Institutions, infrastructures, and data friction – Reforming secondary use of health data in Finland

Author: Ville Aula

Institution: London School of Economics and Political Science, Department of Media and Communications

Abstract

New data-driven ideas of healthcare have increased pressures to reform existing data infrastructures. This paper explores the role of data governing institutions during a reform of both secondary health data infrastructure and related legislation in Finland. The analysis elaborates on recent conceptual work on data journeys and data frictions, connecting them to institutional and regulatory issues. The study employs an interpretative approach, using interview and document data. The results show the stark contrast between the goals of open and big data inspired reforms and the existing institutional realities. The multiple tensions that emerged during the process indicate how data frictions emanate to the institutional level, and how mundane data practices and institutional dynamics are intertwined. The paper argues that in the Finnish case, public institutions acted as sage-guards of public interest, preventing more controversial parts from passing. Finally, it argues that initiating regulatory and infrastructural reforms simultaneously was beneficial for solving the tensions of the initiative and analyzing either side separately would have produced misleading accounts of the overall initiative. The results highlight the benefits of analyzing institutional dynamics and data practices as connected issues.

Introduction

In recent years, ideas of open and big data have greatly influenced thinking about public health data (Keen *et al*., 2013; Stevens *et al*., 2018). However, research on data infrastructures has long underscored how difficult and subject to

contingencies of social and organisational dynamics such changes can be (e.g. Bowker and Star, 1999; Edwards, 2010). New data-driven ideas and infrastructural aspirations thus present a fundamental tension between the ideals of data-drivenness and the realities of healthcare infrastructures.

In this paper I analyse how health and biomedical data infrastructures have been reconfigured on the national level in Finland, a country that has a highly state-driven health-care system and research institutions. It explores how ideas of open and big data were advanced in a series of infrastructuring projects in secondary use of health data during the period from 2014 to 2019, and how these ideas stirred controversy among the existing data governors. For ease of reference I refer to these coordinated measures by the collective name of Secondary Health Data Initiative.

The goal of open data in government context is to proactively open data sets generated through government services and registries for companies and civil society to use them to create public benefit for the society through engagement and innovation (Janssen *et al*., 2012; Zuiderwijk and Janssen, 2014; Zuiderwijk *et al*., 2014). Big data, in turn, stresses the new varieties of data that have become available when digital technologies permeate the society and produce vast amounts of new data, and how new analytical technologies are used to make sense of this data, prompting both positive and cautionary assessments (Amoore

and Piotukh, 2015; boyd and Crawford, 2012; Leonelli 2014; Mayer-Schönberger and Cukier, 2013; Kitchin, 2014).

The Finnish case is relevant to the current debate because it pursued big and open data inspired policies to reform both the infrastructure and the legislation around it. Rather than being only a technical project, it centred around governance and regulatory questions, and the paper shows how they were pushed forward and negotiated hand in hand. Thus, the dynamics of the state institutions that were responsible for the governance of the Finnish health data were as important to the reform as the technological questions of data.

On the conceptual side, the paper draws from recent work on data journeys (Leonelli, 2014; Bates *et al*., 2016) and data friction (Edwards, 2010; Bates, 2018). It elaborates on these concepts by linking them more closely to the existing literature on infrastructuring (e.g. Edwards *et al* 2007; Bowker & Star 1999; Pipek and Wulff, 2009), and explores how they can be used to understand the tensions and frictions emerging during infrastructural and legal reform. Parallel examples of the tensions and challenges in health data infrastructuring have been found in the Danish DAMD database for general practitioners (Langhoff *et al*., 2018; Wadmann and Hoeyer, 2018), in the Swedish LifeGen (Cool, 2016), and in the British care.data initiative (Vezyridis and Timmons, 2017), although they all have their differences. In all these cases, ambitious health data initiatives led to legal

challenges, public outcry, and eventually to scrapping of the initiatives, whereas in Finland the new legislation was passed in 2019 and the reforms are now being rolled out.

At the outset of the research, it was quickly found out that the role of data governing institutions was particularly interesting in the Finnish case, because the initiative was driven forward through extensive expert consultations and negotiations between different actors. To explore the role of public institutions in health data infrastructuring and data frictions and journeys, the paper asks the following research questions:

1. *What is the role of data governing state institutions in reforms that aim to make the journey of data easier and decrease data friction?*
2. *How do data frictions break out on the institutional and regulatory level?*

Based on the empirical results, the paper argues that institutional factors and more mundane data practices are intricately connected. A complex dynamic occurs between them, and more research attention should be given to the institutional factors of health data infrastructures, especially in their public sector settings. Elaborating and expanding on earlier results, the paper shows that frictions and tensions do not emerge only from the movement of data, but also in

the legal negotiation on what would be the ideal state of the health data infrastructure and how it should be governed. The results underscore the role of public institutions acting as safeguards of public interest, and the emergent tensions being legitimate concerns that prevented the more controversial parts of the reform from passing.

Theoretical framework and literature review

Studies on infrastructures stress that they are not built but grown (Edwards *et al* 2007; Edwards *et al*., 2009), and the metaphor of growing has also been used about data (Pink *et al*., 2018). This approach stresses the process of infrastructuring as an interplay of design and use towards the point when something has been fully integrated to the work practices of an organisation (Pipek and Wulff 2009), and the technology itself ceases to be a visible and separate part of those processes (Star and Ruhleder 1996). In other words, neither infrastructures nor data comes into to the world ready-made, but they emerge through an incremental process of enacting, extending, standardizing and embedding technical and social practices in specific contexts for unique needs (Bowker and Star, 1999; Edwards, 2010; Hanseth and Lyytinen, 2004; Hughes 1983). Ribes and Finholt (2009) emphasise that even as infrastructures

are extended from earlier configurations, they are always intended for future use and need to be robust. In this way, infrastructuring in mature sociotechnical environments takes the shape of re-infrastructuring, which applies especially to health data infrastructures (Grisot and Vassilakopoulos, 2017).

