

# Producing and projecting data: Aesthetic practices of government data portals

Big Data &amp; Society

July–December 2019: 1–16

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/2053951719853316

journals.sagepub.com/home/bds

**Helene Ratner<sup>1</sup> and Evelyn Ruppert<sup>2</sup>**

## Abstract

We develop the concept of ‘aesthetic practices’ to capture the work needed for population data to be disseminated via government data portals. Specifically, we look at the Census Hub of the European Statistical System and the Danish Ministry of Education’s Data Warehouse. These portals form part of open government data initiatives, which we understand as governing technologies. We argue that to function as such, aesthetic practices are required so that data produced at dispersed sites can be brought into relation and projected as populations in forms such as bar charts, heat maps and tables. Two examples of aesthetic practices are analysed based on ethnographic studies we have conducted on the production of data for the Hub and Warehouse: metadata and data cleaning. Metadata enables data to come into relation by containing and accounting for (some of) the differences between data. Data cleaning deals with the indeterminacies and absences of data and involves algorithms to determine what values data can obtain so they can be brought into relation. We attend to how both aesthetic practices involve normative decisions that make absent what exceeds them: embodied knowledge that cannot or has not been documented as well as data that cannot meet the forms required of data portals. While these aesthetic practices are necessary to sustain data portals as ‘sites of projection,’ we also bring critical attention to their performative effects for knowing, enacting and governing populations.

## Keywords

Aesthetic practice, data cleaning, data portals, embodied knowledge, metadata, open government data

This article is a part of special theme on Algorithmic Normativities. To see a full list of all articles in this special theme, please click here: [https://journals.sagepub.com/page/bds/collections/algorithmic\\_normativities](https://journals.sagepub.com/page/bds/collections/algorithmic_normativities).

## Introduction

Over the past decade, an open government data (OGD) ‘movement’ (Attard et al., 2015) has resulted in the dissemination of government data, especially through centralised data portals such as data.gov.uk (United Kingdom), opendata.dk (Denmark) and data.europe.eu (European Union). These sites have been critiqued for serving the ‘entrepreneurial goals of enhanced competitive positioning and attracting investment’ (Barns, 2016: 554) and neoliberal objectives of marketising public services and privatising public assets (Bates, 2014). Others have addressed how they de-politicise the role of publics by engaging them as ‘individual auditor–entrepreneurs’ who monitor state activities through data (Birchall, 2016: 2) and require ‘intermediaries’ to translate or mediate the use of OGD (Schrock and Shaffer, 2017).

These accounts bring critical attention to government claims about the commercial and democratic values of open data. We contribute to these critiques by arguing that OGD portals also operate as governing technologies in three senses. First, rather than simply publishing and communicating government data, portals require that data meet a range of technical standards and formats that come to produce the data that are then made open. That is, like the practices of ‘packaging data’ involved in the production of

<sup>1</sup>Danish School of Education, Aarhus University, Copenhagen, Denmark

<sup>2</sup>Department of Sociology, Goldsmiths, University of London, London, UK

### Corresponding author:

Helene Ratner, Danish School of Education, Aarhus University, Tuborgvej 164, 2400 Copenhagen, Denmark.

Email: [helr@edu.au.dk](mailto:helr@edu.au.dk)



centralised research databases such as those studied by Leonelli (2016), practices such as standardisation, labelling, classification and documentation regulate data from dispersed sites so that they can be compared and aggregated. We develop the concept of aesthetic practices to capture what these practices do: bring ‘data into relation’ (Walford, 2013) so they can be disseminated via OGD portals.

Second, while centralised portals such as data.gov.uk are the focus of critiques, numerous topic-specific government portals also disseminate open data. They are typically produced by government ministries and departments that regulate and manage not only dissemination, but the production of data. We analyse two such government data portals: the Census *Hub* of the European Statistical System (ESS), which enables users to access, query and download census population data for all EU member states; and the Danish Ministry of Education’s Data *Warehouse*, which enables users to access, query and download education data for primary and secondary school student populations. We demonstrate that packaging occurs not only after data is produced as studied by Leonelli. It includes aesthetic practices of statisticians that involve standardising how data is produced across dispersed sites so that, for example, data from different countries or schools can be joined up and compared.<sup>1</sup> This is related to a third sense in which we argue portals are governing technologies: the Hub and Warehouse do not simply make data transparent and open but enable comparison of the relative performance of populations as objects of knowledge and governing. Through the Hub, the economic and social performance of EU member states can be compared and evaluated. Through the Warehouse, the performance of Danish schools can be compared and assessed. In both cases, governing interventions can then be identified and initiated. So, while governing populations is the ‘final objective’ as advanced by Foucault (2009), specific aesthetic practices are necessary to ‘translate the imaginings of state officials’ into measures of the population and its activities (Curtis, 2001: 32). Data portals such as the Hub and Warehouse are two sites that participate in such translations and are part and parcel of governing populations.

Our interest is how this governing technology requires practices that can manage differences and uncertainties in data produced at dispersed sites. These practices, we suggest, are necessary for bringing data into relation and disseminated as populations. We address this by first adopting Latour’s (2017) conception of the ‘bifocalism’ of knowledge to consider the separation between *practices* involved in the making of data at *sites of production* and *forms* data need to acquire to be disseminated at *sites of projection*. We argue that aesthetic practices are key to this separation and involve purging data of

inconsistencies, differences and uncertainties so that they can meet the forms required by the bar charts, heat maps and tables of the Warehouse and Hub. As such, while researchers generally refer to data practices to capture that data is not given but made (Latour, 2017; Leonelli et al., 2017; Schaffer, 2017), we distinguish aesthetic practices as specific and distinct practices necessary for data to be brought into relation. They involve not simply making data pristine (Plantin, 2019) but enable it to achieve forms required by sites of projection. It is by bringing critical attention to aesthetic practices that we then attend to their performative effects for knowing, seeing, enacting and governing populations.

In the following sections, we first develop our conception of aesthetic practices. We then describe how this conception emerged from our analysis and comparison of data from two ethnographies we conducted of different day-to-day practices of statisticians in the production of data for the Hub and Warehouse. Through our ethnographic accounts we then identify two aesthetic practices that emerged from our analysis and which we argue work to bring data into relation: the production of metadata (for the Hub) and the cleaning of data (for the Warehouse). Typically defined as ‘data about data’, metadata documents the when, where and how of data generation and is integral to enabling data to be related in databases.<sup>2</sup> We emphasise that metadata accomplishes this by smoothing out and accounting for (some of) the partiality of and differences between data. As we will develop, rather than resolving such differences, metadata enables data to be related in spite of these differences. We analyse data cleaning as another aesthetic practice that attends to the inconsistencies, indeterminacies and absences of data by, amongst other things, using algorithms to determine what values data can obtain. Both aesthetic practices, we argue, make absent what exceeds them: data that cannot achieve the forms required of data portals and embodied knowledge that cannot or has not been documented. In the conclusion, we offer final reflections on our conception and analysis of aesthetic practices in relation to OGD portals and how they operate as governing technologies.

## Aesthetic practices of bringing data into relation

Our conception of aesthetic practices begins with Schaffer’s (2017) reflection on data practices as ‘aesthetic tools’ (17) that elicit, extract and select rather than merely produce data. Aesthetics for him is not a matter of beauty but of the operations that work on data to achieve desired forms and meet certain ends. For instance, Schaffer notes how such tools were historically deployed in analogue mappings of the world,

seeking to manage a deluge of data that threatened to overwhelm the world of learning and culture (19). Schaffer investigates the 'relation between schemes that aim somehow to assemble universal knowledge in a single site and the way these schemes work through ingenious techniques of production, design, and storytelling'. Here, cartographic techniques are understood as aesthetic tools 'that help make worlds as well as picture them' (11–12). This echoes the understanding advanced by Science and Technology Studies (STS) researchers such as John Law (1994) who argue that knowledge practices not only represent but enact realities by making some things absent or present, that is, forms such as maps are not simply representations but performative, not simply reflections but 'world making' (Schaffer, 2017: 21). In relation to emerging digital orders, Schaffer argues that many practices similarly seek to manage a deluge of data through aesthetic tools that render it into forms for knowing worlds.

Some anthropologists also adopt aesthetics to interpret, for example, cultural symbols or myths as forms or patterns instead of as representations of something else (Bateson, 1972). Along these lines, Riles (1998) considers aesthetics as 'distinct from questions of "meaning"' (378) and as a departure from analysing the 'hidden politics of meaning' in knowledge practices (Riles, 2006). Aesthetics, in the ways advanced by these authors, is thus neither about questions of representation – to what extent data represents the entities it speaks for – nor its hidden meanings, but rather how data or knowledge are given material forms (Maurer and Martin, 2012; Riles, 2006).<sup>3</sup> From maps, cartographic images, charts and documents to corporate structures, these different authors consider how various material forms come to shape how worlds are made.

