Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "What do we mean when we talk about data?." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 21–40. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-2>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 14:44 UTC.

Copyright © Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy 2022. Released under a CC BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/3.0/>). You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

What do we mean when we talk about data?

‘The people who get to impose their metaphors on the culture get to define what we consider to be true.’¹

How we speak about complicated things matters. The words we choose (or are encouraged to choose) shape not only how we interact with others but also how we ourselves perceive the things we speak of. Language shapes our emotional responses, how we assimilate and organize new information. This is true in particular in terms of our relationship to digital technology, which changes fast, may be hard for non-experts to understand the function of, is often invisible and is central to our lives. Using metaphors to embed technologies into our lives is a process as old as technology development, and we can certainly find recognition of the importance of this process in work such as David Edge’s 1974 study ‘Technological Metaphor and Social Control’.² More recently (in the context of the internet) Sally Wyatt commented:

Metaphors not only help us to think about the future; they are a resource deployed by a variety of actors to shape the future ... Metaphors can mediate between structure and agency, but it is actors who choose to repeat old metaphors and introduce new ones. Thus, it is important to continue to monitor the metaphors at work to understand exactly what work it is that they are doing.³

This chapter will look into the implications of this phenomenon as they relate to our relationship with both the idea and reality of data in our lives: a complex, interesting and indeed fraught set of interactions driven only in part by the manner in which the fantasy of data seems to have established itself as a contemporary fetish object, touted as ‘the new oil’ or even (as one particularly astonishing 2015 advertising campaign launched by the investment company Winton Capital proposed) as the ‘secret to living happily ever after’.⁴ It is also a danger zone where ephemeral details about our lives and selves may be lost, stolen, sold and ultimately used to exploit us. So fundamental is our

dysfunctional relationship with the concept of data that even though talk of it seems to be everywhere, there is not even consensus about whether the word should be used as a singular (as in 'data is') or plural (the grammatically correct, but less commonly used form, 'data are').

The language used in the European General Data Protection Regulation (GDPR) provides a foundation for beginning to understand what data might be from the point of view of an individual. The GDPR defines three different roles related to data regulation: the data subject (a living person who is either identified or identifiable in data), the data controller (a person, company or other body which decides the purposes and methods of processing personal data) and the data processor (a person, company or other body which processes personal data on behalf of a data controller).⁵ In Western societies, we are generally acculturated to believe that we somehow own what we create: if you bake a cake, you may sell or eat or give it away. It would be generally considered theft for anyone else to do these things to your cake without your permission. If you do harm, you should take responsibility and redress the ill effects of your actions. But this is not the case as pertains to your data: you may feel that you own it, but according to the regulations (and indeed according to many questionable legal instruments you may have signed) you are a data subject, not a controller, or indeed an owner.

This is not merely a bit of legal language intended to control existing relationships between individuals, governments, institutions and companies, however, as what our data 'say' about us is more than a form of personal expression. Marketers on some level realize this, even as they try to ensure we see the positive sides of technology and overlook the threats. As the representative of a company that advocates the move from the 'Internet of Things' (IoT) to the 'Internet of Me' (IoMe) explains: 'Now imagine tech working in your body at the biological level. Your body could express itself on its own, without you having to be in charge, to deliver more happiness, better health, whatever you truly need and want.'⁶ Seen this way, your personal data seems to transcend even the idea of free speech to take the place of free will. But even far short of this techno-utopian vision we can find challenges in the idea of personal data being personal in terms of ownership, rather than representation. Whether or not selling our personal data would be akin to selling our autonomy (as Morozov, argued already in 2014);⁷ the implications of how we view our agency over the record of our actions, decisions and indeed our biology are wide ranging, because at the end of the day, privacy is a public good. This has been established as a legal precedent,⁸ but in the years since that publication plenty of illustrations of precisely how this concept operates in practice have emerged, from revealing

the location of US military bases via data from the Strava fitness app⁹ to the use of openly shared genomic data, gathered with the intention of learning about one's heritage, to solve cold murder cases.¹⁰

The semantic gymnastics according to which we co-create data with a variety of places, devices and platforms will be discussed below. But at the heart of the difficulty we have with our personal data is the fact that very often, it is a by-product of other activities. We undertake a Google search with the intention of answering a question or finding information – we do not intend to produce data and may not even be aware that this is precisely what we are doing. We order from Amazon with the intention of getting stuff, we register with Tinder with the intention of finding a partner, we carry our phone in our pocket so that we may be contacted wherever we are. We create data via each of these mundane activities, just as we create carbon dioxide by breathing, but (also like carbon dioxide) we can neither see nor ourselves reuse the data we create, so we tend to ignore them and any implications they may have for our lives. It is this status as what Shoshona Zuboff calls 'behavioural surplus'¹¹ that makes them so easy for us to ignore, and for companies to exploit. Because we cannot 'see' data, we struggle to understand their size, their presence, their implications. An image search for the word 'data' demonstrates how narrowly our imagination is constrained: numbers and swirls dominate, as do various shades of blue and black.

