

Data governance and the Datasphere

Introduction

In recent years, the term data governance has garnered growing attention. It has moved from being a niche topic, addressed solely as a technical aspect of data sharing projects, or within enterprise Information Communication Technology (ICT) disciplines as a companion to data management, to becoming the overarching container for thinking about both data protection and access to data.

In particular, data governance has emerged as a key framework within which to address both the opportunities and the risks of data collection, sharing and use. This reflects a growing recognition of the importance of data within wider processes of governance, as well as the potential power data has as both a resource for progress, and a catalyst of harm when misused.

However, the relatively rapid convergence of interest from policy makers, technologists, activists and practitioners on 'data governance' comes with some challenges. Different agendas, conceptualizations, concerns, and areas of emphasis collide, and there is not yet a coherent field of data governance research.

By providing an initial mapping of who is writing about data governance, and the kinds of topics being addressed, this paper offers groundwork for a response to the call from De La Chapelle and Porciuncula (2021) for work on data governance that can *"reframe the discussion, harness emerging innovative approaches, and engage in a much needed global, multistakeholder and cross-sectoral debate"*.

To support that reframing, this paper also looks at the emerging conceptual framework of the Datasphere, understood as "the complex system encompassing all types of data and their dynamic interactions with human groups and norms" (de La Chapelle and Porciuncula, 2022). The conceptual shift this introduces invites a move from discussing relatively flat notions of 'data governance', to discussing 'governance of the Datasphere': bringing into focus the interaction of datasets, norms and human groups. This more holistic framework offers an opportunity for the distinct contributions of different data governance writings to be more clearly described, and allows new questions to be raised addressing the respective responsibilities of different stakeholders and the interplay of norms, rather than simply the regulation of data through policy and law.

In short, this paper is part of a response to the 2021 'We Need to Talk About Data' report (De La Chapelle and Porciuncula, 2021) and sits alongside work mapping the state of the Datasphere as initiated with the 2022 '[Datasphere Governance Atlas](#)'. The bibliometric-driven literature review below proceeds with two primary goals:

- To produce a preliminary map of academic and associated writing on the topic of **data governance**, exploring the topics, approaches and silos reflected, and to explore the points of connection and conflict with a Datasphere framework;
- To explore how **developing narratives of the Datasphere** interface with the growing data governance literature.

Given the rapid evolution of the literature (100s of data governance papers have been published in the short period since we closed our data collection to start writing up), it should be seen as a snapshot intervention, and a moment in the conversation, rather than a conclusion thereof.

Overview

This is not a systematic or exhaustive literature review. Rather, we have pursued a number of strategies to gain an overview of, and to distill insight from, relevant literature, with a primary, but not exclusive, focus on academic writings.

After a section on methodology, the paper introduces the growth of writing on data governance, prior work to define the scope of data governance, and general remarks about the different themes now covered within the overarching concept of data governance. The following section uses bibliographic tools to map the authors and topics evident in existing published work on data governance, and to highlight patterns of connection and fragmentation. The paper then turns to the concept of the Datasphere, outlining how it has developed, before setting out the particular contributions that the concept of the Datasphere can bring to strengthening the data governance literature.

This paper is accompanied by a growing annotated bibliography hosted on Zotero, and a set of interactive network diagrams that readers are encouraged to explore to support their own exploration of contemporary work on data governance.

Methodology

This review deploys several overlapping strategies to provide an overview of current writing on data governance. Whilst the analysis that follows draws primarily on the academic literature, analysis of published books (via the Google Books corpus) and gray literature (via a corpus based on the Datasphere Governance Atlas) are used to provide complementary insights.

Published Books / 'Popular literature'	Academic literature	Gray literature (Reports & 'Policy literature')
To look at general trends in use of key data governance-related terminology.	To identify particular clusters of academic production, and the topics addressed in past and present academic work.	To identify topics and terms present in non-academic policy papers and reports.
Source & analysis		
Using the Google NGrams corpus of digitized books to indicate potential trends in popular and	Bibliometric analysis was carried out using dimensions.ai corpus of academic papers, searched for	Publication pages from the websites of organizations listed in the Datasphere Governance Atlas

scientific use of selected terminology, explored through selectively reviewing books identified (Michel <i>et al.</i> , 2011). Searches were made through the web interface, and data extracted.	<p>“data governance” in titles and abstracts. Trends were cross-checked with the same search on Web of Science to identify potential effects of corpus choice on conclusions. Results were analyzed using VOSViewer (van Eck and Waltman, 2010), Biblioshiny (Aria and Cuccurullo, 2017a), and custom scripts to identify key terms.</p> <p>Separately, literature was selectively added to a Zotero database, and abstracts read and coded. Selected articles were reviewed and a research journal was kept to identify themes. Selected citations were followed and included in the Zotero dataset.</p>	(Datasphere Initiative, 2022) were reviewed for publications relating to data governance which were downloaded, mostly in PDF format. Publications were converted to plain text, and selected terms extracted using a custom script. Term co-occurrence was calculated for terms present in the same document. Selected publications were reviewed to support qualitative analysis.
<p style="text-align: center;">Limitations</p> <p>In addition to the general limitation that the corpus used covers only English language literature up to the end of 2021, and the selection of publications for additional review has been based on the judgment of a single report author, each source has particular limits. Additional limitations include:</p>		
The range of books included in the Google NGrams corpus changes over time, which can affect prevalence of certain terms (Pechenick, Danforth and Dodds, 2015). Similar terms or different capitalization (e.g. ‘Data Governance’ and ‘data governance’) are not combined.	Analysis relies upon the Dimensions.ai algorithms to correctly extract and disambiguate paper meta-data, including author names and abstracts. Analysis relies upon country extraction from Biblioshiny to identify sites of research production. Some duplicate papers were detected in the dimensions.ai dataset, which may have some impact on final counts, although this is minor.	Data collection relied on the text search on organization websites to locate data governance related publications, and not all publications could be accessed or have text extracted. Whereas term analysis for academic papers is based on abstracts, policy reports often lack abstracts, so term co-occurrence was calculated based on full text as extracted using a PDF to text tool.

Data governance is a growing field, bringing together formerly distinct areas of focus

Given the proliferation of current work and writing on data governance (the recent Datasphere Governance Atlas (Rossini and Lach, 2022) counts no less than 261 organizations focused to some extent on data governance topics), it can be surprising to note that the term “data governance” has only entered the research and policy lexicon at scale in the last decade. **Use of the term in academic paper titles and abstracts has increased almost five-fold between 2015 and 2021, and looks set to increase even further in 2022.**

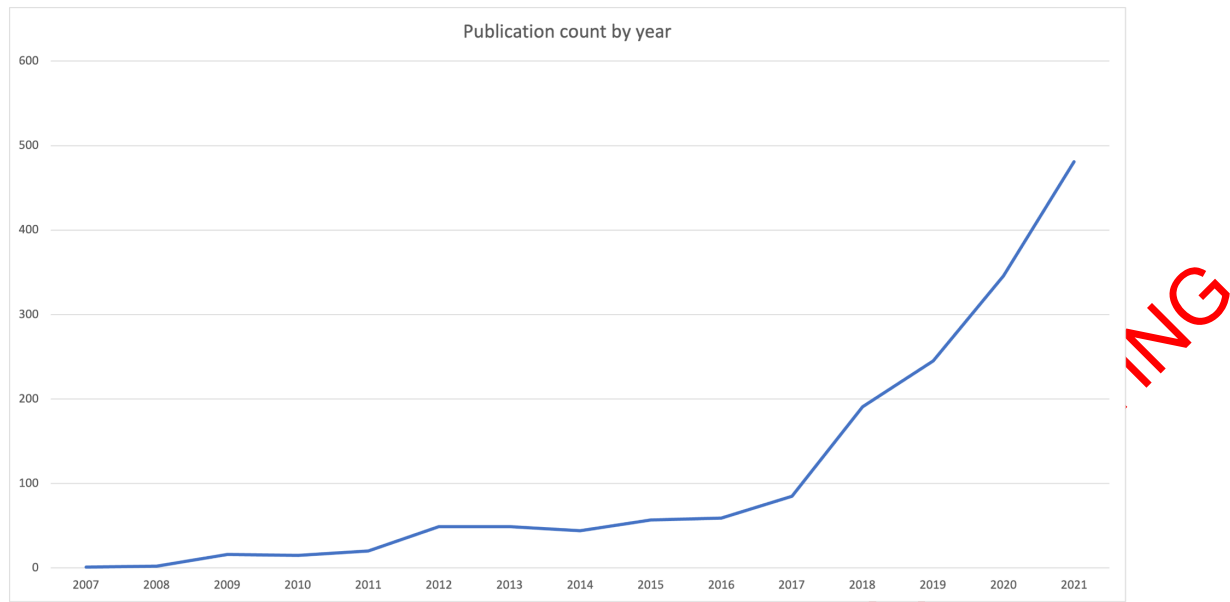


Figure N: Graph showing number of publications recorded by publication year in the dimensions dataset that include the term “data governance” in their title, abstract or keywords.

The rapid development of data governance discourse does not mean that existing debates have been entirely subsumed within data governance. A look at the presence of other terms in the popular literature highlights that readers are much more likely to encounter work on ‘data protection’ or ‘data management’ in books or technical manuals than they are to find discussions of data governance. Even topics like open data, arguably just one particular approach to governing data, have received significantly more direct attention in recent years than data governance has.

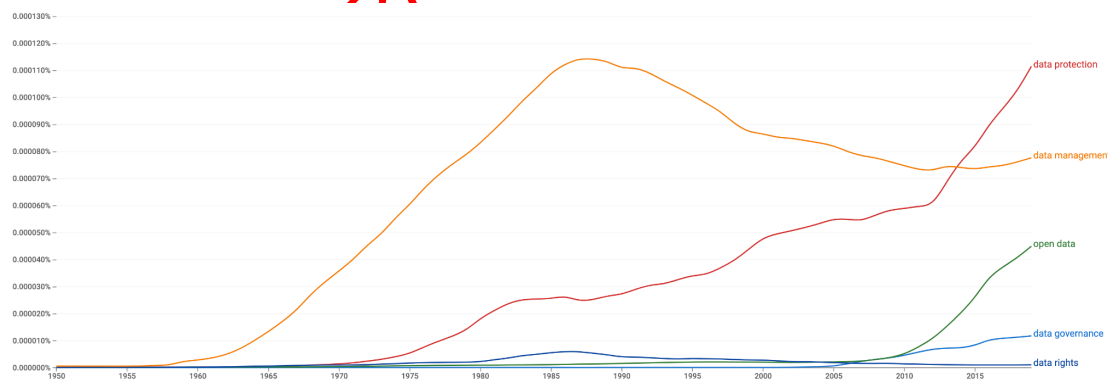


Figure 1: Comparative mentions of terms ‘data protection’, ‘data management’, ‘open data’, ‘data governance’ and ‘data rights’ in Google Books ngrams corpus 1950 - 2019.

These patterns in the popular literature are also broadly mirrored in scientific production, where there are still many more papers published about data protection, data management, or open data, each year than explicitly using the term data governance in their title or abstract. However, in recent years the volume of data governance papers in the academic literature has seen higher year-on-year percentage growth than those focused on either data protection or data management.

Ultimately, the continuous, but incomplete, rise of data governance as a framing term in both research and policy should give pause. Authors are adopting the language of data governance from a range of starting points, and this will color what falls within the scope of their data governance definitions and prescriptions. For instance, as mentioned above, much of the literature on data governance within computing and management considers data governance only within the boundaries of an enterprise, whereas social studies and gray literature frequently explore data governance as a societal issue. At the same time, the ongoing production of work framed in terms of data protection, data management and open data (to name just a few areas) may have much to contribute to thought around the development of norms, policies and practices of governing data, and the Datasphere, even if not directly adopting a data governance language.

The following sections look specifically at literature that uses data governance terminology, but readers should be mindful that this therefore only captures one set of the insights potentially available to inform data governance understanding and action.

Previous literature reviews reveal the diversity of the field

Many of the topics that increasingly fall within the broad frame of data governance were formerly discussed in terms of data protection (Greenleaf, 2012), data management (Panian, 2009; Ladley, 2019), or open data (Davies *et al.*, 2019; Verhulst *et al.*, 2020), each with their own particular agendas around privacy, exploitation of enterprise data assets, and public re-use of data respectively. A shift towards framing these topics within the broader envelope of data governance responds to a recognition of the complexity and trade-offs involved in deciding when, and how, data should be collected, structured, shared, transferred, used, and deleted. Efforts to resolve or reframe these trade-offs and tensions have also given rise to a range of new agendas around data sharing (Micheli *et al.*, 2020) and new models of data ownership and stewardship (Susha, Janssen and Verhulst, 2017; Delacroix and Lawrence, 2019; Lehtiniemi and Haapoja, 2020), which fall within the expanding data governance field. In the gray literature on data governance, a strong normative element is increasingly evident: with the term linked to wider agendas of good governance and global development. As Pisa *et al.* (2020) put it, the ideal of data governance incorporates “*rules about how data is collected, analyzed, used, and shared in a way that protects citizens from abuse while supporting innovation, development, and inclusive growth*”.