Instead of analysing the meaning of data or what they represent, we take up these understandings to consider data projected in bar charts, heat maps and tables as *forms* of knowing, seeing, enacting and governing populations. However, achieving these forms is neither straightforward nor involves the simple application of standards and rules. Rather, accomplishing forms requires bringing data from dispersed sites into relation. As such, accomplishing forms involves myriad decisions and normative judgements, which cannot be settled in advance. It is these decisions and judgements that we attend to rather than the infrastructures, the technical, material, physical and human arrangements in and through which practices operate. While infrastructures are integral to aesthetic practices (Hine, 2006; Larkin, 2013), they are not determining but intrinsically indeterminate, in part due to the fragmentation and partiality of data as well as the friction that arises when data is being brought into relation (Edwards et al., 2011).

The metaphor of friction refers to the energy required, which may be human attention or technical translations, and which is integral to the movement of data:

Every interface between groups and organizations, as well as between machines, represents a point of resistance where data can be garbled, misinterpreted, or lost. In social systems, data friction consumes energy and produces turbulence and heat – that is, conflicts, disagreements, and inexact, unruly processes. (Edwards et al., 2011: 669)

Aesthetic practices are key to resolving such data frictions but they are not what come into view in OGD portals as aesthetic practices get detached from their sites of projection. Nor do they resolve all possible forms of friction that might arise as we will exemplify in the analysis of data cleaning. Rather, aesthetic practices work to resolve frictions that hinder the projection of data. This is due to what Latour (2017) refers to as the bifocalism of knowledge: sites of production, where aesthetic practices take place, are separated from sites of projection, where data come into relation and get disseminated as tables, bar charts, heatmaps, and so forth. By referring to sites of production in the plural we underscore that they are multiple and include myriad dispersed sites, some of which we document below. Critically, aesthetic practices are integral to sustaining a bifurcation between data production and projection such that the knowledge they produce becomes 'confused with the thing, fused with the thing' (Latour, 2017: 177).

The indeterminacies of infrastructures, and the data frictions that inevitably arise, bring to the fore that aesthetic practices always require tacit knowledge. Hine (2006) and Zimmerman (2007) refer to this as informal knowledge and Strathern (2000) describes it as the 'experiential and implicit knowledge crucial to expertise' (313). Others consider tacit knowledge as the skills, know-how and abilities exercised in the making of scientific and expert knowledge that cannot or have not been documented (Collins, 2001; Leahy, 2008; Reay, 2007). Or as Göpfert (2013) has shown, forms are not 'mere décor' (331) but require creative, pragmatic and legal reasonings intrinsic to practices. For these authors, tacit knowledge is recognised as inevitable, necessary and integral when working with data because of the friction, heterogeneity and contingencies that always arise in knowledge practices (Leahy, 2008). As Suchman (2007) has argued 'plans and other formulations of action open out onto a sphere of embodied action and lived experience that extends always beyond their bounds' (21). However, as also suggested by Suchman and argued by Leonelli (2016), abilities,

skills, perceptions and experience cannot be separated into intentional modes of reason and the tacit. For this reason, Leonelli adopts Ryle's (1949) definition of embodied knowledge, which includes all ways of knowing how to produce data. We adopt this definition and the inseparability of ways of knowing but also that understanding plans 'is found in and through, and only in and through' situated practices (Suchman, 2007: 22). Because everything involved in knowing how to produce data cannot be anticipated or documented, we argue that aesthetic practices always include an excess, something beyond knowledge and documentation.<sup>4</sup>

In sum, the concept of aesthetic practices captures the work needed for data to come into relation at sites of projection. This work takes place at sites of production and encompasses tacit knowledge, situated judgments and actions that aim to resolve friction. Sites of projection govern this work through their requirements that data achieve specific material forms of knowing, seeing, enacting and governing populations. For this reason, we argue that practices readying data for projection adhere to a logic of aesthetics rather than that of representation. It is with this framing that we analyse two aesthetic practices involved in the production of data to meet the forms required by our two sites of projection, the Hub and Warehouse. Through a focus on 'unresolved tensions, practical challenges, and creative solutions' such as those studied by Leonelli (2016: 16), we bring attention to the contingencies and normativities of these practices as statisticians work to manage the differences, excesses, absences and indeterminacies of data in order to bring them into relation as populations in the forms required by data portals.

Before doing this, we first outline how our framing and analyses emerged from our discussions of ethnographies we independently conducted on the day-to-day practices of statisticians in the production of data for the Hub and Warehouse.

## A note on methodology

Our article follows from several years of work we did independently on two ethnographic studies of the practices of statisticians as they readied population data for dissemination via the two different OGD portals, the *Hub* and the *Warehouse*. Our ethnographies followed a tradition referred to as the 'practice turn' in contemporary social theory. It is marked by a shift from interpreting social phenomena as structures, systems, life worlds and actions to that of practices. A central precept that unites numerous theories and studies of practices is that they 'are embodied, materially mediated arrays of human activity centrally organized around shared practical understanding' (Schatzki et al., 2001:

10–11). As Gad and Jensen (2014) summarise, in STS the practice turn is exemplified in work such as Latour and Woolgar's (1979) ethnography of how laboratory practices construct scientific knowledge and Mol's (2002) account of how a disease is enacted as multiple through a wide range of medical practices. It is with a commitment to these understandings of practices in the making of social phenomena that we both independently approached our ethnographic fieldwork. We highlight in each empirical section how the practices we analyse required ethnographic methods in order to trace, document and make visible the day-to-day work of statisticians.

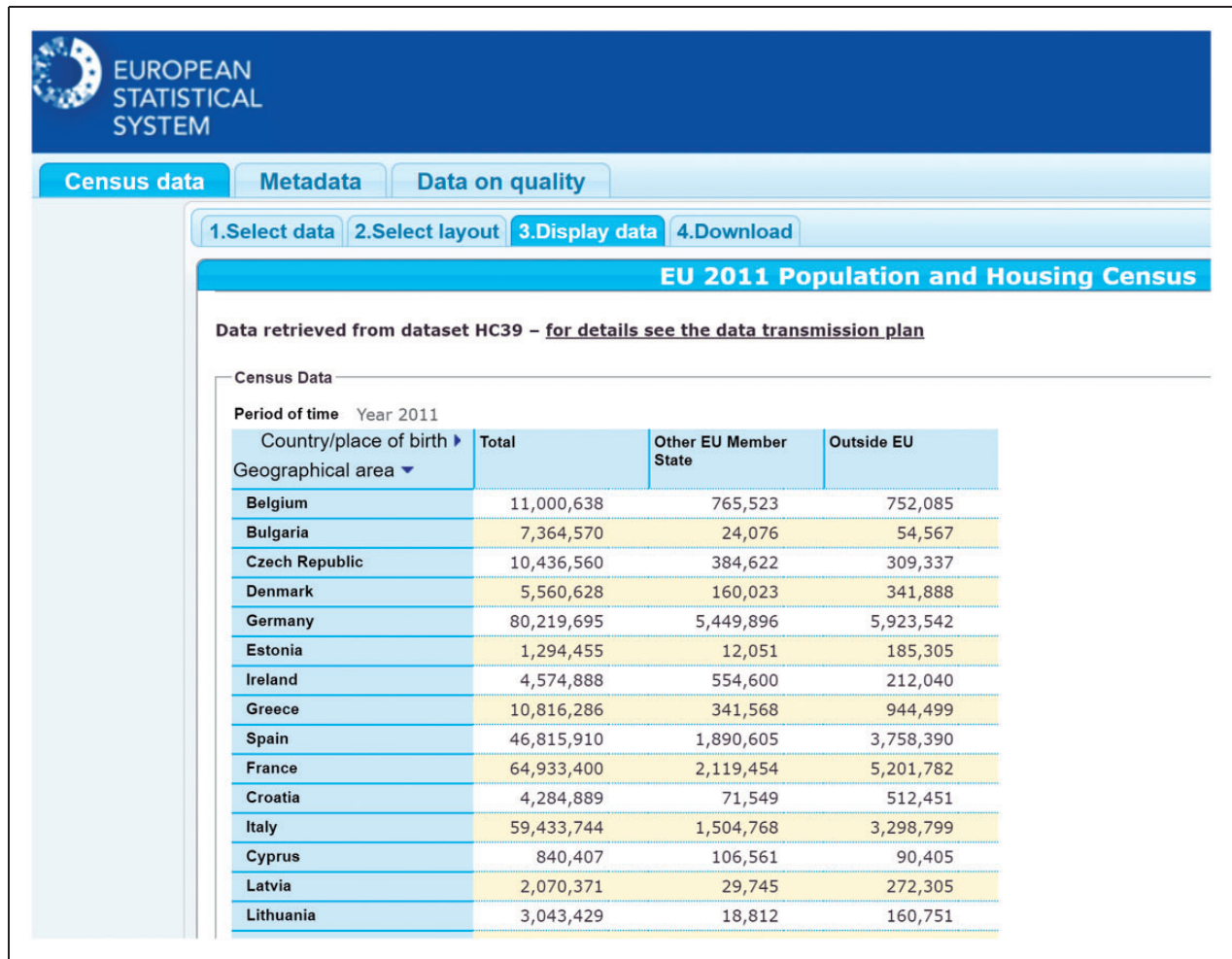
At a workshop in 2017, we discussed our ethnographic findings and found that the populations and specific practices of statisticians we observed were quite different.<sup>5</sup> However, by thinking through the differences between two practices – metadata and data cleaning – we came to identify an organising imperative that they share: to bring population data from dispersed sites into relation in forms required by OGD portals. We then developed the concept of aesthetic practices to capture this commonality. At the same time, we identified a difference in the effects of the aesthetic practices we studied: they enact populations in ways that make them less (homeless people) or more (students) visible. In other words, through the comparison we were able to bring attention to how openness is not singular in its effects but has different consequences for the populations that are enacted. We elaborate this argument in the following sections and here note that it was by considering two different ethnographies of practices involved in making data open that we were able to arrive at both the concept of aesthetic practices and the differential effects of openness.