The demonstrative nature of this example is upheld by research into images used to represent data in two prominent on-line newspapers (*The New York Times* and *The Washington Post*).¹² Images classified there found a reliance on similar forms of abstraction (large-scale numbers, abstract data visualizations, devices, screens, servers, etc.). They also noted the general homogeneity of images beyond these 'curated' media, and also the emphasis on the colour blue. Visually, we have not moved much beyond the compelling image established in the Matrix film series, with a moody Keanu Reeves standing enigmatically in front of scrolling green characters on a black screen, so reminiscent of early monochrome computer monitors

Our vocabulary may not be much more advanced than our visual language for giving shape to data, and indeed the extent to which it is richer may even point towards an even greater weakness. Indeed, the linguistic metaphors we use to refer to data have been the object of a significant body of research, resulting in an interesting overall taxonomy of terms we use to express this seemingly ephemeral, mobile, invisible, but high impact product. The most significant categories of image are those of fluidity (data streams, deluge) and of natural

resources (data as oil, data mining)¹³ which tend to be divorced from the human-constructedness of data,¹⁴ its personal, embodied and spatial elements – it is (in particular within the ‘quantified self’ community) a mirror, a practice and a body.¹⁵ Awati¹⁶ adds the categories of surveillance, industry, food and spatial images, while Davenport¹⁷ pinpoints the idea of data as a by-product, of ‘data exhaust.’ It is interesting how far these metaphors for data are from those Wyatt identified as applying to the internet in 1998: revolution, evolution, salvation, progress, universalism and the ‘American dream’.¹⁸

The language we are encouraged to use in the process of creating this valuable residue stresses the harmlessness of the transactions: we are asked to ‘accept cookies’ and adjust our ‘sharing’ settings as a means to imply positive and safe interactions, which are in fact fraught with complexities behind their innocuous linguistic wrappings. These metaphors, at every level, ‘are thus not only descriptive; they may provide clues to the design intentions of those who use them and, as such, they may help to shape the cognitive framework within which such actors operate. ... they can be used to help the imaginary become real or true.’¹⁹ In other words, metaphors are not merely effective because of their representational capacity (although this helps) but because the users forget that they are metaphors at all, sublimating them instead as ‘embodied cognition’²⁰ or ‘conceptual memory theory’²¹ to become a factor that constructs as well as defines the world. Unfortunately, however, it seems inevitable that the level and nature of this ‘harmlessness’ are defined by – and the profits derived from any inherent risks or costs they may hide being accepted, largely accruing to – corporations selling products and governments engaging in surveillance. These topics will be taken up in more detail in Chapter 6 and 7.

So how should we view our data? Are platforms and services like those provided by Google and Facebook more like utilities we require for daily life, hospitals where we are vulnerable and exposed, cornerstones of democratic exchange? More provocatively, in terms of constitutional protections from warrantless search, is your car like your home, or is your cell phone like your body? Controlling how data are used and how they frame our lives may mean that we need to better control how we speak about them. To many, the resistance to data-driven ‘surveillance capitalism’²² represents a backward-looking denial of the benefits of technology, but one can also view it as a linguistic project to update our metaphors for both our technologies and our identities as they are shaped by them.

No matter how data subjects and producers may imagine and speak of their data, there is no doubt that multibillion dollar companies like Acxiom very much

recognize the value of all of these assets, as does Facebook, Google and a host of other companies, governments and other actors. A series of 2018 advertisements for the credit reporting company Experian took an almost cynical approach to what they deemed the 'data self'. 'Your Data Self is the version of you that companies see when you apply for things like credit cards, loans and mortgages.' It said 'You two should get acquainted.'²³ What data are and mean to this company, however, may be very different from how you might see things – indeed, the activities undertaken by companies in the name of product innovation have at times given rise to a host of unexpected consequences (from clickbait content farms in Macedonia to social unrest in Myanmar). Had these companies been held to the ethical codes of the social science researchers they are acting as, or indeed to those present in other industries, like pharmaceuticals, where products sold might be detrimental to health, the landscape of data privacy and reuse might be very different today.

Needless to say, the most well-documented overstepping of ethical lines by a company seeking to obtain data from and about its users has been the 2014 Facebook emotional contagion testing.²⁴ In this infamous study, the company monitored the effect of subtle tweaks on their algorithm that altered the emotional tone of content viewed by ca. 700,000 of their users. As a result, at the end of a week of this manipulation of their newsfeeds, those same users seemed themselves to post more positively or negatively according to what they had been shown. While it may seem that the question of what responsibility a company has to the users of its service might seem very far from the issues of language at the heart of this chapter, indeed, metaphors are never far from sight. As Hwang and Levy describe the core issues of the case:

The debate around the emotional contagion experiment, for instance, is fundamentally a debate about what metaphor should guide our thinking about what the Facebook News Feed actually is. As Jeff Hancock, a co-author of the paper based on the experiment, has recognized, '[T]here's no stable metaphor that people hold for what the News Feed is.' Proving the point, commentators have deployed a range of conflicting metaphors to argue about whether the experiment crossed the line: the experimental manipulation has been compared to a field study, to an A/B test, to books and television programs, and even to a dime left suggestively in a public phone booth. The controlling metaphor defines the moral burden of the project: If the contagion experiment is like any other routine A/B test, then there is no foul. If the contagion experiment is more, say, like a field study, it implies a greater ethical onus on the researchers' conduct.²⁵

Ethnographer Mary Grey echoes this focus on the language we use to describe what Facebook was doing: ‘If researchers or systems designers are “just” testing a product on end users (aka humans) and another group has access to all that luscious data, whose ethics apply? When does “testing” end and “real research” begin in the complicated world of “The Internet?”’²⁶ The implications of the publication of the research results of Facebook’s experiment clearly resonated for the editorial board of the journal presenting the work as well, leading them to add an annex to the paper defending their decision to present the research, but maintaining the caveat that ‘it is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.’²⁷ The verbal gymnastics around whether or not Facebook is a ‘media company’ or just a ‘tech company’ in the years since this case became public seem to imply that they feel the metaphors are still in their favour. And, indeed, Facebook has only in 2020 set up an independent oversight board²⁸ for its activities – or so it says. Although the Board does include a number of Facebook’s critics and has a statutory independence from the company, its remit is very narrowly proscribed indeed, to cases in which users believe their content has been removed in error or unfairly.