A review of eight past peer-reviewed data governance literature reviews, published between 2016 and early 2022, summarized in the table below, shows this shifting emphasis. While earlier work centered on data governance primarily in terms of data and information management

(Alhassan, Sammon and Daly, 2016; Brous, Janssen and Vilminko-Heikkinen, 2016), work has increasingly addressed data governance as a broader public issue, requiring emphasis on inter-organizational data sharing (Benfeldt Nielsen, 2017; Abraham, Schneider and vom Brocke, 2019), and open data (Bozkurt, Rossmann and Pervez, 2022). Yet, McCaig and Rezaia (2021) argues that the literature ultimately remains “*indicative of a sparse theoretical and empirical knowledge base*” on data governance.

Review	Focus and findings
Data governance activities: an analysis of the literature (Alhassan, Sammon and Daly, 2016)	A systematic review of 31 papers addressing ‘activities in data governance’ published in the Information Science domain was used to identify 110 different data governance activities, primarily in enterprise data governance contexts. The review finds a high volume of research on activities associated with ‘defining’ data governance, but less reporting on the ‘implementation’ or ‘monitoring’ of data governance activities.
Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles (Brous, Janssen and Vilminko-Heikkinen, 2016)	A systematic review of literature relating to data governance principles (covering 35 papers) was used to identify a long list of key data governance topics, synthesized under four categories: ‘Organization’, ‘Alignment’, ‘Compliance’ and ‘Common understanding’. These are presented as principles of data governance and are related to concepts including decision rights and stewardship (Organization); Aligning business an IT and Reducing error of use (Alignment); Accountability, Privacy, Security, Openness and Data quality measurement (Compliance); and Use of standards, metadata management and shared data commons (Common understanding).
A Comprehensive Review of Data Governance Literature (Benfeldt Nielsen, 2017)	A cumulative review covering 62 papers and summarizing the disciplines, methods and concerns of academic works on data governance. The paper finds a strong focus on organization level data governance, with few papers (6 of 62) addressing inter-organizational data governance, and even fewer (2) covering cross-sector data governance issues. The paper concluded with recommendations for data governance research in the context of public organizations, calling for case studies, experiments, action research and analysis of data governance discourse.

Data governance activities: a comparison between scientific and practice-oriented literature (Alhassan, Sammon and Daly, 2018)	Building on Alhassan, Sammon and Daly (2016) this paper provides a systematic review covering 61 papers and finding that practice-oriented publications are more likely to address implementation and monitoring aspects of data governance than scientific publications are.
Data governance: A conceptual framework, structured review, and research agenda (Abraham, Schneider and vom Brocke, 2019)	Drawing on a structured literature review of 145 research papers and practitioner publications published between 2001-2019 the paper puts forward six dimensions of data governance (cross-functional; framework; data as a strategic enterprise asset; decision rights and accountabilities for an organization's decision-making about its data; data policies, standards, and procedures; and compliance monitoring) and outlines 15 future research questions on data governance, including issues of defining data governance responsibilities; facilitating collaboration and retaining control over data in inter-organizational settings; evolution of data governance mechanisms over time; the impact of culture on data governance; and the impacts of data governance on firm performance.
A systematic literature review of data governance and cloud data governance (Al-Ruithe, Benkhelifa and Hameed, 2019a)	A systematic literature review covering studies of data governance in non-cloud and cloud contexts (n=52) identifying six dimensions of 'traditional' data governance (function; structure; organizational; technical; environmental; measuring; and monitoring), and suggesting additional factors that need to be taking into account in a cloud computing context including models of deployment and service delivery, and contractual or other arrangements that set the responsibility of the different actors involved in managing data in the cloud.
A Scoping Review on Data Governance (McGaig and Rezanian, 2021)	Conference paper focusing on papers covering data governance in high rank journals (covering 56 papers), and examining 14 papers in depth to identify themes and methods. This found a lack of unanimity on the term data governance, and a landscape "indicative of a sparse theoretical and empirical knowledge base". Of papers reviewed, 79% focused on data governance concerning medical practice.

A Literature Review of Data Governance and Its Applicability to Smart Cities (Bozkurt, Rossmann and Pervez, 2022)	Conference paper using a text-mining based systematic literature review method (covering 612 papers) to examine the conceptual definition of data governance, structure of current research efforts, and applicability to smart cities. The paper identifies 10 clusters of key terms, covering: big data, data management, data sharing, operation and organization, master data, cloud computing, decision making, corporate governance, health and city and citizen. In the context of urban data, particular emphasis is placed on data quality, data access (incorporating open data), and data management.
---	--

Note: the ‘practice’ literature considered in a number of the papers above was primarily literature from industry associations, intergovernmental organizations, and software vendors or consultants (Abraham, Schneider and vom Brocke, 2019) and includes only limited literature from the civil society, advocacy and think-tank organizations identified by the Datasphere Governance Atlas research process.

A broad working definition of data governance foregrounds both benefits and harms

Given the breadth of contexts in which data governance must be applied, it is not reasonable to expect a single unified definition that can tie together a single field of study. However, common aspects of data governance can still be distilled, and for the purpose of this paper, the following working definition is offered:

Data governance concerns the rules, processes and behaviors related to the collection, management, analysis, use, sharing and disposal of data - personal and/or non-personal.

Good data governance should both promote benefits and minimize harms at each stage of relevant data cycles.

At an organizational level, this generally translates into a focus on internal policies and their implementation, on compliance with external regulation, and on the creation of cross-functional frameworks and responsibilities for managing and extracting value from data as a business asset (Abraham, Schneider and vom Brocke, 2019). At the state level - be it national, regional or international, this may translate into a focus on the development and implementation of policies, standards, laws, regulations, agreements and practices that cover the management of data within countries, and the transfer of data across jurisdictional boundaries (Aaronson, 2021). However, as will be seen in the literature that follows, the organizational literature often pays little attention to the state level, and vice-versa. (See section XX)

A number of authors also highlight that governance of data sits amongst a range of wider practical and governance concerns. Christiane Wendehorst (2020), for example, describes data

governance as one of a number of overlapping frameworks of governance concern in relation to artificial intelligence (AI), considering for instance how the same issue might be alternatively explored through the lens of data governance (considering how datasets are created, managed and used), through a lens of AI systems design (using the language of bias or adequacy of methods), or through a focus on wider societal governance (asking questions about the goals and governance of the wider policy areas to which datasets and AI systems relate).

This highlights the importance of resisting the tendency to treat data as entirely in the abstract: meaningful data is always *about* something, and those somethings are also frequently subject to their own governance regimes, with which any practical data governance will intersect. As is evident in the next section's discussion of academic work on data governance, many researchers have arrived at the topic of data governance because of highly grounded challenges around protecting, managing, or sharing data in relation to a particular field of action.

Current data governance research is highly fragmented

When turning to the academic literature to identify promising concepts, ideas, innovations, and frameworks that may be applied to contemporary data governance policy problems, it is important to have an understanding of how far different projects and papers are part of a coherent research agenda, or - by contrast - how far each publication using the language of data governance may have developed in isolation from other data governance-related work.

The concept of data governance not only brings together academics previously working on distinct issues of data protection, management, and access, but it has also been invoked in disparate academic fields, from health research to work on international trade. In these fields, data governance still can appear more-or-less as a niche sub-field, rather than as cross-cutting field of inquiry in its own right.

This section draws upon a range of bibliometric methods to analyze more than 1,300 publications addressing data governance as indexed by the Dimensions literature database, and a collection of gray literature publications from organizations identified in the Datasphere Governance Atlas.

Computer science and health research dominate academic writing on data governance, but it is discussed in a long-tail of disciplines

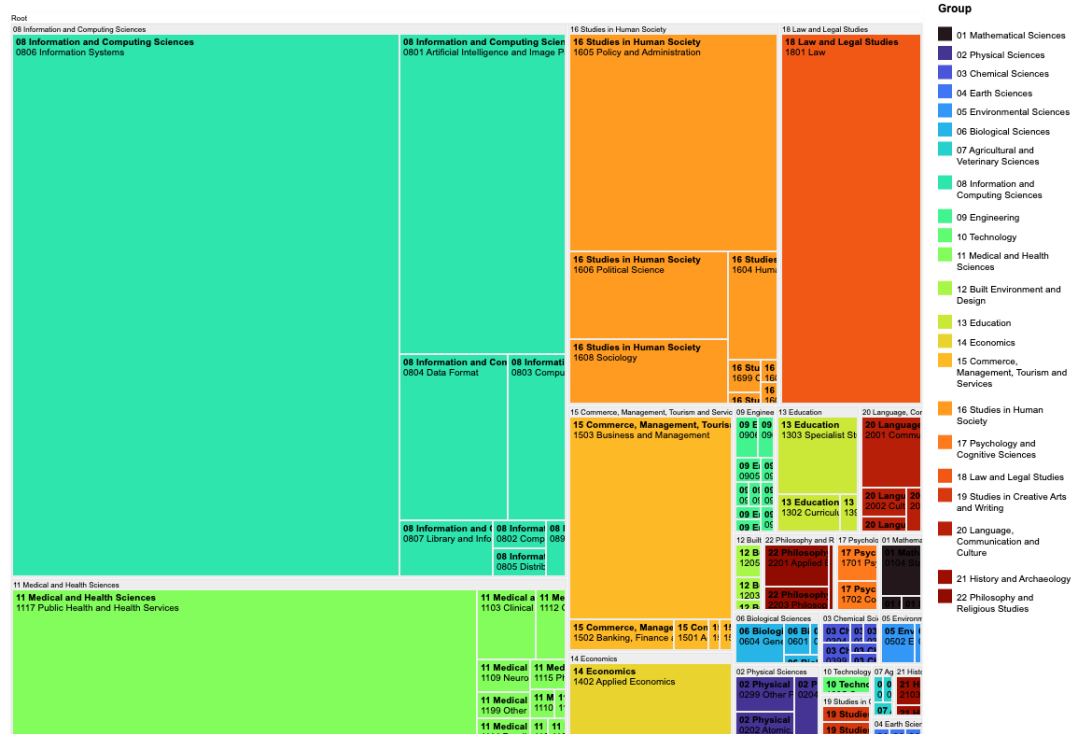


Figure N: Treemap based on Dimensions AI subject classification of papers using “data governance” in title, keywords or abstracts.

The figure above shows a breakdown by field classification of published papers using the term “data governance” in their title, keywords or abstract, as indexed by the Dimensions literature database. The analysis shows that papers on data governance published in Information and Computing Sciences predominate (top-left), followed by Medical and Health Science research (bottom-left), then Studies in Human Society (top-middle), then Law and Legal Studies (top-right), and finally Commerce, Management, Tourism and Services (middle), before a long-tail of other subjects including Economics, Engineering, Education, Built Environment and Philosophy.

Although interest in data governance across all these fields appears to be growing, interest has existed for longer in the Information and Computing Sciences, and Medical and Health Sciences category, with writing on data governance in the Studies in Human Society and Law fields only really starting to gain pace since 2017.

There are distinct clusters of writing on data governance, even within each discipline

Co-citation analysis reveals where authors are drawing upon each other’s work and can be used to indicate particular clusters of interest and focus. Figure N below shows a co-citation analysis

produced using the VOSViewer algorithm (van Eck and Waltman, 2010) which groups together authors with those they most cite and identifies these clusters by color. The resulting diagram reveals both influential researchers, and the particular sub-themes within data governance they are focusing on, as well as showing the interrelations between themes.