The contingency of infrastructuring and its multiple possible futures leaves us with a dilemma that 'new e-infrastructures always imagine them as "future proof" and universal, yet real-world systems are always future-vulnerable and particular' (Edwards *et al*., 2009, p. 371). Furthermore, uncertain future trajectories are loaded into present discourse on technological change, and these expectations are not only performative, but also constitutive of the different futures that can materialise (Borup *et al*., 2006). Infrastructuring thus requires negotiation between competing futures whose veracity and viability is indeterminate (Edwards *et al*., 2009).

Data is created through categories, classification, and standardisation, which are interdependent with the very mundane purposes and practices of the organisations that make use of them (Bowker and Star, 1999; Gitelman and Jackson, 2013). This creates a diverse set of standards and metadata that are specific to their local contexts. Bowker (2000, 668) proposes that 'there is no uniform way of separating off the data objects… from their spatial and temporal

packaging', and results below show that this notion rings true among many health data specialists.

Leonelli (2014) has introduced the concept of data journey to analyse the process of disseminating, decontextualizing, recontextualizing, and reusing data to create knowledge. Bates *et al*. (2016) have elaborated on this to stress the differences across the social and material contexts where data practices take place. In these conceptualizations, Leonelli places more emphasis on the epistemological side of the journey, whereas Bates and her collaborators concentrate on the material side and political economy, both acknowledging their connected nature. According to Leonelli (2013), databases are created for local epistemological needs, and larger research infrastructures must serve a variety of epistemological needs to be fruitful. Yet, as Leonelli (2014) points out, not all fields lend similar possibilities for the journey of data to yield benefits in creating new knowledge. Indeed, Bates *et al* (2016) shows that it is the social element of the data practices that defines what is inbuilt in the minutiae of the materialities of producing, formatting, and using data.

Here, data is treated as a material object that is subject to change, and the social and organisational context of use has direct consequences to this materiality. Recent studies have explored the fragile nature of data, and the need for constant tending and repair of data to make it usable (Pink *et al*., 2018, also Jackson,

2014). After all, data passes through several stages of decision-making and manipulation before it is anything that can be stored in the first place (Wallis *et al*., 2008).

Data journeys lead to data crossing contexts and being moved to places far removed from where it was initially produced or managed. The concept of data friction (Bates, 2018; Edwards, 2010) captures the difficulties and tensions that emerge when combining datasets from different contexts or using data from one context in another one. Similarly, there is science friction between disciplines (Edwards *et al*., 2012) because their differences in needs, practices and culture are reflected in their use of data, and the configurations made for data. However, data frictions are shortcomings only if seamlessness and unhinged flow of data are taken as normative imperatives. In contrast, friction can also originate from legitimate reasons to hinder the movement of data to protect citizen's privacy or national interests (Bates, 2018).

Data friction implies the movement of data across contexts, and Neff *et al.* (2017) underscore how data acts as a medium that can be interpreted through various lenses and contextual cultures (see also Seaver 2015). These contexts have their backgrounds in specific socio-technical arrangements contributing to emergence and stability of 'local data cultures [that] constantly recreate themselves' (Bowker, 2000, p. 653). Even simple definitions and standards can be interpreted in

different ways, which poses a fundamental challenge for any attempt to create ontologies of biomedical data (Bowker and Star, 1999; Rea *et al*, 2012; Ure *et al*., 2009). This hold especially true to health data, because even the most basic clinical, nursing and biomedical data embody complex power dynamics, making the minutiae of biomedical data constitutive of the whole medical practice (Bowker and Berg, 1997).

In combination, data journeys and data friction capture the aspiration for movement and the fundamental challenges of using data across contexts. This tension is at the heart of the novel attempts to reinfrastructure health data landscapes.

The ideas about local data cultures are fundamentally at odds with more celebratory and technically oriented notions that emphasise the possibility of drawing insights and meanings from data without knowing its original context or choices embodied in its production. Leonelli (2013) has shown how it is the local ideas that guide development of databases, whereas the different epistemological and practical underpinnings of big data discourses in health sectors have been excellently captured by Stevens *et al* (2018). This conflict between ideas of either generic and local data continues the long-standing tension between contextuality of producing medical data and making data atomistic and portable for secondary use (Berg and Goorman, 1999).

Let me now turn to the role of institutions. Institutions also play a role in the emergence of data frictions, which has earlier been argued by Bates (2012; Bates and Goodale 2017). Her approach to institutions draws from political economy of platform companies, data protection jurisdiction, open data policies, and state surveillance, and has an emphasis on the macro level differences of institutional relations. Her positive notion of data frictions emerges precisely from the attempts to hinder the flow of data to jurisdictions and corporate environments that might have an adverse effect on individual citizens. Moreover, a sustained critique has been mounted on open data policies for their neoliberal overtones of counting on third parties to unlock new forms of value, handing power and economic gains to private actors (Davies and Bawa, 2012; Longo, 2011; Vezyridis and Timmons, 2017).

In contrast to these studies that emphasise the political economy of data, I concentrate on two levels of analysis: the level of data practices, and the role of institutions as governors of data. The two-fold orientation caters for analysing different processes happening on the level of intra and interorganisational micro-processes of infrastructuring, and on the power-struggle of policy, law, and authority on the more political science perspective of institutional dynamics. My analysis concentrates on the dynamics that happen inside the government, but the political economy of the relationship between public and private actors or

national jurisdictions, which is central to Bates' (2018), is outside the scope of the paper.

In the rest of this paper, I refer to the public institutions that are both policy makers managers of data s as data governors. This concept captures their double role in being not only technical managers and interpreters of law, but an integral part of the policy and regulatory process as experts and initiators of change with their own aspirations and agendas. The data governors act both as arenas for practices to unfold in and as actors in the health data landscape.