Another place in which the linguistic slipperiness of the business models based on data can be seen is in the ubiquitous terms of service prospective users are asked to agree to. Terms of Service statements very often govern what claims a service provider may make to track, use and resell evidence regarding an individual’s behaviour on their platform and beyond. They may purport to being a tool for establishing the rights and responsibilities of both user and service provider; yet, they fail spectacularly to fulfil this function. Social or psychological contracts (based on unwritten, normed expectations, rather than formal, legal instruments) negotiated around the collection and application of technologies based upon big data often demonstrate a core attribute of misunderstanding or miscommunication. In their book *Reengineering Humanity*, Selinger and Frischman bring their perspectives spanning business, law and philosophy together to explore the validity and purpose of these contracts. As such, they raise a large number of concerns about the contents and prevalence of such instruments, deeming them ‘oppressive’ and based – by design – on the ‘irrational’ proposition that users might actually read them. From their length to the heuristics of their design to the nature and object of the consent they extract, these so-called contracts represent not only a dramatic change in the number of legal instruments we are expected to engage with in the course of an adult life

(from a few in the course of decades to that same number in weeks or days or, potentially, minutes) to the ultimate end of a commodification of users through ‘pseudo-relationships.’²⁹ This brings us back to the need for a stable functional or descriptive definition of data to enable the management of activities with regards to this thing we can neither see nor touch, but which can perhaps be useful or harmful. But finding this clear definition to help us navigate through the obscurity of a data-driven world is also anything but simple.

If we turn for a concrete definition to the vast corpus of material produced within the field of Science and Technology Studies (STS), we find the contentious nature of the space occupied by the word data to be seemingly confirmed. In fact, even between two of the most prominent voices in this chorus, we learn that data has a ‘pre-analytical, pre-factual status’ that ‘resists analysis,’³⁰ and yet ‘... may be facts, sources of evidence, or principles of an argument that are used to assert truth or reality.’³¹ This direct contradiction between the (possibly) factual and the pre-factual is par for the course within this particular area of interest and comment, however, as we can easily find as well that data is/are ‘rhetorical,’³² ‘fiction ... illusion,’³³ ‘a sort of actor’³⁴ and/or ‘performative.’³⁵ It/they ‘has no truth,’³⁶ ‘resists analysis’ and ‘cannot be questioned.’³⁷ This polysemy not only makes it hard to understand the place of data in sensemaking processes (the topic of the next chapter) but also seems to lend to data a certain quality of agency, as in Garvey’s statement that data ‘will out.’³⁸ In particular, and equally unusefully, many of the most concrete definitions of data focus less on what they might be and more on what data lack, such as meaning or value, it being unorganized, unprocessed, pre-epistemic, etc.³⁹

Of the many definitions for the word data that exist, one of the most useful is surely that posed by Sabina Leonelli. For her, data is a ‘relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved.’⁴⁰ For Leonelli, data is not a fixed thing in itself, but a ‘portable object,’ and a means for communication that relies heavily on the observer to attribute meaning to it. With this approach, she manages to demonstrate how data itself is like a language, the components of which can act as evidence for different phenomena from different perspectives at different times, something we cannot view properly when it is divorced from its starting and endpoints, its sources and functions. This points towards the issue of data and its context, which the discussion returns to below.

One of the further factors complicating our search for firm ground with regards to speaking about data is the slippery relationship between data and some

of its semantic near neighbours, in particular, ‘information’ and ‘knowledge’ (the three of which, together with wisdom, make up the ‘DIKW pyramid’, ‘one of the fundamental, widely recognized and “taken-for-granted” models in the information and knowledge literatures.’)⁴¹ Like many such broadly applied models, there is a lot of variation in the application of DIKW and in particular in how the terms that comprise it are seen to be related, except for the fact that they are seen to stand in a hierarchical relationship to each other: that is, that each component of the model acts as a sort of raw material for transforming into the next. This fundamental assumption seems to have been a part of the model from the beginning, appearing already in the paper by Ackoff so many cite as having introduced it, in which he states that each of the higher types in the hierarchy ‘includes the categories that fall below it.’⁴² The hierarchical nature of the model is problematic, however, seen from the contemporary perspective of how information objects actually circulate. For example, a philosophical treatise may be said to contain wisdom, or at least a record of knowledge, but when it is digitally imaged, transcribed and federated into a digital collection, it then becomes data again. Interesting in this context is the NASA EOSDIS standard for the description of enhancements to data. Like the discussion of DIKW, the framework proposes a number of additions and transformations that may be added to an original dataset, each of which may cause it to then be considered to have reached another level of preparation and refinement. And yet, when that same, now highly processed data is passed to a new use case or context, it reverts in that context back to having a preparation status of level 0 in the framework.⁴³ This seems also to have been recognized by Carlisle, who states that ‘given the idea that information consists of augmented data, then information should possess all of the characteristics of its contributing data, plus additional characteristics. However, since information fails to include the characteristics of existing in time and space, stating that information emerges as a summation of all contributing data entities seems to be a misnomer.’⁴⁴