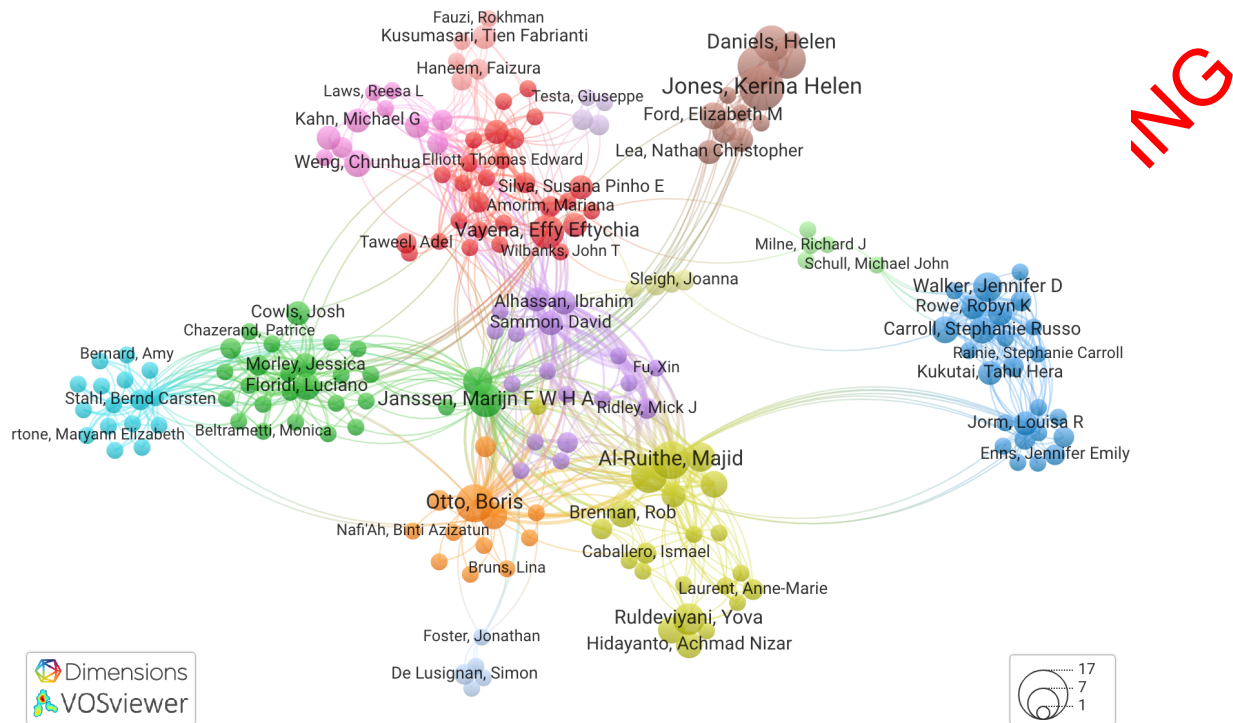
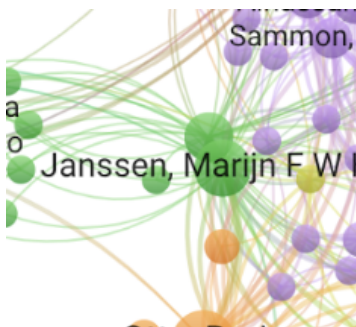


Figure N: Largest connected group for a co-citation analysis of articles with “data governance” in title, keyword or abstracts recorded by Dimensions 2007 - 2021. (Maximum authors: 500)



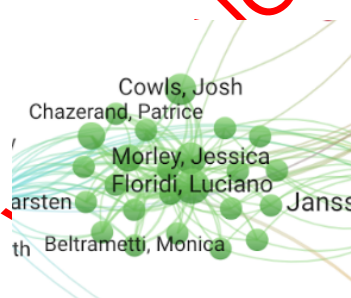
Towards the center of the map, we see a number of relatively highly cited authors. For instance, **Boris Otto** (orange cluster, bottom left) has written widely on data quality management and enterprise use of data in supply chains and, more recently, on ‘data spaces’ as multi-sided platforms for secure data exchange among businesses in a given sector (Weber, Otto and Österle, 2009; Otto, 2011; Hüner, Otto and Österle, 2011; Otto and Jarke, 2019). Central to the work of others in this cluster has been exploring the role of ICT architectures in supporting data governance practice. In this sense, Majid Al-Ruithe (yellow cluster, bottom right) has written on ‘Cloud data governance’, using examples from Saudi Arabia in particular, and producing a literature review in 2018 that identified some of the specific concerns resulting from a move to data storage in the cloud, including that “*Middle Eastern countries and Africa lack compulsory regulatory support for data protection, governance, and privacy*”, as well as technical issues of data migration and security management (Al-Ruithe,

Benkhelifa and Hameed, 2016, 2019b; Al-Ruithe and Benkhelifa, 2017, 2020). The cluster around Al-Ruithe includes several authors focused on using cloud computing, blockchain or big data analytics in sensitive domains, including health and social media data mining. Notably, work from Attard and Brennan (2018) discusses how to focus data management efforts based on an assessment of the exploitable value of data, calling for work on “value-driven data governance”, which should be clearly distinguished from values-driven governance. Others in this cluster have looked at applied cases of data governance implementation, generally within single private sector or governmental institutions (e.g. Aisyah and Ruldeviyani, 2018; Maulina and Ruldeviyani, 2019; Basukie, Wang and Li, 2020).



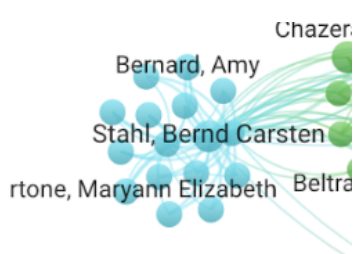
Marijn Janssen and Paul Brous at the Delft University of Technology, feature as connectors, displayed here at the edge of a policy-oriented cluster of authors (green, middle left), acting as a bridge between the enterprise data governance literature, literatures on governance of data in applied contexts such as healthcare or food supply chains, and researchers interested in the social impacts and implications of data and data-driven technologies, including Internet of Things, AI and algorithmic systems (Brous, Herder and Janssen, 2016; Brous, Janssen and Vilminko-Heikkinen, 2016; Brous, Janssen and Herder, 2018;

Janssen *et al.*, 2020; Brous, Janssen and Krans, 2020; Behnke and Janssen, 2020). In ‘Data Governance as Success Factor for Data Science’ Brous, Janssen and Krans (2020) focusses on the importance of data governance addressing data quality, unambiguous data ownership, and legal compliance of ‘data lakes’ to build the trust of decision-makers in the use of data science products built from them. Janssen has also worked extensively with collaborators on e-government, and open data agendas (Janssen, Charalabidis and Zuiderwijk, 2012; Zuiderwijk and Janssen, 2014), although rarely framing this work in terms of data governance. One paper, ‘A conceptual model of decision-making support for opening data’ (Luthfi and Janssen, 2017), explores approaches to balance benefits and risks of opening data using Bayesian belief networks (chosen in part because of their ability to deal with uncertainty) to weigh which fields of a sensitive dataset can, and cannot, be released for public access.



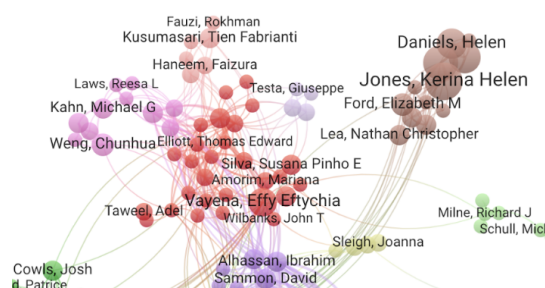
The wider policy cluster (green, middle left) centers on authors associated with the Internet Institute at the University of Oxford and the Alan Turing Institute, including the philosopher **Luciano Floridi**, who has written on data protection (Floridi, 2018) and the ethics of AI (Floridi *et al.*, 2018), and with **Jessica Morley** and others on governance of data and AI in healthcare (Morley *et al.*, 2020). A number of the highly cited studies from authors in this cluster are specifically focused on making policy recommendations that combine both technical and regulatory

interventions.



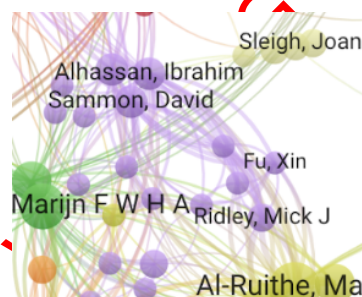
To the left of this cluster on the map (light blue, far left) we see a collection of authors, including **Amy Bernard** and **Bernd Carsten**,

focused on neuroscience data, and in particular on issues of ethics, responsible use of AI, and international data sharing mechanisms to support the creation of global brain data ecosystems (Stahl and Wright, 2018; Fothergill *et al.*, 2019; Eke *et al.*, 2021). The recent literature responds to concerns that a lack of “*clarity surrounding the EU’s GDPR requirements and their varied interpretations have disrupted international data sharing collaborations*” resulting in barriers to potentially life-saving research (Eke *et al.*, 2021). One agenda-setting paper identifies challenges related to ethics, regulation and policies, different definitions of core concepts, language barriers and cultural diversity, all drive the call for a sector-specific International Data Governance (IDG) framework for neuroscience to be developed between researchers, organizations like OECD and UNESCO, and technical societies (Eke *et al.*, 2021).

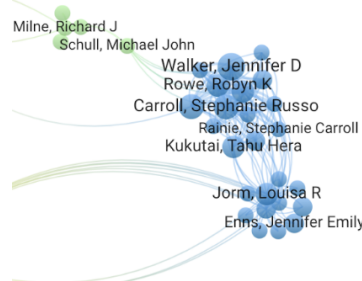


Several further clusters (pink, red and brown, top spread) predominantly look at health data governance, albeit with little cross-citation between authors in each cluster. Data governance is variously the focus of papers, as in work by **Kerina Helen Jones Jones *et al.* (2020)** (brown, top-right) that explores data governance standards for the use of clinical free-text data in research, or seen as an external factor affecting the use of data for some application, as in Cho *et al.* (2021) (pink, top left) who identify ‘data governance–related factors’ as amongst the issues impacting on the quality of ‘person generated wearable device data’ for research. The network around **Effy Eftychia Vayena** of the University of Zurich (red, top center) contains authors looking at both technical mechanisms for cross-site medical data governance (Scheibner *et al.*, 2021), and organizational and policy interventions including data portability, improved mechanisms for informed consent, and participatory governance schemes that involve individuals more directly in data governance (Vayena and Blasimme, 2017). The theme of participation and voice in governance is also picked up in a small cluster around Shah *et al.* (2019) which reports on a cross-country survey of views on post-project sharing of research data.

Turning back to the center and right of the map, we find a small cluster of management and information science researchers (purple, middle), with **Alhassan, Sammon and Daly (2018)** comparing the scientific and practice-oriented literature on data governance, highlighting that the practice literature tends towards a focus on implementation and monitoring, whereas scientific literatures look more at defining data governance activities, producing the kinds of models found in Alhassan, Sammon and Daly (2019).



The dense cluster of authors (blue, mid-far right), with relatively limited citation connections to the wider map, are focused on the concept of **Indigenous Data Sovereignty** and the governance of data for indigenous communities (Taylor



Work from the United States has outsize influence in the academic literature

The biblioshiny package (Aria and Cuccurullo, 2017b) can estimate the country of a paper's author based on stated affiliations. Applying this against the data governance publications extracted from the Dimensions dataset, reveals that the United States and the United Kingdom contribute substantially to work published in English on data governance, followed by China, Germany and Canada. China, Germany and Switzerland are amongst the countries much more likely to be producing single country publications (SCP), rather than publications with authors from multiple countries (MCP).

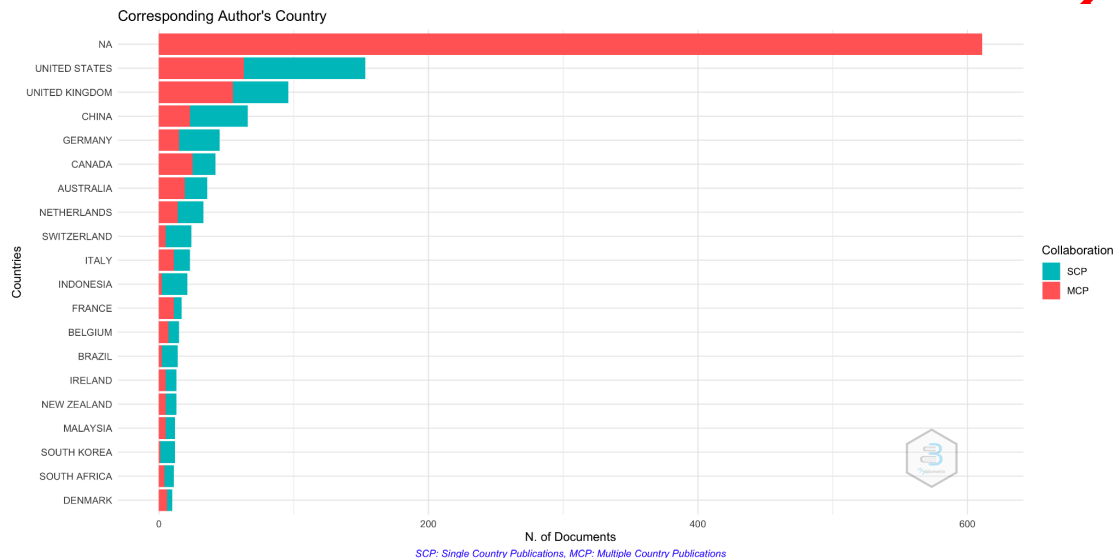


Figure N: Country of corresponding author for papers with “data governance” in titles, keywords and abstracts (extracted from dimensions AI). Note, for almost 50% of papers a country could not be identified from the available data. Source: Dimensions; Analysis: Biblioshiny

It's also possible to map out which countries publish the highest cited papers, providing insights into the countries that may be having the greatest influence on the evolution of the data governance field. *Geographic citation analysis of the Dimensions dataset suggests the highest cited papers were written by authors from the United States, followed by the United Kingdom, Netherlands, Germany, Switzerland and Australia, with comparatively fewer citations for China.* This initially suggests a strong influence of North American and European writing on current data governance debates and may point to gaps in the circulation of academic knowledge and perspectives between regions.