To conceptualize this double role of data governors in data journeys and frictions, I draw from the role of institutions in the literature on infrastructuring. Mayernik (2016) argues that data practices can act as institutional carriers that contribute both to change and stability of institutions and data practices. Shared data practices make datasets mutually more intelligible, manageable, and interoperable within an institution, and adoption of these shared practices also leads to increasing institutional similarity. Conversely, new institutional arrangements and purposes also cause changes to the data practices and deployment of technical solutions, with similarity in one leading to commonality in the other (Iannaci 2010). This approach has a long history in studies on infrastructuring, and can be traced back to what already Star and Ruhleder (1996) called the third order issues of infrastructuring, which are of political nature and

address foundational questions of what is desirable for infrastructuring in the first place (see also Ribes and Finholt 2009; Edwards et al 2007). However, unlike Mayernik (2016) and Iannacci (2010), they left the notion of institutions largely unarticulated.

On the other hand, role of institutions in reconfiguring health data infrastructures has been explored in several studies. For example Currie and Guah (2007) show how the British institutional setting has played a major role in the well-documented and high-profile failures of the health-care related UK National Programme for IT. Others (Keen *et al*., 2013; Carter *et al*., 2014) have noted the same effect in health data initiatives. Building on a similar two-fold approach to institutions as both arenas for action and actors on their own right, Sahay *et al*. (2009) have highlighted the asymmetric relationship between different actors during health care infrastructural change.

My approach elaborates on data journeys and friction by Leonelli, who concentrates on the connection between data practices and infrastructures, and Bates, who stresses the political economy of data policies, regulation, and practiecs. I approach the question from the perspective of data governing public institutions and how data frictions emerge among them during an ambitious infrastructuring initiative in a specific country. This elaboration and the empirical results of my paper are especially salient in countries that have extensive public

healthcare systems and detailed regulation about data governance, of which Finland is a case.

Methodology

In this study I approach infrastructures hermeneutically, trying to understand how the participants themselves see the subject without giving epistemological privilege to any single account (Ezzy, 2002; Hennink *et al*., 2011). The empirical data consists of interviews and documents, and the analysis concentrates on constructing the dynamics of the analysed case to map the different approaches of the participants, concerns raised by different stakeholders, and the different perception of the reforms.

The document materials analysed consisted of legal drafts, project plans, strategies and policy papers, working group minutes, original presentation slides from presentations, reports commissioned from consultants, enterprise architecture descriptions, and other supporting reports. In addition to documents, 17 semi-structured elite interviews (Mikecz, 2012) were conducted with managers and experts who have participated in the analysed case to map their personal insights to the analysed case (Magnusson and Marecek, 2015). The interviewees were recruited through purposive sampling on the merit of the interviewees having played a substantial part in the Secondary Health Data

Initiative. Potential interviewees were identified with desk research and two informal discussions with the Ministry of Social Affairs and Health, and Finnish Innovation Fund SITRA. Interviewees included representatives from the coordinating bodies, participating government institutions, and members of a legislative working group working on the subject.

Interviews were conducted in June 2018 either face-to-face in Helsinki or via Skype. The interviews were conducted in Finnish and were recorded, resulting in total 19 hours of interview data, and detailed real-time notes were taken on the computer during the interviews. A research diary was written after each interview to document key findings and implications for subsequent data collection. Analysis was conducted with the interview notes and memos, and key interviews were re-listened to ensure validity and to collect verbatim quotes.

Agenda for the future of health data

Finland has a predominantly public healthcare system. Primary healthcare services are offered by over two hundred municipalities, but hospitals are governed by regional cooperatives, and most challenging operations are done by five publicly owned university hospitals. GPs are employed directly by the publicly owned hospitals and healthcare stations, but past years have witnessed a rising

amount of private occupational health companies and primary care services being outsourced to private companies. A distinguishing factor in these services is that all public service data is connected to personal identification numbers that are unique to every citizen but used across public services and systems. This makes nearly all information produced in public services linkable at least in principle. Health policy and legislation are unitarily national, and central agencies act as key players in policy-making along with the ministries. Nevertheless, the mandate of central agencies is inscribed into laws and decrees, and the legalistic nature of Finnish public sector renders most changes to the system as complex legal struggles. During the past decades, the key actors had also established numerous health data registries and expanded their use and scope cumulatively.

The Secondary Health Data Initiative saw daylight in 2015. Two different factors led to it. First, the National Institute for Health and Welfare (NIHW), the top government research, development, and policy institute in the field of health and social care, had expanded their national biomedical registries for decades with the legislation remaining static. Because the Finnish healthcare system was decentralised, and thus the health information systems unique to the regional cooperatives, the national registries had played a significant role in making more data available for researchers. In 2014 the Finnish Deputy Chancellor of Justice declared that the regulatory foundation of these registries had become outdated,

and both the legislation and practices needed clarification and correction, but the registries could continue operating (OKV 628/1/2012). This led to the Ministry of Social and Welfare Affairs establishing an expert working group that was given the task of drafting the new legislation to solve the problems (STM011:00/2015). Unlike for example the Danish DAMD, the Finnish case thus had the legal side of the reform as its leading goal, which caused it to be sensitive to contrasting legal interpretations (cf. Wadmann and Hoeyer, 2018).

Second, in 2014 the Finnish government approved two government policy strategies that reframed how Finnish health data was to be used. The 'Making use of social and welfare data' -strategy (Ministry of Social Affairs and Health, 2014), and the "Growth strategy for research and innovation in the health sector" (Ministry of Employment and Economy, 2014) stressed the shift from merely retaining public health data towards considering data as a public asset that should be used more widely both within and outside government.

The new framing was crystallised in the launch of the ISAACUS-project in 2015, which was initiated and funded by the Finnish Innovation Fund SITRA. SITRA is a publicly owned innovation and policy think tank acting directly under the mandate of the Parliament of Finland, making it independent from the executive central government but giving it no direct authority over how health data is managed in Finland.

