be made available for research purposes and services would not be required to produce tailored datasets in response to specific proposals, this could enable timely access to data. However, in the absence of an intermediary, dispute resolution risks being ineffective.

The effectiveness of this approach is dependent on how services define 'public' data. Even if data is openly viewable online, it may still be subject to restrictions under terms of service, data protection laws such as the UK GDPR, or copyright regulations. The compounding effect of these rules may cause services to take a cautious approach to data provision. Additionally, developing and maintaining direct access systems requires significant technical investment from services. We note that some services may have this in place already due to Digital Service Act requirements.

The policy option also has limitations in terms of research utility. Researchers have no control over what data is shared, when it is made available, or how it is structured. This lack of influence can reduce transparency, hinder independent verification, limit the methodological robustness of studies, and limit the utility of

[208] Large data, especially datasets with high dimensionality, carry an increased risk of reidentification. Detailed or unique information, when combined or cross-referenced with additional data sources, can reidentify people, even when direct identifiers are removed.

the data. Finally, because private or more detailed datasets are excluded, research based solely on public data may lack substantive depth and diversity.

Evaluation against stated criteria

Legal and ethical

6.30 Direct access may reduce perceived uncertainty and legal risk by providing formalised data access routes. However, the scope of data is often limited by services' interpretations of 'public' data. Freely viewable content may still face usage restrictions under services' terms of service, the UK GDPR, and copyright protections. Services have reservations regarding user privacy and the absence of explicit user consent, particularly given low user awareness of data-sharing practices. It should be noted that many services are able to give third parties access to users' data or insights derived from them, without concerns over data protection compliance or explicit per-instance user consent. Some have pioneered sophisticated technical build outs to enable this. These data access modalities and procedures are being used to allow for the delivery of user-relevant advertisements.

6.31 Even public data poses reidentification risks when combined with other datasets and could enable targeting or manipulation of individuals or groups. Consequently, services may choose to restrict the granularity of the data they provide, instead offering ranges and estimates. This is especially a risk in a direct access approach that does not require researcher accreditation and where no legal agreement (such as a non-disclosure agreement) is in place to protect the service against legal or reputational repercussions. These concerns can have an impact on data scope, coverage, detail level, and access duration, affecting the quality and reliability of longitudinal or comparative analysis.

Security

6.32 Services control security features of their data-sharing infrastructure, reducing perceived risk of breaches. Unidirectional data provision minimises risks by eliminating inbound data flows or direct queries from external actors.

6.33 The provision of public data may reduce recipient security obligations, with lower requirements for secure storage and oversight mechanisms. For example, data flows one way from the service to the public – services essentially broadcast or make available a predefined subset of their data, reducing the need for authentication systems and bespoke access controls. Secure storage or individualised oversight mechanisms for researchers are also less onerous given the accessibility of the data. Public data still carries a risk of reidentification and other sensitivities, particularly when combined with other datasets. As a result, services may still choose to aggregate, anonymise, or down sample data[209] to mitigate those risks, especially within a general access framework with no accreditation system in place.

[209] Downsampling is a common data processing technique that addresses imbalances in a dataset by removing data from the majority class such that it matches the size of the minority class. Source: Murel, J., 2024. What is Downsampling? What is downsampling? | IBM. [accessed 23 June 2025]

Technical feasibility

6.34 Direct access could require significant upfront technical resource for services to develop and maintain. Researcher costs vary depending on the delivery format and associated technical demands. For example, API access requires technical capacity for querying, extracting and processing the data. Some organisations may already possess these skills as a result of making use of accessible APIs or solutions mandated by Article 40 of the Digital Services Act and its associated technical requirements.

6.35 When data is provided in unextractable formats, such as static dashboards, aggregated reports, or non-machine-readable outputs, limited technical skills may be required to access it. However, the scope and depth of analysis would be constrained by the data format. Ultimately, the cost and accessibility of a direct access option for researchers will vary depending on the format, technical complexity, and level of standardisation adopted by services.

Cost

6.36 The cost of empowering a body to regulate access arrangements would likely be far less than establishing a formal intermediary body.

6.37 Access costs for researchers vary widely. Some services offer access for free, while others charge monthly or per-request fees ranging from hundreds to tens of thousands of pounds, often making them unaffordable. Although some researchers have received exemptions, this can create a two-tier system favouring those with prior agreements over those with valuable research or financial need. If access is regulated, pricing could be controlled or set to zero – placing the cost on services or distributed through a cost-sharing model. Technical complexity may also require hiring engineers or data scientists, and large-scale or long-term projects can incur significant cloud storage and computing costs, especially on teams with limited resources.

6.38 Storing and serving large volumes of data across multiple jurisdictions to fulfil various data access requirements may also require significant service investment before data release. Ongoing updates to tools and compliance with national and regional regulations require dedicated engineering and compliance teams. The complete cost of such data access tools will vary depending on the scale and scope of a direct access-based solution.

Operational effectiveness

6.39 In terms of the obligations on an enforcement body, the operational complexity of a direct access model is significantly less than establishing an intermediary. However, from a whole system point of view, for both researchers and services, the potential lack of standardisation or rule-setting could result in far greater operational complexity. In theory, the absence of intermediary-administered accreditation processes and elimination of data request applications could potentially improve operational efficiency. However, this reduces oversight and introduces new challenges around responsible data use, misuse prevention, and harms mitigation. Strained trust between services and researchers exacerbates operational challenges and reduces confidence that service-provided datasets and resultant insights are an accurate reflection of activity occurring on services.

6.40 Effective scaling requires automation, clear documentation, and consistent standards across services. Ongoing maintenance, user support, and feedback mechanisms would be essential to keep the option effective and responsive over time. The user experience should be simple and intuitive,[210] as poor interface design or unclear guidance could deter potential users, particularly research teams with limited resources.

6.41 The unidirectionality of data provision may limit transparency and research utility where researchers can only passively receive data with little to no control over what is shared, when or how. This may hamper independent verification and can undermine methodological rigour. Researchers cannot query the full universe of content or request custom datasets, limiting scope. Without an independent intermediary to facilitate a feedback mechanism or establish parameters or expectations for data provision, data provided could be of limited utility and this approach risks being ineffective. Additionally, data filtering by service policies may exclude harmful or borderline data that researchers need to understand the scope and evolution of harm and risk on a service. Furthermore, delays in data provision significantly limit the utility of data in sensitive political periods or developing crisis situations. Even with the inclusion of an accreditation function, there is a risk that researcher access could be revoked or restricted without an independent intermediary serving as mediator, resulting in disruption or complete blockage of research.

Strategic effectiveness

6.42 While direct access avoids the risk of access rejection, they are also likely to offer the most limited data modalities and richness of data. Direct access offers an opportunity to standardise how data is offered. However, the strategic effectiveness of a direct access approach depends heavily on how data is presented and structured. Utility of data can be constrained by limited search and filtering capabilities, unextractable formats that hinder trend analysis or comparison over time, and short-term storage that prevents the development of deep archives, complicating longitudinal research. Researchers would also lack visibility into service-side filtering, curation, or content moderation processes.

6.43 Insufficient metadata, especially around content targeting or audience engagement, can further limit analytical depth, with aggregated data often lacking the nuance required for meaningful insights. Additionally, the scope of available data may be narrow, as what constitutes 'public' data remains largely at the discretion of services. Coverage may also vary widely.

6.44 The regulator would likely not have the power to compel services to provide access to a specific dataset or accredit researchers. Instead, their role would be limited to ensuring that services adhere to their own access policies and that those policies comply with legal requirements. Whether data is made available often depends on services' priorities or third-party involvement in shaping what qualifies as relevant to online safety. Services would retain significant discretion to refuse access. This is something researchers have consistently flagged as a barrier.

6.45 While timely access may be possible, this will largely depend on the rules or oversight roles of third parties involved in the governance of any direct access approach. Without clear requirements, timely access remains uncertain.

---

[210] Van Drunen, M.Z, and Noroozian, A., 2024. How to design data access for researchers: A legal and software development perspective, Computer Law & Security Review, 52. https://doi.org/10.1016/j.clsr.2024.105946. [accessed 23 June 2025]

Establish independent intermediaries to oversee new duties on services

6.46 The third policy option proposes the introduction of an independent intermediary body to address many of the challenges currently limiting researcher access to services' data. This policy option draws on recommendations from the EDMO's report on Platform-to-Researcher Data Access, which found that the current conditions for researcher access are suboptimal and that an independent intermediary could play a critical role in trust and access.[211]

6.47 In this section, we explore three institutional forms for an intermediary:

a) Direct access intermediary
b) Notice to service intermediary
c) Repository intermediary

6.48 Each of the three intermediary models differ slightly, but all would retain a similar set of core functions, detailed in the following sub-section. Intermediary models are potentially the most effective at addressing the constraints on information sharing we have identified through our evidence gathering. An intermediary can help foster trust between services and researchers, clarify baseline rules and processes as well manage accreditation processes and mediate disputes between researchers and services. We assess each of the three intermediary models against the same set of criteria as we did for policy options 1 and 2.

6.49 An independent intermediary could act as a neutral, trusted third party between researchers and services. It would have the ability to develop and maintain deep expertise across a range of subject matters relevant to researcher access. This intermediary would provide a means for dispute resolution to occur before any enforcement action is needed. This is more likely to foster trust and collaboration between researchers and services.

6.50 A highly effective intermediary would be a specialised data-centric organisation. Research into how parties organise around data notes that robust data governance is needed to realise the most value from data. Data-focused organisations are typically better equipped to establish this governance and manage data more generally.[212]

[211] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p2. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]

[212] Benfeldt, O., Persson, J.S., and Madsen, S., 2020. Data Governance as a Collective Action Problem. Information Systems Frontier, 22. https://doi.org/10.1007/s10796-019-09923-z. [accessed 23 June 2025]

6.51 This intermediary would have the credibility and contextual knowledge of existing research and research needs to support the development of governance frameworks and normative practices. A recent US-UK report on enabling data sharing noted that the starting point of many data access challenges is "not technical but cultural. Problems posed by data access therefore involve the tools of the social sciences, including norms, codes of conduct, incentives, disincentives, and governance". Technological solutions are years away from widespread implementation, making normative frameworks necessary for any implementation of access. [213]

6.52 A highly effective intermediary would have detailed understanding of data-protection practices and law. It would have the technical expertise to both manage and maintain data flows and any needed technical infrastructure, to assess the technical infrastructure and needs of both services and researchers, as well as to understand emerging trends in service design and usage, security and privacy enhancing technologies.[214] It would also have deep understanding of, both established and emerging, research methodologies and ethics, as well as the current state of online safety research globally and any gaps in that research ecosystem. An intermediary body could conduct several additional activities beneficial to the research ecosystem as outlined in Annex 5.

6.53 Intermediary models could reduce reliance on riskier alternatives, such as scraping, by providing data within a well-defined framework that can help ensure data protection compliance. This reduces legal uncertainty, though researchers may still choose alternative methods should access not be sufficiently timely or the data fit-for-purpose, so the models do not eliminate all associated risk.

6.54 The intermediary could play various roles, including managing governance, establishing standards, providing access, and managing dispute resolution before enforcement is taken.

6.55 Establishing an intermediary, regardless of its form, could be costly and operationally complex. It could require an interdisciplinary team that includes researchers, data scientists, legal experts and governance professionals to manage access, ensure compliance and foster trust. These operational demands and overhead costs related to potential infrastructure and security would require sustained investment and careful planning to ensure the regime is effective.

[213] National Academies of Sciences, Engineering, and Medicine. 2024. Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. https://doi.org/10.17226/27520. [accessed 23 June 2025]

[214] "Establishing an intermediary third-party organisation is one way to ensure that a researcher access framework remains effective, adaptable, and aligned with the evolving nature of online harms and platform technologies." Source: Bartosz, M. and Pavel, V., 2025. Potential unreached: challenges in accessing data for socially beneficial research. Potential unreached: challenges in accessing data for socially beneficial research | Ada Lovelace Institute. [accessed 23 June 2025]

6.56 In the absence of an intermediary, forums within civil society and academia have organised to provide forms of intermediary functions with regards to researcher access to address challenges, such as a lack of common data standards and shared infrastructure.[215] [216] [217] [218]

6.57 The relationship between any future intermediary and regulators with an interest in online safety matters requires careful consideration. In their report on online targeting, our statutory consultee, the Centre for Data Ethics and Innovation, noted that any theoretical online safety regulator may not necessarily be best placed to make the final decisions on researcher access to data on online safety matters. It recommended considering "designating an expert independent third-party organisation to make decisions about data access".[219]

Core functions of an independent intermediary

6.58 In this section, we explore the core functions of an independent intermediary in more detail. We do not give a view about what kind of organisation would be best placed to take on the role. The most important attribute is that the organisation has the capacity and expertise, as described above, to discharge its functions independently and without conflicts of interest.

Accreditation

6.59 The intermediary could accredit researchers, institutions and, where applicable, individual research proposals. Researcher accreditation could be based on factors such as institutional affiliation, data-handling credentials, and previous conduct. Proposal accreditation could be based on factors such as the subject matter, access methodologies, security and data protection processes, and future potential for misuse. If a service denies an accredited researcher's request, they would be required to provide detailed justification.[220] Researchers could appeal to the intermediary to adjudicate, who may suggest alternative data modalities or modified data requests.

6.60 However, given that individual proposals are not assessed, additional safeguards would be needed to limit access to sensitive personal data, such as special category data protected under Article 9 of the UK GDPR (see Annex 1 for more information about the current data protection framework in the UK). Safeguards could be tailored to 'tiers' of access, where access to sensitive personal data would require more robust vetting by the third party.

---

[215] Mattioli, M., 2017. The Data-Pooling Problem, Berkeley Technology Law Journal, 3(1), http://dx.doi.org/10.2139/ssrn.2671939. [accessed 23 June 2025]

[216] For work currently undertaken in this space, see UK Research and Innovation's Social Platforms Data Access Taskforce.