Whether or not the hierarchical nature of relationships between data, information, knowledge and wisdom is considered to be linear or fluid, the fact of there being any hierarchical or even clear distinctions between them seems refuted by the very linguistic practices of those who write about them (the present authors not necessarily excluded!), creating not only tautologies and chiasms in usage (as, for example, in the statement: ‘[A researcher states that] in 2006, the world created 161 exabytes of **data** and forecasts that between 2006 and 2010, the **information** added annually to the digital universe will increase more than six fold].’⁴⁵ Other definition practices are additive, stating for example: ‘Data is

the most basic level; Information adds context; Knowledge adds how to use it; Wisdom adds when and why to use it.⁴⁶ Or, more succinctly, the proposal that the four forms equate to 'knownothing', 'knowwhat', 'knowhow' and 'knowwhy', respectively.⁴⁷ Indeed, perhaps the most defining characteristic underlying how these key terms will be defined is the disciplinary context any particular author happens to bring to the discussion, be that data science, information science or knowledge management (there is, to our knowledge, no discipline that names itself after or approaches the challenge of understanding knowledge creation from the perspective of wisdom).

The word data seems to obscure itself in any number of ways: through its grammatical form, its semantic complements, its malleability. Even its Latin roots have been called into question as a basis for speaking of the phenomena it proposes to represent. Johanna Drucker has therefore proposed that we should be speaking not of data (from the word 'to give') but of *capta* (from the word 'to take'):

Differences in the etymological roots of the terms data and *capta* make the distinction between constructivist and realist approaches clear. *Capta* is 'taken' actively while *data* is assumed to be a 'given' able to be recorded and observed. From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact.⁴⁸

No matter what it may be or how one might tacitly or explicitly define it, data are never neutral, never natural phenomena, never 'given'. It is as when Suzanne Briet writes of the difference between an object and its documentation⁴⁹ or Greg Crane of the 'digital incunabula',⁵⁰ even when there are potentially strong links and relationships between natural phenomena (such as temperature readings or pollutant levels in the air) and their documentations as data, this close relationship does not mean that we can or should ignore their status as 'always already' epistemically marked by the humans who created the measurement protocols, designed the sensors and determined their placement and use (again, an issue we revisit in the next chapter). Instead, if we are to take the position that data function primarily as our personal pre-epistemic 'stuff' (regardless of however processed they may have been by someone else for a different purpose), then we must at least use a term for them that reminds us of their constructedness, and their status within the context into which we import them.

One would hope that where theory fails us, the practices of experts might lead. For this reason, one of the aspects of the cultural underpinnings of software

production the KPLEX project interviews probed was how a group of computer scientists used the word data, how they defined it, and what their own habits of thought and speech might tell us about how to improve those of software users. After being given the chance to present their research in their own words and on their own terms, participants were asked how they would define the word data as it pertained to their own work.

The word data was defined as ‘text’ (WP2 INT1), ‘stored information that I can manipulate, search, query, get some statistics about’ (WP2 INT2), ‘anything that I am analysing, or using to train a system’ (WP2 INT4), ‘any material that you have in hand ... like digital material’ (WP2 INT3), ‘everything that I can use to study a certain subject’ (WP2 INT5), information that could be quantified. ... that you would use (WP2 INT7), any piece of information that ... can be recorded in an index, ‘just evidence’ (WP2 INT9), or ‘any piece of information, literally anything, but if you’re looking for a computer science point of view, any structured bit of information is data’ (WP2 INT8). The tendency to view data as something almost transcendently broad was perhaps best summarized in the statement ‘data exists, it does exist, it just exists in and of itself’ (WP2 INT6).

The clear trend running throughout these examples points towards an epistemic cultural bias towards viewing data, whatever it is, as broadly encompassing, and in terms of its function or utility in the research project, rather than a complex set of information objects that come with biases built in to them, and which might merit a certain amount of meta-reflection. Data, in other words, means ‘stuff’. As a side note, it is also interesting to observe that almost all of the respondents used the word ‘data’ in the singular, a habit of thought that seems to underline this apparent bias against constraining the nature, status and role of data in their work.

Perhaps all the more interesting in this context, then, were those individuals for whom the request to define this key term for their work – and there is no question that the term is central, appearing between 50 and 220 times in each sixty- to seventy-five-minute interview – was met with some discomfort or resistance. Two people began their responses with the very honest disclaimers that they either didn’t ‘have a perfect text definition of data’ (WP2 INT8) or that it was ‘not clearly defined’ for them (WP2 INT13) while another claimed not to use the term to describe any aspect of their work (WP2 INT10). Most striking perhaps in this respect was the following response, in which a computer scientist, who described their general research space as comprising data analytics, knowledge extraction from texts, natural language processing and information retrieval, replied:

I don't think like my opinion is that important. I try to explain what I know. I think of data as ... I just mean that I don't have maybe enough knowledge in the area. I know some things, but there are definitely like way smarter people but I try to give you what I have.

