IJDC | *Conference Pre-print*

Co-Creating Autonomy: Group Data Protection and Individual Self-determination within a Data Commons

Janis Wong
University of St Andrews, Scotland

Tristan Henderson
University of St Andrews, Scotland

Abstract

Recent privacy scandals such as Cambridge Analytica and the Nightingale Project show that data sharing must be carefully managed and regulated to prevent data misuse. Data protection law, legal frameworks, and technological solutions tend to focus on controller responsibilities as opposed to protecting data subjects from the beginning of the data collection process. Using a case study of how data subjects can be better protected during data curation, we propose that a co-created data commons can protect individual autonomy over personal data through collective curation and rebalance power between data subjects and controllers.

Submitted 15 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to University of St Andrews, Jack Cole Building, North Haugh, St Andrews, KY16 9SX, UK. Email: jccw@st-andrews.ac.uk

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Rapid technological innovation in our data-driven society (Pentland, 2013) has changed how data subjects (those about whom personal data are collected), interact with data controllers (those who collect and determine what these data are used for). Privacy scandals such as Cambridge Analytica secretly harvesting 50 million Facebook profiles to build models to influence elections[1], widespread personal data sharing in the Google Nightingale Project[2], and the intrusion of private life[3] and society[4] have made individuals more cautious about the information that they put online. However, data subjects are often left out of the conversation with regards to data protection (protecting data subjects' personal data in relation to the processing of data, where processing refers to any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, storage, or use). While certain laws and technologies attempt to encourage data subject participation and provide them with the ability to control their personal data, this is insufficient as it relies on data subjects having a high-level of understanding of both the law and the resources available for individual redress. Moreover, such redress usually arises after data collection and sharing, after which time the damage may already have been done.

In this paper, we introduce a new framework, the data commons for data protection, in an attempt to improve data subject participation in the data protection process through collaboration and co-creation. The paper is outlined as follows. First, we explore the challenges facing data subjects who feel helpless as a result of the increasingly sizeable data controllers who collect, process, and share their personal data. We identify some of the solutions to this problem and also explore how these may be inadequate. We then introduce the commons as a potential framework for bridging the data protection divide, before detailing data curation as a use case for the data commons, examining what similar frameworks have been established in this space and how our data commons can aid better data protection for data subjects. Finally, we conclude that a co-created data commons can protect individual autonomy over their personal data through collective curation and rebalance power between data subjects and controllers, establishing the requirements of a data commons to help data subjects and explore how this could work in the context of data curation.

The Data Protection Divide

Under existing legislative frameworks and the available technological solutions, data protection focuses on putting responsibilities on data controllers and enforcement. However, this continues to place pressure on individual data subjects to protect their own personal data and seek individual remedies in case of breaches, rather than including them in discussions that can help shape data protection policies and better protect personal data.

[1] Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>, accessed 15 December 2019

[2] I'm the Google whistleblower. The medical data of millions of Americans is at risk <https://www.theguardian.com/commentisfree/2019/nov/14/im-the-google-whistleblower-the-medical-data-of-millions-of-americans-is-at-risk>, accessed 15 December 2019

[3] The House That Spied on Me <https://gizmodo.com/the-house-that-spied-on-me-1822429852>, accessed 15 December 2019

[4] China's "Social Credit System" Has Caused More Than Just Public Shaming (HBO) <https://www.youtube.com/watch?v=Dkw15LkZ_Kw&>, accessed 15 December 2019

The Law and Legal Frameworks

Laws such as the European General Data Protection Regulation (GDPR) and the California Consumer Privacy Act attempt to rebalance power between data subjects and data controllers. As European legislation is already in effect, we focus our work on the regulation and supporting information rights relating to the protection of personal data. The GDPR came into force on the 25th May 2018, introducing significant changes by acknowledging the rise in international processing of big datasets and increased surveillance both by states and private companies. Data subject rights offered by the GDPR include the right of access (Article 15), the right to be forgotten (Article 17), and the right not to be subject to a decision based solely on automated processing (Article 22). The GDPR has also clarified the means for processing data, whereby if personal data are processed for scientific research purposes, there are safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes (Article 89), applying the principle of purpose limitation (Article 5). Information rights such as the right to access to recorded information held by public sector organisations through the Freedom of Information Act and intellectual property rights offered under the Directive on Copyright in the Digital Single Market can support data protection for data subjects in offering greater transparency and alternative remedies to issues relating to personal data.

However, even with legal protection, data subjects continue to be relatively powerless when exercising their rights against the increasingly sizeable and international data controllers (Edwards, 2019). Although people are more aware of their data subject rights, these are not well understood (Norris et al., 2017). Only 15% of EU citizens indicate that they feel completely in control of their personal data (Clusters et al., 2019). Evaluating location-based services, Herrmann et al. (2016) found that individuals do not necessarily know all the inferences that are made using their data and thus do not know how they are used. Importantly, individuals are unaware of, and unable to correct, false inferences, making the collection, transfer, and processing of their location data entirely opaque. With privacy policies written in legalese and privacy-protecting options hidden beneath "dark patterns", data subjects cannot easily find out how their data are reused, aggregated, and anonymised to make decisions about them (Utz et al., 2019). Additionally, laws focusing on placing data protection responsibilities on data controllers and empowering enforcement bodies assume that data controllers understand how to implement those responsibilities and that enforcement is successful. Data protection officers' and authorities' enforcement practices are inconsistent and unclear due to lack of guidance (Norris et al., 2017). Data controllers responding to GDPR Article 20 right to data portability requests provided a large variation of file formats that were not all GDPR compliant and confused the right with other data subject rights (Wong and Henderson, 2019). Kamarinou and others (2016) found inconsistencies in detail and lack of transparency about third-party storage and the processing of personal data of cloud service providers' in their terms and privacy policies. Funding for data protection authorities may also be limited, especially in comparison to the large multinational corporate data controllers. For example, the Irish Data Protection Commission was only given 27% of its requested increase by the Irish Government, totalling €21.1 million, despite increased responsibilities post-GDPR (The Irish Times, 2019).

