From Data to the Creation of Meaning Part 1: Unit of Analysis as Epistemological Problem

by Justin Joque[1]

Abstract

Aligning data and research infrastructure is, as our daily work often reminds us, a difficult process. While data professionals often focus on research lifecycles, incentives, storage and transmission technologies, metadata and data sharing we tend to overlook the epistemological incongruences of diverse research and data practices. All data creation processes, even if unknowingly, make assumptions about the world and what exists as a unique unit that can be analyzed. In attempting to make data meaningful to different audiences, especially across disciplines, we must pay attention to these epistemological assumptions. Failure to do so will inevitably frustrate our attempts to develop meaningful infrastructure for research data and even potentially undermine effective research through misunderstandings of data. Looking at census and zip code data as examples, this paper explores the issue of unit of analysis as an example of such disciplinary epistemological assumptions. The complexities that arise even in these simple examples suggest the importance of addressing the theoretical complexities of dealing with data collections, management and interpretation.

Keywords: Data Profession, Categorization, Harmonization, Epistemology, Data Theory, Critique.

Introduction

There has been a growing interest in the humanities in adopting and developing digital methods from other disciplines, especially including techniques for collecting, describing and modeling large datasets. While there is much interesting work to be done in this regard, these interdisciplinary conversations will be most fruitful if the dialogue is carried out in both directions. By adopting an especially humanistic and critical perspective, this paper and Kristin Partlo's, also published in this issue (see page 12: From Data to the Creation of Meaning Part II: Data Librarian as Translator), are thus an attempt to both historicize and theorize the work of describing and using data to generate new knowledge in the social sciences. In working with data it quickly becomes evident how difficult and problematic it can be to deal with multiple datasets or research questions that do not align with available data, but these difficulties are infrequently explored in a wider context. The difficult relationship between data and the world is often overlooked and not considered in a serious and abstract way. In day-to-day work with data the issues that arise tend to appear as accidental discrepancies rather than as endemic to the relationship between data and world. In this

The difficult relationship between data and the world is often overlooked and not considered in a serious and abstract way.

light, this paper attempts to open the question of what exactly data measure and what relation that measurement has to the world, arguing that this relationship is necessarily problematic and difficult but still incredibly productive.

Ultimately, our attempts to describe the world through data are not processes of merely finding what is really there, but an active epistemological project of description and making the world knowable. All data creation processes, even if unknowingly, make assumptions about the world and what exists as a unique unit that can be analyzed and what 'counts'. In attempting to make data meaningful to different audiences, especially across disciplines, data professionals must pay attention to these epistemological assumptions in their necessary diversity. With the possibilities and daily challenges of describing the world scientifically in mind, I would like, by tracing some of these problems historically as well as in relation to contemporary problems, to suggest that the difficulties we as data professionals face are a necessary difficulty of the relationship between data and the world that we will never be able to fully overcome.

There have been many important contributions to confronting and describing these issues from epistemology to philosophy to science studies, including authors such as Michel Foucault and Bruno Latour.[2] More closely related to our work with social science data, a number of authors working in information science have addressed some of these questions. Ronald Day's (2014) recent work on the history of the documentary tradition, from early twentieth century information science to current work on big data, argues that the very act of describing of things—and the consequent use of this information as evidence—is a constructed, mediated and ideological process. His account of data and documentation in a historical context, suggests that all data and the description of information has always been mediated by technology and culture. Furthermore, as Geoffrey Bowker (2014) has recently observed: big data appears to be able to efface categories in favor of temporary clusters of correlation (for example in the commercial world you no longer "need to know whether someone is male or female, queer or straight, you just need to know his or her patterns of purchases and find similar clusters" (Bowker, 2014: 1796)). Despite this, he argues social categories still produce real effects that need to be reckoned with and described. While algorithms may be able to bypass gender identity, the expression of gender in the real world creates undeniable effects. These categories cannot simply be ignored or replaced with uncategorized click-streams or DNA sequences. With Bowker and Day's arguments, it becomes clear that the categories through which data are identified, collected, aggregated and mediated are social and ideological constructs. Despite this, these categories cannot simply be ignored. We must ultimately account for them and their effects, while also recognizing their constructed nature. All of the data work that social scientists, scientists, marketers and others do inevitability rests on a very complicated process of defining and recognizing categories of things in the world.