(WP2 INT3)

It is perhaps a humanistic bias to expect expertise to include a precision in language around key terminology, and a few respondents did offer alternative words that they would use instead of data in certain situations, such as 'content' (WP2 INT9), or a 'corpus' (WP2 INT13). The gap in the confidence of these researchers between their ability to work with certain kinds of material and talk about it was, however, striking. As a side note, this tendency stands in interesting contrast to the apparent humanistic resistance of the term 'data', which a fascinating Twitter thread instigated by Miriam Posner in 2018 showed to be perceived as narrow, derived, impoverished, simple or monophonic, perhaps even sinister or indicative that a researcher 'doesn't value [a source] or respect its integrity'.⁵¹

As a follow-up to the prompt to suggest their own definitions, the participants were presented with some of the STS definitions of data discussed above. The participant reactions were interestingly passionate. In many cases the definitions were deemed potentially or partially true (WP2 INT3) or interesting (WP2 INT7), but many more participants found them 'negative' (WP2 INT3), wrong (WP2 INT2), 'just nuts' (WP2 INT6) only applicable in certain contexts and under certain assumptions (WP2 INT8), contradictory to commonly held understandings (WP2 INT7), or at least very flawed, likely to be based on a 'lack of the understanding of data' (WP2 INT2) or 'anthropomorphising'. Quite a few responses deemed the entire discussion to be too 'philosophical' (WP2 INT4, INT5, INT6), and several participants did not want to engage the question, characterizing their 'opinion' as 'not important' (WP2 INT3) or 'just not as magical' (WP2 INT6) as the STS perspectives seemed to them. The researchers often struggled very openly to come to grips with the STS definitions, evidencing a significant gap between STS and computer science, as can be seen particularly well in this excerpt:

Yeah, I'm surprised by the (definition that data are] false, data is false. I don't, I think I need more context in that one. ... I take data from, like, another, a completely different system and I claim that to be what it's not, that could be false. But it's not the data that is false, it's what you claim about the data that is false. Data has no truth? I don't think that's true. Data, all data, has, they are truth. Even if it's false or not, for example.

(WP2 INT5)

The cultural practices of data usage as both an object of and input to software development are deeply engrained, in particular as pertains to the deeply embedded, contextual understanding of the term 'data'. Data as research object appear(s) as overdetermined and highly volatile within the communities that use it most. As such, these expert user practices are of little use in the development of language-based tools to inculcate more informed equivalents among the general public of data subjects and producers of data by-products: are the data an input or an output? Human or only machine readable? Are they being used transparently, and for the purpose they were gathered? Mary Grey reminds us that how we weight the ethical implications of our activities is highly dependent on how we see our objects of study, whether we believe ourselves to be working with strings of numbers and letters or whether we can see the people and the social interactions behind them.⁵²

There is one further thing that these interviews do throw into strong relief, however, an issue that stems directly from how data are viewed and their relationship to the speech acts that surround them, specifically the question of **context**. In order to make and share meaning around a term with no truly fixed referent, one must ensure that the word is instead anchored temporarily for a given purpose, and that there is strong communal agreement around these processes and purposes.

The preservation, disruption or alteration of context both is and isn't a linguistic process: perhaps it would be better to characterize it as a communicative one, and certainly in this respect it is a challenge we face within in our daily lives as consumers and circulators of information. What aspects of one's adventure of a Saturday night might one share with one's best friend, as opposed to one's socially conservative grandmother? What language differences might there be between these two accounts? How might you respond differently to an account of a scientific experiment represented in a research paper, or an interview with a politician, or a post on Facebook from an unknown source (the phenomenon known as 'context collapse'⁵³)? Context gates the future actionability and credibility of a data signal received: as Christine Borgman described the relationship, '[data] exist in a context, taking on meaning from that context and from the perspective of the beholder. The degree to which those contexts and meanings can be represented influences the transferability of the data.'⁵⁴ The many layers to the context that enable this transferability are often sublimated and indeed disregarded in its processing, however, regardless of which of the definitions or models one chooses to ascribe to in coming to terms with the concept of data.

The power of data is strongly, and perhaps ironically, linked to its context. In general, the variety of definitions we looked at above imply that one of the key characteristics of data is that it has been divorced from its originating context and the complexity that would come with it: it is this lack of context that would render it pre-epistemic, pre-factual or resistant to analysis. It is also the characteristic of data that the NASA model explicitly recognizes: when the context of data use changes, so also does its complexity, regardless of what transformations or contextual information it may have previously been enhanced with.⁵⁵

The contextual layer for any given piece or collection of data, were it captured in full, would be complex indeed. If we start with the point at which Drucker identified the rift between data and capta,⁵⁶ we can find already a fundamental layer of context, what a historian would probably recognize as resembling the **provenance** of a source, including how a certain intention could create data utterly unrelated to the context later applied to it. Was a Google search about suicide methods motivated by a personal crisis or a writing project? Was a particular location visited in order to walk a dog, buy a doughnut or escape a sense of being followed? Even the most simple scientific data is marked at some level as human produced capta (who designed or manufactured the sensor? Who determined its precise position?), but human behaviour is always complex, and no matter how much data about it is captured, there is always more that is missed. After all, as the refrain reminds us, all models lie (even if some models happen also to be useful).

How and what context is maintained alongside what data have been gathered is complicated by the additional question of what new context has been added, in terms of description, structuring or even of aggregation with other data. This layer denotes a transition from context as a question of provenance to one of **curation** and **information modelling**. Although this former term has its roots in the field of museology, some sort of similar process of filtering, aggregation, refinement, arrangement or enhancement takes place in almost every processing of knowledge. Even the data produced by the Large Hadron Collider undergoes a (human-driven, hard-coded) process of filtering and selection before it is made available to researchers.⁵⁷ It is almost inevitable that the capta we collect, be they observations via a telescope or collider, a walk in the park or perusal of manuscript images, will need to be winnowed, aggregated like-with-like and organized in such a way as to represent a reasonable baseline knowledge organization – even more so in our current scientific and human condition of potential information overload. This is not a process to be taken lightly, however. In the age of fake news and alternate facts, the process of curating a dataset

