Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "Power through datafication." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 135–157. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-7>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 12:21 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.

Power through datafication

The current deluge of data and the epistemic rupture created by the advent of big data can be termed a ‘data revolution’¹ which can be equated with the industrial revolution beginning in the late eighteenth century. The ripples outward from this rupture have been discussed throughout this book. The discourses around big data are often marked by big promises about progress and prosperity driven by technological innovation, the increase in efficiency of scalable businesses and the reduction of their cost, societally beneficial transformations such as open government, transparency and accountability, as well as scientific advances and paradigm shifts. But the creation of data is a costly endeavour. Because they are ultimately a product of human activity and cannot be found ‘out there’ like a natural resource, one has to recall the whole chain of their production to envision the expenses necessary for their creation: the conceptual work of interpreting the chaos before the eyes of human beings in order to determine what should be identified as ‘datum’ relevant for a certain purpose and therefore be collected; the work of individuals to either gather data manually (like an archivist writing archival descriptions) or design machines to automatically collect data, both according to the conceptual scheme established earlier; the process of actual data collection, their storage, processing, cleaning, refining, feeding into information infrastructures (which have to be created to begin with) and their wrangling and analysis in order to determine their heuristic value; finally, the societal process of discussing and estimating the truth value of those data. These are the bare minimum steps necessary for the production of data, which involves the cost of human labour as well as the provision of collecting devices, computers, algorithms, storage space and infrastructures.

The result of data being expensive and resource-intensive assets is that ‘only already powerful institutions – corporations, governments, and elite research universities – have the means to work with them at scale.’² Especially when it comes to the establishment and processing of big data, such institutions are

able to raise the cost of workers and technological resources necessary to deal with them. If we truly stand at the beginning of a new epoch in which big data and new data analytics reshape the way business is done, government is exercised, and knowledge is produced, a critical stance will inevitably ask for the implications this data revolution has for power. This goes beyond Zuboff's idea of how companies exploit 'behavioural surplus'³ and beyond the ability of these actors to make money from human activities; it is about the ability of these same forces to change how we view ourselves, our histories and our identities in the long term.

The tension between the comprehensive promises uttered by the proponents of the new digital age⁴ and the voices taking a more critical stance may remind humanists of the debate conducted in the late eighteenth century between the proponents of Enlightenment and Jean-Jacques Rousseau. The latter questioned inequalities and new forms of social injustice introduced by the new era, taking into perspective the sense of exclusion and injury of the underprivileged as well as their resentment against top-down modernization. He was demanding a new 'social contract' by means of which just social conditions as well as a social order based on morals, virtue and human character are central to politics, rather than the gains of commerce and money. Rousseau's critique of the political philosophers focused on their neglect of the negative consequences of the innovations brought about by the revolutions of that time, questioned the pursuit of self-interest as serving civil society, denounced modernity as a system in which power flows to an elite and exposed the use of language to deceive and exploit others as pretension: 'All these great words of society, of justice, of laws, of mutual defence, of assistance to the weak, of philosophy and of the progress of reason are only lures invented by skilful politicians or by flattering cowards to impose them on the simple.'⁵ These reflections have been termed 'the autocritique of Enlightenment'.⁶

The nascent digital era and the age of Enlightenment can be compared: while the *philosophes* lived in a time of emerging nation-states, globally acting big tech corporations nowadays assume functions of states. An acknowledged definition of a state describes them as compulsory political organizations with a centralized government maintaining a monopoly of the legitimate use of force. These are elements which can be identified with regard to big tech companies: they determine the rules and restrictions for their users and customers which define the terms of participating in the services provided ('to be naturalized'), and they are hierarchically structured and contain the possibility to be excluded from them ('to be expatriated'). Moreover, the revenues generated by the big tech

companies equal the gross domestic product (GDP) of whole nations. Apple, for example, creates yearly revenues which can be compared to the GDP of Vietnam; Amazon's annual sales can be compared to the GDP of Pakistan.⁷ Seen against Max Weber's famous definition of a state as 'that human polity which claims for itself (with success) within a certain territory [...] the monopoly of legitimate physical violence',⁸ such corporations only lack a territory. Territory is however irrelevant in the digital sphere; for the classic nation-states, it has the function of assets that can be distributed to its citizens and access granted to natural resources. In the digital sphere, data are the complement to natural resources; hence – amongst others – the metaphor of 'data as the new oil' to be exploited. These data are created via the multiplicity of digital devices with their built-in recording and processing functions, the instruments typical for surveillance and governance systems (such as produced by videos or scans as in Google Street view or GPS functionalities built in navigation systems), and by the users and customers themselves (such as through tracking of online activity, transactional data or in the form of the content provided by the users like videos, images and texts).⁹ Availability of big data thus puts each big tech company in a position to be 'the superpower that establishes its own values and pursues its own purposes above and beyond the social contracts to which others are bound'.¹⁰

In taking a critical approach to the discursive regimes around big data and its implications for power relationships, we analysed them as a set of interlocking discourses which justify and support new developments and naturalizes their use. Chapter 2 of this book has analysed the metaphors and terms used for data and data collections, and qualified them as a powerful set of arguments which are euphemistic and revealing with regard to the inequalities and power asymmetries inextricably connected with the distribution of data and thus with the extraction of knowledge and profit. The metaphorization of data as 'natural resources' raises questions about the distribution of the profits gained from their exploitation and towards which end large-scale data are established. Such an investigation opens up the possibility to analyse an extractive system which 'creates a profound asymmetry between who is collecting, storing, and analyzing data, and whose data are collected, stored, and analyzed'.¹¹ With regard to the latter and in the cultural sphere, state-funded endeavours like the provision of archival descriptions or metadata created by libraries and museums stand in contrast to projects advanced by private companies, like Google Books. In their euphemistic function, the metaphors and terms used for data and data processing conceal the fact that data can be treated as goods with a surplus value, which can be traded and monetized while their markets are 'driven by