Other legal frameworks have also been considered as a means to protect data subjects. Data trusts have been proposed as a legal framework for data stewardship and data management (Open Data Institute, 2019). A data trust is a legal structure that facilitates the storage and sharing of data through a repeatable framework of terms and mechanisms. Data trusts aim to overcome difficulties for data sharing and assure the credibility, trustworthiness and reliability of the resulting data analysis (Pinsent Masons, Queen Mary University, and BPE Solicitors, 2019). A data trust aims to respect the interests of those with legal rights in the data, ensure that the data is used ethically according to the rules established by the data trust, and collectively manage individual rights and interests. A key feature of data trusts is that they create rules to govern data sharing where a custodian or steward makes decisions on behalf of data users and subjects. However, this may still leave data subjects out of the data protection process. Without direct data subject engagement, decisions are made on their behalf by trustees as opposed to

with them, representing the data trust and not the data subject. Data trusts also maintain data protection enforcement issues as it relies on the trust to respond to such challenges, delegating data protection responsibilities (Open Data Institute, 2019). A data trust could in theory respond to certain data subject rights, but it would be difficult to mandate rights to portability, access, and erasure that rely on the data controller (Delacroix and Lawrence, 2019). An alternative mechanism that takes this into consideration is data collaboratives that focus on harnessing privately held data towards the public good through collaboration between different sectors[5]. However, individuals and groups of data subjects are still excluded from participation where they are only the potential beneficiaries and are not part of designing the data collaborative framework. A report by Pinsent Masons, Queen Mary University, and BPE Solicitors (2019) suggests that both legal reform in data protection and technical considerations should be used to ensure that a suitable framework preserves the core rights of data subjects and balances the society benefits from data sharing against the interests of the data subjects. Although data protection law and other legal mechanisms provide data subjects with data subject rights and legal protection, data subjects are only seen as beneficiaries of these frameworks and not active participants or contributors to the practices for protecting their personal data.

Technological Solutions

New technologies have also attempted to give users the ability to control their own data. Some tools include Databox[6] (a personal data management platform that collates, curates, and mediates access to an individual's personal data by verified and audited third party applications and services) and Solid[7] (a decentralised peer-to-peer network of personal online data stores that allows allow users to have access control and storage location of their own data). Other applications attempt to facilitate data reuse with privacy-by-design built in, such as the Data Transfer Project[8] (an open-source, service-to-service platform that facilitates direct portability of user data), OpenGDPR[9] (an open-source common framework that has a machine-readable specification, allowing data management in a uniform, scalable, and secure manner), and Jumbo Privacy[10] (an application that allows data subjects to backup and remove their data from platforms, and access that data locally). While these tools are useful if they offer controls that limit the processing of personal data according to data subject preferences, it results in the responsibilisation of data protection from data controllers to data subjects. Existing tools assume that data subjects have a high-level understanding of the data subject rights they have, framing privacy as control and placing individual onus on data protection. It also requires data subjects to trust the companies and the technological services they provide. Further, these solutions do not offer means for collaborative data protection where information gathered from individuals could be shared amongst each other. This disenfranchises data subjects from each other and prevents them from co-creating data protection solutions together through their shared experiences.

Finding a Co-Created, Collaborative Solution

While law and technology separately attempt to address some of these concerns, they may inadequately protect personal data because they rely on a high-level of understanding of both the law and the resources available for individual redress, usually after data collection. Focusing on individual protection assumes that data subjects have working knowledge of relevant data protection laws (Mahieu, Asghari, and van Eeten, 2017), access to technology, and that alternatives exist to the companies they wish to break away from (Ausloos and Dewitte, 2018).

[5] Data Collaboratives <http://datacollaboratives.org/>, accessed 15 December 2019
[6] Databox <https://www.databoxproject.uk/>, accessed 15 December 2019
[7] Solid <https://solid.inrupt.com/>, accessed 15 December 2019
[8] The Data Transfer Project <https://datatransferproject.dev/>, accessed 15 December 2019
[9] OpenGDPR <https://github.com/opengdpr/opengdpr>, accessed 15 December 2019
[10] Jumbo Privacy <https://www.jumboprivacy.com/>, accessed 15 December 2019

Individuals are unaware of how their data are being used after it is collected and are disempowered from the data sharing process as they cannot identify the data controllers to exercise their rights against. Without a data breach, individuals do not know who else is affected, what data controllers and processors are using their data, or how to organise collective action to strengthen their argument for recourse. Even when notified, data subjects rely exclusively on data protection authorities to fully enforce the law on data controllers. Data subjects lack a meaningful voice in creating solutions that involve protecting their own personal data. As a result, although legal and technological mechanisms are being implemented to address existing data protection issues in our data-driven society, the focus on individual protection, asymmetry of information, and the power imbalance between data subjects and data controllers make it difficult for data subjects to engage with the data protection process.

The Commons for Protecting Data Subjects

Given the limited ability for data subjects to voice their concerns and participate in the data protection process, we posit that the protection of data from harms resulting from mass data collection, processing, and sharing could be improved by involving data subjects in collaboration and co-creation.