## Metadata as aesthetic practice

Our first site of projection is the ESS Census Hub. The Hub is part of a broader EU open data initiative that includes a central site, the EU Open Data Portal (EU ODP).<sup>6</sup> The EU ODP provides access to data produced and published by EU institutions and bodies and is also projected via multiple portals such as that of the ESS Census Hub.

Figure 1 is the result of a search query of the Hub. Launched in December 2014, the Hub enables users for the first time to access, query and download census population data for all EU member states via a single portal.<sup>7</sup> It is promoted as providing consistently classified, structured, standardised and methodologically comparable data produced by National Statistical Institutes (NSIs) so that a census of Europe can be centrally accessed and projected in tables. Search queries enable users to aggregate and relate population





**Figure 1.** Census Hub query result (source: <https://bit.ly/2DjOrz1>, retrieved 2 November 2017).

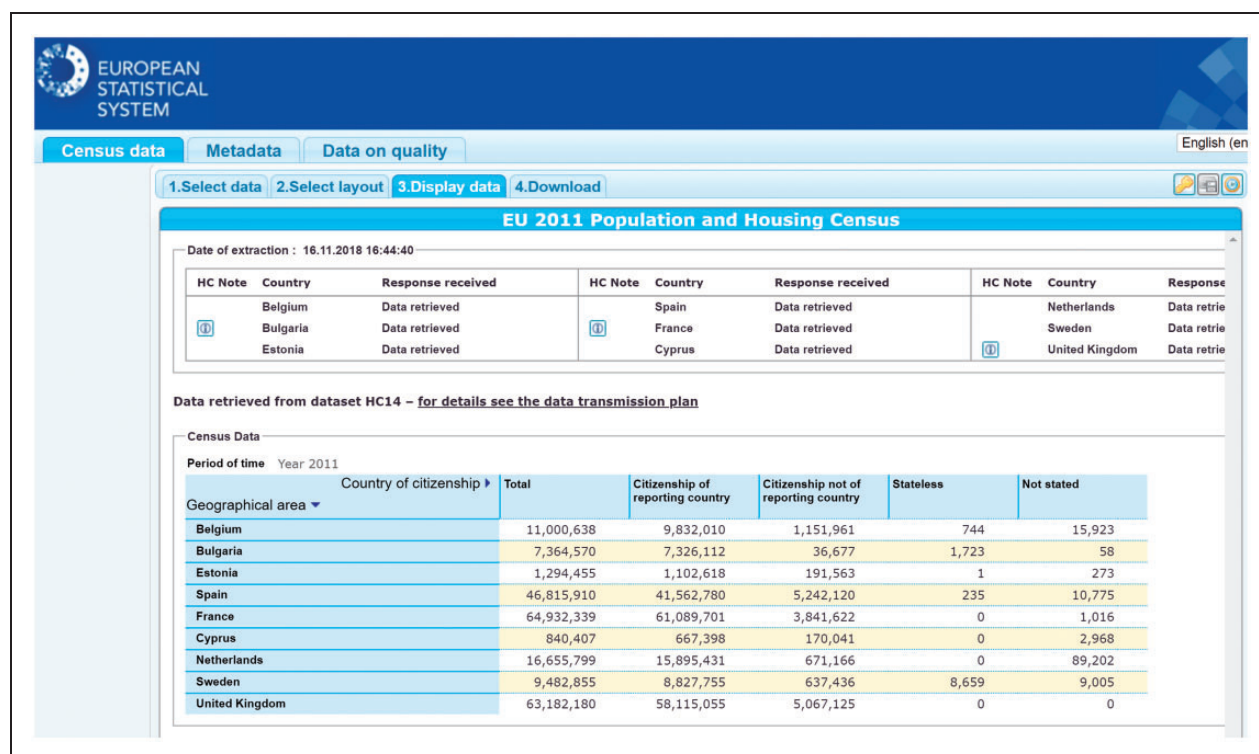
data from different countries according to combinations of three to eight topics (e.g. age, gender, marital status, citizenship) and at varying levels of aggregation.<sup>8</sup> The Hub is the result of regulations on census data and metadata passed by the European Parliament in 2008. The regulations mandated for the first time in EU history, that the NSIs of member states provide 2011 census data in standardised formats so that they could be brought into relation and projected as illustrated in Figure 1.<sup>9</sup>

After the Hub launched in 2014, an ESS task force made up of statisticians from Eurostat and member state NSIs met to assess whether the Hub had achieved its objectives and, based on this, identified regulatory changes for the 2021 enumerations. Their review of the Hub identified irregularities and gaps in the tables that it projected. Our analysis draws on ethnographic observations of their review at quarterly meetings of the task force held at Eurostat between 2014 and 2017.<sup>10</sup> In particular, we focus on their discussions of

irregularities and gaps in the tables that the Hub projected and the regulatory solutions they identified. One solution we analyse is how metadata was identified as a way to resolve gaps in population tables that projected the category of ‘homeless people.’ As we detail below, metadata can be understood as an aesthetic practice mobilised to achieve the desired form of complete tables.

### *Metadata as a container of difference*

The gaps and irregularities that statisticians discussed during their review were evident in projections produced as a result of queries to the Hub. For example, a query that projected data for 2011 on ‘household status’ for 12 member states returned what is illustrated in Figure 2. A category defined as ‘primary homeless people’ appears in only four states, a flag indicates data is ‘temporarily unavailable’ for the UK, and a flag for Sweden – ‘d’ – states that ‘Data on primary homelessness are not



**Figure 2.** Census Hub query result (source: <https://bit.ly/2DjOrzI>, retrieved 2 November 2017).

available' (though it is also not available for seven other states).<sup>11</sup> A tab on metadata leads to a complex table of 21 textual fields. On the theme of comparability, the metadata notes that 'Sweden has done a complete register based Census. This can impair the comparability of the data with Censuses conducted in a traditional or a combined way.' Why and how so are not elaborated. Navigating through the metadata, an entry on household status does not refer to homeless people but states that 'Persons not possible to link to a dwelling cannot form a household, are classified as "Persons not living in a private household, but category not stated"'. In other words, homeless people may be part of this category though the reasons why and their numbers are not provided. Yet, data on all of Europe can be projected and the total 'primary homeless' people reported: 116,510. The number is underpinned by innumerable provisos, missing data, variations in methods and so on that would be practically impossible to assemble and make sense of. Yet, what remains are tables that compare countries but with empty cells and missing data.

Stepping back from the example of Sweden – which is not exceptional – and examining the projection of data for 12 states in the table in Figure 2, variations in the counting of homeless people are impossible to evaluate. They may have been counted – or not – but

by which states, why and how are not evident. Rather, a table with some empty cells is projected and possibly explained by metadata.

The objective of open data and its projection in tables thus made two things visible: gaps in data and variations in data production across member states. These visibilities and variations in how member states defined, counted or did not count homeless people in 2011 became a source of controversy during task force discussions on two census topics: household status and housing arrangements. The 2011 regulations defined household status according to two categories: people living in a private household (as a family, living alone or living with others) and people not living in a private household (in an institution or primary homeless). The category of primary homeless referred to 'persons living in the streets without a shelter that would fall within the scope of living quarters', which excluded what is sometimes defined as secondary homeless: 'persons moving frequently between temporary accommodation'.<sup>12</sup> In 2011, only the 'primary homeless' category was included.