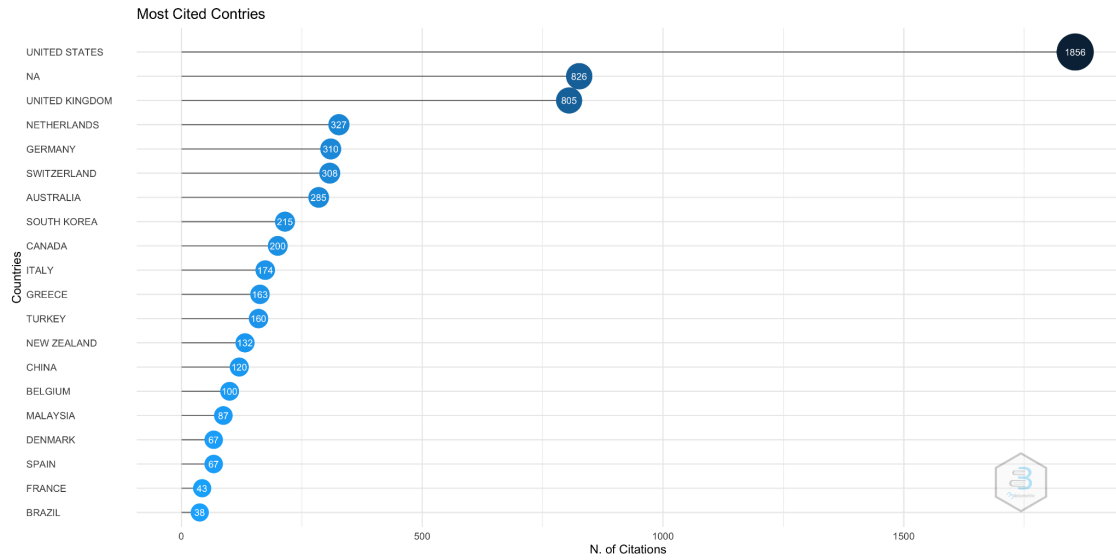


Figure N: Most cited country of publication for papers that include the term “data governance” in their title, keywords or abstract (extracted from dimensions AI). Note, for over 800 papers a country of publication could not be identified from the available data. Source: Dimensions; Analysis: Biblioshiny

Data governance covers a wide range of issues, addressed through different silos

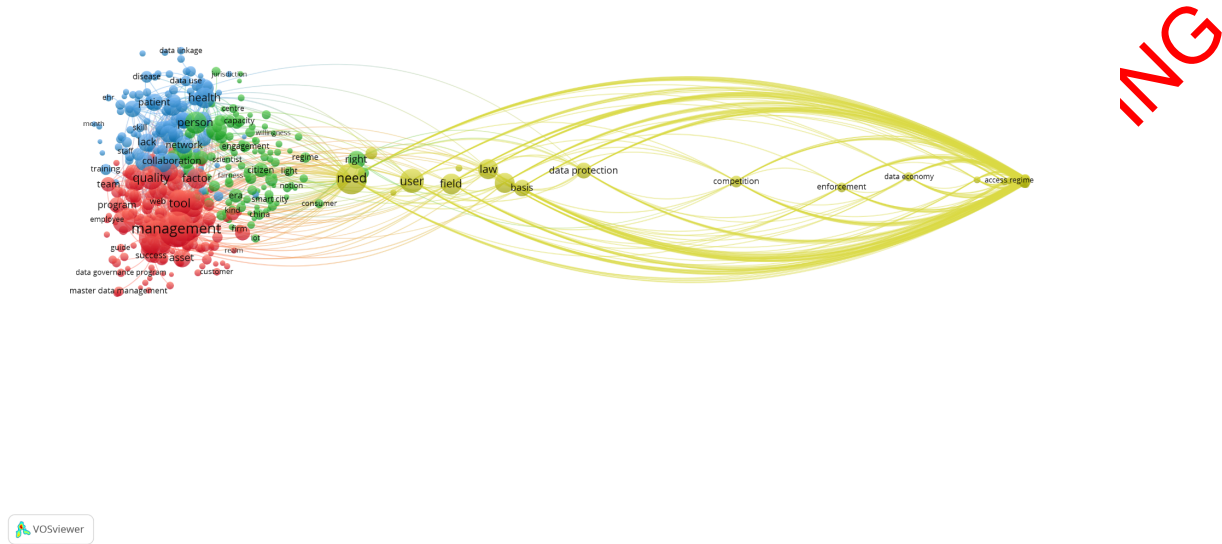


Figure N: Term relationships for title and abstract of papers using the keyword “data governance” plotted with VOSViewer

Figure N shows the co-occurrence of individual terms in titles and abstracts of data governance papers, extracted using the VOSviewer (van Eck and Waltman, 2010) algorithm and visualized using a force-directed network in which terms appear larger the more commonly they are present, and closer together the more commonly they appear together.

The striking ‘fish-like’ diagram that results contains four clusters:

1. the outlier yellow cluster made up of terms related to legal and international policy agendas (law, data protection, competition, access regime, data economy, enforcement etc.);
2. the red cluster centered on data management and quality;
3. the blue cluster covering terms like health, patient and data linkage;
4. and a green cluster containing terms such as citizen, engagement, consumer, smart-city and rights.

A number of preliminary conclusions can be drawn from this arrangement. Firstly, that the legal and regulatory literature more-or-less ‘sits outside’ both information systems literatures, and

applied data governance literatures in fields such as health care. Secondly, issues of data management, the rights of data subjects, and the application of data are tightly linked. Thirdly, zooming into the map and looking at the presence or absence of particular terms, reveals that issues around jurisdiction receive relatively little attention, and novel methods of data governance such as trusts, data stewardship and open data do not appear to currently play an influential role in the current academic literature.

A co-occurrence analysis also can be conducted to identify relations between topics. Where the term map above based on the most commonly occurring terms within data governance paper abstracts reveals the contexts in which data governance is being discussed (e.g. in relation to patient care, or firm management), by looking for the presence of a predefined list of data governance related terms it becomes possible to also see how different data governance topics are connected in the literature. For example, how often are papers discussing issues of open data and data stewardship together? Or how often is data ownership related in the literature to issues of data sovereignty?

Figure N below shows a *term network* where each term is connected to other terms it occurs with, *in papers'* titles and abstracts. Terms cluster closer to terms they occur more commonly with.

WORKING DRAFT PAPER NOT FOR POSTING

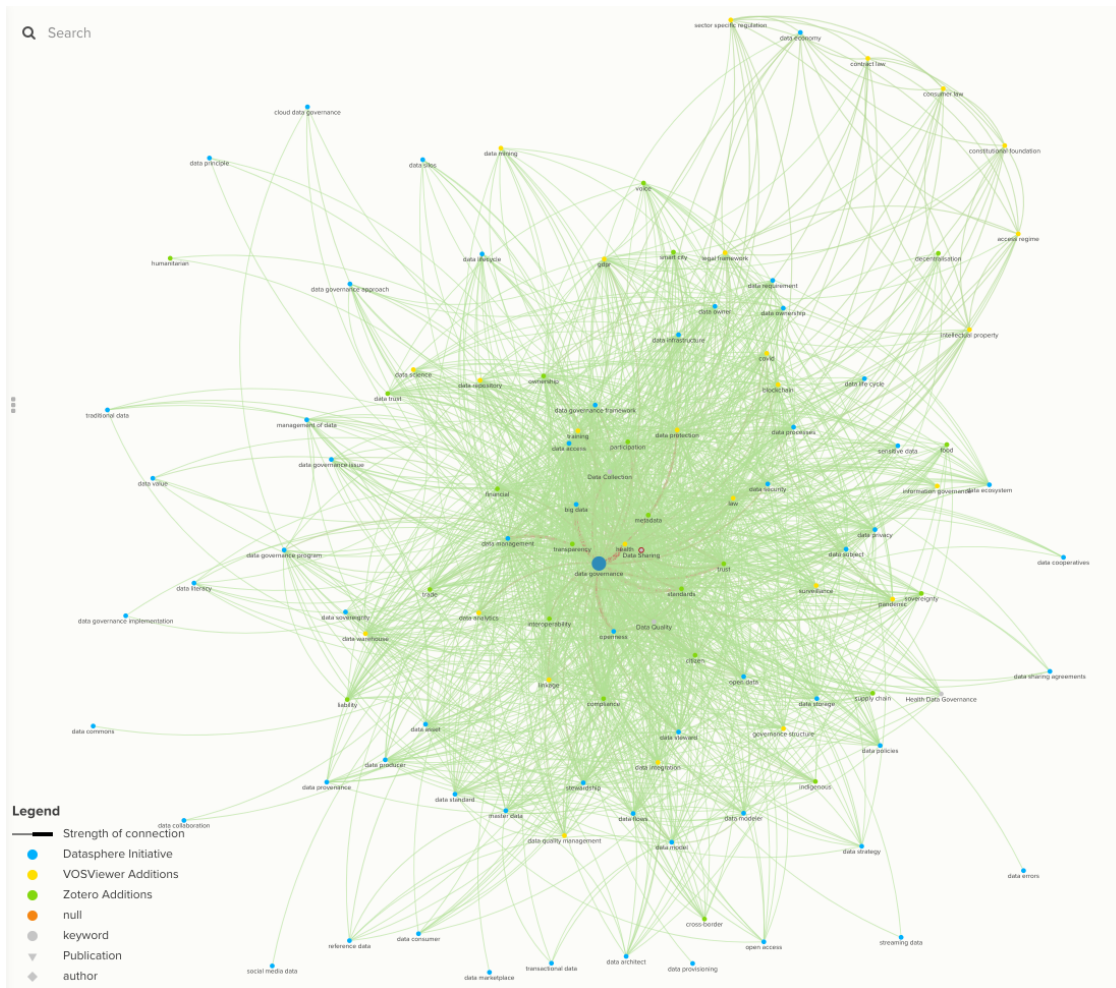


Figure N: Map of selected term co-occurrence in Dimensions titles and abstracts containing the term “Data governance”.

This map reinforces a number of findings from the analysis above: namely that law-related terms cluster together at the edge of the map, but appear peripheral to it, and that there are distinct areas of the literature centered on data governance primarily as a counterpart to data management, as opposed to data governance as a social policy issue.

It also brings into greater focus a number of cross-cutting concepts that play a relatively central role in discussions of data governance across a range of research silos. For instance, the map surfaces transparency, openness, participation, standards and trust as relatively important concepts in a range of contexts.

The density of the map above can make pulling out a particular term relationships difficult. Fortunately, the platform it is generated with allows interactive exploration of each term. For example, the map below (Figure N) looks at terms connected with the topic ‘data sovereignty’, highlighting that papers’ abstracts also address this topic in terms of stewardship and ownership, as well as issues of indigenous data sovereignty, and surveillance. COVID-19, and its contact tracing apps, or disease tracking datasets, appear to have provided a particular framing

ved, but with quite
s of data, the
ce.