A framework that considers individual and group collective action, trust, and cooperation is the commons, developed by Elinor Ostrom in her seminal work 'Governing the Commons' (1990). The commons itself guards a common-pool resource that may be over-exploited to depletion. A common-pool resource (CPR) refers to a natural or man-made resource system that is sufficiently large as to make it costly (but not impossible) to exclude potential beneficiaries from obtaining benefits from its use. Ostrom created a commons framework that depends on human decisions and activities, and management of the CPR according to the norms and rules of the community autonomously (Ostrom, 1990). The commons then represents a CPR for transparency, accountability, citizen participation, and management effectiveness, where 'each stakeholder has an equal interest' (Hess, 2006). A central part of governing the commons is recognising polycentricity in decision making, a complex form of governance with multiple centres of decision-making, each of which operates with some degree of autonomy (Ostrom et al., 1961). The commons framework respects the competitive relationships that may exist when managing a CPR. Its success relies on stakeholders entering into contractual and cooperative undertakings or have recourse to central mechanisms to resolve conflicts (Ostrom, 2010). The norms created by the commons are bottom-up, as illustrated by Ostrom's case studies of Nepalese irrigation systems, Indonesian fisheries, and Japanese mountains. The structure of these commons has enabled communities to find stable and effective ways to define boundaries of a common-pool resource, define the rules for its use, and effectively enforce those rules (Ostrom, 2012).

From these case studies, Ostrom identifies eight design principles that mark a common's success (Ostrom, 1990):

1. **Clearly defined boundaries**: Individuals or households who have rights to withdraw resource units from the CPR must be clearly defined, as must the boundaries of the CPR itself;

2. **Congruence between appropriation and provision rules and local conditions**: Appropriation rules restricting time, place, technology, and/or quantity of resource units are related to local cognitions and to provision rules requiring labour, material, and/or money;

3. **Collective-choice arrangement**: Most individuals affected by the operational rules can participate in modifying the operational rules;

4. **Monitoring**: Monitors, who actively audit CPR conditions and appropriate behaviour, are accountable to the appropriators or are the appropriators;

5. **Graduated sanctions**: Appropriators who violate operational rules are likely to be given assessed graduated sanctions (depending on the seriousness and context of the offence), from other appropriators, by officials accountable to these appropriators, or by both;

6. **Conflict-resolution mechanisms**: Appropriators and their officials have rapid access to low-cost local arenas to resolve conflicts among appropriators or between appropriators and officials;

7. **Minimal recognition of rights to organise**: The rights of appropriators to devise their own institutions are not challenged by external governmental authorities; and

8. **For larger systems, nested enterprises for common-pool resources**: Appropriation, provision, monitoring, enforcement, conflict resolution, and governance activities are organised in multiple layers of nested enterprises.

Ostrom's design principles are important in the process of the common's lifecycle, where the limitations of the commons and regulation of CPR can iterate within changes to stakeholders as part of the collective governance process.

Co-creation and Collective Action within a Data Commons

Using the theory and principles of the commons, we suggest that a commons for data protection, a "data commons", can be created to allow individuals and groups of data subjects as stakeholders to collectively curate, inform, and protect each other through data sharing and the collective exercise of data protection rights.

In a data commons, the common property is personal data from data subjects that are used for a specific purpose, and the framework incorporates existing legal and technological structures as well as data subject input and preferences. As personal data are aggregated and used to generate economic value (Singh and Vipra, 2019), data protection should move the focus away from individuals and towards groups, 'from processes of consumption to those of citizenship and accountability' (Taylor, 2017). Diaconescu and Pitt (2017) identify the need to build 'pro-social socio-technical systems' to better balance transparency and privacy, where identified pathologies stem 'from regulatory choices and associated power struggles'. The data commons aims to help contextualise privacy beyond control and move towards privacy as ability and as a state, enabling a mechanistic expectation that addressing differences will make more people comfortable with the same technologies through relationships of respect (Shklovski, 2019). Our data-driven society has also become one that has privacy dependencies, where one person's privacy is implicated by information revealed by others (Barocas and Levy, 2019). The data commons builds upon existing group theories on the risks involved in public use of anonymised personal data (Floridi, 2017) and the necessity for collective rights (Raz, 1986) both before and after data are collected. A data commons encourages iterations of individual and group data protection objectives that can be different, personalised, and change over time (Making Sense, 2018). Figure 1 shows how data subjects are the focal point of the data commons, while other stakeholders are bound by the data subject's desire for better protection of their personal data. Within the framework, the data subject should also be able to interact with data controllers, data managers, researchers, and civil society for better data protection outcomes. A data commons developed using Ostrom's design principles is useful because of the

vast number of stakeholders that have a diverse set of opinions, problems, and preferences on how the data subjects' personal data are managed.

While some data commons have been established in context of data and research archives, they focus on increasing the distribution of data rather than on data protection. Local and international attempts have been made to further open science and open access initiatives through creating research data commons. For example, the Australian Research Data Commons[11] (ARDC) is a government initiative that merges existing infrastructures to connect digital objects and increases the accessibility of research data. The National Cancer Institute (NCI) also has a Genomic Data Commons (GDC) that is used to accelerate research and discovery by sharing bio-medical data using cloud-based platforms. In Europe, the European Open Science Cloud[12] (EOSC) is a Europe-wide digital infrastructure set up by the European Commission for research, with the aim to simplify the funding channels between projects. The EOSC was inspired by the F.A.I.R. principles, representing Findable, Accessible, Interoperable and Reusable data sharing and aims to become a 'global structure, where as a result of the right standardization, data repositories with relevant data can be used by scientists and other to benefit mankind' (EOSC European Commission, 2019). While these frameworks recognise that