techniques of “predictive optimization” that attempt to generate future value.¹² A side effect of the ‘naturalization’ of data as natural resources is the obfuscation of the status of data as ‘always already’ epistemically marked by those humans who created the measurement protocols, designed the sensors, and determined their placement and use. Wordings such as ‘data warehouse’ imply a deprivation from context and suggest that data are to be seen as products without the rich context described in other chapters of this book (which renders them as pre-epistemic and fosters their proposed ‘factuality’). As another example, terms like ‘artificial neural networks’ naturalize information processing by suggesting that activities of predictive modelling can be compared to the functional aspects and structure of the human brain. This metaphor denies that human and computational information processing proceed in entirely different ways; it supports the vision that an understanding of the world as complex as that of a human being can be achieved and excelled by a computer, and that humanity will ultimately be superseded by a superior artificial intelligence equipped with supernatural powers (‘singularity’). Finally, even those tasks which cannot be processed by computers are covered behind a figurative term: Amazon’s ‘Mechanical Turk’ recalls the eighteenth-century fake chess-playing machine in which a human chess master was hidden. It appears to accomplish tasks automatically, but in truth makes use of human labour with particularly low wages – another sad example of the exploitation brought about by the data industry – or exploitation that belies the narrative of ‘progress’ as it did during the industrial revolution.

This chapter focuses on power asymmetries and inequalities introduced in the cultural sector by way of datafication and the exploitation of shared cultural assets by private capital. It sets out to investigate and discuss three areas where these phenomena can be observed and lead to significant disadvantages and marginalizations: language resources, digitized cultural heritage and skewed distributions of data in academia and the private sector. All three areas serve as examples where cultural identities, the extrapolation of knowledge from cultural heritage for the purpose of interpreting history, and the societal functions of the sciences and humanities are set in opposition to the aims and purposes of tech corporations.

Language as data

Language is not only one of the best-developed means of human articulation, but also forms the basis for some of the finest artistic creations of humanity.

Be it in prose, lyric poetry or drama, in philosophical treatises or scientific reflections, cultural and creative expressions make use of language and thus form one of the objects on which the arts and humanities focus. As a connective feature of identity, the use of language links groups of human beings and forms collectives. Since the invention of logics by the Greeks, it has been a dream of humankind to reduce all reasoning to some kind of calculation – a dream that has not been achieved yet, since computers are still poor at truly understanding the complexity of the semantic information contained in any of the languages of the world. While the ambiguity inherent in artistic products bears witness of the wealth of cultural heritage, the transformation of the ambiguous, polysemic, conflicting and contradictory phenomenon of language into data still presents, from a computational view, one of the most thorny challenges to be solved. Language data is in raw form unstructured text, which makes up a sizeable (but by no means dominant) portion of the so-called big data landscape. Ambiguity of human language is a major challenge, both at the lexical level, such as polysemy, and at the syntactic level insofar as sentences formulated in a specific language may not be understood in an unequivocal way.

Major populations in the world either speak English or Mandarin Chinese, with speaker numbers of up to 1.268 billion and 1.120 billion respectively.¹³ Since predominant language use entails the provision of large datasets, it is not surprising to find that two of the most often used multilingual neural machine translation services – Google Translate and Baidu Translate – are based in either English- or Mandarin-speaking countries.¹⁴ These circumstances result in power asymmetries and inequalities. First of all, current digital ecosystems are resource-hungry and data-hungry; they need big data. Second, computational power, the application of statistical models and technologies like machine-learning methods (which can be united under the umbrella term ‘Artificial Intelligence’), as well as the work of capable developers of algorithms, are necessary for dealing with big data.¹⁵ All these costly ingredients can only be provided by powerful and financially robust institutions. Language as a resource therefore provides an obvious example where big tech companies can be seen as AI Superpowers.

By contrast, in the European Union (EU) there are twenty-four official languages and more than sixty national and regional languages. Many of the latter are spoken by less than half a million speakers, like Breton, Frisian, Icelandic, Irish, Luxembourgish, Maltese and Scots.¹⁶ While for some EU languages like English or Spanish large data volumes are available, it is obvious that the limited coverage of these ‘smaller languages’ equals data scarcity. For the integration of the EU, on the other hand, the even provision of large volumes

of data would be necessary, as this enables machine translation. Machine translation systems need large collections of aligned bilingual text, which are for example provided in parallel corpora such as the Europarl containing the proceedings of the European Parliament in eleven European languages, a resource that wasn't created for the purpose of machine translation. Since language as a resource is context-dependent, data collected in one domain (here parliamentary proceedings or, more commonly, Amazon product reviews) cannot easily be transferred to other domains such as literature, philosophy or film subtitles. Moreover, machine translation is nowadays facilitated by the training of neural networks, which require training data in the magnitude of at least 1 million sentence pairs. Even if there are data in this scale at hand, rare and unseen words won't be translated or have to be complemented manually. These technical requirements have a direct consequence: languages with less than 100,000 speakers do not generate enough data and will therefore not be served by machine translation services. Beyond these data inequalities, the training of neural networks requires powerful hardware with a lot of memory and graphics processing units to parallelize training. Such an infrastructure is costly and can rarely be afforded by small and medium-sized enterprises facilitating machine translation even within the EU, which has been relatively proactive via its policies, funding programmes and research interest groups in promoting machine translation. If these requirements are not fulfilled, the quality of the translation suffers. Moreover, mistakes in input data are exacerbated by neural machine translation systems which are sensitive to faulty data.¹⁷ Data issues – such as errors, noise and inconsistencies in coverage – therefore have a crucial impact on the quality of the services. Furthermore, frequently encountered issues with language resources pertain to the openness of available data – especially where corporations own large volumes of data¹⁸ – as well as Intellectual Property Rights, for example, with digitally available texts in small languages. The unequal availability of language data disadvantages and marginalizes smaller languages, since 'machine learning is only as intelligent as the amount of data it has to train on',¹⁹ and it causes power asymmetries with regard to the companies providing translation services: 'No LT [language technology] company in Europe is even of comparable size to the big players like Google or Baidu.'²⁰ The fact that large amounts of data are available in some languages while others are only represented by a little also results in a Matthew effect²¹ created by ownership: larger corporations have access to large volumes of data, while smaller companies are left behind. While larger corporations are able to extend their services and their quality,