The map of terms around open data (sometimes regarded as an approach to data governance, at least in-so-far as open data models take a strong position with respect to the accessibility and reusability of data) takes in a range of themes. These include policies, ecosystems, repositories and infrastructures, as well as linking to themes of standards, trust, metadata, participation, and liability, and showing connections to thematic areas of finance and trade. As in other cases, at the individual article level, the connections drawn between data governance and open data are diverse. Reis, Viterbo and Bernardini (2018) argues that open data portals need to apply stronger data governance frameworks to manage the limited standardization and low quality of data on data portals; whereas D'Agostino *et al.* (2018) explores open data as a public health

resource and calls for the creation of broader governance and policy frameworks that can support responsible data publication.

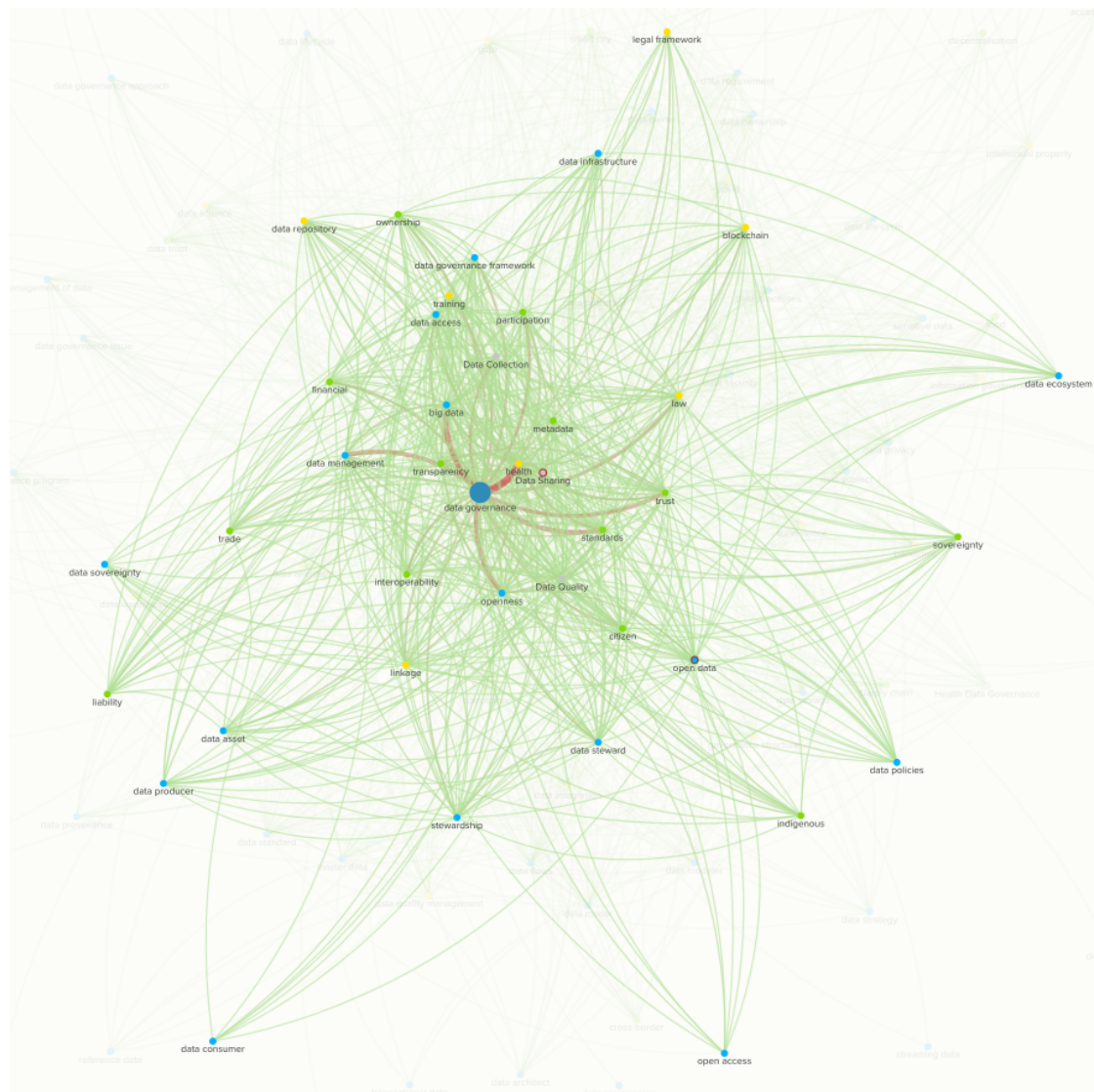


Figure N: Sub-region of term map for 'Open Data'

Analyzing the relationships between all the terms in the maps above is beyond the scope of this paper. However, the full network can be found online at [LINK LINK LINK](#), supporting interactive exploration of key terms.

Methodological note [Box out]

We constructed term co-occurrence networks using three pre-defined term lists (Annex 1). The first, the Datasphere Lexicon (Rossini and Lach, 2022), was developed by the Datasphere Initiative team during the creation of the Datasphere Governance Atlas, as a theoretically informed list of relevant terms. The second list, VOSViewer additions, was created by reviewing the terms extracted from academic papers by the VOSViewer software, and selecting any terms

specific to the topic of data governance (e.g. data science, and access regime) that were not present in the Datasphere Lexicon list, and selecting a number of highly occurring terms that, whilst not exclusively about data governance, have appeared widely as a focus of the recent literature, such as COVID, GDPR and blockchain. The third list was developed through the manual tagging of approximately 100 selected data governance papers using the Zotero reference management software. This introduced a range of terms highlighting particular themes addressed in the literature, including compliance, trust, trade, participation, interoperability and ownership.

A custom script was then used which applied the porter stemmer algorithm (Porter, 1980) to 'fuzzy match' terms (e.g. data market and data marketplace should both be counted as a match) and build a term matrix. This was then visualized using Kumu.io, including edges for any two terms found together in more than one article (i.e. a single co-occurrence of terms is not mapped).

Policy literatures are bringing new terms and concepts into data governance debates

One notable feature in the term map of the academic literature is that emerging forms of data governance, such as data cooperatives, are only weakly connected, with links, for example, to citizens, trust, and data sharing, but to few other widely employed terms. Other emerging governance models, such as data commons, data sharing agreements, and data cooperatives are similarly 'out on a limb' at present, receiving relatively little treatment in the academic data governance literature - albeit each having nascent literatures of their own that are not always framed in terms of data governance.

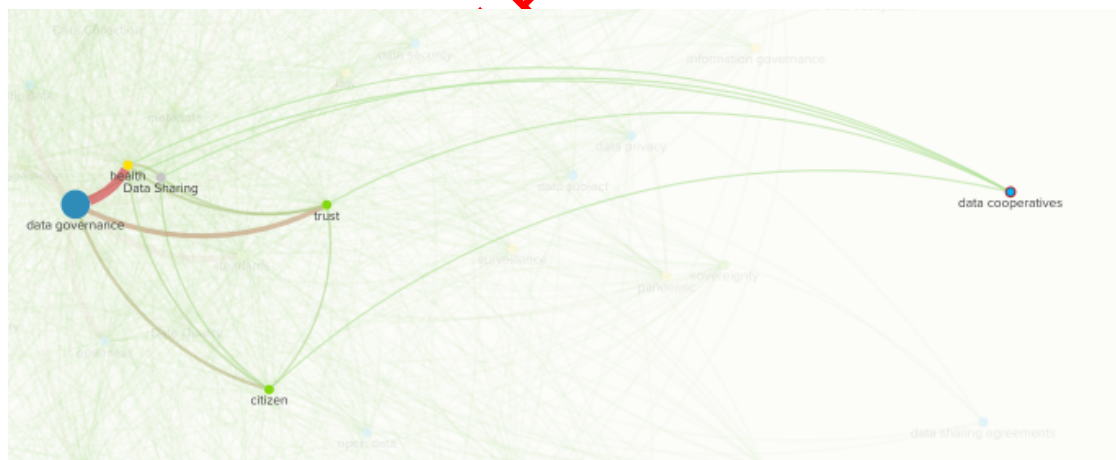


Figure N: Sub-region of term map for 'Data co-operative'

By contrast, when term co-occurrences from the gray literature captured through the Datasphere Governance Atlas are included in this analysis in addition to the academic literature, a much denser network map is generated, where terms such as 'data commons' and 'data cooperatives' are pulled closer in towards the network, although they are still far from central topics of discussion. This indicates that much of the attention on these topics as tools of data

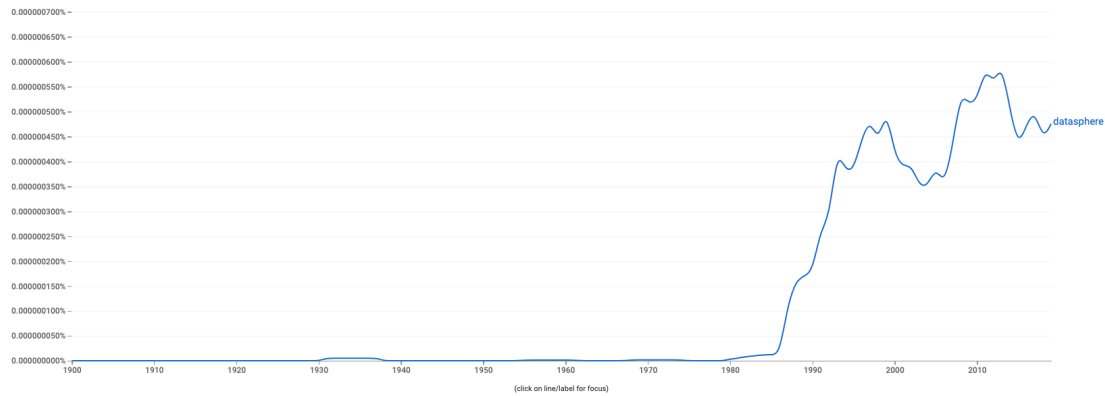


Figure N: Mentions of the term ‘Datasphere’ in Google Ngrams viewer corpus from 1900 to 2019

The evolution of the concept

Google’s NGrams viewer shows appearances of the term “Datasphere” in a corpus of published English books from the last 100 or so years. This reveals that the term was almost unused until the 1980s. Since then, it has been deployed in a variety of contexts: as a product name, conference title, a technique for data mining, a term of science fiction art, or a loosely defined shorthand for a digitizing world. It is in this latter sense that the term was popularized by Douglass Rushkoff, approaching the Datasphere as the “circulatory system for today’s information, ideas and images”, understood as “our new natural environment” (Rushkoff, 1994; Cobb, 2008, p. 39). Rushkoff’s conceptualization, centered on media theory, was deployed to explain how ‘media viruses’ - ideas that capture public attention - rapidly spread. As such, Rushkoff’s Datasphere invokes ideas of information flow, rather than being focused on structured data and its analysis.

Although not using a language of “Datasphere”, Manuel Castells’s work on digital technology as creating a new ‘space of flows’ that transform social relationships to place, can be used to dig deeper into this idea of the Datasphere as a new environment we inhabit. In particular, Castells has explored how the interaction of physical infrastructures, logical network architectures, and exercise of power over information circulation, create patterns of inclusion and exclusion in ‘the network society’ (Castells, 2010; Castells, 2013). Although it can be rhetorically convenient to talk of a universal Datasphere, in practice, localities and lives across the world are differentially digitized, whether as a result of connectivity (nearly 3bn people still lack internet connectivity), or a result of ontology (many people’s lived realities do not fit within the categorizations prescribed in structured data systems, and/or their concerns are not reflected in corporate choices over the data that is worth collecting, transmitting, or storing).

The conceptualization of the Datasphere Initiative

The report “We Need To Talk About Data” (De La Chapelle and Porciuncula, 2021) draws upon a law paper by Bergé, Grumbach and Zeno-Zencovich (2018a) that offered a conceptually

expansive, but digitally focused, description of the Datasphere. Explicitly invoking an analogy with the atmosphere, the lithosphere and the hydrosphere, that paper described how:

“The notion of ‘Datasphere’ proposes a holistic comprehension of all the ‘information’ existing on earth, originating both in natural and socio-economic systems, which can be captured in digital form, flows through networks, and is stored, processed and transformed by machines.”

In this definition, there are parallels with the idea of the Infosphere introduced by Floridi (2007) as “the whole informational environment constituted by all informational entities (thus including informational agents as well), their properties, interactions, processes, and mutual relations”, although where Floridi’s concept includes both digital and “offline and analog spaces of information” alongside digital data, the scope of the Datasphere is more tightly defined, concerned primarily with digital representations of the world that have been “found, collected and organized” (Bergé, Grumbach and Zeno-Zencovich, 2018a). Both Floridi and Bergé et. al. however see these new spheres as ‘spaces we inhabit’: architectures and ecosystems affecting the way daily life is lived. For Bergé in particular, the spatial metaphor of the Datasphere highlights the way in which datafication reconfigures relations between “conventional institutional territories (.e.g. States, towns, international and regional organizations)”, and “gives rise to new territories”.