The ISAACUS-project sought to capitalise on the Finnish health databases that were considered unique in their breadth and scope. This goal contained two parts: establishing a centralised licensing authority for all national databases of welfare data and launching a one-stop-shop service operator that will take care of all the data processing. Especially the one-stop-shop was considered important by the proponents of the project. In the old system, researchers and companies who wanted to access data from multiple central national registries and databases had to apply for it through separate processes in different data governors, which led to complex compounding processes. Throughout the initiative, SITRA and the Ministry collaborated and fed their results into each-others work, and the sub-projects of the ISAACUS acted as test-beds and drivers of change for the legislative work.

Moreover, the concept of 'enabling legislation' was fostered, positing that new legislation should safeguard new forms of data use and only set general goals and guidelines for the future to give room for the users, innovators and practitioners to work new ways of using data. One interviewee explained its implications for both regulation and practice:

> *'Our legislation used to be, like, you can collect this kind of information to this register, and you can use it to do this and that. There was no-one who would open it up and say, hey, you can use all [the data] to this and that.*

This is perhaps the big breakthrough here, that we will open up the next layer to it.' (Interviewee, Ministry of Social Affairs and Health, 7 June 2018)

In sum, the existing landscape of health and biomedical infrastructures was presented with two sets of challenges: how to make more data available for more actors, and how to make the existing infrastructure unified in supporting this.

Data governors

This section analyses the roles of three key public authorities that both manage public health data and act as policy-makers in the field. Table 1. provides a summary of the different natures and roles of the three key institutions in the Secondary Health Data Initiative. The differences among the basic features of the three institutions demonstrates how they all had to do with health-related data, but their orientation also had considerable differences. They also play different roles in their respective ecosystems of data production, management, users, and collaborators. Instead of seeing biomedical and health data as a unified phenomenon, and the data as just generic data that can be served with unified solutions, the data governors represented different views on the nature of different data sets and its implications. The differences give rise to institutionalised data practices and cultures, which in turn are connected to differentiated data infrastructures (Bowker, 2000; Mayernik, 2016).

Table 1. Comparison of the main data governing institutions

Institution Aspect	Statistics Finland	National Institute for Health and Welfare	Social Insurance Instution KELA
Primary purpose	Produces and governs official statistics.	Central authority for research, development, statistics, and oversight in health and welfare. Produces research, governs data, advices policy, steers development.	Administers social benefits and reimbursements, including drug expense compensations.
Varieties of data	Official statistics	Biomedical central registers, official statistics, biobank data, research databases	Social benefits and pensions data, prescription data, reimbursement data
In-house use	Statistics	Research, policy, statistics, RDI	Service provision, statistics, research, policy
Collaboration	Research, policy, international cooperation	Research, statistics, policy, RDI	Research
Key regulation	Act on Statistics Finland (48/1992) Statistical Act (280/2004)	Act on the NIHW (668/2008) Acton on the National Personal data registers for health care (556/1989) STAKES Statistics Law (409/2001) Biobank Act (668/2012)	Act on the Social Insurance Institution KELA (731/2001) Special laws on administration of social benefits, e.g. National Pensions Act (568/2007)
Oversight	Ministry of Finance	Ministry of Social Affairs and Health	Parliament of Finland

Key agencies that governed the Finnish health data were in principle supportive for the project but objected starkly to the practical and jurisdictional factors of the

new centralised system (KELA response to ministerial request for comment, 28.9.2016; Statistics Finland response to ministerial request for comment, 30.9.2016). The differences and the role of representing diverse communities of practice is well evident in the quote below. In this way, the Finnish data governors acted as arbiters of public interest in the Secondary Health Data Initiative, and they raised legitimate concerns relevant to themselves and their collaborators. The following quote captures this perspective:

> *All the governmental actors have their own legal duties, their established practices, and their promises for their customers. And when we talk about the use of data and the responsibility of record keepers, it quickly brings about ambiguity – they don't want to betray their promises to their customers.* (Member of the legislative working group, interview 5 June 2018)

If the innovation-driven approach to biomedical and health big data stressed the use of data, the data governors stressed the perspective of the supply and the origins of data. This difference has important consequences, because the Finnish data governors were both the *de jure* and *de facto* arbiters of who gets to use that data and to what purposes. The proponents of big and open health data would gladly have seen the discretion of data governors dismantled. An interviewee commented on the issue as follows:

> *" In one hand we have the register keepers in their own foxholes where they check that we are responsible for this and the legislation is this, and they appeal to their own roles…. But [the data users], they check it like yes, this solves some of the problems [we've had] and this is a great thing." (*Interviewee Ministry of Social Affairs and Health*, 7 June 2018)*

This quote exhibits the tension between different actors, and how the data governors and their different positions were considered as an obstacle for the reform.

Points of tension

The clash between the new user-driven big health data and the existing practices meant that the parties could not agree on how the new system should work and both the current and the desired state of the system was interpreted in different ways. These differences manifest on the level of data practices, but they also become obstacles in legal reforms that aim for precision in enabling new uses and protecting citizens' privacy. One interviewee summed up this ambiguity by saying that:" They *[stakeholders] read the same law and understand in completely different ways what it says and what it means."* (Interviewee in a data

governing institution, 13 June 2018) This section reviews how the data friction breaks out on levels that are once or twice removed from the original source of data, making them second and third order frictions (Boyce, 2014). Table 2 summarises the tensions, which are then dealt each in turn.

Table 2. Points of tension

Theme	Underlying question	Source of dispute
1. Power to grant access	Who controls the acccess to data?	Power of the new central license authority over the data-governing institutions.
2. Data origins	How is data produced and retained, and how does this influence the data?	The amount of expertise and knowledge that is needed to understand, process, and use data.
3. Purpose of use	What forms of use are allowed for different data sets?	Different interpretations of what is allowed and prohibited by the legislation, and how should overlapping regulation be interpreted.
4. Managing databases	How is data processed, governed, and disseminated in the new model?	The relationship between the one-stop-shop and the existing data governors in managing and processing the data in practice.
5. Competing projects	Is this the right way of pursuing big and open health data?	Diversity of alternative goals, needs, and projects among data governors and stakeholders.