smaller companies won't be able to close the gap in terms of data, hardware and software and will become economically inefficient. In this way, language data is making the big bigger and the small smaller. Smaller languages suffer from the Matthew effect as well: due to their limited size, they are regarded as being susceptible to be deviant from the default, as infrequent use is translated into minimal predictions of use.

As the example of the EU shows, data inequalities have direct consequences not only for companies providing machine translation services, but for the economy in general – and for citizenship. Customers ordering online prefer webshops where information on the products is provided in their native language. In a competitive environment, companies must focus on maximizing their profits; 'the affordability of language technologies is [...] a problem for SMEs that want to translate their online services and cannot assume the cost of doing so'.²² Within Europe as a multilingual market, data scarcity thus presents a major impediment for business across language. Businesses that cannot provide content in local languages may therefore become economic losers, and the uneven distribution of language data may lead to fragmented markets. Moreover, for European citizens, multilingualism presents a key feature of Europeanness. While it may be possible to translate in a concerted effort all documents necessary for administration within the EU (e.g. to ensure mobility and the free exercise of a profession within the EU), constraints in the provision of translation services directly affect participation in the political process. To provide but one example: 'Following the Fukushima incident in 2011 there have been discussions about the dangers of nuclear energy in all European countries. These debates were held in the respective language communities only, there has never been a public European debate about the topic because it is, technically, not yet possible to organize such a debate online'.²³ Within the EU, the lack of language resources and technologies may therefore limit citizens' engagement in collective debates, the development of cross-border solutions and their participation in decision-making processes. Fostering a common European identity while maintaining the cultural and linguistic diversity of Europe therefore presents a substantial challenge. Whereas the officially acknowledged languages of the EU are granted – at least in principle – equal status, it will take a large-scale effort to enable the use of high-quality language technologies by balancing the asymmetries inherent in the available data for all European languages. Within the multilingual landscape of Europe, the mutual recognition of the diversity of linguistic origins and traditions forms a cornerstone of the development of a common European identity.

Cultural heritage

The shift from analogue finding aids to digital archival descriptions provided by cultural heritage institutions (CHIs) has been thoroughly explored in preceding chapters. It is useful to revisit this frame of reference once more, however, as this movement can also be understood in terms of Erving Goffman's 'breach and repair' approach, in that cultural heritage 'practitioners' reactions to the disruption of their practice offer the opportunity to glimpse the norms of archival thinking.²⁴ The digitization of archival material and the provision of digitally available descriptions have consequences for the relationships between cultural heritage practitioners and users/researchers and accordingly practitioners and technicians on the one hand, and for the power structures regarding access to knowledge and knowledge creation on the other hand.

The relationship between cultural heritage practitioners (such as archivists) and the users of CHIs has changed profoundly. While more and more archival descriptions have become available in electronic formats – the Archives Portal Europe currently provides 283,774,614 descriptive units of archives²⁵ – it is still the case that archivists function as the gatekeepers to cultural heritage and, as the experts for the material they preside over, provide the first entry points to it. As the interviews we conducted reveal, archivists assume that a direct and dialogic relationship with users is and will be maintained. As one librarian stated:

And our descriptions here are very, very detailed. For example, I have a manuscript here and I'm describing it. I look through it. [...] But at the end, you know the whole thing. It helps us very much when we have to answer requests. Because on one side you can find all that data in our database, and on the other side, even if there are some technical [reasons a user] can't find it, we can find it.

(WP3 INT1)

But the findability of archival descriptions online also promotes the possibility that users avoid a direct exchange with an archivist or deem it to be superfluous, thus cutting archivists off from their function as gatekeepers to cultural heritage. Furthermore, as cultural heritage practitioners learn the language of technology, a power shift towards the technical personnel takes place in these institutions. 'Practitioners' experiences of the power dynamics at play when other specialists enter heritage spaces suggest that their perspective is the unique contribution most at risk of being squeezed out of the discourse.²⁶ It is obvious that sound personal relationships and mutual recognition become essential in order to avoid power asymmetries within CHIs.

While the establishment of trusted relationships between practitioners and users is changing during the transitional phase of digitization, the establishment of digital archival descriptions pre-forms the way in which meaning and knowledge are created. Archivists themselves notice their own detachment from the material they presided over, as they

are not working on item level anymore, you are trying to subtract the general meaning, the general line from a collection [...] For every new collection that comes in, you can't go in depth, reading every page in detail, you skim through and you seek the major subject. We're not as close to the items anymore.

(WP3 INT3)

Cultural heritage practitioners must decide how to describe an item or collection and to place more emphasis on including the context of the item or on the connections with other items or collections. They do not have the resources to transfer their tacit knowledge to the systems with which they must work.

We decided not to make a characteristic description of the pictures. That means that we only give the original caption of the photograph. [...] These are propaganda captions, in fact, because these pictures were taken during the war by specific agencies ... that is why we decided to transcribe the original caption, but without giving any other information. And also, because we don't have time, we don't have staff to describe individually each picture, so we inventorised pictures with a thematic logic.