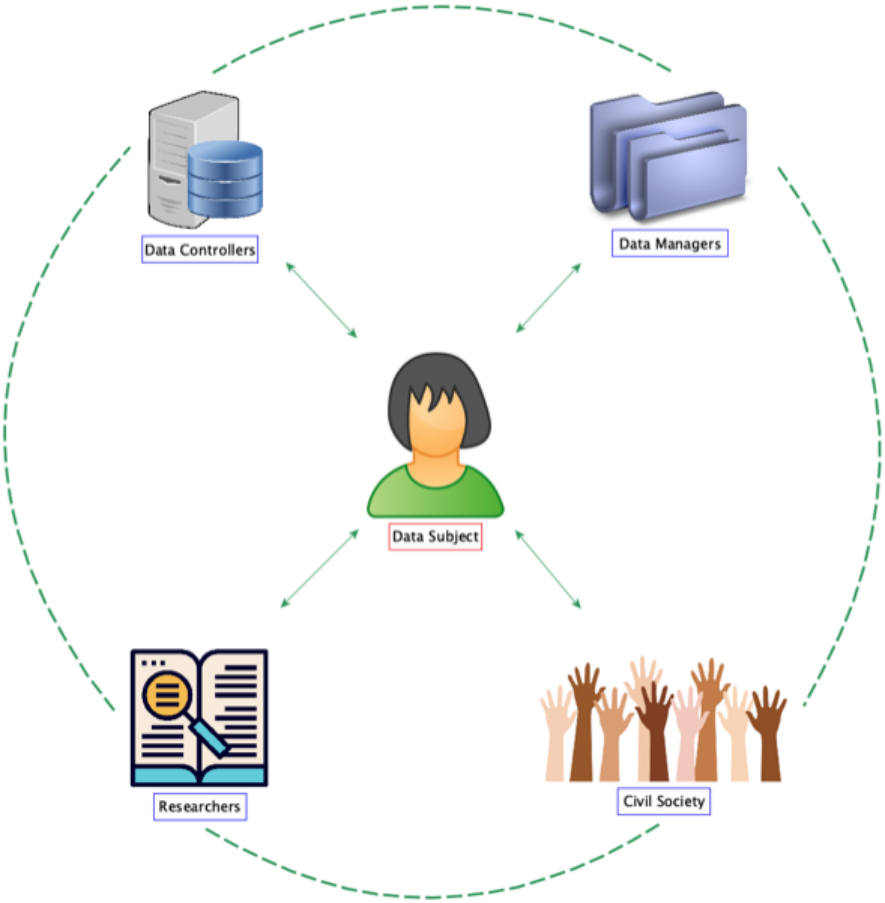

Figure 1. In a data commons (green), the data subject is at the centre. In this framework, the data subject and their personal data is the most important, and other stakeholders are only considered in context of the data subject's data protection. The different stakeholders represent the polycentricity of all the systems which have influence over data subjects.

[11] Australia Research Data Commons <https://ardc.edu.au>, accessed 15 December 2019
[12] European Open Science Cloud <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>, accessed 15 December 2019

the information and knowledge are collectively created, their implementations are hierarchical and top-down without input from archive participants or repository managers. Additionally, existing commons and data commons frameworks do not protect the personal data within them as they prioritise data sharing over data protection, particularly on data curation and reuse. The EU frameworks acknowledge the GDPR as the source for the right to data protection under the law, however it is currently unclear as to how it is implemented.

Other applications of data commons include smart cities. Governments have used commons principles to take more responsibility over its citizen's personal data (Decode European Commission, 2018). These include the Bristol Approach city commons[13], the Barcelona City Council Digital Plan[14] and Data Commons[15], and the Commons Transition Plan for the City of Ghent urban commons[16]. However, smart city frameworks often rely on dynamic consent (Teare, 2019) and informed consent (Mikkelsen et al., 2019). As suggested in the previous discussion, the lack of knowledge and understanding of data protection by data subjects limits their ability to meaningfully consent to the collection, processing, and result of their personal data. Data subjects are also only able to make their decisions based on information that is provided to them and have no information as to the dataset they may be a part of. A data commons framework for data protection can move beyond those of existing research- and smart city-focused commons, applying Ostrom's theory so that data subjects can co-create the data protection responsibilities alongside data controllers, data managers, researchers, and civil society.

Building upon general principles of existing data commons such as the Data Biosphere (Denny et al., 2017) on a modular, community-driven, open, standards-based governance and the National Library of Medicine (NLM) on the necessity for security, searchability, standardisation of metadata, and the management of access control (Brennan, 2018), a data protection-focused data commons can serve as a technical solution within data protection legal structures. Unlike existing research data commons frameworks that focus on the dissemination of data and increased funding opportunities for research, a data commons for data protection focuses on data subjects to further their ability to protect the processing of their personal data. The framework can be used to balance the protection of the rights of data subjects with safeguarding the scientific process and integrity of research results for researchers during the data curation process. A data commons is useful because mechanisms such as licensing for data archives may not be useful for data protection even if they limit forms of data reuse (Guadamuz, 2006). For a data archive data commons, the data subject can better maintain control over their data through the research process (Powell, 2015) and reduces the risks of personal data being misused, with severe repercussions to the data subject. This is especially important when curated data that used to be in the public domain no longer are, with wider ramifications if the data are socially and politically sensitive, such as Twitter data on the 2014 Hong Kong Umbrella Movement (Tromble and Stockmann, 2017). While digital data archives aim to preserve, reuse, and promote ethically sound, methodologically well-grounded research, there continues to be insecurity by researchers about data sharing, where social media data sharing may become hidden and informal (Weller and Kinder-Kurlanda, 2017). A data protection-focused data commons applied to data archives of curated data could also help clarify who the data controllers are from a wide range of archive owners, dataset owners, or participants, identifying who is accountable to and for the publicised data and the resulting reuse outputs. Without a data commons, questions such as 'Who maintains control over curated data?', 'How can data controllers limit who and how collected data is reused?', and 'How can data subjects exercise