Power to grant access

Power to grant access is a dispute about who makes decisions about the openness and governance of any data set. The data governors had different ideas about who should be the ultimate arbiter of the possible uses of data, and what types of data should be part of this piece of legislation. This was at odds with the aim to establish a single authority that would grant all licenses. In other words, the data-governors were reluctant to relinquish the power to grant access to their data and interpret the related regulation. The institutions had different interpretation of the laws and were afraid that without domain-specific expertise data could end up in wrong hands. One central theme in this variety of disputes was whether the remit of the new law applied only to biomedical data, because several institutions it applied to would still have to govern the data-sharing processes for other types of data. An important part of this dispute is also the power play between institutions, because changing their dynamics also means someone winning and losing in their power to control data. The positions of the data-governors were supported by legitimate concerns based on the existing legislation and ideas of public interest that mirror their data practices. In their view decisions on sharing and releasing data always required domain-specific discretion.

Data origins

Tensions about data origins were about how the material form of data and the epistemological needs of its inception influence further use. Data journeys are a useful conceptual framework here, because the origins of data matter when new data sets are derived from existing ones, or they are combined to create new ones.

According to the interviewees, data sets and infrastructures are very diverse both across the institutions and within them. This diversity stretches from the technical standards to the origins of their production and structure of the datasets. On top of the technical barriers to interoperability, which were constantly felt even within the institutions, the heart of this dispute is whether it is possible to understand the datasets and use them effectively if one does not know the details of their origins. The implication is that datasets cannot just be combined and used, but both require extensive manual labour and local expertise, which is learned slowly through exposure to practical work with the datasets. This position indicates that data governors had a highly contextual understanding of data, and that the material properties cannot be understood without tacit knowledge.

Data governors have expertise to process data for their own purposes from different external primary sources and their own infrastructures, and these metadata capabilities are central to data being usable to the members of that

organisation and their collaborators. These skills are essential in making the data journey inside the institutions between different operational arms, or in enabling the data to journey outside its remit to external users. However, no such metadata expertise exists across institutions. In the words of an interviewee:

> *One thing is that from the outside of the house it requires a lot of involvement, with for example legislation, to understand what [our] data is and what it contains…. From the outside, it requires effort to understand this, and I would say that the metadata can never be written in a way that would reveal all the gimmicks in that data. That is just not possible*. (Interviewee in a data governing institution, 13 June 2018)

Another side of the dispute was how much of the meanings of data are lost when constructing secondary databases from clinical data, and can the existing or the new combined databases support the variety of possible uses that the policy-side of the reform was striving for (c.f. Berg and Goorman, 1999; Leonelli, 2014; Neff et al., 2017). The difference between varieties of data, and how data is subject to change along its journey from the medical practice to more permanent databases, is evident in the following quote "This registry data that goes to Statistics Finland and to the Care Register [of the National Institute of Health and Welfare], it is very condensed and processed. They are ready-made data sets" (Interviewee in a data governing institution, 8.6.2018). According to the same interviewee the

difference between nationally centralized register data and hospital data was “like night and day”, and the hospital level data offered more possibilities but simultaneously required more working and technical skills to process. This meant that building a new integrated infrastructure based on the centralized register data would not necessarily help in pursuing big data inspired goals that were prominent in the goals of the Secondary Health Data Initiative.

Purpose of use.

Disputes over purpose of use are about who can use data for what purposes. In the situation preceding the reforms, same data could be used for various purposes, but the use must be sanctioned by law, and fit the data governing institution’s interpretation of it. Most restrictions to data use come from legislation that protects citizens’ privacy, but because different data governors abided to partly different legislation, considerable differences had emerged in interpreting and implementing these laws or proposed changes to them. Because of these ambiguities the Secondary Health Data Initiative aimed both at widening the possible uses of data and providing legal clarity for different old and new forms of use by categorizing different varieties of use. This discussion was a prominent part of the law proposal submitted to the Parliament of Finland, and both the legislative working group and the parliamentary committee for health and welfare

extensively debated the categories and their interpretation Adding new categories of use was important because many existing forms of use were legitimised by treating them as varieties and extensions of scientific research, whereas some other possible uses were not covered by these reinterpretations by data-governors and users.

However, many of the newly proposed categories of use were regulated by other overlapping pieces of legislation, treating some forms of use as special cases of exemption or restriction. Bringing the different forms of data and their possible uses under one legislative umbrella was thus not without problems, because there would still exist different frameworks for different varieties of use. The connection between categories of use and the practical use, and the ambiguity in interpreting the former is evident in the quote below:

> *When a statistical authority says it gathers data for some purposes, then it is that statistical authority's interpretation [that defines] what are the other purposes that the authority can give away the data to. There are specific ideas [about it], and other actors might interpret the situation in a different way than the actor responsible for the register. Our regulation is not that unambiguous or precise.* (Member of the legislative working group, interview 6 June 2018)

Moreover, another interviewee commented on the same ambiguity:

> *This secondary use legislation, its position is sometimes like a general law and then sometimes like a special law [within legal hierarchy], and it becomes the problem of someone applying it that what piece of legislation they should apply and where. (Member of the legislative working group, interview 8 June 2018)*

As a corollary, the new legislation would better reflect the new data practices that had developed in the years since the passing of the old legislation, but it would not completely dispel the ambiguity of what category should different data practices be interpreted against.