should itself be documented and fully proofed for the biases it may enhance, obscure or introduce into the new interpretations this process facilitates. A survey of an interdisciplinary cohort of researchers undertaken by the KPLEX project in order to understand the range of attitudes towards data-driven research and the management practices that underpin it shows a significant lack of consensus regarding how to responsibly balance the limits one's data capture processes and the processability of data for a specific purpose and again it is the terminology used to describe differing processes that tell us most about the risks and gains inherent in them. In particular, the terms 'data cleaning' and 'data manipulation' emerged as descriptors for the same process with very different values implied by them. The latter term was seen as an arbitrary act to make data fit with other data, which seems to have a negative connotation as this email feedback makes clear: 'I doubt it very much that any self-respecting quantitative researcher would admit to manipulating their data! This is what your questions on p. 2 imply'.⁵⁸ Data cleaning is seemingly understood as unscientific when handling data, and less so as a necessary step within the processing of data that assures the accuracy, completeness, consistency (and uniformity) of a dataset. In the social and natural sciences on the other hand, more than half of the survey participants perform some sort of data cleaning.⁵⁹ This difference in perspectives is reminiscent of what Leonelli describes at the 'journeys' taken by data, passing through research contexts and acquiring and losing aspects of its scaffolding as it is decontextualized, recontextualized and reused.⁶⁰

One of the imperatives often driving the processes of data cleaning (or manipulation) is the need to organize data in order to make them processable. In its simplest format, this may involve populating a table that aligns simple answers to the same question from multiple respondents, but in more complex operations, structured, limited sets of descriptors applying standardized sets of terms known as metadata may be used. This is another processing step where the KPLEX survey exposed significant disagreement between fields, in particular as pertained to the standardized vocabulary for describing emotions (EmotionML) which, in spite of being accepted by the regulatory body W3C, was not widely used in research.⁶¹ While standardization does lead to comparability across datasets and ease of processing, the question remains how much context must be removed, changed or obscured by the strong structuring of a standards-driven information architecture. Metadata can become a determining factor in whether or not an item is returned as a positive or relevant item when the structured data is processed, and the 'raw' (or 'native') data shaped by the database or metadata structure and contexts applied to it.

Although the effects of explicitly applied knowledge organization frameworks will be a focus for later chapters in this book, it is worth noting even at this early point that such descriptions within data systems can have an impact on messages received (even when they are not intended to be ‘heard’). Re-contextualization of this sort can undo efforts to anonymize data (in order to protect its subjects), leading to the phenomenon known as ‘dark linking’, in which proxy data from one source can essentially unmask subjects in another. The problem of data as an instrument for the transfer of the fundamental building blocks for understanding is of course also a problem on the receiver’s side of the communication paradigm. The value of knowledge creation processes is often measured according to norms quite distanced from those that give immediate shape to them. Here, no matter how we view this process or the biases it may bring, is where we invariably see the hierarchy perhaps envisioned by the DIKW pyramid converge with that so suspiciously viewed in the KPLEX interviewees (described further in Chapter 3): **narration**. Organized, synthesized, filtered and augmented capta alone will not bring meaning to *homo narrans*. Regardless of what our stories are – the tale of a day’s activities or the findings of a research experiment – our communication will be considered flawed unless we bring our knowledge into a temporal framework, into a language of cause and effect, into a story. There is an obvious difference between a chat between friends over dinner and a scientific paper however, in spite of the fact that each of these narratives might reflect roughly the same content and processes: capta, curation and narrative. For that reason, one also needs to consider **performance**. Each of these facets of context, and the many more possible, reflects a specificity not just of what capta or data are relevant, how they might be filtered and what structure might be placed by an individual around them. We perform for an audience, we share our stories and seek to have them validated, thereby valorizing our own epistemic journey, as well as introducing our narratives back into society, so that they themselves might become a source of capta for others. A lack of attentiveness to this social dimension of knowledge creation is what has led to the emergence of the phenomenon known as *context collapse*, which ‘problematizes the individual’s ability to shift between ... selves and come off as authentic or fake.’⁶² Recognizing the performative aspect of science could also contribute to the improvement of public trust in experts, whose narratives of discovery and evidence should be able to contribute to the knowledge of both lay people and fellow experts, but whose performative range may sometimes be too limited to do so.

The possibility obviously exists to maintain all of these contextual layers of data, of making them a part of how we create services, business models and

indeed knowledge out of the traces of human activity in the digital age. After all, what is a library or a museum if not an attempt to maintain the objects able to transfer meaning as well as the dense web of contexts that enrich and entangle them (as opposed to a ‘data warehouse’ where neatly wrapped components wait to be bought and sold)? A number of factors have driven us towards a different paradigm, however – not only technology, but modernity itself, which brought us the ‘socialization of statistics’⁶³ and the rise of ‘mechanical objectivity’ with its desire to replace human judgement with transcendent ‘rules of method, measurement and work discipline’⁶⁴ which could transcend the limits of human sensory perception. This trajectory can only be seen as continuing in our current day, where data, as a sort of platonic ideal (what Presner might call a ‘data sublime’⁶⁵) for a source of unbiased insight, become portrayed as more accurate and insightful than other information sources.⁶⁶

Just as centuries of mathematics laid the groundwork for the turn towards statistics, this major shift in what was considered objective, credible and robust laid the ground for the knowledge creation that we now see, driven by the availability of massive bodies of data and the algorithmic models able to parse them. If data has become a fetish term, then big data is practically an object of worship, though perhaps one no more strictly defined than its more diminutive precursor.