[13] The Bristol Approach <https://www.bristolapproach.org/>, accessed 15 December 2019
[14] Barcelona City Council Digital Plan <https://ajuntament.barcelona.cat/digital/sites/default/files/LE_MesuradeGovern_EN_9en.pdf>, accessed 15 December 2019
[15] Barcelona Data Commons <https://ajuntament.barcelona.cat/digital/en/blog/ethical-and-responsible-data-management-barcelona-data-commons>, accessed 15 December 2019
[16] Commons Transition <https://commonstransition.org/commons-transition-plan-city-ghent/>, accessed 15 December 2019

their data protection rights when sensitive and identifiable personal data that could potentially be deanonymised is curated?' remain difficult to answer. With a data commons, existing standards and review mechanisms such as research ethics board reviews, funding body requirements, and institutional policies can be integrated into the data commons, acting as the first level of safeguard for data protection for data subjects in the future. Researchers that work in data curation and examine data protection practices on archive data reuse can offer privacy principles for individuals and organisations to adhere to. Although ethics approval may be granted by institutions, if the research data reuse by third parties is granted, the participant as the data subject may not know what stakeholders have access to their personal data and for what purpose. In cases where ethics approval is dubious, researchers may take their work outside of institutions and use data subjects' personal data in commercial ways, as in the case of Cambridge Analytica. This can be mitigated with a data commons where future researchers looking to use the data archive can utilise it to see what data limitations have been discussed and

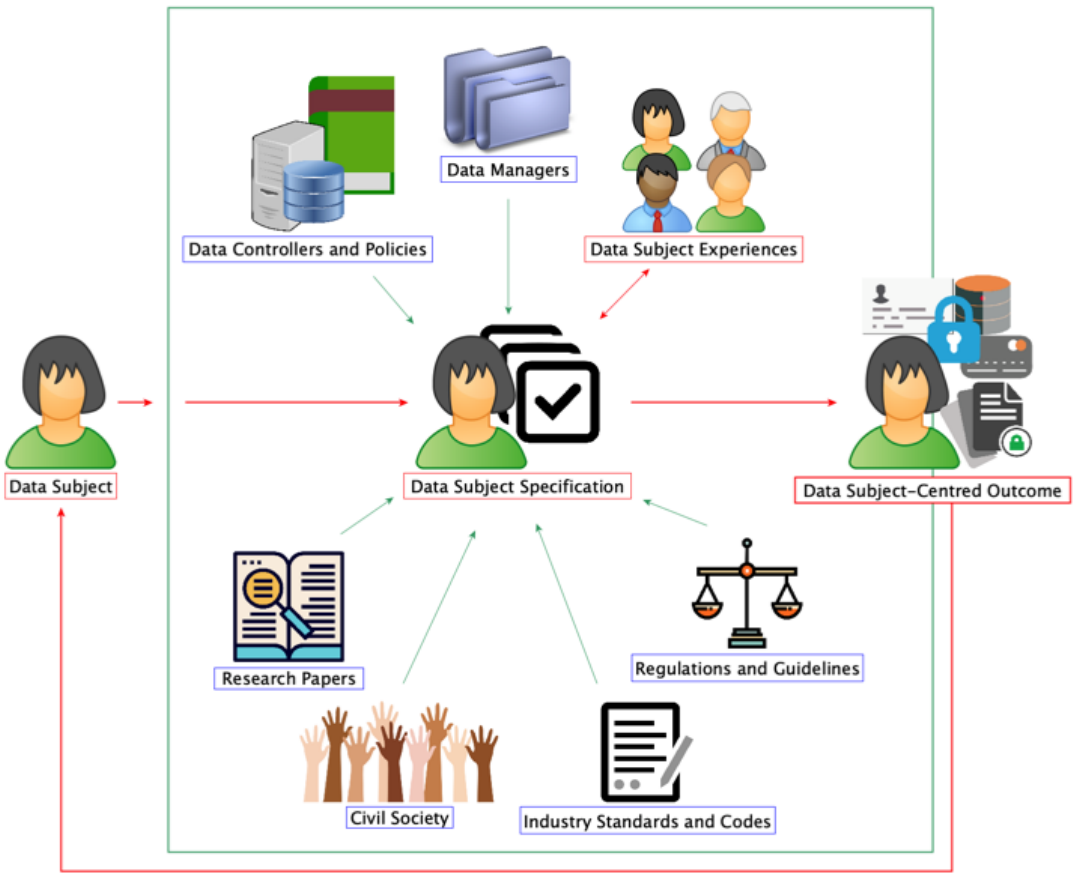

Figure 2. In a data commons (green), the data subject specifies to what extent they would like their data to be protected based on existing conflicts and challenges pre-identified within the data commons for the use case (red). No prior knowledge of existing law, norms, or policies are required. Along with stakeholder information (blue), the data subject specification is then used to inform their data protection outcome that is generated from the system. As the outcome is data subject-centred, decisions ensuring the protection of the data subject's personal data may override existing preferences, policies, or standards set by other stakeholders. Data subjects can return to and review their outcome, add their data subject experiences to the data commons, and participate in the co-creation process at any time.

set by data subjects, proceeding with reuse if those requirements are met. Collective participation by different stakeholders further allows research participants to assess these forms of use and curate the data archive data for themselves, engaging with their own research interests while participating in the community interest. The data commons acts as a new means for data management where the reasons for use, limitations on reuse, and recourse after data are aggregated and anonymised are all contained within one ecosystem.