[217] Wanless, A. and Shapiro, J. N., 2022. A CERN Model for Studying the Information Environment. A CERN Model for Studying the Information Environment | Carnegie Endowment for International Peace. [accessed 23 June 2025]

[218] For work currently undertaken in this space, see Columbia-Hertie Working Group on Building Capacity for Data Access, Analysis, and Accountability.

[219] Centre for Data Ethics and Innovation (CDEI), 2020. Online targeting: Final report and recommendations. Online targeting: Final report and recommendations - GOV.UK. [accessed 23 June 2025]

[220] For further discussion on researchers' perception that access requests under DSA Article 40 can be rejected without detailed explanations, please see: Weizenbaum Institute, 2024. Enabling Research with Publicly Accessible Platform Data: Early DSA Compliance Issues and Suggestions for Improvement, p.7. Enabling Research with Publicly Accessible Platform Data: Early DSA Compliance Issues and Suggestions for Improvement. [accessed 23 June 2025]

Rule-setting

6.61 The intermediary could establish baseline rules and standards for data access. These rules would define what constitutes a 'reasonable' access request, set clear expectations for both services and researchers regarding data access modalities, reduce ambiguity, and promote consistency.

6.62 This role avoids one of the common issues in multi-party data governance arrangements – the risk of data standards "encouraging over-protectionism".[221] Research has indicated that standards development faces a dilemma as at least one party must take the lead on the adoption of joint standards. However, each party is incentivised to wait for another to develop and socialise these standards. This can lead to inaction as a group.[222] [223]

6.63 The rules framework could include specifications on data types, data minimisation principles, use of privacy enhancing technologies, and standards for evaluating proportionality and privacy risk. Transparent, codified rules would help reduce the compliance burden on all parties and aid mediation in the event of a dispute. These would need to be developed so that they remain relevant to future research requests, and in a way that balances researcher access with service interests. If requests fall outside of the existing framework, or there is a dispute regarding access, the independent intermediary could adjudicate on a case-by-case basis.

Mediation

6.64 Mediation may be required at various points, throughout the research project lifecycle.[224] The importance of this role is reflected in numerous Call for Evidence responses.[225] [226] [227] [228]

6.65 A trusted mediator can help resolve disputes[229] without resorting to litigation. Mediation via lawsuit risks a power imbalance between relatively well-resourced services and the relatively legally under-resourced researcher community. The real or perceived risk of legal action and the cost of that process, regardless of the legal outcome, could impact on public participation by researchers. There have been allegations of some services engaging in Strategic Lawsuits Against Public Participation (SLAPPs).[230] Mediation may be necessary to address allegations that the service is using legal action to intimidate researchers or public participation. An established mediation process involving initial protections from these costs

---

[221] Bambauer, J.R., 2011. Tragedy of the Data Commons, Harvard Journal of Law and Technology, 25, http://dx.doi.org/10.2139/ssrn.1789749. [accessed 23 June 2025]
[222] Mattioli, M., 2017. The Data-Pooling Problem, Berkeley Technology Law Journal, 3(1), http://dx.doi.org/10.2139/ssrn.2671939. [accessed 23 June 2025]
[223] Bambauer, J.R., 2011. Tragedy of the Data Commons, Harvard Journal of Law and Technology, 25, http://dx.doi.org/10.2139/ssrn.1789749. [accessed 23 June 2025]
[224] Mediation mechanisms are discussed in further detail in Annex 4.
[225] Name Withheld 1 response to October 2024 Call for Evidence, p.9
[226] National Society for the Prevention of Cruelty to Children (NSPCC) response to October 2024 Call for Evidence, p.12.
[227] Open Data Institute response to October 2024 Call for Evidence, p.9.
[228] [✂]
[229] Disputes and mediation mechanisms are discussed in further detail in Annex 4.
[230] SLAPPs are "legal actions typically brought by corporations or individuals with the intention of harassing, intimidating and financially or psychologically exhausting opponents via improper use of the legal system". Source: Multiple HMG Departments, 2024, Economic Crime and Corporate Transparency Act: strategic lawsuits against public participation (SLAPPs), Economic Crime and Corporate Transparency Act: strategic lawsuits against public participation (SLAPPs) - GOV.UK. [accessed 23 June 2025]

to researchers by a body that is obliged to also consider services' legitimate interests could partially address this concern.

6.66 We recognise that there are a range of additional enhancing functions that could further support any effective researcher access regime and that go beyond the core enabling functions mentioned earlier. We discuss these in greater detail in Annex 5.

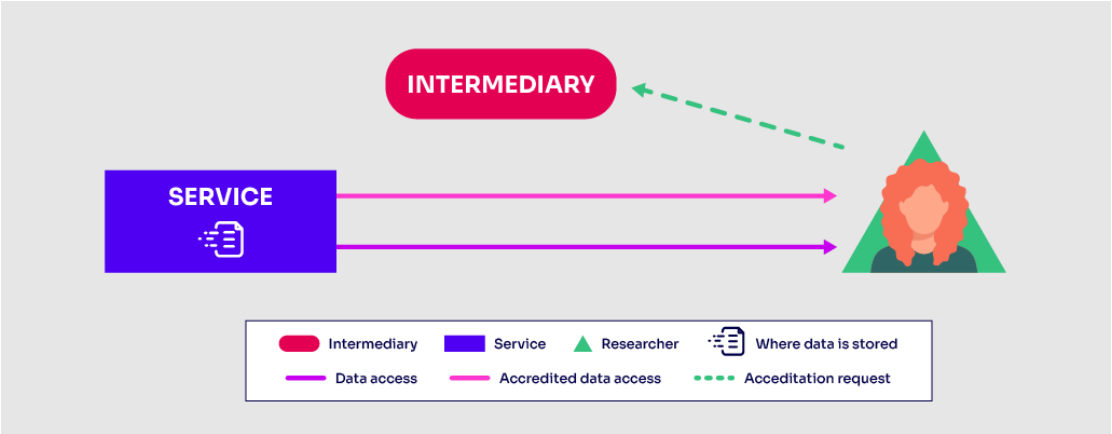
6.67 The following sections introduce and evaluate the three models that make use of an intermediary: Direct access intermediary, Notice to service intermediary and Repository intermediary.

Direct access intermediary

6.68 Under this model, the intermediary provides the access interface while services retain data hosting (similar to a virtual repository using managed APIs). The intermediary would:

- Set and enforce eligibility criteria
- Vet researchers and proposals
- Collaborate with services on data formats and safety protocols
- Facilitate data access for researchers

Figure 2 – A visual representation of the direct access intermediary model



Summary of considerations for direct access intermediary

A direct access with an intermediary model largely mirrors the considerations of those involved in direct access (policy option 2) and would apply to access to the same data. These include allowing services to define what qualifies as 'public' data, the significant technical burden on services to develop and maintain access systems, and the exclusion of private or otherwise sensitive datasets – which can limit the depth and diversity of research.

Unlike direct access without an intermediary (policy option 2), in this model, a neutral intermediary facilitates data access between services and researchers. This presents several advantages: it introduces a mechanism for resolving disputes, centralises the process of accrediting researchers, which may also help filter out potential bad actors. Additionally, the intermediary can establish baseline rules and standards for data access, reducing ambiguity and setting clear expectations for all parties.

The data and data access modalities in scope would be similar to those present in the direct access policy option. However, intermediaries' vetting processes and the setting of data formats and safety protocols could increase the likelihood of successful and safe access for researchers. As some data access would likely involve ultra-low risk data types, researchers could still access data from some services without the need for an intermediary-led accreditation process.

Legal and ethical

6.69 Services would remain liable for the management of the data. This may mean that services' concerns around the potential legal risks of data sharing, identified in the previous section, are addressed to a lesser extent than under models that would see them directly transfer or automate access to the same data. Should services choose to transfer data directly to researchers, researchers would still take on legal and ethical risk by hosting data. Legal risk is not eliminated. Despite best efforts, the intermediary and/or services could still unintentionally undertake actions that leave them liable for any resulting harm.

Security

6.70 Services maintain existing hosting and security obligations without significant additional requirements because most already meet standards such as ISO 27001.[231] Services are also required to meet data protection standards and ensure the necessary security protocols are in place. The intermediary does not host data directly, making it less attractive to bad actors and a lower security risk. Services may need to scale operations if required to collect additional data not currently held.

[231] ISO/IEC, 2022, 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements, International Organization for Standardization, ISO/IEC 27001:2022 - Information security management systems. [accessed 23 June 2025]

Cost

6.71 Whilst we have not undertaken detailed costing of this model, we would expect that the financial burden in this model is placed more on the service than the intermediary. This is because the costs associated with collecting, processing and hosting the data (the role of the services) are higher than the costs associated with providing an interface through which the data is accessed (the role of the intermediary). The intermediary could still incur costs related to hosting APIs. While these costs to the intermediary are lower compared to repository variants – as they would only relate to data that services make available to researchers via a self-hosted-public portal, including those that go through the intermediary – they still represent an increase from the status quo. Intermediaries are not responsible for combining datasets from various services or any other robust cleaning or packaging of data. The costs to services could be significant, depending on the specific role of the intermediary.

Operational effectiveness

6.72 Services assume the operational burden for data hosting and maintenance, while the intermediary's burden is lower compared to other potential models.

6.73 The intermediary has a central interface role, enabling system-wide improvements and maintaining its neutral position between services and researchers. This is unlikely to have an impact on the operational burden on researchers. Under this model, the independent intermediary could still provide the centralised interface through which researchers access the data, despite not hosting the data.

6.74 An intermediary would provide a range of benefits. It would mean there is a single point of contact to which researchers and services could appeal to resolve any dispute and to refine data access modalities and research applications. It could set data standards in light of the research community's needs and requirements, while bearing in mind security, data ethics, and compliance with data protection law. As the single point of contact, it could assess, accredit, and approve researchers, allowing them to access APIs provided by services at a non-commercial rate while assuring non-commercial downstream usage of the accessed data.

Strategic effectiveness

6.75 The intermediary would not have the power to compel services to provide access to a specific dataset. Instead, their role would be limited to ensuring that services adhere to their own access policies, that those policies comply with legal requirements and that researchers and services benefit from a neutral third-party that facilitates efficient direct access.

6.76 Similar to the direct access model without an intermediary, this model is likely to offer limited data access modalities and depth. While it could allow services to standardise their data offerings, its strategic effectiveness depends on how data is structured and presented. Constraints such as poor search tools, unextractable formats, short storage timing and lack of visibility into service-side filtering, curation or content moderation processes could limit research value.

6.77 Limited metadata could further limit analytical depth. Whether data is made available often depends on services' priorities or third-party involvement in shaping what qualifies as relevant to online safety. Additionally, the scope of available data may be narrow, as what constitutes 'public' data remains largely at the discretion of services, potentially leading to inconsistent coverage.

6.78 It is worth noting that services retain significant discretion to refuse access under a direct access model without an intermediary. By accrediting researchers, an intermediary could provide a standardised and transparent basis for access decisions.

6.79 Researchers could have the ability to access insights from service data in a way that reflects and responds to the datasets that exist on the service. The intermediary could provide assurance to researchers of what datasets exist, what constitutes a proportionate request of services, and how research proposals could be better crafted to work with the data that is readily available.

6.80 This model does not presume live direct transfer of data but a medium of enabling researchers to access data that services could choose to make available.[232]

6.81 However, there are some limitations with this approach. Services retain significant discretion in terms of what data is made available and how data access is provided. Because the data is not stored locally with an intermediary, changes to service design, the make-up of internal datasets and the functioning of APIs could disrupt how this data access modality functions. Researchers conducting longitudinal studies could potentially find their research disrupted as a result.

Notice to service intermediary

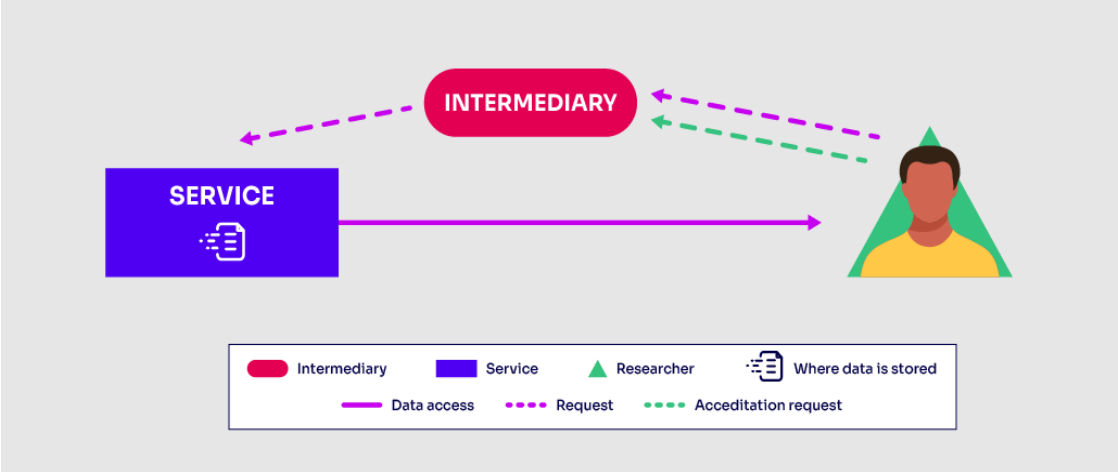
6.82 Under this intermediary model, researchers request access to specific datasets via intermediaries. Notably, this access request could involve public data, private data and/or special category data (which can be public or private). Researchers could notify services of their credentials, research proposal, and preferred data access method (data access modality) via an intermediary. Both the intermediary and the relevant service would assess proposals, with the intermediary acting as a mediator if disputes arise. This is the most similar to the elements of the current model in place under the Digital Services Act (Article 40.4) framework.