As legal scholars, Bergé, Grumbach and Zeno-Zencovich (2018b) have placed particular emphasis on the contrast between European and US data governance regimes. Burk (2005) followed a similar path in a 2005 exploration of ‘Privacy and Property in the Global Datasphere’, where they describe a conflict between a European ‘deontological’ (or rights based) approach to intellectual property and informational privacy, and a US utilitarian framework, relying on market mechanisms to provide privacy protection, and treating intellectual property rights as tools to incentivize production of ‘public good’ knowledge or creative works. Burk (2005) went on to argue that these two dominant models were displacing or overwhelming the development of local, ‘indigenous’ or ground up conceptualizations of both intellectual property and privacy, lamenting the loss of a potential diversity of approaches.

A desire to move outside of a narrow menu of policy options in part motivated the adoption of a refined datasphere terminology in De La Chapelle and Porciuncula (2021), and further developed since then in work by the Datasphere Initiative (DI) which describes the Datasphere as:

“the complex system encompassing all types of data and their dynamic interactions with human groups and norms” (Porciuncula and Chapelle, 2022).

This formula essentially draws attention to the mutual interactions between digital artifacts (datasets), constituencies and social relationships (human groups) and rules and social expectations (norms) - and to the multiplicity of each. It goes further to define each of those elements in the following terms:

- **Digital data**, personal and non-personal, private and public, is organized in datasets of diverse sizes and types, although such classifications have blurred, overlapping and

moving boundaries. Importantly, the same data can be part of multiple datasets or used in different sectors and the infinite potential for recombination and analysis constantly creates new data or metadata.

- **Individuals and human groups** of all sorts generate, collect, store, process, exchange, make accessible or access, analyze, and use data for various purposes. They also include the actors setting governance and management norms of the Datasphere. Distributed across the world, all these actors are interlinked in complex value chains, often with asymmetric power relations.
- **A great variety of norms**, including cultural, legal, and technical ones, set parameters regarding relationships between humans and data, including: high-level principles, international agreements, laws and regulatory frameworks, but also contracts, licenses or terms of service, and even code, standards, and software underpinning technical systems (including that of supporting infrastructures).

At the same time, while noticing that there are asymmetric power relations, it stops short of detailed specification of *how* the interaction of these should be governed.

Notably - the model implies governance of one interconnected Datasphere, not many isolated instances, and does so with the purpose of providing a holistic lens into the evolving complexity of data governance and its impact on the creation of value and well-being for all. That is, the Datasphere is seen as a **single complex system** (Siegenfeld and Bar-Yam, 2020). Or, going further as per (Porciuncula and Chapelle, 2022) the Datasphere is a **complex adaptive system** with emergent dynamics:

“On an ongoing basis and a global scale, the Datasphere engages billions of actors, whose actions are determined by the norms applicable to them, but also by their personal choices, preferences and interests, as well as the information available to them. Such a very large number of interconnected agents with the capacity to individually modify their behavior in relation to the environment and the actions of others constitute what the scientific community labels a “complex adaptive system”. Widely known examples of these types of systems are flocks of birds or schools of fish. This relatively recent yet powerful field of study now finds applications in an extreme variety of domains, including the environment, social dynamics, evolution, brain activity, or markets, to name only a few. We postulate here that the Datasphere, as defined above, is a complex adaptive system, exhibiting the well-documented characteristics⁸ of such systems, including: a large number of interconnected agents, non-linear impacts of their actions, positive and negative feedback loops, unintended consequences, structural unpredictability, emergence and path dependencies.”

Yet, just as it is meaningful to talk of both *the* atmosphere, and also of some local atmospheric conditions, it should be meaningful to talk both of *the* Datasphere, and of how the Datasphere is experienced in relation to some specific places, actors, topics or sectors.

The broad tripartite datasphere definition of the Datasphere Initiative can be usefully compared and contrasted with Kitchin's notion of a 'data assemblage'. Assemblages are the "*contingent, relational and contextual discursive and material practices and relations*" (Kitchin, 2014, p. 25) surrounding some dataset or data infrastructure. Kitchin documents a full range of apparatus that make up these assemblages, including: systems of thought; forms of knowledge; finance; political economy; governalities and legalities; materialities and infrastructures; practices; organizations and institutions; subjectivities and communities; and, finally, places and marketplaces. Across the critical data studies literature, authors highlight that datasets cannot be taken as given but must be seen as enmeshed in particular contexts (Gitelman, 2013), although data is generally still taken as the starting point, or focusing lens, for inquiry.

The Datasphere Initiative's perspective recognizes this concern, but consciously seeks to find connections between contexts, and look for policy and governance strategies that may not necessarily arise from a focus on single embedded local contexts or datasets. The concept of human groups, for example, implicitly points to groups that are potentially co-constituted by, and co-constitutive of, data ecosystems, and that may exist across conventional boundaries of geography and polity. A recognition of both global norms, and a global plurality of norms, points towards governance approaches that have appropriate levels of flexibility and adaptability (or perhaps polycentricity, to use a term from Ostrom (Ostrom, 2010; Benfeldt, 2020)). Similarly, while understanding datasets as part of a wider assemblage can highlight a range of potential issues for, and points of, governance intervention, an approach based solely on descriptive accounts of current political, economic or governance arrangements surrounding certain data artifacts may foreclose attempts to imagine new data governance arrangements, particularly those that become possible when current legacy frameworks of place and territoriality are deprioritized.

Additional perspectives on the Datasphere

There have been a few other uses of datasphere terminology in recent decades that it is useful for us to be aware of. Although these do not entirely parallel the way the Datasphere Initiative uses the term, they were taken into consideration in the development of the Datasphere Initiative's Datasphere concept, and each one offers further perspectives that are useful to keep in mind as we explore data governance in the context of the Datasphere.

Firstly, for Humphreys (2015), the concept of a Datasphere is read in the context of the 'public sphere': driving him to discuss a concern that 'Hobbes' bargain' (Tuck, 2002) that "*we leave public conscience to the state and the state leaves our private conscience alone*" is breaking down. In this reading, the digitisation of daily life leads states, corporations and other institutions to increasingly 'plunder' individuals' data as the boundaries between private and public life blur. At the same time, non-state actors, including individuals, gain new avenues to exercise power, bypassing legitimation through dialogue within the public sphere. The analogy of Datasphere to the public sphere is particularly productive when set against alternative analogies of the Datasphere with the atmosphere or lithosphere. Whilst these 'natural spheres' are inescapable and prior to human activity, albeit greatly affected by it, the public sphere is

only created through people coming together. An individual may also belong to multiple publics, whereas we live within one shared atmosphere.

The question of whether it is useful to discuss one “datasphere”, or many, is prompted by writing from Updegrove (2004) that envisions a world of 6 billion personal dataspheres - containing the birth to death data traces of each individual: interoperable through standards, but ideally within the control of the data subjects they relate to. Ultimately, the ideas Updegrove outlines have been picked up more recently through concepts such as MiData, Vendor Relationship Management, and Personal Data Stores, and there appears to be very little contemporary use of an individual-level conceptualization of the Datasphere. However, use by Béranger (2016) of the concept of a ‘medical datasphere’ highlights that the term may at times be deployed to refer to ‘sub-regions’ of a more universal Datasphere.

Finally, Lucie (2021) offers an arts-led perspective on the role of the human in the Datasphere, drawing attention to the agency of digital representations of the human, and their impact not only on human groups, but also on individual human bodies and experiences. Lucie highlights strategies of “creating messy and disjunctive data” as an antidote to the overwhelming nature of an individual’s digital trace, and proposes “resistance to the all-encompassing and accelerating nature” of datafication. Though Lucie does not directly reference them, her essay draws to mind perspectives from Science and Technology Studies (STS) that foreground the agency of technical artifacts (Felt *et al.*, 2016), as well as work inspired by James C. Scott and others (Scott, 1990, 1998) that explores resistance to state efforts to render populations legible through record keeping and data.

Applying a Datasphere framework to the literature

As the conceptual toolkit of the datasphere is further sharpened, the concept of the Datasphere can support work to both find the connections between, and draw clear distinctions within, currently disparate academic and policy writing on data governance.

Shifting from a discussion of “governing data”, to “governing the Datasphere” involves identifying the particular regions of the Datasphere in focus and acknowledging the relationships between data governance in one region (for example, in relation to the individual or the firm), and data governance in other regions and at other levels (for example, organizational, industry, societal, national or global). By offering the typology of datasets, human groups and norms, the Datasphere framework then invites a clearer specification of the specific focus of any governance research and the factors being taken into account in proposing or evaluating particular governance regimes.

This typology is particularly important when considering the transferability of data governance innovations between settings. For example, how far can the governance approaches developed in the context of healthcare be used to inform data governance arrangements for other settings? In healthcare, datasets frequently contain highly sensitive personal information; the human groups involved may encompass patients, clinicians, researchers, and particular communities disproportionately affected by specific medical conditions; and norms include strong professional ethical frameworks, public attitudes around

health, and cultural attitudes towards the acceptability of health data being used for profit (Skovgaard, Wadmann and Hoeyer, 2019; Shah *et al.*, 2021). Other settings may have data types, norms and stakeholder configurations in common, but may also have distinct arrangements that need to be taken into account. Paying attention to these similarities and differences is important both to explore longer-established work on data protection, data management and open data, and to analyze novel data governance proposals.

Just as conceptualization of the Datasphere as a complex adaptive system involves recognizing the presence of a large number of interconnected agents, whose actions involve a degree of “structural unpredictability, emergence and path dependencies” (Porciuncula and Chapelle, 2022), the broad literature on data governance exhibits some of these properties.

However, as this paper has shown, bibliometric methods offer a starting point to understand some of the emerging structures of the space and may provide an initial map for plotting interventions. By bringing together researchers, practitioners and policy makers from different disciplinary and sub-thematic clusters of work, with a full awareness of where they are addressing common, and context-specific, data governance challenges, there are significant opportunities to both generate and test innovative data governance ideas, and to build a more coherent picture of how the datasphere can be governed.

Regular bibliographic mapping has the potential to become a feedback tool, to identify where stronger networks between researchers and fields are being forged. Further refinement of the methods used in this paper, and the addition of a more explicit datasphere frame to future coding and analysis of the literature can support a more effective interface between work from different fields, regions and areas of study, and may support efforts to bring scholars together around more clearly defined shared problems.

Conclusion

This paper offers a starting point for thinking about academic (and some gray literature) writings on data governance. It provides a high-level overview of research clusters and themes addressed in the literature, and highlights that there is, ultimately, no single data governance field to speak of, but rather a range of distinct fields of work, each responding to thematic or sectoral challenges. While firm, and society-level governance of data are broadly two sides of the same coin, relatively little work has explored issues of cross-boundary data governance, leaving a significant gap to be filled.

A number of limitations of the current work should be noted. The bibliographic data for this paper was gathered at the end of 2021. Since that point, the publication of draft legislation in Europe using the language of data governance (the Data Governance Act) has driven a further growth in published literature using this terminology. These papers, and other publications only indexed in 2022, are not given a full treatment in the analysis above: a limitation that should be kept in mind. Indeed, given the ongoing growth, and dynamic nature of work on data governance, the companion web pages for this report, containing regularly updated analysis, should be consulted to locate the most recent map of the literature.

Furthermore, as noted in the methodology, this paper has also only considered English language literature, and there is a need for future work to consider work on data governance across languages, particularly paying attention to work from China and Latin America. Limitations of time also mean the authors were not able to complete a full manual coding of all data governance papers captured in the research Zotero database, leaving a need for future work that can complete a more comprehensive in-depth scan of emerging data governance approaches and ideas, identifying promising areas of research that are not surfaced by the quantitative, and network-analysis bibliometric approaches primarily employed above.

Deployed carefully, the conceptual framework of the Datasphere has a significant contribution to make to current data governance research and practice, in particular by bringing forward the notion of “governance of the Datasphere” as a systems approach to data governance. By recognizing that distinct concerns drive data governance approaches in different regions of the Datasphere, but also bringing attention to the interconnectedness of different sites of data governance, a Datasphere perspective can help bridge between research silos. In particular, there is work to be done to better connect work on legal approaches to cross-boundary data governance, with current work on organizational data governance practice. The Datasphere Initiative is well placed to catalyze some of the needed connections, and if it can do so successfully, a map of data governance research clusters could well look quite different in a few years from now.

Appendix 1: Terms

We constructed term co-occurrence networks using three pre-defined term lists (below) to identify data governance.