However, many statisticians argued that collecting and providing data on homeless people in the 2011 censuses was very difficult as their data sources and collection methods often did not include them in the population count. This, they noted, resulted in the projection of incomplete tables. Taking that into

consideration, several members expressed concerns about how data on homeless people could then be provided for 2021. For some, their NSIs do not collect data based on the primary/secondary distinction, which led them to argue that the primary homeless category should therefore be deleted. Others – especially from member states that use population registers to conduct their censuses – did not and cannot report this category at all because homeless persons are not included in their registers. Generally, members noted NSI's use various definitions and methods to 'do a count', including data collected by social agencies such as hostels, which also introduces a variety of definitions and categories. They thus recommended that the category of primary homeless be removed and homeless persons – however defined and counted – be subsumed in the generic category of 'Persons not living in a private household, but category not stated', and that 'including homeless people' be added to the description. In other words, the solution to variations in methods and gaps in data was to do away with the category of homeless people so that complete tables can be projected.

A similar issue arose for the topic of housing arrangements, which refers to the type of housing a person occupies. Two main categories were denoted in 2011: 'conventional' housing (e.g. houses, apartments) and 'other' housing (e.g. huts, caravans). In 2011, the latter was further broken down into the categories of 'other housing unit' and 'homeless'. Some members recommended removing the homeless category for the same reasons noted in relation to household status. One member, for example, argued that the category should be excluded as some states do and some do not collect data on homeless people. He argued that amongst other things, this was due to the mix of methods NSIs use such as registers, surveys, or census questionnaires. As such, it does not make sense to require this data, otherwise some states would continue to report zeros and make it look like they do not have homeless people. Another member noted that the numbers of homeless people in most countries are negligible anyway. He added that including the subcategory would give a false impression that NSIs are able to count homeless people. A survey of the ad hoc metadata provided for the 2011 census on housing arrangements indicated that only 17 (of 28) states reported that their census data included primary homeless persons and only 14 states did so for secondary homeless persons. Yet another member said that homeless people have to be reported somewhere. To be consistent with the recommendation on household status, the separate category should thus be deleted and homeless people included as part of the generic category of 'Occupants living in another housing unit' with 'and the homeless' added to the description.

Both recommendations were eventually accepted by the ESS because 'homeless persons must be included in the total population of a country'. However, the ESS decided that member states should still be required to provide a 'best estimate of homeless persons separately' as part of the metadata and optionally break this estimate down into primary and secondary homeless persons. Metadata would also be required to include the specific definition of homeless used (such as who is included, how the estimate was derived and from what sources (e.g. institutional surveys)), and definitions of primary and secondary homeless people, if applicable.<sup>13</sup>

Because of the decision to include homeless people in a generic population category, but still document their numbers in metadata, homeless people will become an implied and unevenly distributed 'absent presence' (Law, 2004) across Europe (due to variations in their inclusion or exclusion in the generic category across member states). That is, to meet the form of complete tables required of the data portal, differences in methods will be relegated to metadata as will the data on homeless people. Bringing data into relation and enacting a European population thus required deferring differences to metadata, which will become not just 'data about data', as metadata is often defined, but data in-and-of-itself. While metadata is a container of methodological differences, when such differences are too great and a desired form is not possible, then data must also be relegated to metadata. In this way, metadata can be considered as a placeholder in that it enables overlooking something by operating as a 'tool of forgetting, of putting to one side' (Riles, 2010: 803). The aesthetic practice of metadata thus also establishes which social relations – such as being part of a household – can explicitly exist as data relations and form part of a population. As Marquardt (2016) argues, homelessness is not only a social issue ignored by governmental data production, but an 'obstacle to conventional ways of data collection on the population' (301).

Yet, while metadata resolved differences by containing missing data and variations in data and methods, it was also contested. Discussions at one meeting reported that metadata was either too long (60 pages or more) or too short (not very informative for users) and some countries did not make full use of footnotes. Various discussions thus took place on how to revise the metadata regulation towards achieving greater standardisation. However, the different methods and practices of member states stood in the way of achieving this. One example was the requirement to report on all data sources, which is problematic for register based countries which may use ten or more data sources and behind those there are about 100 that are used indirectly. Even though the draft regulation defined a data

source,<sup>14</sup> this did not account for indirect sources. That is, numerous sites of production are involved beyond NSIs such as other departments and agencies of member states. A further concern was that the quality of a source must be assessed. As one member reported, registers of external organisations are not harmonised in terms of data definitions, architecture and metadata. They are thus difficult to combine without considerable manual labour, decisions and judgments, and information about the way the data is collected and treated is often not available. What the member's comments highlighted is that it is not only difficult to account for different methods at different sites of production, it is also impossible to account for the judgments and embodied knowledge that data production requires. Metadata could therefore not contain and account for all data friction and so the agreed solution was that the regulation should state that only direct data sources be accounted for and assessed.<sup>15</sup>

Metadata is thus an aesthetic practice negotiated and governed by agreed-to conventions about what and how conditions of production in the making of data can and must be recounted. By doing so, the distinction between what can and cannot be known and recounted is formalised and is part of what Böschen et al. (2010) call 'epistemic cultures of non-knowledge (783).' In this view, aesthetic practices such as that of metadata can be understood as involving 'strategic ignorance' about the known limits of quantification (Scheel and Ustek-Spilda, 2019). Just as the making of data involves explicit decisions about what to make present and absent, metadata also involves decisions about what practices can and must be accounted for and described (Pomerantz, 2015). At the same time, the aesthetic practice of metadata also results in friction. However, as in the case of data, metadata friction is recognised and allowed to exist. In distinction to Edwards et al. (2011), such differences are accepted and what statisticians prioritise is accounting for difference, and being seen to do so, in relation to established protocols and standards. While metadata gives data the capacity to come into relation, containing difference and excess, and making data and metadata achieve the required forms of the data portal, are more important than resolving friction.

In sum, projections of population data in tables on household status and housing arrangements made visible gaps in data, particularly in relation to homeless people. While the causes were difficult, if not impossible to identify by navigating the Hub, the diagnoses of statisticians revealed irresolvable differences in methods of data production. As a result, the form of complete tables required by the site of projection could not be met. Aesthetic practices thus worked to meet those requirements by first including homeless people in a

generic category and then relegating their numbers and methodological differences to metadata. This example of data on homeless people, while seemingly exceptional, involves the aesthetic practice of metadata that is part of bringing all of the different categories of population data into relation. For example, same-sex marriages or consensual partnerships often get folded into opposite-sex categories of population data. These examples thus highlight the inseparability of data relations and social and political relations. To say so is not to suggest that data is a simple reflection. Rather, it suggests that bringing data into relation to achieve a form follows norms and values of dominant cultures such as people being part of a housing arrangement. Just as the social existence of marginalised groups is often socially and politically invisible, so too are they statistically at sites of projection.

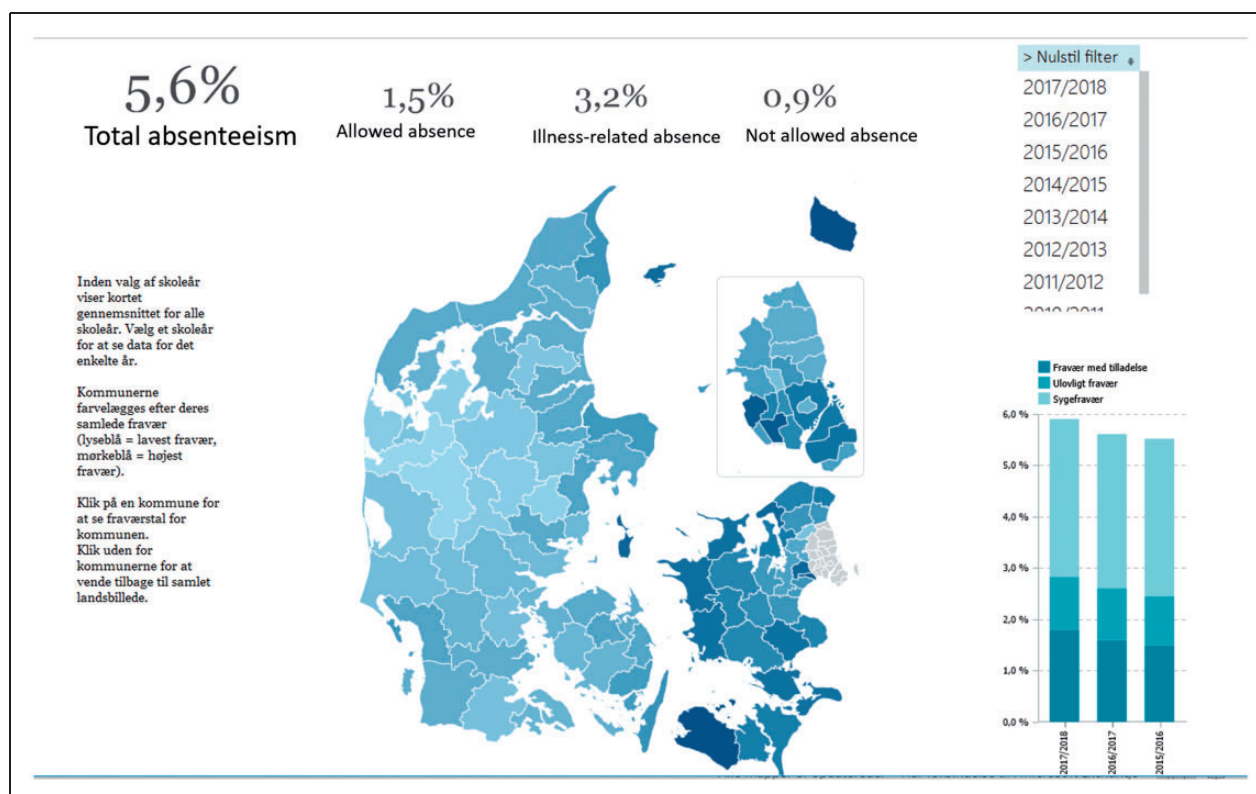
It is in these ways that the projections of an OGD portal can have knowledge effects. The Hub does not simply make visible the relative social performance of EU member states so that they can be compared and evaluated. Through its projections, it enacts versions of populations that make homeless people less visible. The Hub can also have governing consequences as those same projections can inform policy decisions of the EU. One policy that the Hub is intended to inform is the distribution of social cohesion funding, which makes up the lion share of EU spending; in 2014–2020 this amounts to €351.8 bn. However, by rendering one of Europe's most socially excluded groups an absent presence in population statistics, the Hub could lead to resource allocations that do not meet the different degrees and relative needs of homeless people across member states.