6.83 The process could work as follows:

- Researchers submit proposals detailing the required data types and access modality (for example, TREs, federated queries, or direct transfers).
- The intermediary vets researchers and proposals.
- Services assess requests based on their interests (considering things like confidential information, trade secrets, and service security).
- Access is granted or refused through collaborative determination.
- The intermediary mediates disputes when necessary.

[232] National Academies of Sciences, Engineering, and Medicine report notes that this sort of access would enable researchers to use their "own domain-specific standards, vocabulary, [and] semantics" in a way that is context relevant to the services that they query. Source: National Academies of Sciences, Engineering, and Medicine. 2024. Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. https://doi.org/10.17226/27520. [accessed 23 June 2025]

Figure 3 – A visual representation of the notice to service intermediary model



Summary of considerations for notice to service intermediary

The notice to service intermediary model allows researchers to request access to specific datasets, offering a more tailored and flexible approach to data sharing. By using bespoke data usage agreements that are customised to each request, this model can help clarify legal responsibilities and liabilities for all parties involved. It also builds on existing capabilities, as many services already operate data-driven businesses with established security measures, reducing the need to create entirely new frameworks from scratch.

For services already subject to the Digital Services Act or those with voluntary access systems in place, the transition to this model may require only minimal adjustments. However, other services may need to invest in significant new infrastructure to support secure and compliant data access. Importantly, this model expands the scope of research by enabling access to a broader range of data than what is available through existing public portals, allowing researchers to explore more diverse and complex questions.

Despite these advantages, this model introduces greater complexity compared to the other policy options and models discussed. Managing access and ensuring the security of both the service and the data subjects can be resource-intensive, potentially placing a financial and administrative burden on both services and researchers. It may also introduce limitations on the ability to conduct timely exploratory research outside of specific research questions. Each access request would require individual vetting of the proposal and the researcher, which could lead to delays and increased costs.

Additionally, when dealing with more sensitive data, services must address more demanding requirements related to data cleaning, legal compliance, and security. The availability and implementation of privacy-enhancing technologies will also be more critical in this context. An important factor in the success of this model will be the design of the access portal itself – particularly its usability. A well-designed, user-friendly interface will be essential to ensure that researchers can navigate the system efficiently and effectively.

Legal and ethical

6.84 This approach could enable bespoke data usage agreements tailored to specific access requests, which could help clarify liability for all parties involved. Data use agreements (DUAs), also referred to as data sharing agreements or data use licenses, are documents that describe what data are being shared, for what purpose, for how long, and any access restrictions or security protocols that must be followed by the recipient of the data. The approach benefits from the shared GDPR framework within the EU, allowing reference to the experiences of Digital Service Coordinators (DSCs) in implementing Article 40 of the Digital Services Act when addressing legal and ethical challenges.

6.85 Over time, as more agreements are facilitated, the involved parties can develop standardised data access agreements and data modality arrangements (such as templates), reducing the burden for both researchers and services.

6.86 An intermediary could also maintain in-house legal expertise to create standardised drafts for different data access modalities and provide custom drafting for complex research requests.

Security

6.87 Services retain control over security features and data sharing infrastructure, which partially addresses their concerns about security risks. Notice to service models can allow services to refuse requests on security grounds, while data usage agreements can specify security requirements for both access methodology and downstream data use. Since services already operate data-driven businesses with existing security measures, this model builds on established capabilities rather than requiring entirely new security frameworks.

6.88 Data-sharing infrastructure would be set up and run by services, who would therefore have control of security features, alleviating in part their perceived risks concerning security breaches. Any data inventory and data format information would be developed in consultation with and be in the interests of both services and researchers but would have to be developed with the security concerns of the services in mind.

Technical feasibility

6.89 Services and intermediaries need technical infrastructure to process researcher applications, including user-friendly portals with data inventories setting out available datasets and capability. Services could build out the technical infrastructure to process queries in-house, reducing the burden on researchers. Those already subject to Digital Services Act obligations or with existing voluntary access systems may need minimal changes, while others may require substantial new infrastructure development. This creates uneven technical requirements among researchers, where those experienced with Digital Services Act-based access or similar agreements have significant advantages over researchers without such backgrounds. The details of how much change is likely to be required would depend on the specific enablers or responsibilities attached to the model and which services are deemed to be within scope.

6.90 Researchers' technical requirements are also likely to be uneven. Those with experience working in either Digital Services Act-based access, researcher access in other sectors, and/or voluntary service led-researcher access agreements are likely to already have the technical skills, and possibly the technical infrastructure, to enable direct access. On the other hand, researchers without these backgrounds or tools may be at a relative disadvantage.

Cost

6.91 Whilst we have not undertaken detailed costing of this model, we would expect that because services already bear data storage and management costs as part of normal operations and likely have preexisting datasets that could be accessed, this model could potentially be less expensive than directly transferring data. Services maintaining data inventories under Article 40 of the Digital Services Act face reduced additional costs due to existing compliance infrastructure.[233]

6.92 However, the management of the access and security of the service and data subjects could also be a considerable cost to services and/or researchers. The vetting of each proposal and the researcher(s) would incur costs to services and delays to researcher(s) upon each request.

6.93 The intermediary would bear the cost of processing the notice to service requests and retaining and resourcing the necessary expertise to effectively manage the process. Due to the wider data potentially in scope, this could be more resource intensive than direct access models with or without an intermediary.

Operational effectiveness

6.94 This model relies heavily on the intermediary's ability to accurately assess what data services can provide. This includes understanding technical constraints, legal boundaries and operational realities. Failure to align with services' capabilities and data protection obligations could result in delays and disputes.

6.95 Due to the potential diversity of data requests, this model is more complex and likely to involve lengthier negotiations between parties. Effective rule-setting by the intermediary could address some of these challenges.

6.96 Another important factor in the operational effectiveness of any notice to service regime would be the design of any access request portal, particularly with regards to the ease of usage or 'user-friendliness'. Commentators on this subject have noted the need for any access modality to incorporate certain software development practices from services, particularly with respect to their data access modalities focused on end user needs.[234] How these user experience (UX) principles could be implemented or enforced would depend on the specific variant of the notice to service model, the presence and power of any independent intermediary body, and the details of any regulation.

6.97 Data cleaning and organisational efforts typically fall to services or intermediaries due to legal, ethical and security concerns, and the availability of certain PETs.

6.98 Without an intermediary, services must handle researcher vetting, which can be burdensome and inconsistent. Centralised intermediary accreditation reduces service workload but requires significant expertise and resources.

---

[233] Commission Delegated Regulation (EU), supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data pursuant to Article 40 of Regulation (EU) 2022/2065, Ares(2024)7652659, art 6. Delegated Regulation on data access provided for in the Digital Services Act. [accessed 23 June 2025]

[234] Van Drunen, M.Z, and Noroozian, A., 2024. How to design data access for researchers: A legal and software development perspective, Computer Law & Security Review, 52. https://doi.org/10.1016/j.clsr.2024.105946. [accessed 23 June 2025]

6.99 Similar elements of this model are in place under the Digital Services Act (Article 40.4) framework. Services already enabling access under Digital Services Act obligations or with existing voluntary access systems may require minimal operational changes, while others may require substantial new infrastructure development. Researchers already accessing data under the Digital Services Act would likely be familiar with these processes. For further explanation of this regime, see Section 4.

Strategic effectiveness

6.100 This model enables researchers to explore a wide range of topics by expanding access beyond public portal data. Access to richer and more varied datasets and data modalities enables better investigations and supports more nuanced analysis of online safety matters. This could provide stakeholders with more accurate knowledge to inform regulation, policy and further innovative research that advances the state of the art.

6.101 An intermediary can also incentivise more strategic research by prioritising certain questions, enabling longitudinal and 'cross-platform' studies, and setting standards for data access modalities to address the 'collective action issue' present in data dilemmas.[235]

6.102 However, strategic impact depends heavily on the specific model, data access modalities and intermediary, as limited intermediary involvement leaves significant discretion with services to reject access requests, potentially undermining broader policy objectives. Repository intermediary

6.103 In a repository model, an independent intermediary requests specific data from services which are then responsible for granting access to that data. Depending on the format of the repository, it may also host and/or store the data locally or virtually. The intermediary would facilitate the sharing or transfer of requested data to researchers who meet specific eligibility criteria, either through direct transfer or through a data access modality such as a TRE (for more information on TREs, see Annex 3). The intermediary handles managing, securing and processing the data while setting rules for data transfer and data access. The independent intermediary vets and accredits researchers, their project proposals, and the requested data types to determine repository access eligibility. Data formats, safety protocols and other standards are established collaboratively between the intermediary and services to ensure the safety, security and usability of the data.

6.104 Repositories could take a virtual or local format:

- Virtual repository: A data-focused intermediary responsible for centrally hosting access to data but likely not permanently locally storing data within the repository.
- Local repository: A data-focused intermediary responsible for locally stored data within the repository.

6.105 A local repository would host and store data provided by services. Data could be transferred to the repository at regular intervals. The intermediary would work with the services to determine which data would be provided, in what formats and according to which data standards. It would also have the ability to review the data before making it accessible to researchers, based on an assessment of both the researchers themselves and their proposed research.

[235] Benfeldt, O., Persson, J.S., and Madsen, S., 2020. Data Governance as a Collective Action Problem. Information Systems Frontier, 22. https://doi.org/10.1007/s10796-019-09923-z. [accessed 23 June 2025]

6.106 A virtual repository would provide the technical infrastructure needed to enable access to data that remains hosted on the services themselves. It would still have the capacity to undertake some form of vetting process for researchers and their proposals before granting access and would consult with services about which data is made available and accessible through the repository's data access interfaces. Unlike the Direct Access model (Option 2), this approach would likely support access to data that cannot be shared directly due to its more security-focused technical infrastructure.

Figure 4– A visual representation of the local repository model

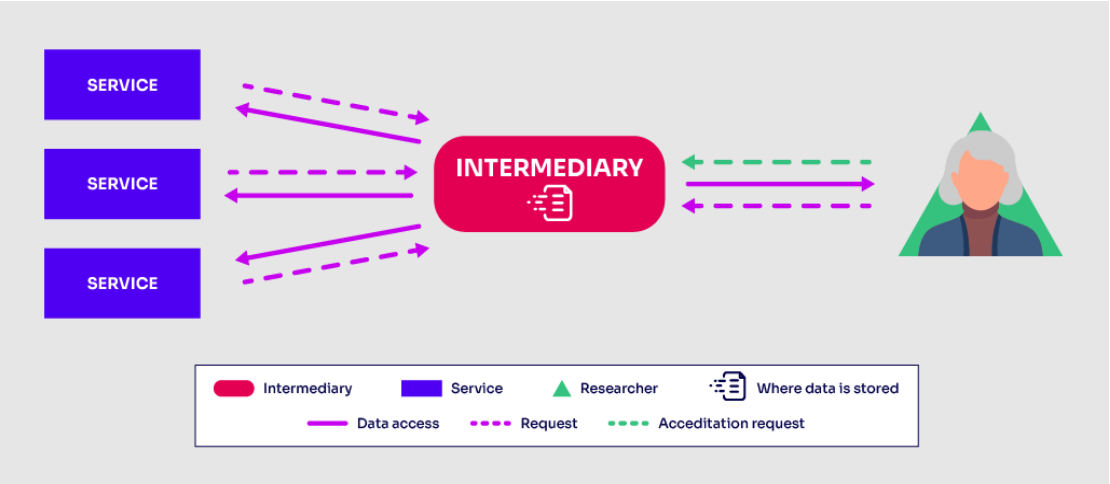
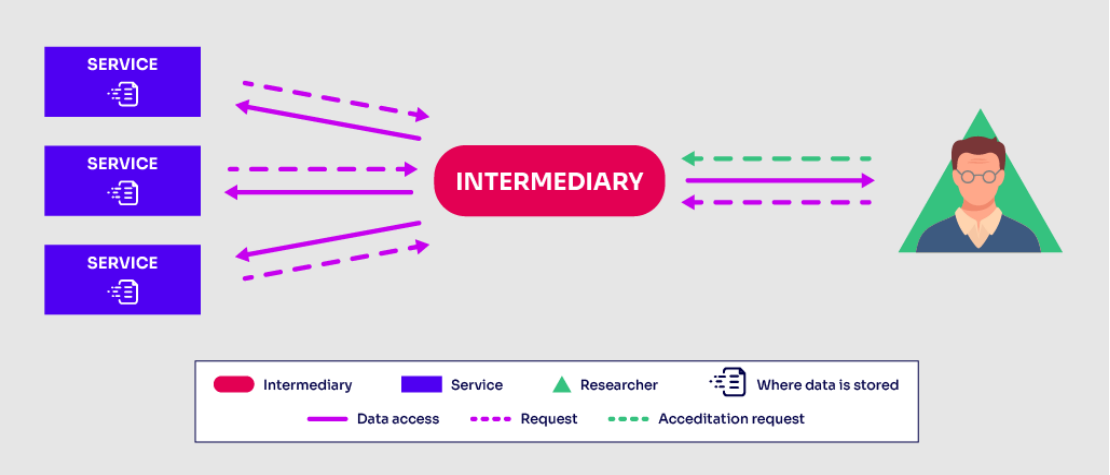


Figure 5– A visual representation of the virtual repository model



Summary of considerations for repository intermediary

The repository models are potentially the most complex and costly models but could offer major advantages through centralisation and provide the greatest flexibility for researchers. They allow for the setting of consistent data standards, streamlined access management, centralised security assurances, and the establishment of a single point of contact between services and researchers. These models support online safety objectives by reliably supplying researchers with data from a wide range of services. By requiring the delivery of data in a standardised format, they also enhance comparability and enable more effective cross-service research, helping to address many current research gaps.