The first, the Datasphere Lexicon, was developed by the Datasphere Initiative team during the creation of the Datasphere Governance Atlas, as a theoretically informed list of relevant terms. The second list, VOSViewer additions, was created by reviewing the terms extracted from academic papers by the VOSViewer software, and selecting any terms specific to the topic of data governance (e.g. data science, and access regime) that were not present in the Datasphere Lexicon list, and selecting a number of highly occurring terms that, whilst not exclusively about data governance, have appeared widely as a focus of the recent literature. The third list was developed through the manual tagging of approximately 100 selected data governance papers using the Zotero reference management software. This introduced a range of terms highlighting particular themes addressed in the literature.

Terms from the Datasphere Lexicon (Datasphere Initiative)	Additional selected terms from VOSViewer clustering (VOSViewer Additions)
big data, biometric data, cloud data governance , cross-border data flows, data access, data accountabilities, data accountability, data agenda, data architect, data asset, data collaboration, data collection, data commons, data consumer, data cooperatives, data economy, data ecosystem, data errors, data exchange standard, data fiduciary, data flows, data governance approach, data	access regime, blockchain, constitutional foundation, consumer law, contract law, covid, data analytics, data integration, data mining, data protection, data quality management, data repository, data science, data warehouse, gdpr, governance structure,

governance council, data governance effectiveness, data governance framework, data governance implementation, data governance issue, data governance leader, data governance office, data governance program, data governance, data guidelines, data inconsistencies, data infrastructure, data life cycle, data lifecycle, data literacy, data management activities, data management activity, data management initiative, data management, data marketplace, data model, data modeler, data owner, data ownership, data policies, data principle, data privacy, data procedure, data processes, data producer, data professional, data provenance, data provisioning, data quality, data representation, data requirement, data retention, data role and responsibilities, data scope, data security, data sharing agreements, data sharing, data silos, data sovereignty, data standard, data steward, data storage, data strategy, data subject, data understanding, data value, datasphere, enterprise data model, enterprise data modeling, information governance, machine-generated data, management of data, master data, metadata or meta data, open access, open data, openness, reference data, sensitive data, social media data, stewardship, streaming data, traditional data, transactional data	<p>health data governance, health, information governance, intellectual property, law, legal framework, linkage, pandemic, sector specific regulation, smart city, surveillance, training</p> <p>Additional selected terms from qualitative reading (Zotero additions)</p> <p>citizen, compliance, cross-border, data fiduciary, data spaces, data trust, decentralization, financial, food, humanitarian, indigenous, interoperability, liability, metadata, MyData, ownership, participation, participatory data governance, safe harbor, smart city, sovereignty, standards, supply chain, trade, transparency, trust, voice</p>
---	--

References

- Aaronson, S.A. (2021) *Data is disruptive: How data sovereignty is challenging data governance*. Hinrich Foundation, p. 26. Available at: <https://www.wita.org/wp-content/uploads/2021/08/Data-is-disruptive-Hinrich-Foundation-white-paper-Susan-Aaronson-August-2021.pdf>.
- Abraham, R., Schneider, J. and vom Brocke, J. (2019) 'Data governance: A conceptual framework, structured review, and research agenda', *International Journal of Information Management*, 49, pp. 424–438. Available at: <https://doi.org/10.1016/j.ijinfomgt.2019.07.008>.
- Aisyah, M. and Ruldeviyani, Y. (2018) 'Designing Data Governance Structure Based On Data Management Body of Knowledge (DMBOK) Framework: A Case Study on Indonesia Deposit Insurance Corporation (IDIC)', in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 307–312. Available at: <https://doi.org/10.1109/icacsis.2018.8618151>.
- Alhassan, I., Sammon, D. and Daly, M. (2016) 'Data governance activities: An analysis of the literature', *Journal of Decision Systems*, 25, pp. 64–75. Available at: <https://doi.org/10.1080/12460125.2016.1187397>.
- Alhassan, I., Sammon, D. and Daly, M. (2018) 'Data governance activities: A comparison between scientific and practice-oriented literature', *Journal of Enterprise Information Management*, 31(2), pp. 300–316. Available at: <https://doi.org/10.1108/JEIM-01-2017-0007>.

Alhassan, I., Sammon, D. and Daly, M. (2019) 'Critical Success Factors for Data Governance: A Theory Building Approach', *Information Systems Management*, 36(2), pp. 98–110. Available at: <https://doi.org/10.1080/10580530.2019.1589670>.

Al-Ruithe, M. and Benkhelifa, E. (2017) 'Analysis and Classification of Barriers and Critical Success Factors for Implementing a Cloud Data Governance Strategy', *Procedia Computer Science*, 113, pp. 223–232. Available at: <https://doi.org/10.1016/j.procs.2017.08.352>.

Al-Ruithe, M. and Benkhelifa, E. (2020) 'Determining the enabling factors for implementing cloud data governance in the Saudi public sector by structural equation modeling', *Future Generation Computer Systems*, 107, pp. 1061–1076. Available at: <https://doi.org/10.1016/j.future.2017.12.057>.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (2016) 'A Conceptual Framework for Designing Data Governance for Cloud Computing', *Procedia Computer Science*, 94, pp. 160–167. Available at: <https://doi.org/10.1016/j.procs.2016.08.025>.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (2019a) 'A systematic literature review of data governance and cloud data governance', *Personal and Ubiquitous Computing*, 23(5-6), pp. 839–859. Available at: <https://doi.org/10.1007/s00779-017-1104-3>.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (2019b) 'A systematic literature review of data governance and cloud data governance', *Personal and Ubiquitous Computing*, 23(5), pp. 839–859. Available at: <https://doi.org/10.1007/s00779-017-1104-3>.

Aria, M. and Cuccurullo, C. (2017a) 'Bibliometrix: An R-tool for comprehensive science mapping analysis', *Journal of Informetrics*, 11(4), pp. 959–975. Available at: <https://doi.org/10.1016/j.joi.2017.08.007>.

Aria, M. and Cuccurullo, C. (2017b) 'Bibliometrix: An R-tool for comprehensive science mapping analysis', *Journal of Informetrics*, 11(4), pp. 959–975. Available at: <https://doi.org/10.1016/j.joi.2017.08.007>.

Attard, J. and Brennan, R. (2018) 'Challenges in Value-Driven Data Governance', in H. Panetto et al. (eds) *On the Move to Meaningful Internet Systems. OTM 2018 Conferences*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 546–554. Available at: https://doi.org/10.1007/978-3-030-02671-4_33.

Basukie, J., Wang, Y. and Li, S. (2020) 'Big data governance and algorithmic management in sharing economy platforms: A case of ridesharing in emerging markets', *Technological Forecasting and Social Change*, 161, p. 120310. Available at: <https://doi.org/10.1016/j.techfore.2020.120310>.

Behnke, K. and Janssen, M.F.W.H.A. (2020) 'Boundary conditions for traceability in food supply chains using blockchain technology', *International Journal of Information Management*, 52, p. 101969. Available at: <https://doi.org/10.1016/j.ijinfomgt.2019.05.025>.

Benfeldt Nielsen, O. (2017) 'A Comprehensive Review of Data Governance Literature', *Selected Papers of the IRIS, Issue Nr 8 (2017)* [Preprint]. Available at: <https://core.ac.uk/download/pdf/301373908.pdf>.

Benfeldt, O. (2020) *Polycentric governance of organizational data ventures: An organizing logic for data governance in the digital era*. PhD thesis. Aalborg Universitetsforlag. Available at: https://vbn.aau.dk/ws/portalfiles/portal/400433240/PHD_OB_E_pdf.pdf.

Béranger, J. (2016) *Big Data and Ethics: The Medical Datasphere*. Elsevier. Available at: <https://books.google.com?id=RYVOCwAAQBAJ>.

Bergé, J.-S., Grumbach, S. and Zeno-Zencovich, V. (2018b) *The 'Datasphere', Data Flows Beyond Control, and the Challenges for Law and Governance*. SSRN Scholarly Paper ID 3185943. Rochester, NY: Social Science Research Network. Available at: <https://papers.ssrn.com/abstract=3185943> (Accessed: 7 December 2021).

Bergé, J.-S., Grumbach, S. and Zeno-Zencovich, V. (2018a) *The 'Datasphere', Data Flows Beyond Control, and the Challenges for Law and Governance*. SSRN Scholarly Paper ID 3185943. Rochester, NY: Social Science Research Network. Available at: <https://papers.ssrn.com/abstract=3185943> (Accessed: 21 October 2021).

Bozkurt, Y., Rossmann, A. and Pervez, Z. (2022) 'A Literature Review of Data Governance and Its Applicability to Smart Cities', in *Proceedings of the 55th Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences*, Hawaii. Available at: <http://hdl.handle.net/10125/79666> (Accessed: 18 May 2022).

Brous, P., Herder, P. and Janssen, M. (2016) 'Governing Asset Management Data Infrastructures', *Procedia Computer Science*, 95, pp. 303–310. Available at: <https://doi.org/10.1016/j.procs.2016.09.339>.

Brous, P., Janssen, M. and Herder, P. (2018) 'Internet of Things adoption for reconfiguring decision-making processes in asset management', *Business Process Management Journal*, 25(3), pp. 495–511. Available at: <https://doi.org/10.1108/BPMJ-11-2017-0328>.

Brous, P., Janssen, M. and Krans, R. (2020) 'Data Governance as Success Factor for Data Science', in *Responsible Design, Implementation and Use of Information and Communication Technology*, pp. 431–442. Available at: https://doi.org/10.1007/978-3-030-44999-5_36.

Brous, P., Janssen, M. and Vilminko-Heikkinen, R. (2016) 'Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles', in H.J. Scholl et al. (eds) *Electronic Government*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 115–125. Available at: https://doi.org/10.1007/978-3-319-44421-5_9.

Burk, D.L. (2005) *Privacy and Property in the Global Datasphere*. SSRN Scholarly Paper ID 716862. Rochester, NY: Social Science Research Network. Available at: <https://doi.org/10.2139/ssrn.716862>.

Carroll, S.R. *et al.* (2020) 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, 19, p. 43. Available at: <https://doi.org/10.5334/dsj-2020-043>.

Carroll, S.R. *et al.* (2021) 'Indigenous Peoples' Data During COVID-19: From External to Internal', *Frontiers in Sociology*, 6. Available at: <https://www.frontiersin.org/article/10.3389/fsoc.2021.617895> (Accessed: 12 April 2022).

Castells, M. (2013) *Communication Power*. Second Edition. Oxford, New York: Oxford University Press.

Castells, M. (2010) *The rise of the network society*. 2nd ed., with a new pref. Chichester, West Sussex ; Malden, MA: Wiley-Blackwell (The information age : Economy, society, and culture, v. 1).

Cho, S. *et al.* (2021) 'Identifying Data Quality Dimensions for Person-Generated Wearable Device Data: Multi-Method Study', *JMIR mHealth and uHealth*, 9(12), p. e31618. Available at: <https://doi.org/10.2196/31618>.

Christiane Wendehorst (2020) *Data Governance Working Group: A Framework Paper for GPAI's work on Data Governance*. Available at: <https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf> (Accessed: 10 January 2022).

Cobb, K. (2008) *The Blackwell Guide to Theology and Popular Culture*. John Wiley & Sons.

Cormack, D., Reid, P. and Kukutai, T. (2019) 'Indigenous data and health: Critical approaches to "race"/ethnicity and Indigenous data governance', *Public Health*, 172, pp. 116–118. Available at: <https://doi.org/10.1016/j.puhe.2019.03.026>.

D'Agostino, M. *et al.* (2018) 'Open data and public health', *Revista Panamericana de Salud Pública*, 42, p. e66. Available at: <https://doi.org/10.26633/RPSP.2018.66>.

Datasphere Initiative (2022) *Datasphere Governance Atlas: 2022*. Datasphere Initiative. Available at: <https://www.thedatasphere.org/wp-content/uploads/2022/04/Datasphere-Governance-Atlas-2022-Datasphere-Initiative.pdf> (Accessed: 18 May 2022).

Davies, T. *et al.* (eds) (2019) *The State of Open Data: Histories and Horizons*.

De La Chapelle, B. and Porciuncula, L. (2021) *We Need to Talk About Data: Framing the Debate Around the Free Flow of Data and Data Sovereignty*. Internet & Jurisdiction Policy Network.

de La Chapelle, B. and Porciuncula, L. (2022) 'Hello Datasphere - towards a systems approach to data'. Datasphere Initiative. Available at: <https://www.thedatasphere.org/datasphere-publish/hello-datasphere/>.