The Ideological Work of Harmonizing Data

The 2013 OECD Global Science Forum report on New Data for Understanding the Human Condition, which provided the context for the most recent IASSIST meeting, makes explicit both the stakes and amount of work that are required to align infrastructure and data in such a way as to support data driven investigations:

> Many of the research issues we face will require social scientists to work in close collaboration with scientific investigators in other disciplines, notably the biomedical and natural sciences, and across national boundaries. In part these changes arise from the need to adopt a multidisciplinary approach in our search for the causal mechanisms underlying and potential spread of communicable diseases, migration and human responses to climate change. But they also derive from a more 'data driven' approach to scientific investigation. The advances we have made in terms of our ability to generate, capture and re-use information on all aspects of human behaviour places us in a sea of data that has the potential to inform and inspire innovative approaches to scientific investigation. (OECD, 2013: preface)

While the underlying idea of this project is desirable and of critical importance to the future of social science data, what I would like to attempt to articulate is precisely the ways in which this is an ideological project and what that means for such a project. I decidedly do not mean ideological in a negative way. Slavoj Žižek, a contemporary Slovenian philosopher, has argued that to claim that one operates outsides of ideology is the ideological maneuver par excellence (Žižek, 1989). For him there is no outside to ideology and in denying that one has an ideological agenda, one merely attempts to obfuscate and naturalize that ideology. Thus, what I mean by ideology is merely how one relates to the world and what epistemological assumptions allow this relation to the world. So, the point is not the standard leftist ideology critique, but rather to attempt to open the question of what is ideologically and politically at stake in such projects. Furthermore, what is at stake in such attempts to harmonize data across disciplines and countries is not merely the enactment of one universal political-ideology, but rather an attempt to deal with a whole spectrum of different and often times competing national, local and individual political-ideological frameworks.

So, what ideological framework do our attempts to manage data, create repositories, create linked data, share data across disciplines, promote the reuse of data, etc. operate within? These questions can most clearly be answered by turning first to what is required to work with disparate datasets. As the OECD report states, to achieve this goal "data must be comparable across cultures, languages and environments. The concepts used to implement and communicate data at the international level should derive from universally recognized methods and standards" (OECD, 2013: 15). It is worth noting here the recommendation that one use universally recognized standards. While we will return to the notion of universally recognized standards, for the moment the important issue is that in order for data to be comparable and useful across datasets, it must of course be describing the same thing or at the very least a related phenomena. A dataset about zoning likely provides little additional information when combined with a dataset about astronomical objects. In our daily practice as data librarians, data producers and scientists, we are constantly confronted with the incommensurability of datasets and the difficulty of working with datasets that we wished were comparable.

Especially in working with spatial data, one often deals with things that do not coincide even if data users expect that they will. For instance, working with zip codes in the United States in a completely accurate way is a difficult, if not impossible, task. A large number of health and survey related datasets are anonymized and aggregated at the zip code level, since this address level data is often attached to the individual surveys.

Despite this common practice zip codes do not define areas. Rather they are lists of addresses that often change between census years and for unpredictable reasons. While the Census Zip Code Tabulation Areas (ZCTAs) are often a workable stand-in for 'zip code level demographics', it is clear that computational comparison between zip code anonymized data and ZCTAs is not necessarily a direct one-to-one relationship.[3] What the available data describe is determined not by the pure desires of our knowledge nor by the requirements of our research questions, but rather by the accidental exigencies of the postal infrastructure.