To implement a data commons, various legal and technological components need to be created for stakeholders to be engaged. Figure 2 shows how a data subject specifies to what extent they would like their data to be protected based on existing conflicts and challenges pre-identified within the data commons for the use case. In addition to data subject preferences, using information such as data controller policies, research papers, and input from civil society, a data subject specification is created and used to inform their data protection outcome that is generated from system. For example, the data subject-centred outcome could ensure that the data subjects' archive data are only to be reused by the researchers and data managers directly associated to the archive within the data commons and not by specific external researchers. The archive researchers would automatically be notified of the data subject's preferences. This allows the data subject to set their own limitations of how their data are used as opposed to it being decided by the archive itself. As the outcome is data subject-centred, decisions ensuring the protection of the data subject's personal data may override existing preferences, policies, or standards set by other stakeholders. Data subjects can return to and review their outcome, add their data subject experiences to the data commons, and participate in the co-creation process at any time. Data controllers can address new risks before collecting data, minimise the potential for data breaches, and meet stakeholder demands. Other stakeholders, such as researchers and civil society, can participate in the data commons to make the data sharing process more transparent, support exercising group rights, and provide information and standards for FAIRsharing (Sansone et al., 2019). These requirements aim to decrease the power imbalance between data subjects and controllers. Mapping out the development of a data commons into Ostrom's CPR design principles, a data commons will be clearly defined based on its use case, where each stakeholders' role is detailed. All data subjects that would like to find out more information about the use case, contribute, or co-create are free to participate in the data commons. Any bad practices, unethical behaviour, and data breaches will be identified by the data commons system, with the remedies updated as stakeholders respond. Data subjects and other stakeholders can collaborate and establish their own norms, such as co-creating data sharing practices which promote data protection by design and facilitate data reuse for data research projects amongst a group of researchers.

Data Commons for Data Curation

We now outline data curation of public data archives as a use case for the data commons, assess how a data commons could address stakeholder issues by increasing accountability for data subjects' personal data, encourage collaborative curation, and allow for data protection to be an iterative process.

Using the Umbrella Movement as the use case for a data curation data commons, created at the beginning, the framework allows potential research participants as data subjects to see what and how their data will be shared with data controllers and researchers, raising and addressing any concerns respectively. Applying Ostrom's design principles, the data commons for data curation will have clearly defined boundaries as to the kinds of data and metadata it will archive and when such archival will cease. In this example, data subjects would like to publicise their experiences from the Umbrella Movement on a platform in the public domain. To participate, the data subject can identify the most applicable data commons by searching for keywords such as social media, data curation, research data, and data reuse. The identified data commons would include information about data controller policies on personal data, research and

archiving, other data subjects' experiences and outcomes from exercising their data subject and information rights, recent news and scandals on data controllers, and expert and researcher findings from their work based on relevant topics and tags. This allows the data subject to identify what settings there are for preferences such as limiting the audience, how information can be published in public and in private, whether data can be deleted, how published information could be used (by who, how, and the process), what intellectual property policies are for the published information, and how other data subjects felt about the platform's responses to information rights based on their experiences.

Without a data commons, the data subject would have to search for this information independently, looking for forums for information. After identifying the most relevant framework, Twitter for example, based on different stakeholder knowledge, the data commons uses the information and prompts the data subject to select a few preferences by answering questions based on the conflicts and challenges that have arisen from them. These questions could include 'Do you want your data or posts to be publicly archives? Available to researchers?', 'If your data or posts were deleted at a later date would you want to notify researchers of such request, for example to not include your data in future studies?', 'Do you want a mechanism to hide all or some of your data and posts?'.

Based on the data subject's responses to those questions, the data commons chooses platform settings and data actions for the data subject that best aligns to their aims. These may override platform policies based on data subject requests. For example, during the Umbrella Movement, Twitter decides to change its historical archive policy regarding the removal of deleted tweets. This would be automatically reflected in the data commons through technical means, notifying the stakeholders in the system. Although Twitter automatically removes deleted Tweets from its data archive, if a data subject would like for that information to be kept in certain pieces of work, researchers have a right to retain such data until further notice by the data subject. Requests by data subjects could be specific such as any Tweets that include the term 'Umbrella Movement' can be kept while ones with 'universal suffrage' or 'Hong Kong independence' can be removed.

Researchers would be notified of these preferences. The system can then be updated to match personal preferences with secondary resources to create a more comprehensive picture of what preferences data subjects would collectively like with regards to data curation, sharing, and reuse. Experts such as Tromble and Stockmann can advise data subjects on what they can do in light of new policy changes as well as data controllers on how to address any concerns raised.

Even without misuse and with GDPR Article 17 right to erasure, certain forms of data archival and curation can make removing personal data difficult, particularly if such data has already been reused in research. If Twitter suffered from a data breach and released the personal data of data subjects during the Umbrella Movement, in a data commons, the breach can be addressed by supporting data subjects in exercising the right to erasure and sending notifications to data controllers to request their data be removed. Researchers who have used the affected datasets and data would also be prompted of the breach and be required to issue corrections in their work and remove identifiable data in relation to the data subject from their data archives should the right to erasure be exercised. Automatic detection of subsequent attacks caused by the data breach can also prompt the system to alert and support data subjects to exercise their data subjects rights if they haven't already as well as look for new alternative platforms that better support data protection practices.