Database management

Another practical question of making decision about data is who processes the data to be accessed. This issue is crystallised in the role that was to be played the new centralised one-stop-shop service operator. The initial plans of a completely unified infrastructure had to be rolled back already in the early stages of the reform, because it proved to be too ambitious to reorganise the data management processes at once. The data governors felt that their expertise about the data could not be substituted in any near future and disregarding this would happen at the peril of the new one-stop-shop.

Instead of creating a new unified infrastructure and model for data management at every level of the data governors, the reform aimed at installing a new layer of

interoperability that would have a unified user-interface for external users. However, beyond this layer the data governors would still have their separate data management systems and procedures. Managing and processing the data in practice was deemed to be a major obstacle even in the new system, and it had to be carried out by the separate data governors. All interviewees agreed that at the time of the interviews was still unclear how the new unified processes would work, and what would be required from the data governors.

On the other hand, the interviewees unanimously welcomed the newly founded cooperation between the data-governors and felt positively about the advances that had been made. Regardless of the direct products of the Secondary Health Data Initiative, it heralded a new stage of collaboration and laid down foundations for future work and progress between the data governors. Some interviewees voiced that collaboration had proceeded quite well on the practical level, and the problems had occurred more on the legal and managerial level. This notion was implied mostly by the interviewees that had a more technical profession. It implies a difference between institutions as frames to develop common data practices and institutions as actors influencing the legal framework. An interviewee commented on this difference as follows: *"Sure the people in different projects did collaborate, but when we put the honchos there [in the table], they were at loggerheads…. On the personal level [the data expertise] goes over*

organisational borders" (Interviewee, SITRA 14 June 2018). The collaboration and integration on the level of data practices was also seen to carry fruit and enduring value regardless of the fate of the legal reform.

Competing projects

As the Secondary Health Data Initiative continued, it became more apparent that several other initiatives were also addressing the same issue of health and biomedical big data infrastructures, but in slightly different ways. Several different projects in policy, legislation, and infrastructure were initiated in narrow subfields such as genomic research and healthcare management. These different initiatives served different needs, and many of them attempted to establish specialist secondary databases or to transform the re-use of primary data in their original source institutions. In sum, they stated alternative goals for infrastructuring. Because future of infrastructure is always indeterminate and enacted (Edwards *et al*., 2009), they effectively posed a challenge for the Secondary Health Data Initiative by contesting it as the best option to advance big and open health data (c.f. Leonelli 2013).

The interviews suggest, as already noted in some of the quotes above, that there is a palpable difference between re-use of primary data and purpose-made secondary data.

> *I think that the data valuable for medical research comes from the [hospital specific] data pools, but [also] traditional [centralized] register data has had a big role, I don't want to belittle it. [--] In my world the register data is static, and the data pool is dynamic, and we have more 'real-world' data in the data pool. (Member of the legislative working group, interview 12 June.2018.*

The proposed model of a new integrative layer for external users would mainly serve the needs of the latter, whereas many believed that the most important advances will be done in former and require altogether different solutions. The main difference between them is whether the benefits of new strategies towards health data are expected to accrue from new forms of use (the big data perspective), or new users (the open data perspective). The difference between the two is built into the following quote:

> *[Stakeholders were asking] are we constructing big national data pools, and will the model be that we dump all these data masses to a national operator. And do we think that this operator would then serve backwards all the original parties that yielded their data. Or to what extent that data should and also must be analysed as part of their everyday operations…. I slightly contest*

whether it is sensible to gather all those data-masses and try to pile the them into a national pool. (Member of the legislative working group, interview 5th June 2018)

Some interviewees suggested that the Secondary Health Data Initiative was overshadowed by the fact that they anticipated advances in big health data to unfold regardless of the new laws passing or the centralised data-platforms or licencing authorities being established. *"I think these are great visions for now, but the world will, and our research will go forward, and all our techniques, before this has been set up…. And this is no criticism, this is how you take things forward."* (Member of the legislative working group, interview 12 June 2018). The quote underscores how some of the interviewees perceived that big and open data have been the drivers of the recent development in the field, and that legislation was bound to lag.

Discussion

The results and their implications can be summarized in four points. First, the analysis shows how the initial big and open health data inspired goals for integrated infrastructuring were at odds with the existing institutionalised data practices and the interests of some of the data-governors. In the language of data journeys and frictions, the goal of the initiative was to make the flow of data easier between the data governors and their respective collaborators, which meant that

the practical work of the projects was focused on reducing the friction between databases, infrastructures, and institutional remits. In this case, the differences among institutions and infrastructures were also codified in laws and regulation, making the tensions emerging from the infrastructuring process also present on the legal level.

The results thus show that the earlier work on data friction by Bates and Edwards is a fruitful starting point but would not capture all the factors that led to the tensions emerging among the data governors and with the proponents of the initiative. Although especially Bates (2018) does hint towards the importance of institutional and regulatory aspects of data friction, my results elaborate on how they emerge as an ongoing struggle during an infrastructural and legal reform, and how the regulation itself might be wrought with different interpretations that mirror the interconnection between data practices and institutions. On the other hand, the relationship between institutions and infrastructures theorized by Mayernik (2016) and Iannacci (2010) offers a good starting point, but their work alone does not address frictions and tensions in data journeys and practices. Combining these two approaches allows the elaboration on data frictions and extension of the institutional aspects of data practices, shedding light on how the challenges of infrastructuring and reducing data frictions were fought on the level of changing the regulation and legislation.

Second, the analysis underscores the difference between data practices and institutional factors. Although these two are connected, the analysis indicates that data practices can be more readily changed than the cross-institutional issues of governance, regulation, and power dynamics. Concentrating on the microsocial and technical side of infrastructuring could thus have led into the analysis mistakenly downplaying the amount of tensions between the data governors, whereas the most painstaking problems of data use were visible only on the institutional and regulatory level. On the other hand, concentrating only on the legal and institutional level would have missed the success on the level of integrating data practices and infrastructures.