(WP3 INT4)

In this example, presenting the original context of the holdings was prioritized over making connections with other collections. In a wider sense, the anticipated future uses of these archival descriptions therefore elevate one meaning above the exploration of complex alternatives.

Many archives are still in a transitional period where only a fraction of their holdings descriptions are available online, while the archival material itself still requires a visit to the archive: 'You can access only the part of the documents that are tagged that it's okay to be online. It's a pyramid on the website, you will have only a small part of the description and the documents. In the reading room, you will have this box in addition' (WP3 INT2). The material presented online currently presents only a small fraction, with clear implications for the material buried at deeper levels of accessibility. A decision to choose one collection for digitization and online access over another therefore privileges that knowledge directly in that it increases its chances of being studied by researchers, as has

been explicated in Chapter 5. With regard to an automated analysis of digitized historical sources during the transitional period we are currently in, such a partial approach to digital discoverability and the exploitation of the data provided online would exclude those sources which have not been and will never be digitized or contained in archival descriptions, and such analyses would render at best a distorted knowledge of our shared past.

As has been discussed in Chapter 5, cultural heritage practitioners have already observed the formation of search habits on the side of the users, who imitate the keyword-search approach to knowledge established by Google, thereby simplifying their access to complexity and losing sight of what might not be available in digital format. Cultural heritage practitioners are moreover well aware of their lack of experience in what could be done with the collection descriptions by means of aggregation:

Data-linking is one of the limitations we have to take into account, and it's one of the primary factors in terms of restricting data, because even if you've removed all the direct identifiers, maybe indirect information that could be used to identify them, even something that seems innocuous like the type of building material used in a house could, in some cases, be used to identify the specific house in a region because they use certain types of material ... and the identifiers that are used, even if they're a numeric ID, could be linked to an existing dataset. It could be linked to the personal data that people have stored elsewhere than they're supposed to.

(WP3 INT5)

Such data aggregation could, for example, be performed by institutions or companies which have the means to establish large databases of citation networks, which could then be aligned with archival descriptions containing anonymized information on persons and their written exchanges with other historical personalities, thus enabling de-anonymization. In the interviews conducted within the KPLEX project, archivists proved to be well aware of the fact that the digital material they expose becomes vulnerable to misuse:

Certainly, the machine processing is, I suppose, going to be even bigger soon than it is now, and [...] artificial intelligence has the potential to draw new conclusions from a large amount of data, particularly unstructured data, which [...] until quite recent years have resisted the broader analysis [...] if automated tools are able to make links between those datasets and then [...] infer conclusions about the people, if it's identified, then there's a significant danger to them.

(WP3 INT5)

Research methods may become increasingly mediated by automated tools through which machine learning algorithms will play an increased role not only in analysis but also in discoverability through linked data. Currently, cultural heritage practitioners mediate as gatekeepers of material that is vulnerable to misuse. However, online discoverability and access to digital data may expose a dark side, where potential threats are magnified beyond those that can be seen unaided, by the use of AI and big data tools. It is as yet unclear to which results such automated data analyses will lead, but certainly only a few researchers and institutions are capable of capitalizing on big data on cultural heritage and exploiting the possibilities that lie therein. Whereas the capacity of historians to process large amounts of information is limited (the case of which presents a well-documented and understood confirmation bias), information processing performed by computers can be scaled to enormous quantities of available data. The interpretation of cultural heritage is a complex process endangered by the threat that some forms of knowledge will not properly be translated by computational methods.

Take the development of dominant patterns of thought as an example: while it would be absolutely interesting to investigate the evolution of such concepts like racism or empathy over the course of several centuries and on the basis of large collections of data, it has to be questioned whether such information can adequately be treated by machines incapable of navigating the ambivalences and polysemies inherent in the data which are analysed, since they would have to be able to appropriately place such data in their historical context and consider the shifts and caesuras characteristic of historical development. For example, the encounter of Columbus and his men with the indigenous people they met during their 'discovery' of the Americas can be understood according to the attitudes prevailing in the context of their time; it would form a completely different approach to interpret them in terms of the paradigms of racism and empathy which have been developed centuries later. Historical knowledge is always situated knowledge insofar as it is itself placed in time and assesses and reevaluates past events from a certain point of view. It follows then that power shifts or revolutions result in shifts in the interpretation of history, as has happened, for example, after the end of the cold war and the fall of the iron curtain or during the phase of decolonization. After such caesuras, a reinterpretation and reevaluation of historical events take place which exploit the ambivalences and polysemies typical for historical sources.

Seen against this background, it is highly questionable whether research results achieved by computational means will ever be capable of providing

insights adequate to the historical record beyond such highly formalized data collections and processing methods like social network analysis. Furthermore, processes like identity formation and the provision of answers as to how a collective's identity or sense of belonging to one state or one nation arises do not rely on historical evidence in a self-evident way, as the well-researched examples of invented traditions and imagined communities show.²⁷ The rewriting of the grand narratives and foundational myths of nations has repeatedly taken place, embedded into complex processes like the emergence of nation-states or the fall of empires, and they are themselves part of societal processes and large-scale discursive shifts in which the community of historians is in constant exchange with each other in order to adjust their insights and epistemics with the society in which they live and whose questions they are supposed to answer. It is precisely this knowledge of context and this experience of being-in-a-situation which separates the work of a historian from that of machine intelligence, which always must begin at the level of objectivity and rationality where the facts have already been produced and the data are already at hand. The question is therefore what the power shift currently on the horizon with regard to knowledge creation in the realm of cultural heritage will look like. With regard to the power structures at stake here, the tension between knowledge creation processes performed by government-financed historians and private data analysis laboratories owned by tech corporations is crucial. This tension pertains to questions of the legitimate production of knowledge, the credibility accounted to the results²⁸ and the mechanisms of producing expert consensus in scientific communities.²⁹