## Data cleaning as aesthetic practice

The Data Warehouse, created by the Danish Ministry of Education in 2014, is our second OGD portal.<sup>16</sup> The Warehouse was initially coined to 'provide parents, school boards, school principals and other stakeholders a comprehensive and user friendly overview of how schools are performing on a range of relevant parameters' (Danish Ministry of Education, 2016).<sup>17</sup>

Figure 3 is the result of a search query to the Warehouse. The Warehouse allows users to access pre-defined reports and interactive maps, which compare and benchmark schools against municipal and national averages (providing data on, for example, well-being, final exam grades and students' absenteeism). Through these forms of projection, the Warehouse invites users to bring together data and render municipalities, schools, student populations (e.g. boys and girls) commensurable and comparable. For example, the Warehouse projects data as





**Figure 3.** Data Warehouse query result (source: <https://bit.ly/2PWThK8>, retrieved 20 November 2018 and translated by the authors).

interactive heat maps, which bring data from across municipalities into relation. Figure 3, on students' absenteeism, is such a projection. Through different intensities of blue (with darker colours indicating the highest ratio of absence), each municipality is made visible in relation to the rest of Denmark according to different degrees of student absenteeism. By clicking different places on the heat map, users can change the data relation: the relation between different categories of absence (allowed absence/illness-related absence/not allowed absence), in different school years, or even zoom in on a municipality and see all schools within that municipality in relation to each other. The Warehouse thus works as a governing technology in the sense that it projects schools' performance benchmarked against municipal and national averages and through centrally defined quality indicators. These projections in turn can come to shape how schools and their local stakeholders understand and intervene in educational quality.

Our analysis draws on ethnographic observations (2015–2017) at the ministerial Agency for IT and Learning (Styrelsen for IT og Læring – STIL), which is responsible for developing and maintaining the Warehouse. Data production is distributed across multiple sites, including schools and Statistics Denmark,

which provides background data on, for example, parents' level of education and students' Danish or non-Danish heritage. This allows the Warehouse to render student populations comparable vis-a-vis what the statisticians describe as socio-economic background factors. Like the Hub, the production of data for the Warehouse is regulated by legislation that specifies how data is to be produced and reported by schools. However, again like the Hub, our analysis of the Warehouse concerns how such regulations are not sufficient to bring data into relation and ready it for projection. As our ethnography of the work of STIL statisticians demonstrates, another aesthetic practice is required to address absences and indeterminacies in the data that statisticians refer to as 'cleaning the data.'<sup>18</sup> This involves algorithmic and human interactions with data, tasks that are considered routine and typically not documented.<sup>19</sup> This is because, as we outline below, it involves embodied skills and experience working with data and software, which are assumed competencies of statisticians.

While our ethnography involved observations and interviews with several members of STIL, we focus on data cleaning through an account given by the statistician 'Jonas', a middle-aged statistician with a background in political science and quantitative analysis.

Jonas had worked at STIL for five years when the field-work commenced. He describes himself as a ‘number person’, working on data collection and data analysis. In an interview focusing on data cleaning, Jonas walked Ratner through the process of cleaning data in front of his computer. Learning about data cleaning through such an interview conveyed to us the intricacies of this aesthetic practice.

When asked about data cleaning, Jonas reflected on how this was oriented to the requirements of the Warehouse:

Data has to be capable of enduring the data warehouse’s disaggregation to the level of school and grade [students’ year group]. We aim to not have holes in data, we can’t have a missing school or grade in the data warehouse. (...) In that sense, we don’t have to simply take into account statistical uncertainty. We need detailed data from all institutions although data cleaning can never remove all holes. (Jonas, 1 May 2017)

Cleaning data, in his account, is about preparing it for projection at different levels of aggregation such as that in Figure 3, where a dashboard provides users the option of disaggregating data at the level of municipalities and schools. Because such projections are the desired form, STIL requires data sets to be as complete as possible. In the following section we describe how the aesthetic practice of data cleaning achieves this through an interplay between the patterns revealed by algorithms, statistical software and human judgments about which data to delete, correct or keep in its original form (Helgesson, 2010; Plantin, 2019; Walford, 2013).

### *Data cleaning as managing absence, inaccuracy and indeterminacy*

Important aspects of data quality, as Jonas noted above, are to secure a close to 100% response rate and to have both nationwide and detailed data. While all primary and lower secondary schools are required to submit data to STIL, there is no guarantee that they do so. This calls for a practice of filling in absences in data.<sup>20</sup> First, STIL uses SAS software<sup>21</sup> to check for ‘holes’ in data. They do this by looking at response rates to find schools that did not submit, and by making frequency tables and variable cross checks to find missing data within datasets. Frequency tables that bring data into relation from different schools in this way serve as referents for identifying missing data. That is, the completeness and quality of data from any individual school is assessed in relation to data from all schools. This aesthetic practice of looking at patterns and holes in data may prompt the statisticians to

contact schools to obtain the ‘missing data.’ This human involvement, however, is also a question of resources as there is a limit to how many times they can contact a school for missing data. As Jonas explained, ‘Data isn’t better than what comes in. We have to decide whether to correct data, to delete it or leave it be’ (Jonas, 1 May 2017). In this way, pragmatism is part of his exercise of embodied knowledge.

Data cleaning also involves managing inaccuracies by sorting data so that it is ‘correct.’ Jonas explained:

The political wish is to facilitate easy measurement and comparison of institutions. Yet, we know that schools feel exposed and worry about misrepresentation. Knowing that schools are held accountable through our data, it is extremely important for us to have correct data. (Jonas, 1 May 2017)

Such correcting of data begins when schools attempt to submit data to STIL. Most categories of data are delimited by so-called automated validation rules, an algorithm coded by STIL’s technicians that delimits the values that data can take. For instance, when reporting the number of students in a class, which by law is delimited to be a maximum of 28, the algorithm automatically rejects values deemed to be unrealistic and asks the school to resubmit data. If the deviation is smaller but realistic, the submitter receives an automatic error message asking them to check data. If the submitter does not act on this message, the submitted values are ready for manual cleaning. The sending of an automatic email is thus aimed at attracting the submitter’s attention towards data and thereby improve data quality. When accepting and rejecting data values, the algorithm determines what values data can obtain. With these actions, as well as through generating automatic error messages, the algorithm also determines which human actor should take the next look at the data: the submitter or a statistician working in STIL.

There are also inaccuracies that escape the algorithm’s automatic sorting procedure and these might lead to indeterminacies in data that cannot be settled. For example, like other governmental bodies in Denmark, STIL uses the civil registration number, a unique identifier for all people with a Danish residence permit, to correct data. The number begins with the date of birth (six digits), followed by a serial number (four digits) with the last number indicating the gender (even is female and uneven is male). The information about age and gender is sometimes used to infer trust in the civil registration number as a whole. Jonas gave an example from his last practice of cleaning of data:

I found one person aged -1. This is obviously a mistake in the civil registration number. But then, can we then

trust the gender for this person? Does it mean that this civil registration number is fundamentally wrong? Do we trust this number? Or do we prefer to delete it? (Jonas, 1 May 2017)

This issue of trust is due to the civil registration number's property of being both a unique identifier of a person and data (age and gender) about that person. When the number is treated as information about age and this is deemed inaccurate, it is impossible to determine the quality of the number's other informational capacities.