Furthermore, similar discrepancies arise, especially when attempting to compare data internationally, not for infrastructural reasons but for more directly socio-political reasons. The Canadian Census definition of a family includes same sex couples whereas the United States definition does not include them, even in states where they are legally married (US Census, 2012). While these discrepancies arise repeatedly in working with data, the most important point is that they are not merely contingent. Despite UN recommendations and other attempts to harmonize these types of data, a family is not at all a given unit. The Canadian definition makes reference to the Canberra report, which distinguishes the nuclear family from the economic family (Statistics Canada, 2013). This notion of an economic family suggests how politicized the naming of a unit of analysis potentially is. One could describe an entire Marxist critique of the 'economic family', but for the time being that will have to be left aside. The point is that the family and questions of how it should function, what constitutes its boundaries, even how children should be socialized, etc. have been the site of political contestations for centuries if not millennia. So, one cannot simply state 'what a family is', without making at least some political-ideological assumptions about the world and how it functions or should function. Thus, it is not only a problem of categorization and the delimitation of categories, but also the difficulty of delimiting the things that are counted themselves; it is a problem of defining the unit of analysis even before the problem of analysis or categorization. The question here is not what type of family is this, but what counts as part of a family. Where, even before one may try to categorize a family, does the family itself begin and end?

At least in the United States, the Census' primary aim is not a sociological one, but is purely political. It is designed to count the population for purposes of political representation and the demographic data that are produced are a secondary result. This political function has affected the way in which individuals are counted from the infamous 3/5th compromise to current debates about where prisoners should be counted. There is a concern now that with such large prison populations in the United States coming from inner-cities and being held in rural areas that there is now a noticeable process of exporting representation (and counts) from inner-cities to elsewhere. At first glance some of these issues may appear accidental or chosen for expediency, but the results they produce are contentious and thus carry political weight even when the initial decision may not have.

Moreover, as we have seen in the United States and to an even greater degree in Canada the entire process of counting itself has become political (perhaps more accurately the always-political nature of counting has gained renewed political attention). Replacing the Census with a voluntary form has seriously called into question the validity and accuracy of the most recent National Household Survey. All of these political/ideological issues intervene to complicate any attempts at harmonization, cross-national studies and long-term comparisons. Furthermore, the complications come not only from directly political questions but also as suggested earlier from a combination of political and infrastructural problems. I think it is worthwhile, in our work with social data, to think of these as two types of classification problems: on the one hand political decisions and on the other infrastructural problems, such as the need to change zip codes to make the postal system more efficient. Of course these two categories overlap, but still they require different strategies to deal with them in terms of data harmonization.

Historical Attempts at Harmonizing the World

Ultimately, all of these complications are not merely technical data issues but rather directly political, ideological and also infrastructural problems. Thus, the work required to overcome or at the very least deal with these issues is then not merely technical work but is political and epistemological work in its own right. It is here where it becomes apparent what ideological position underlies many of the attempts to internationally harmonize data. To suggest this in a larger context: in hoping to harmonize data across and between these political as well as infrastructural differences there exists a universalizing ideology that well predates our current work with data going back to Linnaeus or likely even further to Aristotle. Attempts at international data harmonization can be seen as the most recent iteration of a long set of historical attempts to universally describe the world. While I lack both the space and the general expertise to trace these attempts at universal description historically in any sort of comprehensive manner, it is worthwhile to mention a couple examples from this history to at least begin to situate the type of work data professionals do with data in this larger tradition.

Charles Godfray, Chair of Zoology at Oxford University, summarized the relationship between Linnaeus' work and data well, saying "I like to think Linnaeus faced the first bioinformatics crisis: the problem of organizing information about the increasing number of species that were being discovered in the eighteenth century, and he developed solutions using the best technologies available at the time" (Paterlini, 2007: 814). Linnaeus' attempts to classify the living world provide an early example of these attempts to create a harmonized structure for defining things. While Linnaeus' work focused primarily on the biological, others have attempted even more comprehensive structures to describe the entire world that may more directly resemble modern attempts to describe in rigorous fashion all 'social entities'.