In deciding the best platform and settings for the data subject's purpose of broadcasting the Umbrella Movement, further advice is also provided on how data from the data subject can be best protected. This includes: setting up an account with a disposable email, having an anonymous platform username, setting up tools that can automatically delete the data subject's posts, links to how to exercise information rights on the platform, and the successes and failures of other data subjects in this regard. This information is saved in the data commons and is accessible by the data subject at any time. Any information that the data subject has gathered can also be put into the data commons.

By applying Ostrom's principles to a data curation data commons, established methodologies and organisational structures are built into the data protection-focused framework while enabling individuals and groups whose data form part of the datasets to determine how their personal data are used. A data commons framework enables data protection because it operates as a polycentric system, working in tandem with data protection law and policy, data subjects and their rights, data controllers, data managers, and researchers to develop a better understanding of how personal data can be protected. The data commons simplifies the data protection rights procedure by including information, instructions, and templates on how rights should be collectively exercised, giving data subjects and opportunity to engage with and shape data protection practices that govern how their personal data is protected.

Conclusion and Future Work

To overcome the limitations of laws and technologies in protecting group data, we propose a co-created data commons to maintain individual autonomy of personal data. Identifying requirements for the data commons based on Ostrom's framework on the commons, the data commons supports more accountable data protection practices, collaborative data management, and data sharing for the benefit of data subjects and data controllers. Applying the data commons to the use case of data curation, and specifically to the collection and repurposing of Twitter data of the 2014 Hong Kong Umbrella Movement, we have shown how this framework could assist data subjects in limiting and preventing the misuse of identifiable public, sensitive personal information. The data commons encourages the co-creation of data protection for data subjects while also allowing them to participate in shaping how other stakeholders manage their personal data.

Future work can use the data commons requirements established in this paper to build a prototype of a data commons for data archival and data curation. In order to assess whether a data commons is useful to data subjects, technical and non-technical requirements should be developed to identify what stakeholders should be involved and what data could be incorporated into the system. Surveys and interviews could be conducted to data subjects to identify the issues that the data commons could prioritise in helping them achieve the data protection they want. Experts can also provide input on what they believe are factors that support a successful data commons based on their knowledge of other commons frameworks. Further, the data commons should be built based on the principles of the GDPR such as data protection by design and by default (GDPR Article 25), where users can be anonymous in the data commons and any data included should be pseudonymised unless explicitly allowed to be identifiable by the data subject. With the use of Ostrom's commons framework to develop a data commons, a prototype could be built to test the feasibility of the system in tackling stakeholder issues.

To conclude, in this paper, we established a framework for a co-created data commons that can rebalance power between data subjects and controllers. By including wider stakeholder participation such as researchers and civil society, the polycentric system places the data subject in the centre, supporting data subjects from the beginning of the data protection process, prior to any data being collected. Acknowledging the legal mechanisms and technological tools available to data subjects in protecting their personal data, the data commons not only incorporates existing data protection safeguards but also takes into consideration the needs of the data subject as the fundamental means to protect individual autonomy over their personal data through collective action and co-creation.

References

Ausloos, J. and Dewitte, P. (2018). Shattering one-way mirrors — data subject access rights in practice. *International Data Privacy Law,* 8(1). 4–28. doi:10.1093/idpl/ipy001

Barocas, S, and Levy, K. (2019). Privacy Dependencies. *Washington Law Review (Forthcoming).* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3447384

Brennan, P. (2018). *What makes a data commons work.* Accessed from https://nlmdirector.nlm.nih.gov/2018/04/24/what-makes-a-data-commons-work, 15 December 2019.

California Legislative Information. 'The California Consumer Privacy Act of 2018 Assembly Bill 375' [2018] Retrieved from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

César J., Debussche, J., and Van Asbroeck, B. (2017). *White Paper - Data ownership in the context of the European data economy: proposal for a new right.* Accessed from https://www.twobirds.com/en/news/articles/2017/global/data-ownership-in-the-context-of-the-european-data-economy, 15 December 2019.

Custers, B., Sears, A.M., Dechesne, F., Georgieva, I., Tani, T., van der Hof S. (2019). Conclusions. In: *EU Personal Data Protection in Policy and Practice. Information Technology and Law Series*. 29. T.M.C. Asser Press, The Hague. doi: 10.1007/978-94-6265-282-8_10

Decode European Commission. (2018). *Reclaiming the Smart City: Personal data, trust, and the new commons*. Retrieved from https://media.nesta.org.uk/documents/DECODE-2018_report-smart-cities.pdf

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.* doi:10.1093/idpl/ipz014

Denny, J., Glazer, D., Grossman, R. L., Paten, B., and Philippakis, A. (2017). *A Data Biosphere for Biomedical Research.* Accessed from https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d, 15 December 2019.

Diaconescu, A. and Pitt, J. (2017). Technological Impacts in Socio-Technical Communities: Values and Pathologies. *IEEE Technology and Society Magazine*. 36(3). 63–71. doi:10.1109/MTS.2017.2728780

Edwards, L. (2019). Data Protection: Enter the General Data Protection Regulation. In *Law, Policy and The Internet.* 77-117. Oxford, United Kingdom: Hart Publishing.