During the conversation with Jonas, he used terms such as 'strange' and 'mystic' when talking about what he called 'inaccurate' or 'unrealistic' data. He described how he could become completely absorbed by 'a mystery' in a civil registration number, producing a tension between his own desire for correcting data and the need for this to be done:

One student with a strange civil registration number doesn't make a difference [...] in a data set of 700,000 students. It doesn't make a difference but when you've just discovered it it's very difficult not to correct it, it's like, ooh, we can't have this! (Jonas, 1 May 2017)

Data cleaning, as the statistician reflected, produces a sense of obligation to correct data even when it is no longer necessary for purposes of disaggregation in the Warehouse.

Indeterminacy also became an issue when repurposing data. This was particularly the case with the registration of numbers of periods in a grade (year group) and students in a class (group of students being taught together). Rather than schools reporting this data separately, it is automatically extracted from their digital schedules. For schools, schedules are primarily a planning tool, organising the daily rhythms of students and teachers, including a flexible and dynamic division of students into classes. For STIL, schedules contain information about the number of periods per subject per grade and information about size of classes (both regulated by law) and the data is only extracted once a year. Schools thus enter data on an ongoing basis for the purpose of scheduling, which then becomes input for STIL's annual informational need to register whether schools live up to the legislation on the minimum number of subject periods per grade and maximum number of students in a class.

This could cause data friction especially in the early school years as many schools have variable classes:

I mean, data is just data but it is collected with a different idea about how to use these registrations (...) They [schools] know that they can only have 28

students in a class but they practice ongoing school start where they take in students slowly, make [temporary] classes across different grades (...) with students from 1st, 2nd and 3rd grades in the same class. In reality they are no more than 30 students in a class but when we see data in a statistical form it looks as if there are 40 students. How can we handle that? It's challenging when data originally has been registered for a different purpose. (Jonas, 1 May 2017)

Repurposing data means that statisticians sometimes have to guess whether a registered class refers to the category of 'class' or 'grade' as some schools do not use these categories in a consistent manner. Often, these schools would contact STIL's statisticians to have their data corrected. 'A few schools report that our data is wrong. But it's the numbers they registered themselves. But if they register it differently, it hassles their scheduling' (Jonas, 1 May 2017). This is a source of data friction as the schools cannot simply change their reporting of numbers without it having consequences for their own planning purposes. And the statisticians, in turn, need the schools to report the correct numbers in order to change their data. This speaks to Edwards et al.'s (2011) suggestion of a 'process' view of data that emphasises 'friction' rather than flow. Rather than a fixed and interoperable product, interlocutors often have to work-out data over the phone or email (667). Data cleaning is not simply about cleaning data but also negotiating and sorting out with submitters how to understand and use that data. This friction is recurring and made one of the senior statisticians complain about the 'difference between school reality and statistical reality' (Senior statistician, 1 June 2017). The school reality, according to the same statistician, refers to schools' pedagogical experiments with the dynamic organisation of classes over the flux of a school year. However, for statistical purposes, STIL has to follow the legal definition of a class and the ongoing flux of students entering or leaving classes has to be reported as a fixed class. However, when data travels from schools to STIL, these different objectives and requirements generate friction. Rather than changing the 'school reality', the aesthetic practice of data cleaning resolves this friction so that data can be projected in the forms required of the Warehouse. This, however, becomes a new source of friction for schools as it interferes with their scheduling. In that sense, while data cleaning resolves the friction that hinders STIL's projection of data, it generates new friction at the schools. Aesthetic practices thus do not resolve all possible forms of friction and can even become a new source of friction elsewhere.

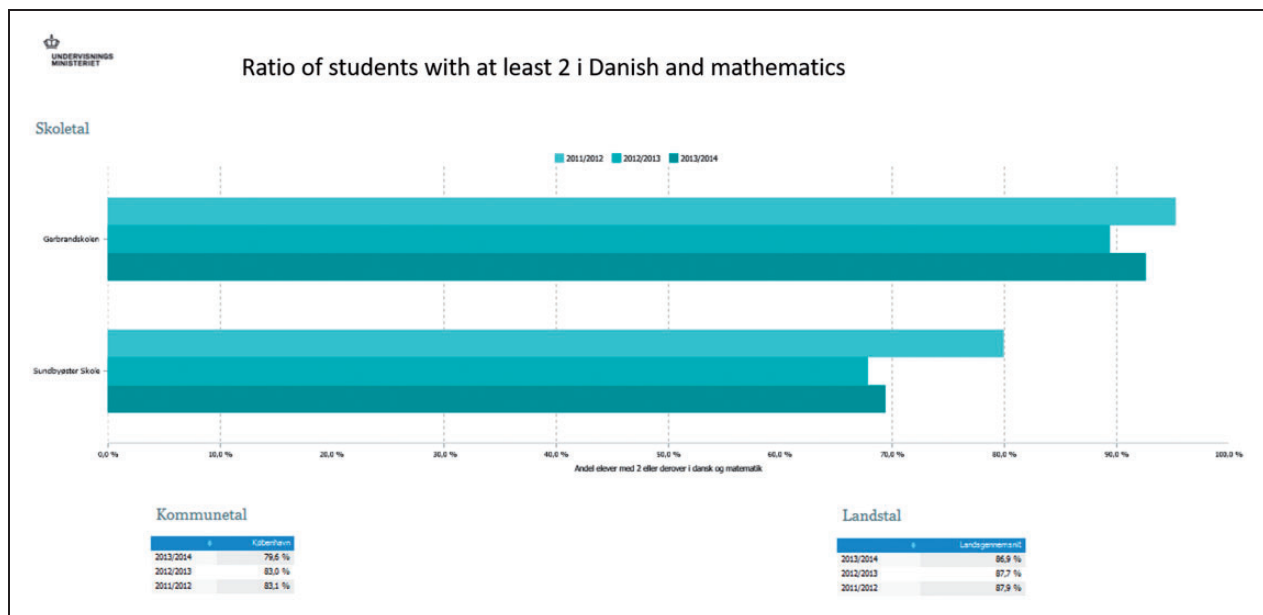
As an aesthetic practice, cleaning involves discovering and managing absence, inaccuracy, and indeterminacy

in datasets. Many different elements, statisticians, data, SAS, algorithms, schools, communication lines such as email and telephone calls participate in these practices. There are no official protocols for which statistical functions to perform on data, and different statisticians hold, as Jonas explained, different personal coding preferences. The small variations in how the functions are programmed are made possible by statisticians' experience with data and their embodied knowledge of the software. The importance of embodied knowledge, what Zimmerman (2007) refers to as a 'non-codified feel for data', especially emerges when the results of that coding are interpreted, for example, the estimation of frequency tables and the decisions of how to act on missing data as discussed above.

Through the aesthetic practice of data cleaning, data comes to be deleted or recognised as correct and ready for projection in the form of bar charts and heat maps of the Warehouse. Once projected, the absences, inaccuracies and indeterminacies discussed above are no longer visible as the frictions at the site of production disappear at the site of projection. For example, Figure 4 is a bar chart that projects the ratio of students with at least a grade point of two in Danish and math. One cannot see, for instance, that gender is identified from civil registration numbers and the indeterminacy that might come with this (including the civil registration number's gender binarism), even though disaggregation in terms of gender is possible. Rather, gender appears as a variable that can be queried and, for example, boys at different schools can be compared and in relation to a national average. Neither is it visible to

what extent the statisticians managed to fill in all absences. The idea of complete data, however, is enacted with the posting of municipal and national percentages, with percentages containing the 'implication of completion, or wholeness' (Guyer, 2014: 156). The Warehouse thus separates the projection of data from the uncertainties and intricacies of their production: the data holes, frequencies and mysteries made visible by SAS and statisticians that must be managed in order to bring data into relation.

The aesthetic practice of data cleaning illustrates two major points. First, data cleaning at sites of production such as STIL manages the unruliness of singular data points by putting them into relation with each other. The requirements of forms are further imposed by predefined quality indicators through which the Warehouse projects populations. While data cleaning is about adding, deleting or correcting data, it is never complete and stable. Yet, at some point, statisticians decide that data is sufficiently ready for projection in the forms of bar charts and heat maps. This is possible only after a combination of algorithmic sorting, iterative SAS analysis, manual acting on inaccuracies, personal communication with schools, etc. Second, as an OGD portal, the Warehouse is also a technology of governing that renders student populations visible through comparisons and benchmarks. One potential effect is to foster a competitive environment by exposing schools' relative performance, as Jonas noted. As a consequence, centrally defined indicators and targets can come to shape how schools, auditors and parents evaluate student populations according to skills and



**Figure 4.** Data Warehouse query result (source: uddannelsesstatistik.dk, retrieved 26 November 2018).<sup>22</sup>



competencies and in turn demand and implement changes in educational practices.