Other philosophers, not to mention bibliographers, scientists, etc. at the time and since have attempted to make universal languages of description with the hope that all information could be knowable, retrievable and computable. More closely to our own time and work, the work of Paul Otlet, a Belgian working on Information Science prior to World War II, is indicative. As part of his major contributions to information science, he developed and advocated for what he called Universal Documentation. He claimed in a 1907 text, "Through its collections and its various repertories [Universal Documentation] would truly become a 'World Memory.' This would not be limited to recording facts, but would automatically and instantly permit their retrieval. It would be a vast intellectual mechanism designed to capture and condense scattered and diffuse information and then to distribute it everywhere it is needed" (Otlet, 1990 [1907]: 110). As Ronald Day has begun to do, one could trace from Paul Otlet through

to current work on Big Data a fascinating variety of attempts to universally describe all knowledge in order to make it instantly retrievable (Day, 2014).

Despite these lofty aims, these universal systems often fall flat. Jorge Luis Borges, in a short essay on John Wilkins, a natural philosopher who attempted, prior to Linneaus, to create a universal language and classificatory scheme, offers a rather humorous and concise criticism of these attempts at the creation of universal systems of classification, citing:

> A certain Chinese encyclopaedia entitled 'Celestial Empire of benevolent Knowledge'. In its remote pages it is written that the animals are divided into: (a) belonging to the emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies. (Borges, 1964[1952]: 103)

Borges' taxonomy instantly suggests the difficulty and arbitrariness of any such classification system. As Michel Foucault says of Borges' taxonomy, "we apprehend in one great leap, the thing that, by means of the fable, is demonstrated as the exotic charm of another system of thought, is the limitation of our own, the stark impossibility of thinking that" (Foucault, 1970: xv). In short what Borges suggests, if one extrapolates slightly, is the impossibility of ever fully accounting for Zip Codes, families or other social categories in a comprehensive and totalizing manner.

Thus, what I hope to suggest is that our attempts to harmonize data and work internationally require coherent and agreed upon systems for naming and delimiting things; this is not a problem that is unique to the current epoch of 'data.' There is a long and fraught history of various universal attempts to name all things, often at least in the early years of such attempts under the belief that the Abrahamic God created a well-organized and knowable world. While some of these attempts have had invaluable impact, especially those that have been limited to certain fields such as biological taxonomy, many in the light of history appear almost comical. Furthermore, while none of these attempts have ever succeeded in creating a completely universal classification system or language, major breakthroughs were made in their pursuits. For instance, Leibniz attempted to create a system for describing and calculating the answer to all questions, including philosophical inquiries, and in the process made a major breakthrough in binary calculation. Likewise, while Otlet's more utopian dreams never came to fruition he made major contributions to information science and practice.

The Work of Data

Returning to the question of ideology, it is now possible to point towards what is at stake in these attempts. All of these attempts to harmonize and create general descriptive languages are founded on a universalizing logic that in its most utopian dimensions believes that these political and infrastructural differences that stand in the way of classification are accidental and can ultimately be overcome. It is an ideology that believes that the world itself can successfully be homogenized and through erasing difference be made completely knowable. It is in short, in our time, a neo-liberal dream of flattening and connecting the entire globe. While most individuals working with social science data rarely, if ever, make such utopian claims, I think it is beneficial to consider the more totalizing historical antecedents to such work. Those who work with data on a daily basis engage in a certain soft-utopianism that is entirely defensible and more often than not productive and beneficial, but placing that work in this larger historical context is helpful for thinking through the opportunities, challenges and risks of that work.