The results of the study highlight the differences between practical developments and the institutional and regulatory change. Many practical developments were enabled by reinterpreting the existing regulation in new ways, but because different data governors had different interpretations of the existing laws, they were adopting new practices in different paces or rejecting them altogether as unlawful. By simply extending and reinterpreting categories of use found in the existing regulation the users of data were able to legitimate new data practices without wider scrutiny. Similar practice-driven developments have been earlier witnessed in Denmark (Wadmann and Hoeyer, 2018), and in Sweden (Cool,

2016), which in both cases led into considerable problems of legitimacy and eventual scrapping of these practices.

This difference leads to the third point. The findings indicate that the Finnish public institutions were both *de facto* and *de jure* arbiters of what constitutes the public interest to be defended in the case of their own data and their respective regimes of data. During the lifecycle of the reforms only minimal public or media interest was given to them, this being the case also during the lengthy parliamentary processing in 2018 and 2019. Unlike in for example the British care.data initiative, it was not the citizens and doctors that formed the front line of resisting the celebratory open and big data policies, but the data-governing institutions. Debating open and big health data in secondary health data unfolded primarily as a bureaucratic process inside the government instead of a public and political upheaval.

In the Finnish case, legal reform was part of the initiative from its outset. This is important because similar projects are often primarily driven by their technological prospects. Although big and open data were important drivers of the Secondary Health Data Initiative, it was also sparked by the Chancellor of Justice decreeing that the existing legal frameworks of biomedical and health registers was untenable. Finding the balance between new forms of use and ethical and legal propriety was thus a key question in the initiative. I argue that a key factor in the

success of the initiative was that the data governors not only asserted and negotiated their different interests and interpretations throughout the regulatory process, but the same actors also engaged in practical collaboration on the more technical level of infrastructuring. This co-evolution of regulation and infrastructures equips the initiative with more solid foundations for future developments than for example the example of the Danish health data infrastructure in which legal unclarity and mission creep destabilised the otherwise already operational system (Langhoff *et al*., 2018).

The role of the data governors was similar to what Wadmann and Hoeyer (2018) have called for as political mechanisms of accountability and deliberation between users and suppliers of data. This role also reinforces Bates' (2018) argument that sometimes data friction can be a good thing, because it signals legitimate differences and contrasting interests in data policies and management. I thus argue that although concentration on data friction might indicate an inadvertent glorification of frictionless flow of data as the norm, the conceptual language of data journeys and flows should be understood as an analytical framework that requires normative judgements about them to be made explicit. What constitutes an unwanted friction hindering the use of data for one actor might be an important safeguard of public sector legitimacy and propriety to another. Moreover, in the Finnish case the data governors have become eager

in defending their normative judgement of what constitutes a ‘right’ kind of data journey and friction from their perspective.

The fourth implication is that furthering the use of big and open data is not only about negotiating a balance of the local and generic qualities of the data and its use but striking a balance between different institutional contexts of data. These results corroborate earlier findings of the importance of context and its effects on secondary use and database integration (Berg and Goorman,1999; Bowker and Berg, 2000; Leonelli 2013). Existing data governors have worked extensively to make their systems seamless and data to journey more easily within their own remit and designated collaborators, but these efforts then lead to differentiated infrastructures between these networks. In this way, data practices act as institutional carriers, as has been proposed by Mayernik (2016).

Moreover, in the highly regulated public sector setting data practices are inscribed into the infrastructure, organisational processes, nature of collaborative networks, laws, and strategic goals. The balance between local and generic is different depending on what is the intent of using data, and the right infrastructure to support this use also depends on the needs and goals of using data. No data infrastructure can cater for all needs, and no single regulatory framework can solve all problems. The Finnish Secondary Health Data Initiative had to face this

problem right from the start, and the tensions analysed in this paper indicate the topics that are likely to emerge in similar projects globally.

Conclusions

The paper explored the role of public authorities in a Finnish health data legal and infrastructural reform, concentrating on how data frictions break out on the levels of data practices, institutional remits, and regulation. It elaborated on earlier work on data friction and infrastructures and highlighted the importance of incorporating institutional and regulatory considerations to this discussion especially in public sector context that has extensive and complex regulation. The results show that the Finnish governors of health data played an important part in the existing frictions in data flow, and these frictions also emanated to the legal level. Moreover, collaboration between actors was more readily achieved on the practical level of technical infrastructure, which also helped the regulatory and institutional side to succeed. The study highlights the role public institutions can have as safeguards of public interest during big and open data inspired reforms.

References

Amoore, L., & Piotukh, V. (2015). Life beyond big data: governing with little analytics. *Economy and Society*, *44*(3), 1–26. https://doi.org/10.1080/03085147.2015.1043793

Bates, J. (2018). The politics of data friction. *Journal of Documentation*, *74*(2), 412–429. https://doi.org/10.1108/JD-05-2017-0080

Bates, J. (2012). "This is what modern deregulation looks like": co-optation and contestation in the shaping of the UK's Open Government Data Initiative. *The Journal of Community Informatics*, *8*(2), 1–20.

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, Vol. 3, pp. 1–12. https://doi.org/10.1177/2053951716654502

Berg, M., & Goorman, E. (1999). The contextual nature of medical information. *International Journal of Medical Informatics*, *56*(1–3), 51–60.

Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. *Technology Analysis & Strategic Management*, *18*(3–4), 285–298. https://doi.org/10.1080/09537320600777002

Boyce, A. M. (2016). Outbreaks and the management of 'second-order friction': Repurposing materials and data from the health care and food systems for public health surveillance. *Science & Technology Studies*, *29*(1), 52–69.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics*, *41*(5), 404–409. https://doi.org/10.1136/medethics-2014-102374

Cool, A. (2015). Detaching data from the state: Biobanking and building Big Data in Sweden. *BioSocieties*, *11*(3), 277–295. https://doi.org/10.1057/biosoc.2015.25

Currie, W. L., & Guah, M. W. (2007). Conflicting institutional logics: a national programme for IT in the organisational field of healthcare. *Journal of Information Technology*, *22*(3), 235–247.