An advantage of a local repository is its reduced reliance on any individual service. If a service withdraws access, the data already held in the repository remains available, helping to prevent data loss. This model also eases the burden on services by enabling regular transfers at agreed intervals, using pre-agreed standards and formats. However, this regular transfer schedule can limit the ability to conduct real-time research, and any data quality issues present in a given transfer will persist until at least the next scheduled transfer.

A virtual repository offers researchers the benefit of a centralised access process without requiring permanent local storage of data, reducing risk for the intermediary. Virtual repositories can more readily facilitate real-time access to data by enabling researchers to query information hosted on services via intermediary interfaces without the delays associated with periodic data transfers. However, this model is more vulnerable to disruptions caused by changes in service design or operations, which could adversely impact longitudinal research.

Virtual and local repositories require significant resource to develop, maintain and secure, although virtual repositories generally involve lower infrastructure and operational demands. Hosting data in local repositories introduces additional security and data protection risks, as it requires storing and managing sensitive information on-site, increasing the potential for breaches. In contrast, virtual repositories reduce these risks by keeping data within the original service environments, though they still require strong access controls and secure data transmission protocols.

Despite the procedures an independent intermediary may implement, assuming it is equipped with adequate security to protect services' data, there remains an inherent risk of downstream security breaches stemming from how researchers use or interpret the data. To mitigate these risks, a model would require advanced, customised security infrastructure. Its overall effectiveness would largely depend on whether researchers access data directly or through privacy enhancing technologies. Additionally, services would need to dedicate staff to support a repository's operations. This includes tasks such as data collection, processing, and regular transfers, all while ensuring compliance with security protocols, data protection regulations, quality assurance, and formatting standards.

Legal and ethical

6.107 The intermediary assumes legal responsibility as both data controller and processor for data protection purposes, potentially alleviating services' concerns about the legal risk from data sharing, while placing that legal risk on the intermediary itself. While this may help alleviate some of the concerns services have raised around the potential legal risks of data sharing, concerns about downstream use of the data may remain. No matter the procedures such an independent intermediary may put in place, there would always be a risk of potential downstream security breaches stemming from researchers' use and interpretation of the data.

6.108 Success requires codifying rules around data access and use, establishing clear roles and responsibilities for services, intermediaries and researchers. The intermediary could be empowered to set access parameters according to legal principles or alternatively issue guidance for parties to enter data-sharing agreements themselves. Such frameworks must account for multi-jurisdictional legal complexities while avoiding disproportionate application process complexity. As with notice to service, data usage agreements may address some downstream usage concerns.

Security

6.109 The repository enables implementation of security measures and data protection safeguards at scale, but its centralised nature creates reidentification risks and presents an attractive target for malicious actors. The intermediary must implement PETs and data protection best practices, while collaborating with services and researchers to establish security requirements.

6.110 Due to the centralised nature of the repository, there may be additional concerns around the risk of reidentification of data subjects, as data from multiple sources is hosted and provided in one place. This can mean both inadvertent reidentification in the course of legitimate research and purposeful reidentification by malicious actors. Due to this, the independent intermediary may need to take additional steps to proactively reduce the chance of reidentification. This would likely involve the use of various PETs and setting rules around data standards (for example, requesting that data is pseudonymised by default). While the independent intermediary would assume responsibility for implementing these measures, it would need to collaborate with services and researchers as well as rely on the relevant data protection framework to establish the details of what security measures must be in place. We discuss PETs in more detail at Annex 3 of this report.

6.111 Security effectiveness depends significantly on whether data is transferred directly to researchers or accessed through a secure data access modality. Secure access modalities reduce breach avenues and mitigates researchers' resource constraints for secure data management, while direct transfer requires additional vetting of researchers' security capabilities and clear liability frameworks for data misuse. These could include the use of PETs, data protection best practices as set out in existing frameworks such as the zero trust architecture design principles,[236] the Caldicott principles,[237] the 5 Safes framework,[238] and/or

[236] National Cyber Security Centre, 2021. Zero trust architecture design principles. Zero trust architecture design principles - NCSC.GOV.UK. [accessed 23 June 2025]
[237] National Data Guardian, 2020. The Caldicott Principles, The Caldicott Principles. [accessed 23 June 2025]
[238] UK Data Service, What is the Five Safes framework? What is the Five Safes framework? [accessed 23 June 2025]

data deletion requirements. Clear data use agreements can clarify these requirements and outline possible penalties for failure to comply, requiring researchers to agree to these terms before accessing and/or receiving the data.

6.112 While both local and virtual repositories carry risks associated with centralised researcher access modalities, a local repository carries a higher security burden. This is primarily because they store data on-site for extended periods, even when specific access modalities are not in use. As a result, there is increased scope for bad actors to access data in an unauthorised manner beyond posing as a legitimate researcher.

Technical feasibility

6.113 This model requires significant technical infrastructure development and maintenance from the intermediary, with less utilisation of existing technologies compared to other policy option or models. However, once established, it places relatively low technical burdens on researchers and services.

6.114 A repository could need to build bespoke technical infrastructure to receive, process, host and provide the data, including potentially systems for data transfers from services. The repository model would require significant technical infrastructure to support data transfers from the services to the repository, and possibly from the repository to the researchers. In the case of a local repository, using transfers of pre-agreed data types in pre-agreed data formats, the technical burden can be reduced with technical requirements established upfront with minimal modification needs.

6.115 For any repository the intermediary would likely need to standardise data formats to enable researchers to use diverse datasets effectively for comparative research. Given the centralised nature of the repository, and the existing challenge of researchers struggling to compare data from different services, the independent intermediary would likely also need to standardise data formats as much as possible to ensure researchers are able to use a wide array of different datasets for the same research.

Cost

6.116 Whilst we have not undertaken detailed costing of this model, we would expect that building and maintaining a repository could involve substantial startup and running costs, primarily related to security measures and, in the case of local repositories, the centralised storage of large data volumes. There may be cost minimisation strategies including standardising data deposit intervals and formats to enable operational planning and reduce processing variations. Nonetheless, the cost of implementing adequate security to protect services' data could be high.

6.117 The intermediary benefits from economies of scale through centralisation, making PETs deployment less expensive overall than requiring individual services to build separate systems. Services and researchers face comparatively low financial burdens since the model utilises existing data collection and provision infrastructure, though some costs remain for safe data management. The model does not resolve resource constraints for researchers facing complex vetting processes, though this challenge is not unique to this approach. However, a full analysis of potential costs is beyond the scope of this report.

6.118 A local or virtual repository model would likely impose comparatively low financial burdens on services and researchers. This is because a repository model typically does not require any specific infrastructure or additional mechanisms beyond those already used by services to collect or provide data (for example, to comply with the Digital Services Act), or those

used by researchers to access or receive data from various sources. An exception to this could be made if the independent intermediary were to recoup their costs by, for example, imposing an access fee on researchers. However, some financial burden on both services and researchers may be unavoidable, primarily due to the costs associated with safely and securely collecting, managing, receiving, and analysing data.

6.119 A repository model would not, by itself, resolve the issue of certain researchers lacking the resources to meet the requirements of complex vetting and application processes. Depending on the sensitivity of the data, the independent intermediary would likely require researchers to demonstrate their ability to safely and securely access, manage, analyse and/or host the data. These requirements and implied costs might be a barrier to this for some researchers. However, this is not unique to repository models.

Operational effectiveness

6.120 The intermediary facilitates relationships between services and researchers, potentially mitigating historical tensions and inspiring greater collaboration. However, repository models lack an inherently independent mediation mechanism, as the repository itself is often the party involved in disputes – whether with researchers requesting access or services granting access. This dual role can become problematic, particularly in disputes between the intermediary and either a service or a researcher where impartial resolution may be difficult to achieve. As such, the repository would need a mechanism to appoint or invite another party to mediate disputes. The intermediary would assume most of the data access process work, requiring expert personnel and resources for receiving, processing, hosting and managing large data volumes.

6.121 Additional expertise is needed for developing access eligibility criteria, data format and security requirements, quality standards, researcher accreditation, and application review processes. Consequently, the independent intermediary should have significant control over the operational effectiveness and usability of the model given its ownership of the repository. The centralised control enables system-wide improvements to operational effectiveness and usability at scale, unlike case-by-case models without a central infrastructure.

6.122 Services would need personnel for collecting, processing and transferring the data to the repository at regular intervals, while meeting security, data protection, data quality and data formats and standards. Many services will already have some personnel and resources in place from existing data-sharing obligations such as Article 40 of the Digital Services Act.

6.123 Researchers require sufficient resources for receiving, processing, analysing and possibly hosting repository data, with many already possessing relevant expertise from the repository. Again, it should be noted that researchers in this field may already be at least somewhat resourced for these activities.

Strategic effectiveness

6.124 The repository model supports online safety objectives by reliably providing researchers with access to data from a range of services, according to established data quality and data protection standards.

6.125 A single local or virtual repository providing access to data from multiple services could allow for the creation of a technical standard that would enable comparative cross-service research, which is difficult to carry out under current conditions.

6.126 In other fields of inquiry, the importance of interoperability of datasets to enable research insights has been well established.

Case study: The Rubin Observatory

The Rubin Observatory is an astronomy and astrophysics research centre based in Chile. It has adopted the International Virtual Observatory Alliance (IVOA) technical standard for virtual observatories for astronomical datasets.[239] This means that if the IVOA has defined an interface or standard model, the Rubin Observatory will use that, instead of developing their own standard. By following the International Virtual Observatory Alliance (IVOA) standard, the Rubin Observatory sets an example for the field, which in turn helps smaller projects adopt the latest technology and standards.

The Rubin Observatory makes available a suite of data products and tools to enable researchers not physically based at the observatory to process the data outputs and contribute insights based on these.[240]

High-level, generic metadata exchange is important, even as individual datasets use their own domain-specific standards, vocabulary, and semantics. The Rubin Observatory is committed to open science, ensuring its code is publicly accessible on GitHub. By making data widely accessible, the observatory hopes to engage the public in its exploration of the universe.[241]

6.127 The model could enhance research replicability by enabling validation of previous findings, while high-level data inventories help researchers understand available information before submitting access requests, resulting in better-informed proposals. Depending on required data provisions, researchers may access non-public data including service-side processes like recommendation or moderation algorithms, though services may cite reasons for not supplying such sensitive information.

6.128 Local repositories offer an advantage for research replication, as they can retain received datasets and safeguard them from deletion, ensuring availability for future verification or reproduction of findings. However, regular data transfer schedules can prevent real-time research, and delay correction of data quality issues until the next scheduled transfer.

6.129 Virtual repositories excel in responsiveness; when incomplete or incorrect datasets are corrected at the service level, the updated data can become immediately accessible to researchers without requiring the intermediary to amend or update the repository's data holdings. Virtual repositories can more easily support real-time access by allowing researchers to query data hosted on services via intermediary interfaces, avoiding delays from scheduled data transfers.

---

[239] Desai, V., M. Allen, and C. Arviset, et al., 2019. A science platform network to facilitate astrophysics in the 2020s. Bulletin of the American Astronomical Society. 51(7). https://baas.aas.org/pub/2020n7i146. [accessed 23 June 2025]

[240] Rubin Observatory, Data products, pipelines, and services. Data products, pipelines, and services | Rubin Observatory. [accessed 23 June 2025]

[241] National Academies of Sciences, Engineering, and Medicine. 2024. Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. https://doi.org/10.17226/27520. [accessed 23 June 2025]

6.130 The model does not guarantee resolution of data quality challenges such as limited access to audiovisual or gaming data, inconsistent data from filtering and moderation, and lack of timely data. The intermediary may establish data quality standards for services, although regular transfer intervals approach could create trade-offs between operational efficiency and real-time research capabilities.

6.131 The intermediary's control over researcher access decisions may provide greater transparency than service-led processes, including clear rationales for access decisions and independent review assurance. Services may gain confidence that approved projects are legitimate and researchers are properly vetted, potentially increasing trust and alleviating conflict of interest concerns in this historically challenging data-sharing space.

Case study: OpenSAFELY

OpenSAFELY is a secure analytics platform that provides researchers with an interface to access NHS patient records enabling analysis by medical researchers.[242] It was built around the 5 Safes Framework.[243]

Created during the COVID-19 pandemic, it has provided access to over 58 million patients' full pseudonymised primary care NHS records. The platform was developed with security, privacy and cost to researchers in mind in a way that prevents "researchers ever needing direct access to the disclosive underlying data to run analyses".[244]

It provides something akin to the suggested repository model's possible frontend and access mechanisms. In this case, the data is stored separately and accessed remotely via the software platform.

OpenSAFELY provides an example of the use of synthetic data (referred to by OpenSAFELY as 'dummy datasets') as a means of privacy preserving access to sensitive data. Researchers are able to develop specific and nuanced analytics code against dummy datasets, provide refine their code and researcher parameters based on realistic results.

When the refinements are set, researchers then provide this to OpenSAFELY who run the code against their data and then provide the results to researchers. Only summary tables and graphs are released from the system, after manual review by the OpenSAFELY team.

The analytics software used by OpenSAFELY is open to security review, scientific review, and reuse. The software and documentation is also made available on GitHub.[245]

---

[242] OpenSafely, OpenSAFELY: About OpenSAFELY. [accessed 23 June 2025]
[243] Department of Health and Social Care, 2022. Secure data environment for NHS health and social care data - policy guidelines. Secure data environment for NHS health and social care data - policy guidelines - GOV.UK https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines. [accessed 23 June 2025]
[244] Goldacre, B., Bacon, S., Hulme, W., 2020. What is OpenSAFELY? What is OpenSAFELY? | Bennett Institute for Applied Data Science. [accessed 23 June 2025]
[245] OpenSAFELY, GitHub - opensafely/documentation: Documentation for the OpenSAFELY platform. [accessed 23 June 2025]

Despite the privacy and security issues that access to medical records has traditionally been viewed as risking, the OpenSAFELY methodology has been endorsed by various organisations including medical privacy advocacy group MedConfidential.[246]

Since the height of the COVID-19 Pandemic, the secure access platform has expanded its access remit beyond pandemic-related reasons as of 2025.[247]

Implementation considerations for our policy options

Governance arrangements: fourth-party oversight

6.132 To enhance the effectiveness of any intermediary model, a distinct fourth-party organisation could provide strategic oversight of the independent intermediary's activities. Under this governance structure, the intermediary would focus on operational functions under each of the model options while the fourth party handles higher-level governance and dispute resolution. The independent intermediary could be a more specialised data-centric organisation than the fourth party, focused solely on providing researchers with access to data in a safe, secure, legal, ethical, and effective manner.