The tendency to view knowledge objects at every level of epistemic preparation and contextualization as products, rather than human-driven processes, leads to a large number of barriers to the systematic exploration of how knowledge is created by professionals, but also by individuals in their personal lives and contexts. Greater precision of language among the professionals regarding the nature of personal information stored and processed – *capta* become *capital*, as it were – could only have a positive effect on the empowerment of individuals in the face of fake news, meaningless consent protocols, filter bubbles and all of the related issues where something perhaps given as data becomes curated into a very different kind of story than we might have expected. The cost of not addressing these gaps is already becoming apparent, as values that have long been a safe haven within democratic systems, like unbiased judgement and free speech, seem somehow to have become suspect. Rehabilitating the idea of the epistemic narrative upon solid ground would be a strong first step in addressing this state of affairs, and of creating both experts and citizens who can speak more clearly about their knowledge environments, processes, outputs and goals. These are the processes we turn to in the following chapters.

Notes

- 1 Lakoff, George and Mark Johnson, *Metaphors We Live By* (Chicago: University of Chicago Press, 2008).
- 2 Edge, David, 'Technological Metaphor and Social Control', *New Literary History*, 6/1 (1974), 135–47.
- 3 Wyatt, Sally, 'Danger! Metaphors at Work in Economics, Geophysics, and the Internet', *Science, Technology, & Human Values*, 29/2 (2004), 242–61.
- 4 Edmond, Jennifer and Georgina Nugent Folan, 'Digitising Cultural Complexity: Representing Rich Cultural Data in a Big Data Environment', in *Ways of Being in a Digital Age – A Review Conference* (Liverpool, United Kingdom, 2017), <https://hal.archives-ouvertes.fr/hal-01629459>.
- 5 'Art. 4 GDPR – Definitions', *GDPR.Eu*, 2018, <https://gdpr.eu/article-4-definitions/>
- 6 Aspell, Ryan, 'The Internet of Me: When the Consumer Becomes the Electronics', *Wired* (8 January 2018), <https://www.wired.com/brandlab/2018/01/internet-consumer-becomes-electronics/>.
- 7 Morozov, Evgeny, 'Selling Your Bulk Online Data Really Means Selling Your Autonomy', *The New Republic* (14 May 2014), <https://newrepublic.com/article/117703/selling-personal-data-big-techs-war-meaning-life>.
- 8 Fairfield, Joshua A.T. and Christopher Engel, 'Privacy as a Public Good', *Duke Law Journal*, 65/3 (2015), 385–457.
- 9 Ruser, Nathan, *Twitter*, 2018, <https://twitter.com/Nrg8000/status/957318498102865920>.
- 10 Murphy, Heather, 'How an Unlikely Family History Website Transformed Cold Case Investigations – *The New York Times*', 2018, <https://www.nytimes.com/2018/10/15/science/gedmatch-genealogy-cold-cases.html>.
- 11 Zuboff, Shoshona, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontiers of Power* (London: Profile, 2019).
- 12 Pentzold, Christian, Cornelia Brantner and Lena Fölsche, 'Imagining Big Data: Illustrations of “Big Data” in US News Articles, 2010–2016', *New Media & Society*, 21/1 (2019), 139–67.
- 13 Lupton, Deborah, 'Personal Data Metaphors and Imagery', *This Sociological Life* (2018), <https://simplysociology.wordpress.com/2018/10/10/personal-data-metaphors-and-imagery/>.
- 14 Levy, Tim and Karen Hwang, '“The Cloud” and Other Dangerous Metaphors', *The Atlantic* (2015), <https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/>.
- 15 Watson, Sara M., 'Data Is the New “___”', *DIS Magazine*, <http://dismagazine.com/discussion/73298/sara-m-watson-metaphors-of-big-data/>.

- 16 Awati, Kailash and Simon Buckingham Shum, 'Big Data Metaphors We Live By', *Towards Data Science*, <https://towardsdatascience.com/big-data-metaphors-we-live-by-98d3fa44ebf8>.
- 17 Davenport, Thomas H., 'Who Owns Your Data Exhaust?', *WSJ Magazine* (2013), <https://www.wsj.com/articles/BL-CIOB-3319>.
- 18 Wyatt, Sally, 'Danger! Metaphors at Work in Economics, Geophysiology, and the Internet', *Science, Technology, & Human Values*, 29/2 (2004), 251.
- 19 Ibid., 244.
- 20 Watson, 'Data Is the New "___"'.
 21 Puschmann, Cornelius and Jean Burgess, 'Big Data, Big Questions| Metaphors of Big Data', *International Journal of Communication*, 8/0 (2014), 20.
- 22 Zuboff, Shoshana, *The Age of Surveillance Capitalism* (London: Profile Books, 2019).
- 23 McGovern, Laura, 'Experian: Meet Your Data Self', *Laura McGovern* (2018), <https://www.lauramcgovern.com/Experian>.
- 24 Kramer, Adam, Jamie Guillory and Jeffrey Hancock, 'Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks', *Proceedings of the National Academy of Sciences of the United States of America*, 111 (2014).
- 25 Levy and Hwang, "The Cloud" and Other Dangerous Metaphors'.
- 26 Gray, Mary L., 'When Science, Customer Service, and Human Subjects Research Collide. Now What?', *Ethnography Matters* (2014), <https://ethnographymatters.net/blog/2014/07/07/when-science-customer-service-and-human-subjects-research-collide-now-what/>.
- 27 Kramer, Adam D. I., Jamie E. Guillory and Jeffrey T. Hancock, 'Experimental Evidence of Massive-scale Emotional Contagion through Social Networks', *Proceedings of the National Academy of Sciences*, 111/24 (2014), 8788–90.
- 28 'Governance | Oversight Board', <https://www.oversightboard.com/governance/>
- 29 Frichmann, Brett and Evan Selinger, *Re-Engineering Humanity* (Cambridge: Cambridge University Press, 2018), 60–6.
- 30 Rosenberg, D., 'Data before the Fact', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 18.
- 31 Borgman, C.L., *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: MIT Press, 2015), 17.
- 32 Rosenberg, 'Data before the Fact', 18.
- 33 Ibid., 3–4.
- 34 Ribes, D., and S. J. Jackson, 'Data Bite Man: The Work of Sustaining a Long Term Study', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 148.
- 35 Raley, R, 'Dataveillance and Counterveillance', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 128.
- 36 Rosenberg, 'Data before the Fact', 18.