The academic field

'Knowledge is power', according to Sir Francis Bacon. Even though data are not powerful in and of themselves, scientific disciplines realize meaning and value from data; control over data therefore also introduces asymmetries into the academic field, as we were able to observe within the KPLEX project in our study of the research topic 'emotions', in which we analysed data collection, production, differing methodological epistemic cultures and research organization in a broad range of scientific disciplines.³⁰ In the field of academic emotion research, such asymmetries can best be illustrated by juxtaposing the experimental approach applied by neuroscience and research conducted in the field by anthropologists. Neuroscience has built a huge research structure in order to explore research participants' individual emotional responses to the stimuli presented in a fMRI scanner. This research setting includes the maintenance of an expensive

laboratory designed for this specific purpose, as well as the provision of specialized technical personnel and administrative staff. The output of the costly machinery amounts to several gigabytes of data per research participant. On the other side of the disciplinary spectrum, anthropological emotion research represents a quite different methodological approach: a solitary ethnographer conducts research within a culture that is initially alien to the researcher and may confront her with overwhelming complexity. Here subjectivities and relationality come into play, elements that are (supposedly) missing in the technical environments of controlled experiments. The output in terms of data consists of the field notes taken by the researcher, dialogues conducted during the field trip, analyses reflecting the positionality of the researcher and so on. The contrast between laboratory and field thus illuminates the extreme poles of differing 'epistemic cultures'³¹ at play in emotion research.

The availability of large datasets in the academic field necessitates employing data scientists alongside the subject matter experts. This division of labour inevitably produces a power shift with regard to the question as to who is the most legitimate producer of knowledge – the domain expert or the statistician analysing the data. Interdisciplinary research settings create their own difficulties of organizing research in larger research units and influence the relationships established between disciplines and researchers: 'So interdisciplinarity [...] always comes with dependencies in a way, between the people, new dependencies, and so you have again power relations within an interdisciplinary project that are different to purely disciplinary projects' (WP4 INT12). Such circumstances may impose structural limitations on the research project insofar as they lead to conflicts and debates about the methodological approach within the research team, as one researcher reported:

And to me one of the kind of biggest lessons learned during this project was that being a humanities social science person and starting to work with quantitative people. I thought that quantitative people are kind of the same (laughing). You know, I thought that statistics people and natural language processing and machine learning that they kind of agree. But there's huge differences between them. So, the first year was basically/I thought that the big divide would be me with them or us with them. But it ended up that they among themselves were fighting. Like really fighting because statistics people, for instance, socially, their data was problematic in MANY, many ways because they didn't know where it came from. So, actually their ideas aligned very well with my ideas that we need to do some sort of ethnography around the production of the data.

(WP4 INT11)

Another of our interviewees reported from the experiences made within a larger research unit where neuroscientists collaborated with social scientists and humanists. Since each discipline brought their own epistemic approach as well as requirements as to knowledge complexity and differentiation of results, each discipline searched for legitimacy, recognition and autonomy which resulted in a competitive struggle within the academic field and brought the tensions created by asymmetric power distributions and monopolies to the surface. In the example of emotion research, the tensions arose between the limitations imposed by the use of a fMRI scanner, the output of which is measured according to the variables of valence and arousal, and the complex and differentiated conceptions of human emotions characteristic of the social sciences and humanities:

In this cluster I worked together with a psychologist, a developmental psychologist. But now: no [more]. [...] And this was a deliberate decision. Because especially during this [...] big excellence cluster we reach the/I would say we came to our limits. And this has to do with research methodologies. And so, one problem was, or an example with persons like [...], he's also a brain researcher. And he was very open-minded and very interested in our kind of work. And we spend hours and hours and nights together and conceptualised very complex projects. And he said: 'You are so right, this is so important, but we cannot do this. We have this brain scanner. And it's too complex. We have to design a common project with two variables.' To be able to/that it fits into this scan thing. This was the end of [...] collaboration. I said: 'It's, sorry, but it's too stupid for us. I cannot do it!' Yeah, it's [...] in my own discipline. If I reduce this complexity to these two things you can see in your scan. [...] And we all made this experience and said: 'So, it's enough!'

(WP4 INT15)

As the example shows, the multiplicity of epistemic approaches within a single large research endeavour fostered the entrenchment of academic disciplines from each other and contributed to mutually exclusive approaches within the ivory tower. Most interestingly, the cessation of interdisciplinary collaboration within emotion research resulted out of the lack of differentiation possible within the large datasets provided by a fMRI scanner:

My own thought on this has been: How can neuroscientists do these brain scans and not consider that the person that they are scanning is an individual person with a biography and cultural embeddedness and that sort of thing? You are not looking at a brain; you are looking at someone's brain! You can aggregate, I suppose, lots of data from a lot of different peoples' brains and arrive at some

sort of average, you know, values, but still you are dealing with brains that are embedded in a certain time and space and society and you should take account of that. But neuroscientists cannot and do not want to. [...] what I would really like to see is the effects of practice on the brain. From the perspective of practice theory there is an assumption that doing something over and over does in fact change the body, the physical body.

(WP4 INT6)

The uneven distribution of data does not only pertain to processes of distinction between academic disciplines; when asked about big data in a conventional sense, cultural heritage professionals proved to be quite critical about the strategies employed by big tech companies, where data are collected and processed by specially designed research units. They underlined here that the commercial aims pursued with the aggregation of big data fundamentally differ from academic research purposes.