6.133 Under this governance structure, the independent intermediary would retain responsibility for researcher accreditation and proposal assessment, notice to service request decisions, repository operations where applicable, setting eligibility criteria and security standards, and day-to-day data access management. Meanwhile, the fourth party would provide strategic oversight and guidance of the intermediary, handle dispute resolution between services, researchers and the intermediary, develop high-level policy and monitor performance and accountability. The division of some of the supplementary functions discussed in further detail in Annex 5, such as guidance production, research gap analysis, and stakeholder engagement, would depend on the specific regime parameters and respective organisational capabilities. An 'arm's length' relationship between the intermediary and fourth party could improve security, internal processes, and stakeholder engagement while providing a clear escalation path for disputes, allowing the intermediary to concentrate on efficient access management rather than complex mediation.

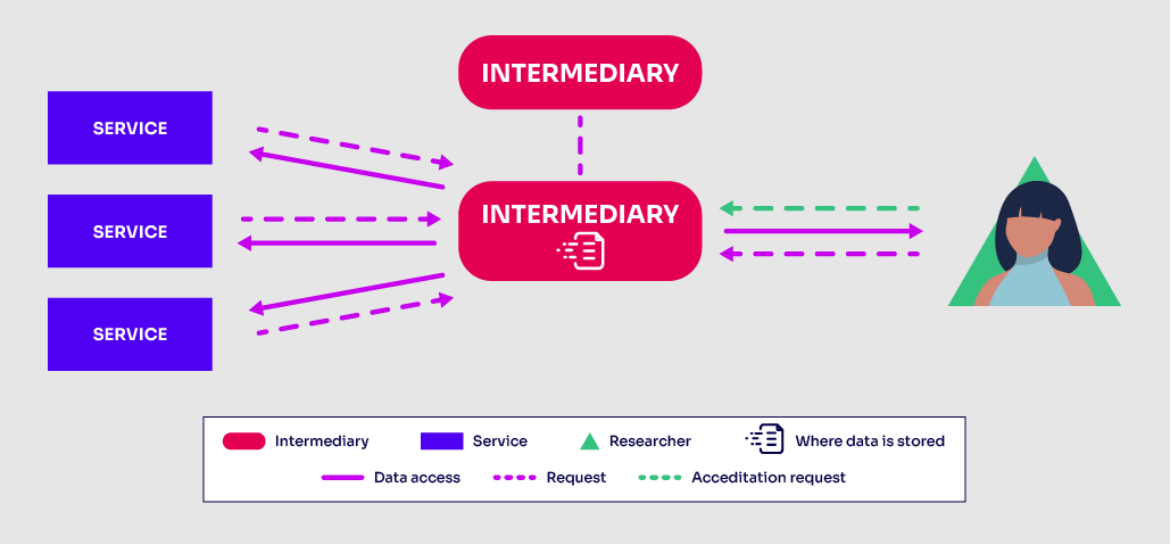
6.134 The additional layer of governance creates overhead costs and complexity that must be weighed against the benefits of specialised focus[248] and enhanced dispute resolution capabilities. The structure works best when clear role boundaries prevent overlap and confusion between the two organisations' stakeholder engagement for both parties.

[246] medConfidential, 2020-06-10-RoG-openSAFELY.pdfhttps://medconfidential.org/wp-content/uploads/2020/11/2020-06-10-RoG-openSAFELY.pdf. [accessed 23 June 2025]
[247] Armstrong, S., 2025. NHS has finally agreed to share GP patient data for research—this is why. NHS has finally agreed to share GP patient data for research—this is why | The BMJ. [accessed 23 June 2025]

[248] Benfeldt, O. N., 2017. O.B., 2017, A Comprehensive Review of Data Governance Literature, Selected Papers of the IRIS, 8(3). https://aisel.aisnet.org/iris2017/3. [accessed 23 June 2025]

6.135 For example, under the Digital Services Act (Article 40), the European Commission acts as a form of fourth party carrying out certain functions in the place of the Digital Service Coordinators. The Commission will establish and manage a Digital Access Portal to ensure that the data access process is "transparent and consistent across Digital Service Coordinators".[249] The Commission also holds an oversight role in ensuring compliance, conducting investigations and enforcing the Digital Service Act more broadly, with particular responsibility for Article 40.[250]

Figure 6 – A visual representation of the fourth-party oversight governance arrangement



[249] Commission Delegated Regulation (EU), supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data pursuant to Article 40 of Regulation (EU) 2022/2065, Ares(2024)7652659, s.3, arts 3, 4, 5.Delegated Regulation on data access provided for in the Digital Services Act. [accessed 23 June 2025]

[250] European Commission (EC), 2025. Supervision of the designated very large online platforms and search engines under DSA. Supervision of the designated very large online platforms and search engines under DSA | Shaping Europe's digital future. [accessed 23 June 2025]

7. Conclusions and reflections

7.1 Research into online safety matters provides substantial value by demonstrating how different online harms manifest across the digital ecosystem and enabling a broad set of stakeholders to interrogate the ways services combat those harms.

7.2 Robust and evidence-based research, especially when well-publicised, can contribute to increased public awareness of problematic behaviours and empower users to make more informed decisions about their digital habits. Such research also encourages policymaker and regulators' action and informs regulated services' policy and product design.

7.3 Both civil society and academic researchers have highlighted the need for greater access than is currently available, demonstrating that their specific needs differ and showing how various researchers create value from their work in different ways. Their input and our own research have identified potential means to unlock greater access while minimising risks to privacy and security, though implementing such methodologies and data access modalities in practice presents significant challenges.

7.4 Research requirements vary considerably in their timing and sensitivity. Some studies require timely data access, where any reduction in processing time creates risk that inappropriate or unsafe data is accessed either accidentally or by bad faith actors. Other research requires access to less time-sensitive data but is either sensitive in nature or is derived from sensitive data that risks reconstruction.

7.5 In response to constraints identified through our consultation, we have presented three policy options that could form the basis of a new researcher access regime, should the UK Government decide to establish one through further legislation.

7.6 The first option involves clarifying existing rules by maintaining the current regulatory framework without introducing new measures. Clearer guidance on data protection and privacy laws could support data donations and data scraping within existing provisions. This approach recognises that services, data donors and researchers already must comply with relevant laws but seeks to address uncertainty around exercising rights and understanding liabilities in research contexts.

7.7 The second option creates new regulatory duties for services through a systems and process-based approach, establishing requirements for data access and creating obligations on services to permit access under certain conditions. This direct access policy option, similar to the Digital Services Act (Article 40) approach, would require services to set up and maintain public portals for data access, with enforcement action taken against services that deny access meeting specific conditions without valid exemptions.

7.8 The third policy option encompasses three distinct intermediary models involving independent third parties to facilitate access: a general access portal providing standardised datasets from services with broad but limited access to pre-processed data, a notice to service model where researchers notify services of their data needs through an intermediary with both parties assessing proposals, and a repository model where intermediaries directly host data and facilitate sharing to approved researchers either through direct transfer or within TREs.

7.9 These policy options can be enhanced through cross-cutting implementation approaches including governance arrangements where a fourth-party organisation provides strategic oversight of intermediaries, and privacy-preserving technologies such as differential privacy, synthetic data, and TREs, which can be integrated with any intermediary model to address specific privacy and security concerns.

7.10 Each policy option or model and the possible implementation of it requires different degrees of technical infrastructure development, maintenance and oversight from services and independent researchers. In general, models that include an intermediary are likely to involve higher initial costs but could lead to long-term savings, with this trend more pronounced as intermediary roles and responsibilities increase. Intermediary models also offer the potential to enhance transparency in access decisions, assure proportionality, and support dispute resolution, while potentially reducing the burden on services by managing request triage and addressing legal and security concerns.

7.11 The policy options and models vary in their alignment with other jurisdictions. Direct Access and notice to service models share similarities with the Digital Services Act Article 40 provisions and delegated act, though different legal contexts create nuanced differences. Repository intermediaries, including the virtual repositories, have similarities with access modalities used in other jurisdictions and sectors, including in the UK but not specifically for these types of services. The degree of alignment with other regimes may be a factor in the difficulty and costs of operationalising any access regime.

7.12 It is evident that no single model is likely to meet the full range of researcher needs. A layered, flexible approach – combining legal clarity, technical safeguards, and independent oversight – offers the best chance of enabling responsible, timely and useful information access. The policy options and models outlined in this report do not need to be considered in isolation and could be regarded as complementary. Elements from different models, combined with enabling measures, may present more effective means of facilitating researcher access depending on policy objectives. Where needed, responsibilities for management of a researcher access regime could be shared between technical and governance-focused bodies to manage complexity, support trust-building, and reduce burdens on any single institution. Meaningful data access in support of making people safer will also require a shift in culture – one built on trust, transparency, and a shared commitment to ethical standards.

7.13 These policy options and models offer varying approaches to balancing access facilitation with privacy protection, dispute resolution, cost distribution, and operational effectiveness. Policy option and model selection should take into account specific research community needs, service capabilities, available resources, and broader policy objectives for online safety research.

7.14 Several operationalisation issues remain beyond the scope of this report. These include a precise definition of 'independent researcher' and eligibility criteria, such as institutional affiliations or commercial interest parameters. There are open questions regarding geographic scope considerations for both researchers and accessible data. For example, whether researchers should be limited to those based in or affiliated with UK institutions or whether research is limited to that which relates to impacts on UK users. Additionally, whether data in scope of access be limited to that which relates to UK users or is of a UK origin must be considered.

7.15 The distinction between public and private data also requires clarification, given that data can have different meanings and context based on services' nature and design, user expectations, and services' operational changes.

7.16 These outstanding questions will require careful considerations during any regime's detailed design and implementation phases. Specifics of this distinction can have implications with regard to legal liability, with regard to how data is obtained and for what purpose it is subsequently used. It is beyond the purposes of this report to take a definitive view of the specifics of this distinction. We have maintained an agnostic approach to these definitional questions, presenting policy options and models and enablers in such a way that they could operate regardless of how these issues are ultimately resolved.

A1. Legal framework[251]

Data (Use and Access) Act 2025

A1.1 The Data (Use and Access) Bill was introduced to the UK Parliament on 23 October 2024. It contained provisions that allow for the creation of a new framework for researchers to access information held by regulated online services. This Bill received Royal Assent on 19 June 2025, becoming the Data (Use and Access) Act 2025 (the "Data (Use and Access) Act").[252]

A1.2 Section 125 of the Data (Use and Access) Act inserts a new section into the Online Safety Act that allows the Secretary of State to make regulations that "require providers of regulated services to provide information for purposes related to the carrying out of independent research into online safety matters." It also removes Ofcom's duty to produce guidance under section 162(7) to (10) of the Online Safety Act.

A1.3 These regulations could be used to determine the scope of a new researcher access regime, and may include: details of the procedure to be followed in the making and determination of applications, the requirements and contents of "researcher access notices", the regulated services that may be required to share information, which researchers can make applications, the form in which information is to be shared, safeguards for handling information, fees payable by applicants and the enforcement of any new regime. The Secretary of State will also be required to consult with Ofcom and other appropriate bodies before making these regulations.

A1.4 The UK Government has indicated that Ofcom's report will provide an evidence base to inform the design of any future access framework.[253]

Data protection framework in the UK[254]

GDPR (EU and UK)

A1.5 The EU General Data Protection Regulation (GDPR) is a framework that governs the processing of personal data within the EU. The GDPR applies to all EU members states, organisations established in the EU that process personal data and organisations established outside of the EU that process the personal data of people located in the EU.[255]

A1.6 Following the UK's exit from the EU, the United Kingdom General Data Protection (UK GDPR) was introduced to retain the GDPR within UK domestic law. While the key principles, rights, and obligations largely remain the same, the UK has the independence to keep the framework under review. The Data Protection Act 2018 (the "Data Protection Act")

[251] This Annex is a summary of the legal framework only. For further information about how to comply with data protection law, see the ICO's UK GDPR guidance and resources.
[252] UK Parliament, 2025, Data (Use and Access) Act 2025.
[253] UK Parliament, 2024, Data (Use and Access).
[254] This section is not a comprehensive overview of data protection law but rather focuses on areas that are particularly relevant to research.
[255] As part of offering goods or services or monitoring the behaviour of people located in the EU.

complements the UK GDPR by providing additional provisions and UK-specific exemptions. It also sets out further rules and exemptions, particularly those related to law enforcement and national security. Both the UK GDPR and the Data Protection Act apply to UK-based organisations, as well as those outside the UK that process the personal data of UK residents.

A1.7 Data protection laws are designed to safeguard personal information and ensure it is handled lawfully, fairly and responsibly. These laws are essential for upholding privacy, fostering trust among individuals and stakeholders, and mitigating the risk of reputational damage or possible enforcement action.

A1.8 The Data (Use and Access) Act (which received Royal Assent on 19 June 2025) will make changes to data protection law, including clarifications to the definition of "research" and some of the other rules related to processing for research purposes. Please refer to the ICO for more information.