## Conclusions

Making government data transparent and open is claimed to counter corruption and empower citizens through new opportunities ‘to actively participate in governance processes, such as decision-taking and policy-making, rather than sporadically voting in an election every number of years’ (Rojas et al. 2014, cited in Attard et al., 2015: 400). These promises, in turn, are countered by concerns of neoliberal configurations that ‘encourage individual rather than collective political agency (...) [and produce a new] auditor-entrepreneurial-consumer subjectivity’ (Birchall et al., 2015: 187). Of course, the (anti-) democratic effects of OGD are important to scrutinise but there is a tendency to take for granted the data of OGD initiatives. In response to these critiques, we have instead explored OGD as governing technologies in three senses. First, data is not simply made. Rather, OGD portals govern the production of data and shape the forms that data can take as well as what data can be made ‘open’. Second, by focusing on sites of production, we have shown that to operate as governing technologies, data portals require aesthetic practices such as metadata and cleaning data. These aesthetic practices manage the differences, excesses, absences and indeterminacies of data, and they are a precondition for OGD portals. Third, as sites of projection, data portals govern in the sense that they make visible the performance of populations through benchmarks and comparisons that ‘value’ (Muniesa, 2011) and enact populations in particular ways.

Taking data for granted is perhaps in part due to the bifurcation between sites of production and their myriad aesthetic practices, from sites of projection where populations are rendered knowable and governable. It is through such a separation, attained and maintained by aesthetic practices, that projections of OGD are made possible, including their normative and political consequences. For the Hub, one consequence is that homeless people disappear from projections as they are made an absent presence in generic categories and/or subsumed in metadata. The objective of bringing data into relation to achieve a form means doing away with and containing the social differences of homeless people. For the Warehouse, one consequence is the generation of a competitive quasi-market where benchmarks, comparisons, socioeconomic performance rankings and heat maps make student populations visible and comparable across a range of indicators. As with homeless people, projections also contain absences such as educational

qualities not visible in performance indicators. Yet, they come to inform not only how schools see and intervene in educational quality but also how parents choose schools for their children. In both instances, decisions are limited to measurable indicators that statistics can provide based on data that has been readied by aesthetic practices. In this way, data relations intervene in not only what is known but also the relations between schools, their respective student populations and their local stakeholders.

Through the comparison of the production and projection of population data for students and homeless people, we have also brought attention to how openness is not singular in its effects but has different consequences for the populations that are enacted. For students, openness involves managing frictions such as indeterminacies and absences of data so that detailed comparisons can be projected. By adding, deleting or correcting data, projections across a wide range of comprehensive and varied parameters concerning student populations are made possible and can be expanded over time (Ratner and Gad, 2018). Openness thus leads to enactments of student populations that make them as visible as possible. For homeless populations, data frictions are too great and irresolvable and are instead contained in metadata so that complete population tables can be projected. By subsuming homeless people in a generic category and then relegating their numbers to metadata, projections render homeless people an absent presence. Openness thus leads to enactments of populations that make homeless people less visible.

At the same time, in both cases, the excesses, uncertainties, inaccuracies and absences addressed by the aesthetic practices of metadata and data cleaning are forgotten once data is projected. Here, data is treated *as* knowledge, for which no version of openness or transparency could possibly account. Rather, projections produce ‘anaesthesia’, a term Anna Munster (2013) defines as a form of forgetting by flattening experience and relational processes involved in the making of knowledge. Our ambition, thus, is not to call for an impossible version of openness but rather draw attention to those aesthetic practices that projections flatten: the embodied knowledge and normativities that bring data into relation so they can be projected and made open in forms required by data portals. Aesthetic practices are uncertain processes, requiring mundane knowledge as well as the labour-some management of frictions. Attending to aesthetic practices elucidates the contingencies of normativities rather than letting them retreat into the background. That includes the role of algorithms, which are part of iterative and recursive relations between human and automated procedures. In this regard, the understanding of aesthetic

practices is a way to ‘resist fetishizing’ algorithms as technical objects (Dourish, 2016) apart from other operations. Importantly, our analysis also suggests that data is not simply constructed but *aestheticised* to bring it into relation and achieve the *forms* that sites of projection demand. In doing so, aesthetic practices have consequences for the populations enacted and made open and transparent. At a time when OGD is increasingly discussed, it is thus important to not only scrutinise the societal effects of open data such as citizen empowerment or the privatisation of public assets. It is also necessary to attend to aesthetic practices such as those involving algorithms that make data open and the power and consequences of the bifurcation between data production and projection sites on which OGD relies.

### Acknowledgements

We are grateful for the criticisms and suggestions of four anonymous peer reviewers and the editors of *Big Data & Society*. We also wish to acknowledge the feedback and leadership of the two Guest Editors of this special theme, Francis Lee and Lotta Björklund Larsen. Our article has benefitted from the discussions and input of participants at two workshops that the Guest Editors organised and which led to this special theme. Finally, we want to thank statisticians who participated in and made the research leading to this article possible.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research leading to this publication received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 615588, Peopling Europe: How data make a people (ARITHMUS). Principal investigator, Evelyn Ruppert, Goldsmiths, University of London.

### Notes

1. Leonelli provides what she terms a relational understanding of scientific research data. Compared to our concept of bringing data into relation (i.e. joining up data), she is concerned with the operations that curators undertake so that data can be shared, retrieved and re-used by researchers via infrastructures and databases.
2. There are many different meanings of metadata and data cleaning. As we will discuss later, we focus on that which is specific to the epistemic community we are studying, that of government statisticians.

3. Riles (2000) and Maurer (2005) are further interested in destabilising boundaries between the conceptual and empirical. In Riles (2000), for instance, mats collected by Fiji women and the brackets of a UN document function as analytic entry points and in this way bring into question anthropological modes of theorising. Although an important line of inquiry, our endeavour in this paper is different. Our interest is how Riles and Maurer use aesthetics to obviate questions of representation and meaning and instead attend to forms and their materiality.
4. van de Port (2016) conceives of excess as the ‘beyond’ of all representation and ‘the-rest-of-what-is’ to capture the writings of philosophers and cultural analysts such as Slavoj Žižek and Alain Badiou on the Lacanian concept of ‘the Real’: ‘the symbolic orders that promise us to make sense of ourselves and the world fail to capture the experience of ourselves and the world in its entirety’ (181).
5. The workshop was organised by the guest editors of this special theme, Francis Lee, Uppsala University and Lotta Björklund Larsen, Stockholm University.
6. See <https://data.europa.eu>. As stated on the website, the portal provides access to data from the European Union (EU) institutions and other EU bodies. It aims to help put data ‘to innovative use and unlock their economic potential’ and ‘make the EU institutions and other bodies more open and accountable.’
7. The ESS is a partnership between Eurostat (the statistical agency of the EU) and the NSIs of the 28 member states.
8. Topics are the convention for what is sometimes also referred to as variables: e.g. age, sex, nationality.
9. See <http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census>. In the 1990s, the ESS reached a ‘gentlemen’s agreement’ that provided guidelines for standardising national population data (definitions, classifications, categories) so that it would be comparable across EU states. For the 2001 round of enumerations, Eurostat assembled this data into tables and disseminated it in pre-defined cross-tabulations on key population topics (e.g. sex, gender and citizenship). In 2008, data was for the first time regulated so that it could be related according to different combinations of three to eight census topics (e.g. age, sex, nationality) (Eurostat, 2011).
10. These meetings typically took place over two to three days at the Eurostat offices in Luxembourg. Ruppert attended these meetings as an observer and took detailed fieldnotes of discussions. The analysis in this article draws principally from meetings held in 2015 and 2016. Notes along with meeting documents (reports and agendas) were then stored in an Nvivo 10 (qualitative data analysis software) database. The documents were then coded and analysed in relation to the topics of metadata and homelessness. The narrative in this section summarises and draws examples from that analysis. This fieldwork was part of a larger collaborative research project called ARITHMUS, funded by the European Research Council (Peopling Europe: How data make a people; [www.arithmus.eu](http://www.arithmus.eu)). It involved six researchers: Evelyn

