Forum: Ph. Reick: On the Bias of Big Data: A Response to Malte Rehbein

by Philipp Reick

In a recent discussion about the significance of structured databases for historical research, Malte Rehbein provided an instructive critique of historians' usage of large datasets.[1] Focusing on a widely acclaimed project by a group of scholars around art historian Maximilian Schich, Rehbein convincingly argues that even the most visually sophisticated chart depends on critical engagement with its data. Schich et al.'s „Charting Culture" draws on birth and death place information of noteworthy persons which the authors collected from the three databases Freebase.com, the General Artist Lexicon, and the Getty Union List of Artist Names. Distinguishing place of birth and place of death by blue and red color, the resulting map highlights migration patterns over two millennia. In the words of the authors, it thus allows us to see a „network of cultural centers connected via birth and death of more than 150,000 notable individuals."[2] Though the short animation that resulted from their research is fascinating to look at, the assumptions drawn here indeed indicate a problematic usage of historical data.[3] In particular Rehbein points out that it is unclear what exactly we learn from a chart that claims to illustrate the evolution of human culture on a global scale over a period of two thousand years if the underlying data stem predominantly from Western sources.

The authors are quite aware that historical datasets are not free of bias. In the supplementary material, they acknowledge several aspects relating to the underrepresentation of regions outside of Europe and North America. What they do not discuss, however, is the fact that their chart relies upon data of *notable* individuals only. This limitation is the reason why „Charting Culture" also draws a socially distorted map of cultural exchange. Illustrated here is the changing cultural attraction of cities and regions for educated, famous, or propertied individuals, yet we learn nothing about the migration of uneducated, ordinary, or unpropertied individuals and their impact on cultural exchange. The chart, in other words, is problematic both with respect to the underlying Eurocentrism of its data as well as the utter neglect of class. This raises the more fundamental question whether historical analyses relying on large databases that are easily available today run the risk of marginalizing lower-class experiences.

For decades social and cultural historians have struggled to shift attention away both from the „great men" as well as the large-scale structural forces that supposedly determine the course of history. In so doing they displayed an enormous creativity in finding new types of sources or in interpreting existing bodies of sources differently. The wish to learn more about the everyday lives of peasants, working women, or slaves was deeply influenced by the hope to restore historical agency to those who were deprived of it. This very agency is threatened by research that relies exclusively on databases which either ignore marginalized groups altogether or which treats them in the same way they had been regarded by the historical data collectors in the first place – that is, as a statistical entity.

Take for instance the correlation of international migration and cultural exchange implied in „Charting Culture". That Schich et al. focus on data about notable individuals cannot be explained by the fact that we generally lack data of „non-notable people" traveling the trans-Atlantic world of the seventeenth and eighteenth century. Meticulously documented in accounting records or logbooks, we do possess data for thousands of slaves who were transported across the ocean. Yet though the international slave trade surely had enormous cultural implication, it cannot be integrated easily into a chart measuring vol-

[1] Forum: M. Rehbein: Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht, in: H-Soz-Kult, 27.11.2015, <http://www.hsozkult.de/debate/id/diskussionen-2905>.

[2] Maximilan Schich et al., A Network Framework of Cultural History, in: Science 345, 6196 (2014), pp. 558–562.

[3] See <http://www.nature.com/nature/videoarchive/charting-culture/index.html> (29.11.2015).

untary migration.[4] For not even records of birth and death are as socially neutral as they might seem. The animation that is based on the findings of Schich et al. for instance informs us that the English-born John Washington, great-grandfather of the first president of the United States, died in the new colony of Virginia.[5] This tells us a lot about the individual experience of Washington himself as well as about the class he belonged to. Given that Washington settled in Virginia voluntarily, we must assume that he was prepared to take the risk of starting a new life abroad – and that he had the financial resources to do so. When, on the other hand, a chart depicts the forced migration of a seventeenth century slave, we learn nothing about her or his desires or socio-economic background. Instead of drawing conclusions about what might have attracted the slave to a particular place, such data primarily allows us to make assumptions about the demand for slave labor or the availability of capital to pay for it. And this discrepancy is inherent in the data itself. After all, such data simply epitomizes the historical marginalization of large groups of people. When Schich et al. conclude that New York City today is „a clear death attractor but gave birth to more notable individuals than it attracted around 1920,"[6] what they are really saying is that the city was apparently less attractive in the 1920s for those people who their crowd-sourced and expert-curated databases regard as „notable." Yet to conclude that New York in 1920 had less cultural attraction would mean to belittle the cultural impact of tens of thousands of African American migrants who might not appear in databases of notable people yet who nevertheless were the driving forces behind the Harlem Renaissance. As long as we are unable to remedy such glaring social bias in our data, we should avoid general assumptions about cultural interaction across time and space as suggested in „Charting Culture."

Although this criticism points towards inequalities inherent in the data itself, much of the social bias in data-based historical research of course stems from discrepancies in the selection of what is digitized and what is not. Think for instance of Pro Quest's „Historical Newspapers", a crucial digital archive for North American newspapers. Providing full text search options for flagship papers like the *New York Times* or *Washington Post*, this widely used database features not one title that catered to a working-class readership, despite the fact that trade union and workers' newspapers mushroomed in the second half of the nineteenth century.[7] Or take its German equivalent, ZEFYS, that provides mostly digital copies but also several full texts of historical newspapers in the German language. Among the 180 digitized titles offered by ZEFYS, users search in vain for papers documenting the rich history of early organized labor, social-democratic politics, or working-class thought in the nineteenth century.[8] Apparently it requires the initiatives of private corporate bodies to digitize fundamental working-class papers such as the German *Vorwärts*.[9] If historical sources of disadvantaged social groups continue to be marginalized precisely in those large databases that provide user-friendly and full text access, then their voices will remain underrepresented in both teaching and research. By implication this means that students today more than ever require training in a broad range of methods and practices, from reading Sütterlin to allocating non-digitized sources. And it is our responsibility to provide this kind of knowledge.

An overview of all contributions to this discussion can be found here: <http://www.hsozkult.de/text/id/texte-2890>.

[4] See Anne Farrow, The Logbooks: Connecticut's Slave Ships and Human Memory, Middletown 2014; Marcus Rediker, The Slave Ship: A Human History, New York, 2007.

[5] See <http://www.nature.com/nature/videoarchive/charting-culture/index.html> (29.11.2015).

[6] Maximilian Schich et al., Supplementary Materials for A Network Framework of Cultural History, in: Science 345, 558 (2014), <http://www.sciencemag.org/content/345/6196/558/suppl/DC1> (29.11.2015).

[7] See <http://www.proquest.com/products-services/pq-hist-news.html> (29.11.2015).

[8] See <http://zefys.staatsbibliothek-berlin.de/list/> (29.11.2015).

[9] See the digitization project by the Friedrich-Ebert-Stiftung that started earlier in 2015, <http://www.vorwaerts.de/artikel/fes-digitalisiert-vorwaerts> (29.11.2015).