Research-related processing

A1.9 The UK GDPR governs data processing for research purposes, and like its EU counterpart, encourages a broad approach to the concept of 'research'.[256] While these purposes are not defined in legislation, some detail is provided in the introductory recitals to the UK GDPR (which are not legally binding).[257] The ICO has provided guidance which provides information on how the research provisions work and sets out its understanding of the provisions' key terms. It explains how the provisions relate to the data protection principles and grounds for processing and details the exemptions set out in the provisions.[258]

A1.10 The three research-related purposes[259] are as follows:

- archiving in the public interest: where organisations manage and preserve records identified as having potentially enduring public value;
- scientific or historical research: includes research carried out in traditional academic settings (such as academic research in social sciences, humanities and the arts), research carried out in commercial settings, and technological development, innovation and demonstration; and
- statistical research: refers to activities where the processing's primary aim or purpose is to produce statistical outputs.

A1.11 In order to use the research provisions, organisations and researchers need to have appropriate safeguards[260] in place. These protect the rights and freedoms of the people whose personal data is being processed. These safeguards take the form of technical and organisational measures to ensure respect for the principle of data minimisation. Where possible, organisations and researchers should carry out research using anonymous information. This information is not personal data and data protection law does not apply. Where it is not possible to use anonymised data, organisations and researchers should consider whether it is possible to pseudonymise the data. Pseudonymous data is still personal data and data protection law applies. Organisations and researchers cannot rely on

[256] For further information, see ICO guidance on research-related processing.
[257] For the relevant provisions in the UK GDPR, see Recitals 158-160, 162.
[258] For further information, see ICO guidance on the research provisions.
[259] The ICO guidance on the research provisions includes indicative criteria for each of these categories.
[260] For further information, see ICO guidance on appropriate safeguards.

the research provisions if the processing is likely to cause someone substantial damage or distress, or the processing is being carried out for the purpose of measures or decisions with respect to particular people (unless the research is approved medical research).

Data protection principles for research

A1.12 Article 5 of the UK GDPR sets out seven data protection principles.[261] Two of these principles – purpose limitation and storage limitation – contain specific provisions about research related data processing.[262] These provisions allow researchers to reuse existing personal data and keep it indefinitely, provided that there is a research related purpose and appropriate safeguards are in place.[263]

Relevant lawful basis for processing

A1.13 One of the principles in Article 5 of the UK GDPR is that personal data shall be processed lawfully. Article 6 of the UK GDPR provides that processing shall be lawful only if and to the extent that a lawful basis applies.[264] In the context of research-related processing, the most relevant bases are likely to be:

- Public task – that the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority.[265]
- Legitimate interests – that the processing is necessary to pursue legitimate interests unless these interests are overridden by the need to protect an individual's personal data.[266]

A1.14 Which lawful basis is most applicable depends on whether the researcher is an organisation or a public body. For further guidance, see the ICO's guide to lawful basis and guidance on principles and grounds for processing.

Special category data

A1.15 Special category data is personal data that is more protected due to its sensitive nature.[267] Article 9 of the UK GDPR defines special data category as personal data about a person's:

- race or ethnic origin;
- political opinions;
- religious or philosophical beliefs;
- trade union membership;
- genetic data;
- biometric data for the purpose of uniquely identifying a natural person;
- health data; and
- data concerning a natural person's sex life or sexual orientation.

---

[261] For the relevant provision in the UK GDPR, see Article 5.
[262] For further information, see ICO guidance on purpose limitation and storage limitation.
[263] For further information, see ICO guidance on principles and grounds for processing.
[264] For the relevant provision in the UK GDPR, see Article 6.
[265] For further information, see ICO guidance on public task.
[266] For further information, see ICO guidance on legitimate interests.
[267] For further information, see ICO guidance on special category data.

A1.16 Special category data can only be processed if one of the specific conditions in Article 9 of the UK GDPR are met, along with an Article 6 lawful basis. Article 9(2)(j) provides that special category data can be processed if it is necessary for a research-related purpose.[268] Paragraph 4 of Schedule 1 to the Data Protection Act sets out additional requirements for researchers relying on the condition in Article 9(2)(j).[269] They are that the processing of special category data must be necessary for archiving purposes, scientific or historical research purposes or statistical purposes, and must be subject to appropriate safeguards and in the public interest.[270] Discussion of special category data following stakeholder feedback is contained throughout the report.

Overview of research provisions and data subject rights

A1.17 People have specific rights over their personal data. Most of these rights have exemptions available when processing data for research-related purposes. These exemptions may apply to the following rights:

- the right to be informed;
- the right of access;
- the right to rectification;
- the right to erasure;
- the right to restrict processing;
- the right to data portability; and
- the right to object.

A1.18 For some of these rights, there is a built-in exception for research. For others, Schedule 2 to the Data Protection Act sets out a separate exemption. Further information, including the matters that must be taken into account when applying the exemptions, can be found in the ICO's guidance on exemptions.

A1.19 This Annex focusses on the rights to data portability and access as both are current methods for researcher access (see Section 4).

Data portability rights

A1.20 Article 20 of the UK GDPR provides that individuals have the right to obtain personal data they have provided to a controller in a structured, commonly used and machine-readable format.[271] A controller is a legal entity or individual who exercises overall control over personal data and is responsible for processing activities.[272] For example, a controller may be a service who processes the personal data of their users. Individuals also have the right to request that a data controller transmits the data to another data controller.

[268] For the relevant provision in the UK GDPR, see Article 9(2)(j).
[269] For the relevant provision in the Data Protection Act 2018, see paragraph 4 of Schedule 1.
[270] For the relevant provisions about safeguards, see Article 89(1) in the UK GDPR and section 19 of the Data Protection Act. For further information, see the ICO's guidance on principles and grounds for processing.
[271] For the relevant provisions in the UK GDPR, see Article 20.
[272] For the relevant provision in the UK GDPR, see Article 4(7). For further information, see ICO guidance on controllers and processors.

A1.21 This right only applies when:

- the lawful basis for processing is consent or the performance of a contract; and
- the processing is being carried out by automated means.

A1.22 These bases are unlikely to arise in the context of researchers seeking access to data from services, as they generally apply to individuals seeking data from organisations who are providing a service (to allow that individual to port their own data to other services). As such, there is no general exemption to data portability for research purposes, although paragraph 28 of Schedule 2 to the Data Protection Act provides an exemption for data holders if the processing is for archiving purposes in the public interest, and that the application the data portability rights would prevent or seriously impair this purpose.[273] For further discussion of data portability rights and data donations, see Section 4, Section 5 and Annex 2.

A1.23 Article 15 of the UK GDPR provides that individuals have the right to obtain a copy of their personal data and other supplementary information.[274] Individuals can exercise this right by submitting a subject access request (SAR) to the data controller. Individuals can make SARs verbally or in writing. Third parties can also make a SAR on behalf of an individual where authorised to do so.[275] Paragraphs 27 and 28 of Schedule 2 to the Data Protection Act provides exemptions for data holders, if personal data is processed for scientific or historical research or statistical purposes, or for archiving purposes in the public interest.[276] Data holders must show that giving effect to the right of access would prevent or seriously impair these purposes, and any exemption must also be necessary, proportionate and applied on a case-by-case basis.

A1.24 The exemptions also only apply:[277]

- if the processing is subject to appropriate safeguards for people's rights and freedoms;
- if the processing is not likely to cause someone substantial damage or substantial distress; and
- if organisations and researchers do not use the processing for measures or decisions about particular people, except for approved medical research.

A1.25 Paragraph 27 of Schedule 2 to the Data Protection Act sets out a further condition on the exemption for scientific or historical research or statistics. It requires anonymisation of research results or any resulting statistics. This condition does not apply to archiving in the public interest. For further discussion of access rights and data donations, see Section 4, Section 5 and Annex 2.

[273] For the relevant provision in the Data Protection Act 2018, see paragraph 28 of Schedule 2.
[274] For the relevant provision in the UK GDPR, see Article 15.
[275] For further information, see ICO guidance on subject access requests.
[276] For the relevant provisions in the Data Protection Act 2018, see paragraphs 27 and 28 of Schedule 2.
[277] For further information, see ICO guidance on appropriate safeguards.

A2. Limitations of currently used access models

A2.1 In section 4 of this report, we describe a wide range of the methods that researchers currently use to access and collect data from services and about users of services for the purposes of online safety-related research. We discuss some of the limitations of these methods in more detail in the section below.

Application programming interfaces (APIs)

A2.2 Historically, researchers have accessed service data through application programming interfaces (APIs)—computing interfaces that enable automated, standardised data retrieval at scale. While APIs have played a key role in supporting research, they come with several limitations.

A2.3 These include restrictions on data volume, incomplete or inconsistent datasets, and limited transparency around changes to the API itself. Access conditions may also be unclear or subject to change. Moreover, APIs typically require specialist technical skills to navigate, including the ability to store, clean, and analyse potentially noisy data. This can limit the pool of researchers able to make effective use of them, despite the availability of some online tools designed to support API use. [278]

A2.4 In addition to technical barriers, APIs can be financially prohibitive. Many services charge high fees for access, making them unaffordable for some researchers or institutions.

A2.5 Finally, data quality remains a concern; API-provided data may lack essential metadata and can be difficult to validate.

Voluntary research partnerships

A2.6 Some researchers have noted that services tend to favour certain types of researchers when forming voluntary partnerships—typically those who are more senior, well-known, or affiliated with prestigious and well-funded institutions in the Global North. As a result, these partnerships may unintentionally reinforce existing inequalities by deprioritising more junior researchers and those from less prestigious or less well-resourced institutions. [279]

A2.7 In some cases, the terms of these partnerships may also raise concerns about research independence. Depending on the nature of the agreement, the data provider may exert significant control over what data researchers can access, the questions they are allowed to pursue, and the findings they are permitted to publish. Even when such concerns are ultimately unsubstantiated, they can still affect perceptions of the research's credibility and integrity.

[279] Name Withheld 1 response to October 2024 Call for Evidence, p.2.

A2.8 Data quality is another challenge. Information obtained through voluntary partnerships may be difficult to validate and may lack the necessary metadata to support robust analysis.

Ad Libraries

A2.9 Researchers have expressed mixed views on the utility of service-operated ad libraries. Some ad libraries are considered unsuitable for online safety-related research due to several limitations, including a lack of cross-service interoperability, unclear criteria for ad inclusion, and restricted search functionality.[280] [281] [282]

A2.10 Data quality is also a concern. Information from ad libraries can be difficult to validate and may lack the necessary metadata to support rigorous analysis. Additionally, the data is often not granular enough to enable robust or meaningful research.

Transparency reports

A2.11 While transparency reports can be useful for the general public and civil society organisations seeking to understand how services develop and enforce user protection policies, they are often seen as having limited value for online safety-related research. This is primarily because researchers typically have little to no influence over what information is included in these reports. [283] In many cases, the data provided lacks the granularity or scope needed to inform future research questions or generate meaningful insights. [284] [285]

A2.12 Questions also remain about the accuracy of the data and researchers' ability to independently validate the claims made in these reports. Furthermore, the infrequent publication of transparency reports, often on an annual basis, means that the information shared is not timely enough to support research into current service practices.

Data scraping

A2.13 Data scraping may be permissible when conducted in the public interest. However, there is significant legal ambiguity around what qualifies as "public interest" in this context. This uncertainty creates challenges for researchers seeking to rely on scraping as a legitimate method for data collection.

A2.14 While some recent lawsuits targeting researchers who used scraping have been dismissed, the lack of legal clarity—particularly around services' terms of service, which often prohibit

[280] Center for Countering Digital Hate response to October 2024 Call for Evidence, p.3.
[281] Global Witness response to October 2024 Call for Evidence, p.4.
[282] Reset.Tech response to October 2024 Call for Evidence, p.5.
[283] Center for Democracy & Technology. Transparency Reports. Transparency Reports - CDT. [accessed 23 June 2025]
[284] Open Data Institute, 2024. Exploring global challenges of regulating researcher access to platform data. Exploring global challenges of regulating researcher access to platform data | The ODI. [accessed 23 June 2025]
[285] The British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence, p.3.

scraping—continues to pose risks. The threat of expensive legal battles and the ethical uncertainty surrounding scraping can have a chilling effect on research. [286]

A2.15 In addition to legal concerns, scraping requires significant technical expertise. [287] Researchers must be able to set up scraping technologies and manage the collection and preparation of large volumes of data. This includes handling potentially unstructured or noisy data and ensuring it is stored and processed securely.

A2.16 Data quality is another key issue. Scraped data can be difficult to validate and may not include the necessary metadata to support robust and reproducible research.

A2.17 There are ongoing efforts to clarify the legal basis for scraping in the context of research. For example, the Knight-Georgetown Institute's *Gold Standard for Publicly Available Platform Data* project is exploring best practices for ethical and lawful data access. [288] Similarly, the European Digital Media Observatory's report on *Platform-to-Researcher Data Access* outlines the "special, more permissive regime for processing of personal data for scientific research" under the GDPR. [289] This includes potential legal bases such as consent that is "freely given, specific, informed and unambiguous," legitimate interests, and public task. [290] These initiatives highlight the need for clearer regulatory guidance to support responsible and legally sound research practices.

Avatar research

A2.18 Avatar research can offer valuable insights into how certain user traits may influence the likelihood of encountering specific types of content. However, this method cannot, on its own, explain why a given trait functions as a risk factor. Researchers also face challenges when conducting avatar studies across jurisdictions, particularly in understanding how services use location data to shape recommendations. [291] Additionally, it is relevant to note that the most prominent examples of avatar-based research involve children. [292]

A2.19 This methodology is also highly resource-intensive—requiring significant time, labour, and financial investment—while often producing limited insights. As such, it is not easily scalable for most researchers. While automation can help address scalability, it introduces new complications. Automated avatars or bots may be flagged by service systems designed to detect malicious behaviour, potentially undermining the research.