- Ruppert (PI), Baki Cakici, Francisca Grommé, Stephan Scheel, Ville Takala and Funda Ustek-Spilda.
11. Other possible flags include: break in time series; not available; confidential; definition differs; estimated; forecast; see metadata; not significant; provisional; revised; Eurostat estimate; low reliability; not applicable.
  12. Living quarters were further defined in the technical specifications for another topic: 'Type of living quarters'.
  13. The proposed wording stated 'metadata shall report the number of all primary homeless persons and the number of all secondary homeless persons as well as provide a description of the methodology and data sources used to produce the data on homeless persons.'
  14. The proposed wording stated: "'data source' means the set of data records for statistical units and/or events related to statistical units which forms a basis for the production of census data about one or more specified topics for a specified target population.'
  15. The draft wording was later amended to: 'data source means the set of data records for statistical units and/or events related to statistical units which *directly* forms a basis for the production of census data about one or more specified topics for a specified target population.'
  16. See <https://uddannelsesstatistik.dk/>.
  17. Danish educational institutions have been required to publish performance data on their websites since the passing of the 'Statute on transparency and openness in education' in 2003 (Danish Ministry of Education, 2018: §1). The Ministry of Education made these data centrally available in a 'data bank' in 2009. Compared to the user friendly Warehouse, statisticians describe the data bank as a 'technical interface'.
  18. For that reason, STIL statisticians hesitated to use 'data warehouse', the name of the data portal, as a technical term as its vision of repositories of automatically integrated data did not fit their reality of manually cleaning and readying data for the Warehouse.
  19. While procedures for data cleaning are not described on the data portal, many statisticians do save some of the data cleaning codes (e.g. for checking gender through the civil registration number) in the statistical software suite SAS so it can be recycled when the next batch of data (typically provided once a year) is to be projected. The calculation of quality indicators, after data has been cleaned, is documented here: <https://uddannelsesstatistik.dk/Sider/Indhold/Beskrivelse%20af%20n%C3%B8gletallene%20i%20LIS.pdf>.
  20. Not to be confused with the SAS function of imputing values.
  21. SAS (previously 'Statistical Analysis System') is a software suite for statistical analysis, used by the Ministry of Education.
  22. Bar charts relate and compare two schools' performance for the past three years with a school's performance in relation to the national and municipal averages. As such, the dashboard projects multiple data relations: as comparisons between selected schools, as benchmarks of individual schools in relation to national or local government averages, and as an individual school's performance in relation to itself over time.

## ORCID iD

Helene Ratner  <https://orcid.org/0000-0002-0842-4049>

## References

- Attard J, Orlandi F, Scerni S, et al. (2015) A systematic review of open government data initiatives. *Government Information Quarterly* 32(4): 399–418.
- Barns S (2016) Mine your data: Open data, digital strategies and entrepreneurial governance by code. *Urban Geography* 37(4): 554–571.
- Bates J (2014) The strategic importance of information policy for the contemporary neoliberal state: The case of Open Government Data in the United Kingdom. *Government Information Quarterly* 31(3): 388–395.
- Bateson G (1972) *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chicago: The University of Chicago Press.
- Birchall C (2016) Shareveillance: Subjectivity between open and closed data. *Big Data & Society* 3(2): 1–12.
- Birchall C, Hansen HK, Christensen LT, et al. (2015) 'Data.gov-in-a-box': Delimiting transparency. *European Journal of Social Theory* 18(2): 185–202.
- Böschen S, Kastenhofer K, Rust I, et al. (2010) Scientific nonknowledge and its political dynamics: The cases of agri-biotechnology and mobile phoning. *Science, Technology & Human Values* 35(6): 783–811.
- Collins HM (2001) Tacit knowledge, trust and the Q of sapphire. *Social Studies of Science* 31(1): 71–85.
- Curtis B (2001) *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875*. Toronto: University of Toronto Press.
- Danish Ministry of Education (2016) Vejledning i brug af åbenhedsinitiativet. Available at: <https://uddannelsesstatistik.dk/grundskolen/Sider/Vejledning.aspx> (accessed 29 November 2018).
- Danish Ministry of Education (2018) *Bekendtgørelse af lov om gennemsigtighed og åbenhed i uddannelserne m.v.* Pub. L. No. LBK nr 810.
- Dourish P (2016) Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3(2): 1–11.
- Edwards PN, Mayernik MS, Batcheller AL, et al. (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5): 667–690.
- Eurostat (2011) *EU Legislation on the 2011 Population and Housing Censuses. Explanatory Notes*. Luxembourg: Eurostat.
- Foucault M (2009) In: Senellart M, Ewald F, Fontana A, et al. (eds.) *Security, Territory, Population: Lectures at the Collège de France 1977–1978* (Trans. Burchell G). UK: Palgrave Macmillan.
- Gad C and Jensen CB (2014) The promises of practice. *The Sociological Review* 62(4): 698–718.
- Göpfert M (2013) Bureaucratic aesthetics: Report writing in the Nigérien gendarmerie. *American Ethnologist* 40(2): 324–334.
- Guyer JI (2014) Percentages and perchance: Archaic forms in the twenty-first century. *Distinktion: Journal of Social Theory* 15(2): 155–173.

- Helgesson C-F (2010) From dirty data to credible scientific evidence: Some practices used to clean data in large randomised clinical trials. In: Will, C and Moreira, T (eds) *Medical Proofs, Social Experiments: Clinical Trials in Context*. UK: Ashgate, pp.49–64.
- Hine C (2006) Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36(2): 269–298.
- Larkin B (2013) The politics and poetics of infrastructure. *Annual Review of Anthropology* 42(1): 327–343.
- Latour B (2017) Gaia or knowledge without spheres. In: Schaffer S, Tresch J and Gagliardi P (eds) *Aesthetics of Universal Knowledge*. pp.169–191. UK: Palgrave Macmillan.
- Latour B and Woolgar S (1979) *Laboratory Life: The Construction of Scientific Facts*. London: Sage.
- Law J (1994) *Organising Modernity: Social Order and Social Theory*. Oxford: Blackwell.
- Law J (2004) *After Method: Mess in Social Science Research*. London: Routledge.
- Leahy E (2008) Overseeing research practice: The case of data editing. *Science, Technology, & Human Values* 33(5): 605–630.
- Leonelli S (2016) *Data-centric Biology*. Chicago: The University of Chicago Press.
- Leonelli S, Rappert B and Davies G (2017) Data shadows knowledge, openness, and absence. *Science, Technology & Human Values* 42(2): 191–202.
- Marquardt N (2016) Counting the countless: Statistics on homelessness and the spatial ontology of political numbers. *Environment and Planning D: Society and Space* 34(2): 301–318.
- Maurer B and Martin SJ (2012) Accidents of equity and the aesthetics of Chinese offshore incorporation. *American Ethnologist* 39(3): 527–544.
- Maurer B (2005) *Mutual Life, Limited: Islamic Banking, Alternative Currencies, Lateral Reason*. Princeton, NJ: Princeton University Press.
- Mol A (2002) *The Body Multiple: Ontology in Medical Practice*. Durham, NC: Duke University Press.
- Muniesa F (2011) A flank movement in the understanding of valuation. *The Sociological Review* 59: 24–38.
- Munster A (2013) *An Aesthesia of Networks: Conjunctive Experience in Art and Technology*. Cambridge, MA: MIT Press.
- Plantin J-C (2019) Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values* 44(1): 52–73.
- Pomerantz J (ed.) (2015) *Metadata*. Cambridge, MA: MIT Press.
- Ratner H and Gad C (2018) Data warehousing organization: Infrastructural experimentation with educational governance. *Organization* 00(0): 1–16. <https://doi.org/10.1177/1350508418808233>.
- Reay M (2007) Academic knowledge and expert authority in American economics. *Sociological Perspectives* 50(1): 101–129.
- Riles A (1998) Infinity within the Brackets. *American Ethnologist* 25(3): 378–398.
- Riles A (2000) *The Network Inside Out*. Michigan: University of Michigan Press.
- Riles A (2006) *Documents: Artifacts of Modern Knowledge*. Ann Arbor: University of Michigan Press.
- Riles A (2010) Collateral expertise: Legal knowledge in the global financial markets. *Current Anthropology* 51(6): 795–818.
- Ryle G (1949) *The Concept of Mind*. Chicago: The University Of Chicago Press.
- Schaffer S (2017) Introduction. In: Schaffer S, Tresch J and Gagliardi P (eds) *Aesthetics of Universal Knowledge*. UK: Palgrave Macmillan, pp. 11–28.
- Schatzki TR, Knorr-Cetina KD and von Savigny E (eds) (2001) *The Practice Turn in Contemporary Theory*. London: Routledge.
- Scheel S and Ustek-Spilda F (2019) On the politics of expertise and ignorance in the field of migration management. *Environment and Planning D: Society and Space* 00(0): 1–19. <https://doi.org/10.1177/0263775819843677>.
- Schrock A and Shaffer G (2017) Data ideologies of an interested public: A study of grassroots open government data intermediaries. *Big Data & Society* 4(1): 1–10.
- Strathern M (2000) The tyranny of transparency. *British Educational Research Journal* 26: 309–321.
- Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd ed. Cambridge: Cambridge University Press.
- van de Port M (2016) Baroque as tension: Introducing turmoil and turbulence in the academic text. In: Law J and Ruppert E (eds) *Modes of Knowing: Resources from the Baroque*. UK: Mattering Press, pp. 165–198.
- Walford AC (2013) *Transforming data: An ethnography of scientific data from the Brazilian Amazon*. Doctoral thesis, ITU of Copenhagen, Copenhagen.
- Zimmerman A (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7(1–2): 5–16.