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), pp. 236–252. Available at: <https://doi.org/10.1093/idpl/ipz014>.

Eke, D.O. *et al.* (2021) 'International data governance for neuroscience', *Neuron*, 0(0). Available at: <https://doi.org/10.1016/j.neuron.2021.11.017>.

Felt, U. *et al.* (eds) (2016) *The Handbook of Science and Technology Studies*. 4th edn. Cambridge, MA, USA: MIT Press.

Floridi, L. (2007) 'A Look into the Future Impact of ICT on Our Lives', *The Information Society*, 23(1), pp. 59–64. Available at: <https://doi.org/10.1080/01972240601059094>.

Floridi, L. *et al.* (2018) 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines*, 28(4), pp. 689–707. Available at: <https://doi.org/10.1007/s11023-018-9482-5>.

Floridi, L. (2018) 'Soft ethics, the governance of the digital and the General Data Protection Regulation', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), p. 20180081. Available at: <https://doi.org/10.1098/rsta.2018.0081>.

Fothergill, B.T. *et al.* (2019) 'Responsible Data Governance of Neuroscience Big Data', *Frontiers in Neuroinformatics*, 13, p. 28. Available at: <https://doi.org/10.3389/fninf.2019.00028>.

Gitelman, L. (ed.) (2013) *"Raw Data" Is an Oxymoron*. Cambridge, MA, USA: MIT Press (Infrastructures).

Greenleaf, G. (2012) *Global Data Privacy Laws: 89 Countries, and Accelerating*. SSRN Scholarly Paper ID 2000034. Rochester, NY: Social Science Research Network. Available at: <https://papers.ssrn.com/abstract=2000034> (Accessed: 2 March 2022).

Humphreys, S. (2015) 'Conscience in the Datasphere', *Humanity: An International Journal of Human Rights, Humanitarianism, and Development*, 6(3), pp. 361–386. Available at: <https://doi.org/10.1353/hum.2015.0029>.

Hüner, K.M., Otto, B. and Österle, H. (2011) 'Collaborative management of business metadata', *International Journal of Information Management*, 31(4), pp. 366–373. Available at: <https://doi.org/10.1016/j.ijinfomgt.2010.12.002>.

Janssen, M. *et al.* (2020) 'Data governance: Organizing data for trustworthy Artificial Intelligence', *Government Information Quarterly*, 37(3), p. 101493. Available at: <https://doi.org/10.1016/j.giq.2020.101493>.

Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012) 'Benefits, Adoption Barriers and Myths of Open Data and Open Government', *Information Systems Management*, 29(4), pp. 258–268. Available at: <https://doi.org/10.1080/10580530.2012.716740>.

Jones, K.H. *et al.* (2020) 'Toward the Development of Data Governance Standards for Using Clinical Free-Text Data in Health Research: Position Paper', *Journal of Medical Internet Research*, 22(6), p. e16760. Available at: <https://doi.org/10.2196/16760>.

Kitchin, R. (2014) *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Ladley, J. (2019) *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. Academic Press. Available at: <https://books.google.com?id=AkW9DwAAQBAJ>.

Lehtiniemi, T. and Haapoja, J. (2020) 'Data agency at stake: MyData activism and alternative frames of equal participation', *New Media & Society*, 22(1), pp. 87–104. Available at: <https://doi.org/10.1177/1461444819861955>.

Lucie, S. (2021) 'Human Objecthood in the Datasphere: The Body as Resistant Material in Twenty-First-Century Media', *Theatre Journal*, 73(3), pp. 319–337. Available at: <https://doi.org/10.1353/tj.2021.0070>.

Luthfi, A. and Janssen, M. (2017) 'A Conceptual Model of Decision-Making Support for Opening Data', in S.K. Katsikas and V. Zorkadis (eds) *E-Democracy – Privacy-Preserving, Secure, Intelligent E-Government Services*. Cham: Springer International Publishing, pp. 95–105. Available at: https://doi.org/10.1007/978-3-319-71117-1_7.

Mann, M., Mitchell, P. and Foth, M. (2021) *Between Surveillance and Technological Solutionism: A Critique of Privacy-Preserving Apps for COVID-19 Contact-Tracing*. SSRN Scholarly Paper 3717370. Rochester, NY: Social Science Research Network. Available at: <https://doi.org/10.2139/ssrn.3717370>.

Maulina, J. and Ruldeviyani, Y. (2019) 'Data Governance and Data Architecture for the Ministry of Foreign Affairs of the Republic of Indonesia', in *2019 International Conference on Information Management and Technology (ICIMTech)*, pp. 409–414. Available at: <https://doi.org/10.1109/icimtech.2019.8843766>.

McCaig, M. and Rezaia, D. (2021) 'A Scoping Review on Data Governance', *SSRN Electronic Journal* [Preprint]. Available at: <https://doi.org/10.2139/ssrn.3882450>.

Michel, J.-B. *et al.* (2011) 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science*, 331(6014), pp. 176–182. Available at: <https://doi.org/10.1126/science.1199644>.

Micheli, M. *et al.* (2020) 'Emerging models of data governance in the age of datafication', *Big Data & Society*, 7(2), p. 2053951720948087. Available at: <https://doi.org/10.1177/2053951720948087>.

Milne, R., Sorbie, A. and Dixon-Woods, M. (2021) 'What can data trusts for health research learn from participatory governance in biobanks?', *Journal of Medical Ethics* [Preprint]. Available at: <https://doi.org/10.1136/medethics-2020-107020>.

Morley, J. *et al.* (2020) 'The ethics of AI in health care: A mapping review', *Social Science & Medicine*, 260, p. 113172. Available at: <https://doi.org/10.1016/j.socscimed.2020.113172>.

Ostrom, E. (2010) 'Beyond Markets and States: Polycentric Governance of Complex Economic Systems', *American Economic Review*, 100(3), pp. 641–672. Available at: <https://doi.org/10.1257/aer.100.3.641>.

Otto, B. (2011) 'Data Governance', *Business & Information Systems Engineering*, 3(4), pp. 241–244. Available at: <https://doi.org/10.1007/s12599-011-0162-8>.

Otto, B. and Jarke, M. (2019) 'Designing a multi-sided data platform: Findings from the International Data Spaces case', *Electronic Markets*, 29(4), pp. 561–580. Available at: <https://doi.org/10.1007/s12525-019-00362-x>.

Panian, Z. (2009) 'Recent advances in data management', *WSEAS Transactions on Computers*, 9(7), pp. 1061–1071.

Paprica, P.A. *et al.* (2020) 'Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on A Working Meeting of Fifteen Canadian Organizations and Initiatives', *International Journal of Population Data Science*, 5(1, 1). Available at: <https://doi.org/10.23889/ijpds.v5i1.1353>.

Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2015) 'Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution', *PLOS ONE*, 10(10), p. e0137041. Available at: <https://doi.org/10.1371/journal.pone.0137041>.

Pisa, M. *et al.* (2020) 'Governing data for development: Trends, challenges, and opportunities'. Available at: <https://www.cgdev.org/sites/default/files/governing-data-development-trends-challenges-and-opportunities.pdf>.

Porciuncula, L. and Chapelle, B.D.L. (2022) *Hello Datasphere — Towards a Systems Approach to Data Governance*. The Datasphere Initiative. Available at: <https://www.thedatasphere.org/news/hello-datasphere-towards-a-systems-approach-to-data-governance/> (Accessed: 8 April 2022).

Porter, M.F. (1980) 'An algorithm for suffix stripping', *Program* [Preprint].

Rainie, S.G. *et al.* (2017) 'Data as a Strategic Resource: Self-determination, Governance, and the Data Challenge for Indigenous Nations in the United States', *The International Indigenous Policy Journal*, 8(2, 2). Available at: <https://doi.org/10.18584/iipj.2017.8.2.1>.

Reis, J.R., Viterbo, J. and Bernardini, F. (2018) 'A rationale for data governance as an approach to tackle recurrent drawbacks in open data portals', in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. New York, NY, USA: Association for Computing Machinery (Dg.o '18), pp. 1–9. Available at: <https://doi.org/10.1145/3209281.3209354>.

Rossini, C. and Lach, B. (2022) *Datasphere Governance Atlas 2022: Mapping organizations in the data governance ecosystem*. Datasphere Initiative. Available at: <https://www.thedatasphere.org/wp-content/uploads/2022/04/Datasphere-Governance-Atlas-2022-Datasphere-Initiative.pdf> (Accessed: 10 September 2022).

Rushkoff, D. (1994) *Media Virus!: Hidden Agendas in Popular Culture*. Ballantine Books. Available at: <https://books.google.com?id=2hpiAAAAMAAJ>.

Scheibner, J. *et al.* (2021) 'Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis', *Journal of Medical Internet Research*, 23(2), p. e25120. Available at: <https://doi.org/10.2196/25120>.

Scott, J.C. (1990) *Domination and the arts of resistance: Hidden transcripts*. New Haven: Yale University Pr.

Scott, J.C. (1998) *Seeing like a state: How certain schemes to improve the human condition have failed*. Nachdr. New Haven, Conn.: Yale Univ. Press (Yale agrarian studies).

Shah, N. *et al.* (2019) 'Sharing data for future research engaging participants' views about data governance beyond the original project: A DIRECT Study', *Genetics in Medicine*, 21(5), pp. 1131–1138. Available at: <https://doi.org/10.1038/s41436-018-0299-7>.

Shah, N. *et al.* (2021) 'Governing Health Data Across Changing Contexts: A Focus Group Study of Citizen's Views in England, Iceland, and Sweden', *International Journal of Medical Informatics*, p. 104623. Available at: <https://doi.org/10.1016/j.ijmedinf.2021.104623>.

Siegenfeld, A.F. and Bar-Yam, Y. (2020) 'An Introduction to Complex Systems Science and Its Applications', *Complexity*, 2020, p. e6105872. Available at: <https://doi.org/10.1155/2020/6105872>.

Skovgaard, L.L., Wadmann, S. and Hoeyer, K. (2019) 'A review of attitudes towards the reuse of health data among people in the European Union: The primacy of purpose and the common good', *Health Policy*, 123(6), pp. 564–571. Available at: <https://doi.org/10.1016/j.healthpol.2019.03.012>.

Stahl, B.C. and Wright, D. (2018) 'Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation', *IEEE Security Privacy*, 16(3), pp. 26–33. Available at: <https://doi.org/10.1109/MSP.2018.2701164>.

Susha, I., Janssen, M. and Verhulst, S. (2017) 'Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development', *Hawaii International Conference on System Sciences 2017 (HICSS-50)* [Preprint]. Available at: https://aisel.aisnet.org/hicss-50/eg/open_data_in_government/4.

Taylor, J. and Kukutai, T. (2016) *Indigenous Data Sovereignty*. ANU Press. Available at: <https://doi.org/10.22459/CAEPR38.11.2016>.

Tsosie, R. (2019) 'Tribal Data Governance and Informational Privacy: Constructing "Indigenous Data Sovereignty"', *Montana Law Review*, 80, p. 41.

Tuck, R. (2002) *Hobbes: A Very Short Introduction: 64*. Illustrated edition. Oxford ; New York: OUP Oxford.

Updegrove, A. (2004) 'Introducing the Personal Datasphere', *Consortium Standards Bulletin*, p. 4. Available at:

<https://www.consortiuminfo.org/wp-content/uploads/2019/11/feb04-trends-1.pdf>.

van Eck, N.J. and Waltman, L. (2010) 'Software survey: VOSviewer, a computer program for bibliometric mapping', *Scientometrics*, 84(2), pp. 523–538. Available at:

<https://doi.org/10.1007/s11192-009-0146-3>.

Vayena, E. and Blasimme, A. (2017) 'Biomedical Big Data: New Models of Control Over Access, Use and Governance', *Journal of Bioethical Inquiry*, 14(4), pp. 501–513. Available at:

<https://doi.org/10.1007/s11673-017-9809-6>.

Verhulst, S. *et al.* (2020) 'The emergence of a third wave of open data'. Available at:

<https://apo.org.au/node/311570> (Accessed: 8 April 2022).

Weber, K., Otto, B. and Österle, H. (2009) 'One Size Does Not Fit All—A Contingency Approach to Data Governance', *Journal of Data and Information Quality*, 1(1), pp. 4:1–4:27. Available at:

<https://doi.org/10.1145/1515693.1515696>.

Zuiderwijk, A. and Janssen, M. (2014) 'Open data policies, their implementation and impact: A framework for comparison', *Government Information Quarterly*, 31(1), pp. 17–29. Available at:

<https://doi.org/10.1016/j.giq.2013.04.003>.