Davies, T. G., & Bawa, Z. A. (2012). The promises and perils of open government data. *The Journal of Community Informatics*, *8*(2), 7–13. Retrieved from http://ci-journal.net/index.php/ciej/article/view/929

Edwards, P. N. (2010). *A vast machine [electronic resource] : computer models, climate data, and the politics of global warming*. Cambridge, Mass.: MIT Press.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding Infrastructure: Dynamics, Tension, and Design - Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructure."*

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

Edwards, P., Bowker, G. C., Jackson, S., & Williams, R. (2009). Introduction: An Agenda for Infrastructure Studies. *Journal of the Association for Information Systems*, *10*(5), 364–374. https://doi.org/10.17705/1jais.00200

Ezzy, D. (2002). *Qualitative analysis : practice and innovation*. London: Routledge.

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 1–14). Cambridge, MA: MIT Press.

Grisot, M., & Vassilakopoulou, P. (2017). Re-infrastructuring for eHealth: Dealing with turns in infrastructure development. *Computer Supported Cooperative Work (CSCW)*, *26*(1–2), 7–31.

Hanseth, O., & Lyytinen, K. (2004). Theorizing about the design of Information Infrastructures: design kernel theories and principles. In *Sprouts: Working papers on information environments, systems and organizations*.

Hennink, M. M., Bailey, A., & Hutter, I. (2011). *Qualitative research methods*. London: SAGE.

Hughes, T. P. (1983). *Networks of power : electrification in Western society, 1880-1930*. Baltimore : Johns Hopkins University Press.

Iannacci, F. (2010). When is an information infrastructure? Investigating the emergence of public sector information infrastructures. *European Journal of Information Systems*, *19*(1), 35–48.

Jackson, S. J. (2014). Rethinkin repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 221–239). Cambridge: MIT Press.

Jackson, S. J., Edwards, P. N., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday*, *12*(6).

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268.

Keen, J., Calinescu, R., Paige, R., & Rooksby, J. (2013). Big data+ politics= open data: The case of health care data in England. *Policy & Internet*, *5*(2), 228–243.

Kitchin, R. (2014). *The data revolution : big data, open data, data infrastructures and their consequences*. Los Angeles, CA: SAGE Publications Ltd.

Langhoff, T. O., Amstrup, M. H., Mørck, P., & Bjørn, P. (2018). Infrastructures for healthcare: From synergy to reverse synergy. *Health Informatics Journal*, *24*(1), 43–53. https://doi.org/10.1177/1460458216654288

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 1–11.

Leonelli, S. (2013). Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research. *BioSocieties*, *8*(4), 449–465. https://doi.org/10.1057/biosoc.2013.23

Longo, J. (2011). # OpenData: Digital-era governance thoroughbred or new public management Trojan horse? *Public Policy & Governance Review*, *2*(2), 38–51.

Magnusson, E., & Marecek, J. (2015). *Doing interview-based qualitative research : a learner's guide*. Cambridge: Cambridge University Press.

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973–993.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data : a revolution that will transform how we live, work and think*. London: John Murray.

Meyer, E. T. (2015). *Knowledge machines : digital transformations of the sciences and humanities* (R. Schroeder, Ed.). Cambridge, Massachusetts : The MIT Press.

Mikecz, R. (2012). Interviewing Elites: Addressing Methodological Issues. *Qualitative Inquiry*, *18*(6), 482–493. https://doi.org/10.1177/1077800412442818

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data*, *5*(2), 85–97. https://doi.org/10.1089/big.2016.0050

Pink, S., Ruckenstein, M., Willim, R., & Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, Vol. 5, pp. 1–13. https://doi.org/10.1177/2053951717753228

Pipek, V., & Wulf, V. (2009). Infrastructuring: Toward an integrated perspective on the design and use of information technology. *Journal of the Association for Information Systems*, *10*(5), 447–473.

Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E., … Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics*, *45*(4), 763–771. https://doi.org/10.1016/j.jbi.2012.01.009

Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development*. *Journal of the Association for Information Systems*, *10*(5), 375–398. https://doi.org/10.17705/1jais.00199

Sahay, S., Monteiro, E., & Aanestad, M. (2009). Configurable Politics and Asymmetric Integration: Health e-Infrastructures in India*. *Journal of the Association for Information Systems*, *10*(5), 399–414. https://doi.org/10.17705/1jais.00198

Seaver, N. (2015). The nice thing about context is that everyone has it. *Media, Culture & Society*, *37*(7), 1101–1109.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Stevens, M., Wehrens, R., & de Bont, A. (2018). Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis. *Big Data & Society*, *5*(2), 1–21.

Ulriksen, G.-H., Pedersen, R., & Ellingsen, G. (2017). Infrastructuring in Healthcare through the OpenEHR Architecture. *The Journal of Collaborative Computing and Work Practices*, *26*(1), 33–69. https://doi.org/10.1007/s10606-017-9269-x

Ure, J., Procter, R., Lin, Y., Hartswood, M., Anderson, S., Lloyd, S., … Ho, K. (2009). The development of data infrastructures for ehealth: a socio-technical perspective. *Journal of the Association for Information Systems*, *10*(5), 415-.

Vezyridis, P., & Timmons, S. (2017). Understanding the care.data conundrum: New information flows for economic growth. *Big Data & Society*, *4*(1), 1–12. https://doi.org/10.1177/2053951716688490

Wadmann, S., & Hoeyer, K. (2018). Dangers of the digital fit: Rethinking seamlessness and social sustainability in data-intensive healthcare. *Big Data & Society*, Vol. 5. https://doi.org/10.1177/2053951717752964

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. *International Journal of Digital Curation*, *3*(1), 114–126. https://doi.org/10.2218/ijdc.v3i1.46

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