---

[286] Keller, J. R., Moriniere, S. and Tinsman, C., 2024. What is 'public data'? And who should be allowed to collect and use it? What is 'public data'? And who should be allowed to collect and use it? | by Jared Robert Keller | Canvas | Medium. [accessed 23 June 2025]
[287] Ada Lovelace Institute response to October 2024 Call for Evidence, p.10.
[288] Knight-Georgetown Institute, Publicly Available Platform Data Expert Working Group – Knight-Georgetown Institute. [accessed 23 June 2025]
[289] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p5. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]
[290] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p29. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]
[291] Center for Countering Digital Hate response to October 2024 Call for Evidence, p.3. link.
[292] Ofcom, Avatar Methodology: A pilot study

A2.20 Many services' terms of service explicitly prohibit avatar research or similar practices, creating ambiguity around what is permitted. This can expose researchers to legal risk and ethical scrutiny.

Purchasing data and access from commercial entities

A2.21 Some researchers obtain data through commercial services, such as social listening tools. While these tools can support exploratory research, they come with several limitations. Upfront access costs can be prohibitively expensive, creating a barrier for many researchers. Additionally, these tools often impose data access restrictions and are susceptible to sampling bias.

A2.22 There is frequently a lack of transparency around how data is sourced, which can affect the reliability and credibility of the information. This opacity also raises potential legal and ethical concerns, such as possible violations of GDPR or terms of service.

A2.23 The data itself is not always well-suited to rigorous research. It is often voluminous, noisy, and unstructured—containing irrelevant or low-quality information that can lead to contextual misunderstandings or misinterpretation.

A2.24 Moreover, the historical practices of some data brokers — such as collecting data without meaningful consent or failing to disclose acquisition methods — further complicate the ethical landscape. These concerns underscore the need for caution and due diligence when purchasing data for research purposes.

Research forums and consortia

A2.25 Research forums and consortia serve primarily as collaborative spaces for knowledge exchange and coordination, rather than as mechanisms for direct data access. Furthermore, Research forums and consortia are typically not data-centric, meaning their primary purpose is not to enable or increase researchers' access to data.

A2.26 Activities such as the pooling of non-data resources and knowledge can be secondary enablers of data access but are unlikely to directly contribute to such access. Even if researchers did want to share data with other researchers in the forums or consortia, legal and ethical barriers would likely prevent this. While data providers might be more willing to collaborate with and provide access to researchers in the context of a forum or consortium, rather than on an individual basis, these partnerships are subject to the limitations outlined in our section on voluntary research partnerships.

Data donations

A2.27 While data donation can offer researchers access to otherwise unavailable datasets, it presents a number of significant legal, ethical, technical, and operational challenges.

A2.28 A data donation model would require clear guidelines around the rights, roles, and responsibilities of the different parties involved, as well as the conditions under which donations should take place. In some cases, data donors may provide information that includes details about individuals other than themselves. Researchers would need to understand how to handle such data to ensure the rights and privacy of all users are respected.

A2.29 There are also ethical considerations around incentivisation. Some researchers may choose to financially compensate data donors to encourage participation. While compensating research participants is not uncommon or inherently problematic, it may raise ethical concerns in certain contexts, particularly if it influences who participates or how much data is donated.

A2.30 Some respondents to our Call for Evidence noted that services have previously prohibited the use of third-party software designed for data donation. Without a legal framework requiring services to allow and support such tools, researchers may face both legal and practical challenges if the use of these tools is suddenly restricted or penalised. Because data donations occur outside of formal data-sharing partnerships, there would also need to be clear agreements outlining responsibilities and liabilities in the event of data misuse or security breaches.

A2.31 Data donations also open up the possibility of researchers being inadvertently provided with illegal content they may not have proper certification legal exemption to possess, such as child sexual abuse material (CSAM) or terror content.

A2.32 From a technical standpoint, data donors receive their data in whatever format the service provides or in the format collected by monitoring software. This means that existing challenges around data standardisation and cross-service comparability would remain. In some cases, researchers may need to develop and maintain their own data collection and management tools. Donors, meanwhile, would need the technical skills to request, retrieve, and share their data, which not all potential participants possess. Some data donation methods are more technically demanding than others.

A2.33 Respondents also highlighted bias as another concern. Donated data may reflect only the behaviours and experiences of individuals who are both willing and able to participate. This self-selection can result in unrepresentative samples, limiting the generalisability of research findings. This issue is particularly acute in online safety research, where topics are often sensitive, controversial, or personal. Individuals with relevant experiences may be less inclined to share their data, further skewing the dataset and potentially omitting critical perspectives.

A2.34 Additionally, data donors may alter their online behaviour, consciously or unconsciously, if they know they are being observed. This behavioural distortion is unlikely to be resolved even with improved ease of use or incentives. The consent provided by data donors is also non-transferable, meaning researchers cannot use the data for purposes beyond those originally specified. While broader consent could be requested, this may discourage participation and introduce further bias.

A2.35 Incentivisation can also affect data quality. If donors are compensated per donation or based on the volume of data, they may be encouraged to contribute more data, even if it is of lower quality. Researchers have also raised concerns about receiving datasets with inconsistencies, errors, or anomalies due to services' data collection, filtering, or moderation practices. Additionally, the data may not be timely, limiting its usefulness for research on current service behaviours.

A2.36 Services may also contest the findings of studies based on donated data, citing a lack of context or the absence of direct collaboration with researchers. The burden on data donors, including the time, effort, and technical skills required to collect and share data, combined with limited incentives, makes this methodology difficult to scale. Data donations are also

limited to the point in time at which the data is captured. Establishing ongoing relationships with donors to receive updated data is only feasible for some researchers and cannot be guaranteed at scale.

A2.37 Finally, two of the three main data donation methods -- those involving the exercise of data portability and access rights -- are limited to EU, EEA, and UK residents under GDPR. This restricts participation to individuals in these jurisdictions, excluding potentially significant groups of data subjects and impacting the quantity and diversity of data available for research.

A2.38 We also discuss these limitations in Chapter 4 under Regulatory clarification, Legal and ethical challenges, Operational challenges, and Data quality challenges.

Voluntary widgets

A2.39 A voluntary widget for data collection is a tool that individuals can choose to install or activate to share specific types of data with researchers, typically for academic or public interest purposes. These widgets often operate through browser extensions or apps and are designed to collect data directly from users' interactions with services, with their informed consent.

A2.40 Because data subjects are exercising their rights to request and then voluntarily donating their data, this methodology can also reduce the occurrence of legal risks. Data obtained in this way can provide detailed, individual-level insight into a range of topics relevant to online safety.

A2.41 However, this methodology introduces risks of data bias due to being skewed towards certain demographics or types of individuals who choose to opt-in to sharing arrangements. This means that data samples may not be representative, which can have a negative impact on the quality of the research. Data donations also do not include inferred data, such as personal characteristics of the individual data subject inferred based on their online interactions.[293]

A2.42 Researchers may pair donated data with other data about the same data subjects, such as survey data, in order to build a more comprehensive picture of the data subjects including their wider attitudes and habits. This may not be possible when using conventional datasets that do not allow for direct engagement with the data subjects.

A2.43 The requirement of voluntary opt-in can be a barrier to studying certain online behaviours – such as illegal practices or controversial topics – where data donation is more unlikely, making it challenging to make generalised observations about specific phenomena.[294] Given that participants' consent is usually not portable, the data can also only be used for the original donation purpose and cannot be 're-donated' for use by other researchers, or even for use by the same researcher on a different project.

[293] Skatova, A. and, Goulding, J., 2019. Psychology of personal data donation. PLoS One, 20(14). doi: 10.1371/journal.pone.0224240. [accessed 23 June 2025]
[294] Dommett, K., Orben, A. and Zendle, D. response to October 2024 Call for Evidence, p.4.

A2.44 Participants may also require financial compensation, which can have an impact on scalability or lead to high costs for the researcher. Further challenges relate to the timeliness of donated data, particularly where data subjects rely on subject access requests and data portability rights, rather than tracking widgets. In these instances, the data is only captured up to the point of the request, and the processing and analysis of the data can introduce further delays, making it difficult to research real-time events and behaviours. Additionally, data may be inconsistent if donated by multiple participants who make their requests at different times. Issues of data inconsistencies may be mitigated through the data subjects making repeated and coordinated requests, but this poses practical challenges at scale.

Data trusts

A2.45 A data trust is a legal and governance framework that enables individuals to delegate the management of their personal data to a trusted intermediary, who oversees how that data is accessed and used.

A2.46 While the exact nature of each data trust is determined at the founding stage, it can generally be assumed that setting up and maintaining a data trust is resource-intensive, requiring significant effort and knowledge. While data trusts can help users share their data while being assured of the necessary safeguards, the resources needed for setting up the trust and ensuring its terms of reference remain up to date with the latest practice and technology, may mean it is a burdensome mechanism to operate. Additionally, the lack of major adoption of this sort of body means that not many people are familiar with, or even aware of this type of data-sharing mechanism. As the trustee is actively managing access and providing safeguards, it is also not clear how such bodies would be financially sustained.[295]

Data cooperatives

A2.47 A data cooperative is a member-owned organisation where individuals voluntarily pool their data and collectively decide how it is accessed and used. Unlike traditional data-sharing models, data cooperatives prioritise democratic governance, giving members a direct say in decisions about data management, privacy, and the purposes for which their data is used.

A2.48 However, it should be noted that the purpose of a cooperate is primarily to benefit its members, rather than the wider public. As such, data cooperatives may not be the most suitable mechanism to facilitate researcher access to data. Data cooperatives also require a highly participatory model of involvement from the data subjects, which may be a barrier for scalability. Operational costs may also serve as a barrier to scalability, as the cooperative would require funding for administration, technological build and maintenance, day-to-day data management and legal considerations (including data protection concerns), among other resource-intensive work. As with data trusts, this concept is relatively novel and not widely used by researchers.

[295] Ada Lovelace Institute, 2021. Exploring legal mechanisms for data stewardship. Exploring legal mechanisms for data stewardship | Ada Lovelace Institute. [accessed 23 June 2025]

A3. Privacy enhancing technologies

A3.1 In section 5 of this report, we discuss the various security constraints that restrict the sharing of information between researchers and services. In section 6 of this report, we use security considerations as one of our criteria for evaluating the proposed policy options and models and draw out the relevant security implications of each model. In the section below, we highlight the various privacy enhancing technologies (PETs) that can play a role in addressing security constraints and requirements and consider their respective advantages and drawbacks.

A3.2 Additional privacy-preserving interventions may be necessary when policy options and/or models do not robustly address legal, ethical, or security concerns. PETs offer technical solutions to protect the privacy and security of sensitive data. The term 'PETs' encompasses a broad range of technologies that share a common feature: they enable access to data without revealing the underlying raw information. As a result, trust in the data owners becomes important as they must ensure that the data made available through PETs is accurate, replicable, and reflective of the original statistical distributions.

A3.3 Multiple stakeholders have highlighted that PETs create an opportunity for research on regulated services data in a manner that addresses the privacy and security risks often associated with data access.[296] [297] [298] [299] [300] [301]

A3.4 These measures should be deployed proportionately to the assessed level of research and/or data risk and should be appropriate for the research purpose, recognising that some interventions may distort data to the point of disutility. Some stakeholders have cautioned that practical application and use of PETs in the context of online safety research is not fully explored or matured.[302]

A3.5 These PETs are policy option and model-agnostic and can be flexibly combined with most access models. They can be used at various points in the access process to potentially facilitate greater and different forms of access and protect various data types. Multiple technologies can be implemented simultaneously at different stages of the research process,

[296] The Royal Society, 2023. From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis. From privacy to partnership | The Royal Society. [accessed 23 June 2025]
[297] Meta Platforms Inc. (Meta) response to October 2024 Call for Evidence.
[298] OpenMined response to October 2024 Call for Evidence.
[299] Smart Data Research UK response to October 2024 Call for Evidence.
[300] The British and Irish Law Education Technology Association (BILETA) response to October 2024 Call for Evidence.
[301] Name Withheld 1 response to October 2024 Call for Evidence.
[302] Open Data Institute response to October 2024 Call for Evidence.

as demonstrated by projects like CARRIER, which utilised secure multi-party computation, homomorphic encryption, and federated learning.[303] [304]

A3.6 Some PETs require significant computational resource and infrastructure, and highly skilled technical staff. This can make implementing them costly, which creates a further barrier for widespread use across both services and research institutions.

Differential privacy

A3.7 Differential privacy is one of the most widely used PETs. Differential privacy is a mathematical standard of privacy where a dataset is defined as differentially private if individual identification from outputs becomes impossible.[305]

A3.8 The technology adds 'noise' to the dataset by replacing participants' answers with random results until it satisfies this definition.[306] This comes with a necessary trade-off of accuracy. The noise may influence the analysis in ways that lead to misleading conclusions. This is especially impactful where researchers are seeking to generalise and when the sample size is small as more noise is added to help preserve privacy. The lack of replicability poses particular limitations for social scientific research that seeks to establish data representativeness and generalisability of findings.[307] It has also been argued that differential privacy could hamper exploratory research because it may limit researchers' ability to accurately look at small subsections of data and to understand the data before deciding which analyses to run.[308]

A3.9 If individuals are impossible to identify, the data may potentially be released as open-access, significantly benefitting researchers. The method is already used in numerous online safety applications as it is less computationally intensive compared to other PETs.