EOSC European Commission. (2019). *European Open Science Cloud (EOSC) strategic implementation plan*. doi:10.2777/202370

European Data Protection Board (2019). *Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects.* Retrieved from https://edpb.europa.eu/sites/edpb/files/consultation/edpb_draft_guidelines-art_6-1-b-final_public_consultation_version_en.pdf

European Union. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Retrieved from https://eur-lex.europa.eu/eli/dir/2019/790/oj

European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0001.01.ENG

Floridi, L. (2017). Group Privacy: A Defence and an Interpretation In *Group Privacy: New Challenges of Data Technologies*. 83–100. Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-46608-8_5

Freedom of Information Act (2000). Retrieved from https://www.legislation.gov.uk/ukpga/2000/36/contents

Guadamuz, A. (2006). Open science: open source licences for scientific research. *North Carolina Journal of Law and Technology,* 7 (2). 321-366. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=886906

Herrmann, M., Hildebrandt, M., Tielemans, L., and Diaz, C (2016). Privacy in Location-Based Serivces: An Interdisciplinary Approach. *SCRIPTed,* 13 (2). Retrieved from https://script-ed.org/?p=3151

Hess, C. (2006). Research on the Commons, Common-Pool Resources, and Common Property. *Indiana University Digital Library of the Commons.* Retrieved from http://dlc.dlib.indiana.edu/dlc/contentguidelines

Kamarinou, D., Millard, C., and Kuan Hon, W. (2016). Cloud Privacy: An Empirical Study of 20 Cloud Providers' Terms and Privacy Policies–Part II. *International Data Privacy Law*, 6(3). 170–194. doi:10.1093/idpl/ipw004

Kemp, R. (2019). *Data trusts and frameworks are gaining traction and on the cusp of widespread adoption*. Accessed from https://www.lexology.com/library/detail.aspx?g=28b042fa-1027-4c6c-b1a8-70250691c226, 15 December 2019.

Mahieu, R., Asghari, H., and van Eeten, M. (2017). Collectively Exercising the Right of Access: Individual Effort, Societal Effect. *GigaNet (Global Internet Governance Academic Network) Annual Symposium 2017.* doi:10.14763/2018.3.927

Making Sense. (2018). *Citizen Sensing: A Toolkit.* Retrieved from http://making-sense.eu/wp-content/uploads/2018/01/Citizen-Sensing-A-Toolkit.pdf

Mikkelsen, R.B., Gjerris, M., Waldemar, G. et al. (2019). Broad consent for biobanks is best – provided it is also deep. *BMC Med Ethics* 20, 71. doi:10.1186/s12910-019-0414-6

Norris, C., de Hert, P., L'Hoiry, X., and Galetta, A. (2017). *The Unaccountable State of Surveillance*. Vol. 34. Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-47573-8

Open Data Institute. (2019). *Data trusts: lessons from three pilots*. Retrieved from https://theodi.org/article/odi-data-trusts-report/

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, 20(4). 550-557. doi:10.1016/j.gloenvcha.2010.07.004

Ostrom, E. (2012). *The Future of the Commons: Beyond Market Failure & Government Regulations*. London, UK: Institute of Economic Affairs.

Ostrom, V., Tiebout, CM., Warren, R. (1961). The organization of government in metropolitan areas: a theoretical inquiry. *American Political Science Review*, 55. 831-842.

Pentland, A. (2013). The Data-Driven Society. *Scientific American*, 309 (4), 78–83. doi:10.1038/scientificamerican1013-78

Pinsent Masons, Queen Mary University and BPE Solicitors (2019) *Data trusts: legal and governance considerations.* Accessed from https://www.bpe.co.uk/media/177005/24779-general-legal-report-on-data-trusts-digitalv5-lr-final.pdf, 15 December 2019.

Powell, A. (2015). Open culture and innovation: integrating knowledge across boundaries. *Media, Culture and Society*. 37 (3). 376-393. doi:10.1177/0163443714567169

Raz, J. (1986). *The Morality of Freedom*. Oxford, UK: Oxford University Press.

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L. and Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature biotechnology*. 37. 358. doi:10.1038/s41587-019-0080-8

Shklovski, I. (2019). Privacy as ability or a state: An argument for a relational view. Accessed from https://blogit.itu.dk/virteuproject/2019/10/28/privacy-as-ability-or-a-state-an-argument-for-a-relational-view/, 15 December 2019.

Singh, P. J. and Vipra, J. (2019). Economic rights over data: necessity and a framework for community data ownership. *Development.* doi:10.1057/s41301-019-00212-5

Taylor, L. (2017). Conclusion: What Do We Know About Group Privacy? In *Group Privacy: New Challenges of Data Technologies*. 225–237. Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-46608- 8_12

Teare, H. (2019). *What is Dynamic Consent?* Accessed from https://www.hra.nhs.uk/about-us/news-updates/what-dynamic-consent/, 15 December 2019.

The Irish Times (2019). *Is Ireland breaching EU rules by underfunding data regulator?* Accessed from https://www.irishtimes.com/business/technology/is-ireland-breaching-eu-rules-by-underfunding-data-regulator-1.4047897, 15 December 2019.

Tromble, R. and Stockmann, D. (2017). Lost Umbrellas: Bias and the Right to Be Forgotten in Social Media Research. In *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts.* 75-94. Oxford, UK: Peter Lang. doi:10.3726/b11077

Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T. (2019). (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19).* ACM, New York, NY, USA. 973-990. doi:10.1145/3319535.3354212

Weller, K. and Kinder-Kurlanda, K. (2017). To Share or Not to Share? Ethical Challenges in Sharing Social Media-based Research Data. In *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts.* 115-132. Oxford, UK: Peter Lang. doi:10.3726/b11077

Wong, J. and Henderson, T. (2019). The right to data portability in practice: exploring the implications of the technologically neutral GDPR. *International Data Privacy Law.* 9 (3). 173–191. doi:10.1093/idpl/ipz008