Thus, I do not at all mean to simply deride and criticize the very real and critical work that data professionals do to harmonize and integrate diverse datasets. Rather, I hope to have suggested two things. First, that this sort of normalization of data across datasets, time and place is work. It is labor in a very real and measurable way. It is not simply a process of finding the 'true name' of things or the proper unit of analysis to delimit these things once and for all. It requires constant revision and integration of political change. Second, I would wager, though I of course do not have the data to back up such claims, it is impossible to completely harmonize everything. Additional types of data will be produced faster than anyone can ever deal with them, political differences and infrastructural exigencies will always intervene to guarantee that at the very least the texture and nuances of our data will be lost or at least smoothed away in combining and defining data that have been produced by diverse sources. We will never describe and know everything. Instead we will always be engaged in a continuous process of discovery, rediscovery and translation between heterogeneous and at least partially incompatible places and times. The OECD report referenced above begins by commenting on how poorly anticipated the Arab Spring was and attributes this failure at least in part to the lack of data collected about new modes of communication. While of course it is a wholly worthwhile and valuable endeavor to know the world around us and the results of the Arab Spring are incredibly complicated, I must say that I am heartened by the fact that humanity and the world can still surprise us.

Acknowledgements

I am grateful for the support and feedback from my colleagues at the University of Michigan, especially Nicole Scholtz who is always willing to engage in conversations about the more theoretical implications of the work we do with data. Most importantly, this paper never would have been possible without Kristin Partlo agreeing to explore these issues and present our initial findings together at IASSIST 2014.

References

Borges L. (1964) Other Inquisitions 1937-1952. Translated from the Spanish by Simms R. Austin: University of Texas Press. (Originally published in 1952)

Bowker, G. (2014) The Theory/Data Thing. International Journal of Communication. [Online] 8:1795–1799. Available from: http://ijoc.org/index.php/ijoc/article/view/2190/1156 [Accessed: 27 July 2014]

Day R. (2014). Indexing it All: The Modern Documentary Subsuming of the Subject and its Mediation of the Real. In iConference 2014 Proceedings 565–576. [Online] doi:10.9776/14140. Available from: http://hdl.handle.net/2142/47318 [Accessed: 27 March 2014]

Foucault M. (1994 [1970]) The Order of Things: An Archaeology of the Human Sciences. Translated from the French. New York: Random House. (Originally published 1966)

OECD (2013) New Data for Understanding the Human Condition: International Perspectives. OECD Global Science Forum Report. [Online] Available from: http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.htm [Accessed: 2 June 2014]

Otlet P. (1990) "The Systematic Organization of Documentation and the Development of the International Institute of Bibliography" In: Rayward W. (ed and trans) International Organization and

Dissemination of Knowledge: Selected Essays of Paul Otlet. Amsterdam: Elsevier. (Originally published 1907)
Partlo K. (2014) From Data to the Creation of Meaning Part II: Data Librarian as Translator. IASSIST Quarterly [Online] 38(2). Available from: http://iassistdata.org/iq/issue/38/2. [Accessed 4 March 2015]
Paterlini M. (2007) There Shall be Order. EMBO Reports. [Online] 8(9): 814–816. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1973966/ [Accessed: 3 June 2014]
Statistics Canada (2013) Economic Family. [Online] Available from: http://www.statcan.gc.ca/concepts/definitions/fam-econ-eng.htm [Accessed: 1 June 2014]
United States Census (2012) American Community Survey and Puerto Rico Community Survey 2012 Subject Definitions. [Online] Available from: http://www.census.gov/acs/www/Downloads/data_documentation/SubjectDefinitions/2012_ACSSubjectDefinitions.pdf [Accessed: 1 June 2014]
Žižek S. (1989) The Sublime Object of Ideology. New York: Verso.

NOTES

1. Justin Joque is the Visualization Librarian at the University of Michigan in Ann Arbor, Michigan, USA. He can be reached by email at: joque@umich.edu. This paper was presented at the 2014 IASSIST conference in Toronto, Ontario, Canada on 4 June, Session 3J, along with its companion paper by Kristin Partlo, "From Data to the Creation of Meaning Part II: Data Librarian as Translator."
2. For example: Foucault M(1994 [1970]) The Order of Things. Latour, B (1993). The Pasteurization of France. Harvard University Press.
3. More information about ZCTAs can be found here: https://www.census.gov/geo/reference/zctas.html