- 37 Rosenthal, Jesse, 'Introduction: "Narrative against Data"', *Genre*, 50/1 (2017), 1.
- 38 Garvey, E.G., "'Facts and FACTS:" Abolitionists' Database Innovations', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 90.
- 39 Rowley, Jennifer, 'The Wisdom Hierarchy: Representations of the DIKW Hierarchy', *Journal of Information Science*, 33/2 (2007), 171.
- 40 Leonelli, S., 'What Difference Does Quantity Make? On the Epistemology of Big Data in Biology', *Big Data & Society*, 1/1 (2014), 180.
- 41 Rowley, 'The Wisdom Hierarchy: Representations of the DIKW Hierarchy', 163–4.
- 42 Ackoff, R.L., 'From Data to Wisdom', *Journal of Applied Systems Analysis*, 16 (1989), 3.
- 43 Science Mission Directorate, 'Data Processing Levels', *NASA Science*, <https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products>.
- 44 Carlisle, J.P., 'Continuing the DIKW Hierarchy Conversation', Spring (presented at the MWAIS 2015 Proceedings Midwest, 2015), 5.
- 45 Jifa, G. and Z. Lingling, 'Data, DIKW, Big Data and Data Science', *Procedia Computer Science*, 31 (2014), 815.
- 46 Jifa and Lingling, 'Data, DIKW, Big Data and Data Science', *Procedia*, 814.
- 47 Zeleny, Milan, 'Management Support Systems: Towards Integrated Knowledge Management', *Human Systems Management*, 7/1 (1987), 59–70.
- 48 Drucker, Johanna, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly*, 005/1 (2011).
- 49 Briet, Suzanne, Ronald E. Day, Laurent Martinet and Hermina G.B. Anghelescu, *What Is Documentation? English Translation of the Classic French Text* (Lanham, MD: Scarecrow Press, 2006).
- 50 Crane, Gregory, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel et al., 'Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries', in *Research and Advanced Technology for Digital Libraries*, ed. Costantino Julio Gonzalo, M. Thanos, Felisa Verdejo and Rafael C. Carrasco, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer, 2006), 353–66.
- 51 Posner, Miriam, *Twitter*, 2018, <https://twitter.com/miriamkp/status/1057706465866133504>.
- 52 Gray, 'When Science, Customer Service, and Human Subjects Research Collide. Now What?'
- 53 Marwick, Alice E. and danah boyd, 'I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience', *New Media and Society*, 13.1.2011.
- 54 Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*.
- 55 Science Mission Directorate, 'Data Processing Levels', *NASA Science*.
- 56 Drucker, 'Humanities Approaches to Graphical Display', 17.

- 57 DARIAH-EU, Sally Wyatt: 'What Are We Talking about When We Talk about Data in the Humanities?' (DARIAH 2019), 2019 <https://www.youtube.com/watch?v=fmziX4lHBoc&feature=youtu.be>.
- 58 Edmond, Jennifer, Thomas Stodulka, Elisabeth Huber and Jörg Lehmann, *KPLEX Report on Data, Knowledge Organization and Epistemics* (31 March 2018), 1–118, https://kplexproject.files.wordpress.com/2018/06/k-plex_wp4_report-data-knowledge-organisation-epistemics.pdf.
- 59 Stodulka, Thomas, 'Emotion Work, Ethnography, and Survival Strategies on the Streets of Yogyakarta', *Medical Anthropology*, 34/1 (2015), 84–97.
- 60 Leonelli, 'What Difference Does Quantity Make? On the Epistemology of Big Data in Biology'.
- 61 Stodulka, 'Emotion Work, Ethnography, and Survival Strategies on the Streets of Yogyakarta', *Medical Anthropology*, 34 (2014), 84–97.
- 62 Marwick, Alice E. and danah boyd, 'I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience', *New Media & Society*, 13/1 (2011), 114–33.
- 63 Fienberg, Stephen E., 'A Brief History of Statistics in Three and One-half Chapters: A Review Essay', ed. by Lorraine J. Daston, Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty et al., *Statistical Science*, 7/2 (1992), 208–25.
- 64 Daston, Lorraine J. and Peter Louis Galison, *Objectivity* (New York: Princeton University Press, 2015), 118.
- 65 Presner, Todd, 'The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive', in *Probing the Ethics of Holocaust Culture* (Cambridge: Harvard University Press, 2016), 175–202.
- 66 Lupton, Deborah, 'The Thirteen Ps of Big Data', *This Sociological Life* (2015), <https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/>.