Just gathering data because you can is something not very reasonable for me, right? And many of the big data sets, for example service providers have or Google has might be usable for specific reasons, the purchase behaviour of people or whatever, but it is not desirable for me if I want to pursue my own research which is then the speech related data. [...] So, the big data is a different kind of data, data types, right? But of course, these companies or entities, organizations also, they have a big advantage if they have the access to this kind of data. [...] They also are very active in research, but not necessarily with externalising all of the data, right? But with participating in the knowledge exploration, basically in the scientific process, yes, but not really sharing the data.

(WP4 INT5)

The competition between economic enterprises controlling big data and the academic field results in a terminological differentiation between on the one hand very large datasets (such as video collections or aggregated fMRI data from the neurosciences) that were termed 'research data', because they were conceived according to an elaborated epistemology and methodology and collected guided by scientific reasoning; and 'big data' on the other hand, a term designating data that were not collected for specific research purposes and thus regarded to be antecedent to the epistemological process on the other hand. The reflections and comments uttered by scientific researchers reflected the debate initiated by the article on the 'end of theory' published by Chris Anderson,³² but focused on the implicit epistemological assumptions with regard to research driven by big data, rather than an overt discussion of

questions of power, legitimacy or financial support. The claim that there is no need for a priori theory was criticized as an unscientific approach, since specific research questions within a determined research framework were seen to be essential for a steered, methodologically controlled research process, which was contrasted with an investigation conducted in an exploratory manner open for serendipitous findings:

I mean this is the big game of big data, of how you/or why everyone who says I do big data speaks of it and gets money and stuff, because it's a very cheap way to correlate datasets and then find new things. But the question remains – if you don't have a question then why do you look for an answer?

(WP4 INT12)

The idea that big data can speak for themselves free of human bias was regarded as misleading, insofar as such a claim ignores the design of data collection tools and the values, assumptions, biases and interests driving the design process. Moreover, a humanistic approach was outlined that called for a balanced approach to biases in big data:

There is a group of researchers who are kind of like: big data is objective, or it gives us objective views to the world. And of course, we all know that that's not the case. [...] And I think the interventions that we can do is to demonstrate that there are so many different ways of understanding what is knowledge. What's the truth value of knowledge? Where it's coming from? How there're always biases. And we have to acknowledge them. That sometimes human bias is a very good thing. You know, we come from a country with a lot of gender politics. That's mostly a bias. But we think that this is an important bias to take gender into account and sometimes favour women. Or sometimes favour minorities. So, these are like human biases that are needed for society building and advancement.

(WP4 INT11)

Finally, the claim that big data strives for exhaustivity and enables one to see everything within a specific domain was directly contradicted in our interviews, which is not surprising since this claim, when applied to the academic field, calls the idea of a differentiation into several specialized scientific disciplines into question. With regard to big data collected online, one researcher remarked that such data may allow knowledge creation about human behaviour, but does not provide insight into motivation and emotion:

So, for example there is, within big data there is a lot of behavioural tracking. And I go back to booking.com example, so they are definitely doing big data analysis. And what they're doing is [...] they answer the question: What do I have to do in order to make the most profit out of the website? And for that their approach is pretty much optimal [...]. If their question would be: What does make my customers feel the best? Then their type of research probably wouldn't be as good as comparable methods within neuroscience for example or [...] qualitative research just asking the users in-depth-interviews or something like that. Because with the behavioural data you don't get any information about how they feel. You just get information of whether they stick on the website or whether they move away from it.

(WP4 INT13)

This line of argumentation echoes on the one hand the dispute about the representativity of data which were collected only online versus those data collected with social sciences methods (surveys, public-opinion polls);³³ on the other hand, it points to the discursive tensions elicited by the suggestive and powerful set of arguments brought forward in the discussion triggered by Chris Anderson's article, which can be described as a discursive regime in Foucault's sense. Moreover, the competition between academia and tech giants holding big data comes into view: the academics interviewed in the KPLEX project expressed their dismay that the production of knowledge from big data occurs unmoored from science. Research fields that are claimed to be traditionally a part of academic endeavours (such as questions pertaining to whole populations and to the characteristics of society) come into opposition with the explanatory power proposed to be derivable from big data and the claims to representativeness driven by business interests. Moreover, the balance of knowledge creation in society becomes destabilized in favour of the big tech companies: 'This unprecedented concentration of knowledge produces an equally unprecedented concentration of power: asymmetries that must be understood as the *unauthorized privatization of the division of learning in society*'.³⁴ In other words, the creation of knowledge out of big data beyond academia and the skewed distribution of data between scientific and business institutions results in power struggles between these two autonomous institutions about the legitimacy of providing answers to societally relevant questions, as well as about their relationship to the third major institution holding big datasets, namely governments.

Conclusion

Language data, digitized cultural heritage and academic endeavours are but three examples of fields where power asymmetries and inequalities can be observed and brought into dialogue with those discourses of power over big data, computing capacity, statistical models and technologies, and algorithm experts exerted by the big tech companies. If we recall the reflections brought forward by Jean-Jacques Rousseau with respect to the Enlightenment's vision of the new era to come and his discussion of the sources of inequalities (competence, ownership, wealth creating and directing new needs), and if we apply these reflections to the new digital era to come, we must discuss the consequences of these asymmetries and inequalities for society. While governments are slow in establishing data infrastructures and providing analytic possibilities for their citizens, and while academia is restricted by the various interests which drive research forward, big tech corporations exert power on their users and customers while they silently exploit the data collectively contributed by them for the purposes of private corporations. The power of these companies rests on the exploitation of volunteered data (such as in the case of Wikipedia or the Open Street Map), in the form of data provided as content, expertise, opinions and knowledge (such as through data created in social media, in applications quantifying the self, on crowdsourcing platforms, or in citizen science), and in cultural heritage data in general (such as in the printed books digitized by Google Books).³⁵ This authority is exerted in an opaque, algorithmically implemented way, as in the manipulation of opinions via social media, which stands in contrast to free opinion formation, in the direction of consumers' choices via search engines and the creation of needs through recommendation systems, and, generally, in behaviour modification, since 'automated machine processes not only *know* our behavior but also *shape* our behavior at scale'.³⁶