[303] Buckley, D., 2023. Statistics Netherlands: Developing privacy-preserving cardiovascular risk prediction models from distributed clinical and socioeconomic data. 12. Statistics Netherlands: Developing privacy-preserving cardiovascular risk prediction models from distributed clinical and socioeconomic data - UN GWG on Big Data - Privacy Preserving Techniques Wiki - UN Statistics Wiki. [accessed 23 June 2025]
[304] To better understand which PETs are appropriate for different data-sharing modalities, see page 83 of the European Digital Media Observatory report advocating for categorising data by risk: D (lowest), C and B (medium), and A (high risk).
[305] Cummings, R., Desfontaines, D., and Evans, D., et al., 2024. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harvard Data Science Review, 6(1). https://doi.org/10.1162/99608f92.d3197524. [accessed 23 June 2025]
[306] Cummings, R., Desfontaines, D., and Evans, D., et al., 2024. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harvard Data Science Review, 6(1). https://doi.org/10.1162/99608f92.d3197524. [accessed 23 June 2025]
[307] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]
[308] Cummings, R., Desfontaines, D., and Evans, D., et al., 2024. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harvard Data Science Review, 6(1). https://doi.org/10.1162/99608f92.d3197524. [accessed 23 June 2025]

Synthetic data

A3.10 Synthetic data (sometimes referred to as 'dummy data') is data created algorithmically to replicate the structure and characteristics of real data. For example, the synthetic data could have similar distributions and averages, without the specific elements of the original dataset.[309]

A3.11 This approach protects data at the point of release and computation, potentially allowing researchers more flexibility in releasing findings. However, synthetic data is less helpful addressing outliers because anomalies are typically not captured when trying to replicate a distribution of data in a privacy-preserving way. This method also adds noise to the original dataset, which may influence the analysis in ways that results in misleading conclusions. This requires trust that the original data is accurate, as it is difficult to ascertain the level of privacy without referring to the original data.

Trusted Research Environments (TREs)

A3.12 Although there is not one clear definition for TREs, they are generally defined as secure spaces with enhanced security and privacy measures where researchers can access sensitive data.[310] These environments provide analytical computing and data storage and a secure access environment as part of a managed service for researchers.

A3.13 TREs can provide the infrastructure for a full service by following all Safes of the Five Safes Framework or provide only the 'safe setting' of the Five Safes framework by providing a secure lab environment.[311] Universities and public sector organisations such as hospital trusts, national labs and government agencies commonly use TREs, though they are used less in private sector organisations.[312]

A3.14 Clean rooms are a type of TRE which offer environments where raw data cannot be exported. Clean rooms can be either virtual, where data is accessed remotely, or a physical space, where a researcher visits a space in-person.[313] While the EDMO recommends clean rooms for analysing high-risk or high-sensitivity data,[314] they discourage their use for low and medium risk cases because researchers' inability to view underlying data prevents exploratory work or code debugging. Accessing physical clean rooms can also impose significant costs on researchers depending on their location.

---

[309] Jordon, J., Szpruch, L., and Houssiau, F. et al., 2022. Synthetic Data -- what, why and how? [2205.03257] Synthetic Data -- what, why and how?. [accessed 23 June 2025]
[310] DARE UK Consortium, 2021. Data Research Infrastructure Landscape: A review of the UK data research infrastructure, p.6. UK Data Research Infrastructure Landscape. [accessed 23 June 2025]
[311] DARE UK Consortium, 2023. UK Sensitive Data Research Infrastructure: a Landscape Review. UK Sensitive Data Research Infrastructure: a Landscape Review. [accessed 23 June 2025]
[312] DARE UK Consortium, 2023. UK Sensitive Data Research Infrastructure: a Landscape Review, p.10. UK Sensitive Data Research Infrastructure: a Landscape Review. [accessed 23 June 2025]
[313] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]
[314] European Digital Media Observatory Working Group, 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access, p.88. Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf. [accessed 23 June 2025]

A3.15 In some cases, cloud computing was cited to have better security and governance than on-premises computing. TREs can also face technical challenges with data storage, including irregular computing demands and the need to establish systems to allow computing resource to grow and shrink with demand.

Trusted Execution Environments (TEEs)

A3.16 One of the most widely used PETs are Trusted Execution Environments (TEEs). TEEs are secure areas within a computer's processor that keep data and code safe while they are being used. They work separately from the main system and protect information even if the rest of the device is compromised. Data is usually encrypted when it goes in or out of the TEE.[315] Compared to other PETs, TEEs can be more accessible and can require little specialist knowledge, equipment and computational power. They also do not distort the original dataset. TEEs have disadvantages in that there is often little standardisation of data (which limits interoperability), system design choices may introduce vulnerabilities, and there is potential for side-channel attacks.[316]

Homomorphic encryption

A3.17 Homomorphic encryption allows analysis to be conducted on an encrypted dataset without decrypting it.[317] Homomorphic encryption can be 'partial', (where one function, such as division or multiplication, can be conducted), 'full' (where all functions are allowed), or 'somewhat', which lies in between. Partial and somewhat homomorphic encryptions are limited to a small number of calculations, restricting the type of research that can be conducted, while full homomorphic encryption is costly. Homomorphic encryption has advantages in that it can potentially leverage quantum computing technologies to provide data privacy guarantees. [318]

[315] Sommerhalder, M., 2023. Trusted Execution Environment. In: Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) Trends in Data Protection and Encryption Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-33386-6_18. [accessed 23 June 2025]

[316] Shepherd, C., and Markantonakis, K., 2024. Deployment Issues, Attacks, and Other Issues. In: Trusted Execution Environments. Springer, Cham. https://doi.org/10.1007/978-3-031-55561-9_9. [accessed 23 June 2025]

[317] Munjal, K., and Bhatia, R., 2023. A systematic review of homomorphic encryption and its contributions in healthcare industry. Complex Intell. Syst. 9. https://doi.org/10.1007/s40747-022-00756-z. [accessed 23 June 2025]

[318] The Royal Society, 2023. From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis. From privacy to partnership | The Royal Society. [accessed 23 June 2025]

Secure multi-party computation

A3.18 Secure multi-party computation attempts to mitigate privacy and security concerns by computing analysis of a shared dataset between multiple parties without centralising data or enabling data sharing between participants.[319] This helps to secure the data against external attacks because accessing data held by one party is not enough to access a shared dataset, and so it cannot be breached with time or computational power. Since a shared dataset cannot be accessed by one party alone, a breach would need to compromise multiple researchers simultaneously.

A3.19 Secure multi-party computation has advantages because it requires little trust between parties, as the information is theoretically secure. It enables research to be performed across datasets, even where the data providers themselves do not want to directly share the data amongst themselves.

A3.20 A disadvantage of secure multi-party computation is that it is extremely computationally expensive, and as a novel technique it is still undergoing development as a PET. The process of setting it up is complex and may require expert training or a trusted third party. Therefore, it may be intensive in terms of both computation and human resources.

Federated learning

A3.21 Federated learning is a PET which is applicable in cases where researchers are using machine learning. Federated learning can enhance privacy because it involves collaborative training of machine learning models across multiple devices, and potentially on multiple datasets, without having to connect those devices or datasets to each other.[320] This is advantageous because machine learning models can be developed without centralising data, which can provide increased security. This technique also avoids introducing noise to the original dataset.

A3.22 However, federated learning is highly complex and requires specialist expertise to manage. It is also computationally highly intensive and therefore expensive to use. While solving some privacy concerns around sensitive datasets, federated learning can introduce new risks for attackers to trace vulnerabilities through multiple iterations and communications.[321]

[319] Merino, L.H. and Cabrero-Holgueras, J., 2023. Secure Multi-Party Computation. In: Mulder, V., Mermoud, A., Lenders, V., and Tellenbach, B. (eds) Trends in Data Protection and Encryption Technologies. Springer, Cham. https://doi.org/10.1007/978-3- 031-33386-6_17. [accessed 23 June 2025]
[320] Qinbin, L., Wen, Z., and Wu, Z. et al., 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. IEEE Transactions on Knowledge and Data Engineering, 35(4). https://doi.org/10.1109/TKDE.2021.3124599. [accessed 23 June 2025]
[321] Mothukuri, V., Parizi, R. M., and Pouriyeh, S. et al., 2021. A survey on security and privacy of federated learning, Future Generation Computer Systems, 115. https://doi.org/10.1016/j.future.2020.10.007. [accessed 23 June 2025]

A4. Further considerations for mediation

Mediation stages

A4.1 In section 6 of this report, we discuss the role of mediation as a core function of an independent intermediary. We consider independent mediation as a mechanism to resolve disputes between researchers and services, prevent disputes from resulting in legal action, and contribute to addressing the potential power imbalance between researchers and services. We outline some additional considerations regarding the stages of this process in the section below.

A4.2 Mediation may be required at various points, including but not limited to:

- Request stage: Concerns about research status, data access modality, or data types in scope. The data request scope, proportionality, and privacy and security implications – in addition to concerns around perceived conflict of interest from researchers, legitimate interests of users and services' rights to security and/or confidentiality – may also require mediation.
- Delivery stage: Disputes regarding allegations of incomplete, biased, irrelevant, manipulated, or otherwise corrupted or unusable data. Disputes may also arise if the data access modality is perceived as disproportionately limited relative to the assessed risk, data types in scope, research proposal, and other safety enablers (such as data usage agreements).
- Publication stage: Contesting of researchers' findings either prior to publication (if pre-disclosed) or post-publication. Disputes may be based on methodological concerns, data interpretation, and/or managing potential reputational impacts.[322]
- In response to allegations of contract breach: Alleged violation of non-disclosure agreements, misuse or commercial exploitation of data, or other possible breaches. Mediation may be required to determine appropriate redress, such as suspension from the researcher access regime or legal recourse.

[322] Within cybersecurity research, it is common practice for independent researchers to pre-disclose findings. This provides services with the opportunity to verify and make changes to operations without alerting bad actors to vulnerabilities. In some cases, researchers may feel that pre-disclosure could be a threat to the independence of their work, though this could be ameliorated somewhat by not making final publication dependent on the response.

A5. Supplementary functions for an independent intermediary

A5.1 In section 6 of this report, we discuss the role of independent intermediaries in facilitating greater access to information, including the core functions necessary to perform this role effectively. We recognise there are a range of additional functions which could be considered as part of this role and discuss these in the section below.

Guidance and process clarification

A5.2 To clarify the data access process for all parties, the independent intermediary could produce clear guidance for both services and researchers, covering eligibility criteria, data types available for request, access modalities, how data protection frameworks apply, and how to evaluate proposals against eligibility criteria. This guidance would also clarify which services are in scope, what data researchers can and cannot request, and which researchers and research projects are eligible to apply for access. The intermediary could decide to provide further guidance should it deem it necessary.

Proposal management and transparency

A5.3 The independent intermediary could maintain records of all data access requests submitted by researchers, the outcomes (access granted, access refused, or amendments to the request required), and the decision rationales. Requests and outcomes would be published (in full, where safe, or in aggregate form) to increase transparency and help researchers develop better proposals.

Stakeholder engagement

A5.4 The independent intermediary could act as a central body, communicating with both services and researchers to understand their experiences with any researcher access regime. They would regularly engage with services and researchers to identify systemic issues, areas needing clarification, and required improvements to the access framework. The stakeholder engagement function would focus on gaining insights from all parties on an ongoing basis, rather than resolving specific issues, which would be handled via a mediation function. The intermediary could increase the legitimacy of the access regime, inspire greater trust, and collect vital information from stakeholders.

Identifying gaps in research

A5.5 By monitoring the broader online safety research ecosystem, the intermediary could report on understudied areas and research concentrations, helping researchers identify opportunities. This information would also be useful for the intermediary, or any other party who may consider prioritising access requests for research on understudied topics or topics of strategic importance to policymakers.

Technological horizon-scanning

A5.6 The intermediary could undertake regular horizon-scanning activities to stay up to date on technical developments related to data privacy, methods, and other new options and opportunities for enabling and improving researcher access. The independent intermediary could also monitor any new developments which may lead to risks or vulnerabilities, such as the use of novel technologies by malicious actors.

System oversight and reporting

A5.7 The intermediary would monitor the state of researcher access and communicate important findings to stakeholders. This could include proposal success rates, gaps or convergences in the online safety research ecosystem, service and researcher feedback, horizon-scanning risks or opportunities, and recommended system improvements.[323]

[323] Other stakeholders have similarly called for this capacity. See, for example, the following excerpt: "An intermediary body could also monitor and analyse the data access projects it supports, tracking which data policy research uses most, the main areas of study, the reasons for access denials, the safeguards and data protection measures that work, those that don't, and those that need to change as platforms change". Source: Maj, B., and Pavel, V., 2025. Potential unreached: challenges in accessing data for socially beneficial research. Potential unreached: challenges in accessing data for socially beneficial research | Ada Lovelace Institute. [accessed 23 June 2025]

A6. Further considerations for real-time access

A6.1 We discuss issues aspects of real-time access in various sections of this report. We set out some further considerations in the section below.

System oversight and reporting

A6.2 Enabling real-time access via a vetted or approved manner or having an intermediary body or other authority set out a clear standard for scraping in the public interest (including limits on what can be scraped and what the data can be used for) could disincentivise unrestricted web scraping. To enable effective access to more sensitive real-time (or near real-time data), an independent intermediary may consider additional vetting for researchers who are interested in such data.

A6.3 Real-time data access faces technical challenges, including the in-memory computing required. Real-time data analysis requires synchronisation, in which data records are kept continuously up to date. There is a tension between real-time data synchronisation and data privacy. Each place demands on system architecture and require computing power, therefore requiring additional resources. Quality of data presents a challenge where large volumes of real-time data are produced, and data cleaning must be conducted quickly. On the one hand, this may mean that researchers are provided with such large amounts of data that it becomes unmanageable. On the other hand, data may be cleaned using quick decisions, which alters the quality of data and therefore affects the ability to conduct analysis later.[324] These challenges may be particularly pertinent where data is semi-structured or unstructured.[325]

[324] Zheng, Z., Wang, P., Liu, J., et al., 2015. Real-time big data processing framework: challenges and solutions. Applied Mathematics and Information Sciences, 9(6). http://dx.doi.org/10.12785/amis/090646. [accessed 23 June 2025]

[325] Mehmood, E., and Anees, T., 2020. Challenges and solutions for processing real-time big data stream: a systematic literature review. IEEE Access, 8. doi: 10.1109/ACCESS.2020.3005268. accessed 23 June 2025]

E03395491

978-1-5286-5878-2