But the notion that big tech corporations can be compared to states opens up the question of whether users and consumers and the data they contribute should rightly be granted the revenues enjoyed by such global corporations. Instead of reducing human beings to users and consumers, they should rather be regarded as citizens endowed with political rights which are suitable 'to protect their liberty, and not to enslave them',³⁷ and who can actively participate in the spread of the revenues gained and profit from the benefits achieved. Following the example of Rousseau, we can observe that the big tech companies turn shared cultural assets – be it in the form of collectively created large language data, in

the form of cultural heritage which can be regarded as cultural commons or those data forming the basis of knowledge creation – into corporate capital. This observation reveals the differences between private corporations and nation-states: cultural heritage represents a public good, and opinions and aesthetic choices contributed via social media, crowdsourced platforms or recommender systems should be regarded as cultural commons as well.

Rather than serving as a resource to be exploited by tech giants – like the labour provided by a Mechanical Turk – and thus encouraging the cultural colonialism of private corporations, such big data should be treated as public commodities. This presents a tricky challenge for regulation and an issue for antitrust laws, since markets can be assumed to exist even if the services provided by corporations are free of charge, for example, in the case where users pay with their data: ‘The problem for regulators is that standard anti-monopoly frameworks do not apply in a world where the costs to consumers (mainly in the form of data and privacy) are thoroughly non-transparent.’³⁸ But these markets have two layers, the first one being the commons which have a societal value, and the second one formed by these commons being datafied, which enables corporations to skim their economic value. The idea of cultural commons therefore needs to be revitalized and reformulated with respect to the monetary value of data, where free access to such big data is granted in order to provide the possibility to realize the chances and opportunities which lie therein for the common good. Free access to such data could either take the form of a ‘data-sharing mandate’,³⁹ or data should stay in public ownership just as non-digital cultural commons. Google Books serves as a good example here, since technically no objection can be raised against a corporation claiming to have digitized 40 million books⁴⁰ as long as all the content is openly accessible. But Google Books is a revenue-generating project and not a library.⁴¹ Libraries, by contrast, are funded by the public sector and provide their data, metadata schemes and APIs openly to everybody interested in using such services. As this example shows, data need to be understood as an essential part of infrastructures, which should be owned by the public just as artificial intelligence built on top of them; conceptualized in this way, data would form public commodities which can be used by several agents simultaneously, but private corporations would have to pay for their use.⁴² In the sense of Rousseau, such a reformulation of the idea of cultural commons would level out inequalities and forms of social injustice brought about by the current data revolution and constitute a new social contract with the full participation of citizens as members of a polity.

Notes

- 1 Kitchin, Rob, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: Sage, 2014).
- 2 D'Ignazio, Catherine and Lauren F. Klein, *Data Feminism* (Cambridge, MA/ London: MIT Press, 2020), 41.
- 3 Zuboff, Shoshana, *The Age of Surveillance Capitalism* (London: Profile, 2019), 74.
- 4 See here for example the publications by Bostrom, Nick, *Superintelligence. Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Brynjolfsson, Erik and Andrew McAfee, *The Second Machine Age. Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York: W. W. Norton & Company Inc, 2014); Tegmark, Max, *Life 3.0. Being Human in the Age of Artificial Intelligence* (London: Allen Lane, 2017).
- 5 Rousseau, Jean-Jacques, *Œuvres Complètes (Fragments Politiques)* (Paris: Pléiade, 1964), 475, own translation.
- 6 Hulliung, Mark, *The Autocritique of Enlightenment: Rousseau and the Philosophes* (Cambridge, MA: Harvard University Press, 1994).
- 7 Numbers were provided in July 2020: Daly, Kyle, '4 Numbers That Show the Combined Power of Facebook, Google, Apple and Amazon,' *Axios*, <https://www.axios.com/big-techs-power-in-4-numbers-de8a5bc3-65b6-4064-a7cb-3466c68b2ea0.html>.
- 8 Weber, Max, *Wirtschaft und Gesellschaft: Grundriß der verstehenden Soziologie*, 5th edn (Tübingen: J.C.B. Mohr (Paul Siebeck), 1980), 822, own translation. Compare here: Hay, Colin, *Routledge Encyclopedia of International Political Economy* (New York: Routledge, 2001), 1469–74. Lesser-known alternative definitions of states do without the notion of territory by focusing on the functions of states with regard to sovereignty, welfare and security. See Lehmkuhl, Ursula and Thomas Risse, *Regieren ohne Staat? Governance in Räumen begrenzter Staatlichkeit* (Baden-Baden: Nomos, 2007), 2–10.
- 9 Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, 87–98.
- 10 Zuboff, *The Age of Surveillance Capitalism*, 82.
- 11 D'Ignazio and Klein, *Data Feminism*, 45.
- 12 Raley, R., 'Dataveillance and Counterveillance,' in 'Raw Data' Is an Oxymoron (Cambridge, MA: MIT Press, 2013), 124.
- 13 'List of Languages by Total Number of Speakers,' *Wikipedia* (2021), https://en.wikipedia.org/w/index.php?title=List_of_languages_by_total_number_of_speakers&oldid=1006949475.
- 14 Google Translate supports translation between 108 languages, with 100 billion words translated per day. The most common translations are between English and

- Spanish, Arabic, Russian, Portuguese and Indonesian. See ‘Ten Years of Google Translate’, *Google* (2016), <https://blog.google/products/translate/ten-years-of-google-translate/>; ‘Google Translate Adds Five Languages’, *Google*, 2020, <https://blog.google/products/translate/five-new-languages/>.
- Baidu Translate supports translation between 203 languages, with 100 billion characters (in modern Chinese, most words are compounds written with two or more characters) translated per day. The most common translations are between Mandarin Chinese and English, Japanese, Korean, Spanish, Thai, French and Arabic. See Baidu Research, *Twitter* (2020), <https://twitter.com/baiduresearch/status/1248316242689019904>; ‘百度百科’, https://vhsagi.smartapps.baidu.com/pages/lemma/lemma?lemmaTitle=%E7%99%BE%E5%BA%A6%E7%BF%BB%E8%AF%91&lemmaId=10607809&from=bottomBarShare&hostname=baiduboxa_pp&_webfr=1.
- 15 Lee, Kai-Fu, *AI Superpowers. China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018), 14.
 - 16 See the report: European Parliament. Directorate General for Parliamentary Research Services. *Language Equality in the Digital Age: Towards a Human Language Project* (Brussels: STOA, 2017), <https://data.europa.eu/doi/10.2861/136527>.
 - 17 Edmond, Jennifer, Rihards Kalnins, Andrejs Vasiljevs and Mārcis Pinnis, ‘KPLEX Report on Language and Culture as Data’ (2018), <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:114127>.
 - 18 Google, for example, uses the texts scanned by the Google Books project for machine translation purposes.
 - 19 Zuboff, *The Age of Surveillance Capitalism*, 188.
 - 20 SRIA Editorial Team, *Language Technologies for Multilingual Europe – Towards a Human Language Project* (2017), <http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V1.0-final.pdf>, 13.
 - 21 Merton, Robert K., ‘The Matthew Effect in Science’, *Science*, 159/3810 (1968), 56–63.
 - 22 European Parliament. Directorate General for Parliamentary Research Services. *Language Equality in the Digital Age: Towards a Human Language Project*, 38.
 - 23 Language Technologies for Multilingual Europe – Towards a Human Language Project, 24.
 - 24 Horsley, Nicola, ‘Intellectual Autonomy after Artificial Intelligence: The Future of Memory Institutions and Historical Research’, in *Big Data – A New Medium?* (Abingdon: Routledge, 2020), 130–44.
 - 25 ‘Archives Portal Europe’, <https://www.archivesportaleurope.net/home>.
 - 26 Edmond, Jennifer, Mike Priddy and Nicola Horsley, ‘KPLEX Report on Historical Data as Sources’ (2018), <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:114127>, 77.

- 27 Anderson, Benedict, *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (London: Verso, 1991); Hobsbawm, Eric and Terence Ranger, *The Invention of Tradition* (Cambridge, UK: Cambridge University Press, 1983).
- 28 Haraway, Donna, 'Modest Witness: Feminist Diffractions in Science Studies', in *The Disunity of Science. Boundaries, Contexts, Power* (Stanford, CA: Stanford University Press), 432.
- 29 Oreskes, Naomi, *Why Trust Science?* (Princeton: Princeton University Press, 2019), 57.
- 30 Edmond, Jennifer, Thomas Stodulka, Elisabeth Huber and Jörg Lehmann, 'KPLEX Report on Data, Knowledge Organization and Epistemics' (2018), <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:114127>.
- 31 Knorr-Cetina, Karin, *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, MA: Harvard University Press, 1999).
- 32 Anderson, Chris, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired* (2008), <https://www.wired.com/2008/06/pb-theory/>. For a thorough discussion of the fallacies inherent in basic assumptions with regard to big data, see Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, 133–7.
- 33 See, for example, the debate conducted in Germany between the two public opinion poll institutes Forsa and Civey, the latter of which does collect its data only online; compare here: Pausch, Robert and Fritz Zimmermann, 'Umfrageinstitute: Kampf Der Torten', *ZEIT ONLINE* (2020), <https://www.zeit.de/2020/07/umfrageinstitute-meinungsforscher-sozialforschung-civey-forsa>.
- 34 Zuboff, *The Age of Surveillance Capitalism*, 192. Original emphasis.
- 35 Google estimated in 2010 that there were about 130 million distinct titles in the world; see Jackson, Joab, 'Google: 129 Million Different Books Have Been Published', *PC World*, https://www.pcworld.com/article/202803/google_129_million_different_books_have_been_published.html, and it stated that it intended to scan all of them. As of October 2019, Google celebrated fifteen years of Google Books and provided the number of scanned books as more than 40 million titles in over 400 languages; see Google, '15 Years of Google Books', 2019, <https://www.blog.google/products/search/15-years-google-books/>.
- 36 Zuboff, *The Age of Surveillance Capitalism*, 8.
- 37 Rousseau, Jean-Jacques, *The Social Contract and Discourses* (London and Toronto: J.M. Dent and Sons, 1913), <https://oll.libertyfund.org/title/cole-the-social-contract-and-discourses>.
- 38 Rogoff, Kenneth, 'Big Tech Is a Big Problem', *Project Syndicate* (2018), <https://www.project-syndicate.org/commentary/regulating-big-tech-companies-by-kenneth-rogoff-2018-07?barrier=accesspaylog>.

- 39 Mayer-Schönberger, Victor and Thomas Ramge, *Reinventing Capitalism in the Age of Big Data* (London: John Murray, 2018), 167.
- 40 Google, '15 Years of Google Books' (2019), <https://www.blog.google/products/search/15-years-google-books/>.
- 41 Vaidhyathan, Siva, *The Googlization of Everything (and Why We Should Worry)*, (Berkeley: The University of California Press, 2012), 149–73.
- 42 Morozov, Evgeny, 'Socialize the Data Centres!', *New Left Review*, 91 (2015), <https://newleftreview.org/II/91/evgeny-morozov-socialize-the-data-centres>, 45–66.