Analog information overload: Censuses conducted in the 19th and early 20th centuries generated huge amounts of paper that required manual sorting.

Stacking **Data**

Big data isn't an entirely new phenomenon, as far as historians of science are concerned. Even in the 18th and 19th centuries, scholars, scientists and state authorities collected huge quantities of data, and analyzing all this raw material posed a challenge back then just as it does today. A group led by **Elena Aronova**, **Christine von Oertzen** and **David Sepkoski** at the **Max Planck Institute for the History of Science** in Berlin looks at the methods used in the past – many of them unexpected – and examines how changes in data handling has ultimately brought about changes in science and society.

TEXT **TINA HEIDBORN**

Prussia, mid-19th century: At the census bureau in Berlin, a tabulator reads out the enumeration lists of the current census. The counting staff, seated around a large table, listen attentively; each of them is responsible for a separate category. When the operation is complete, the marks they made in their section of a big interim table form are counted and the resulting numbers noted in a statistical table for publication. This marking process was not only very time consuming, it was also costly and error prone.

Some twenty years later: The scene is a private apartment in the Prenzlauer Berg district of Berlin, where the wife of a statistics employee is tallying up the counting cards of the current census. The cards were delivered in large wooden crates of 5,000 or 10,000 units by the Prussian Statistical Bureau. In this middle-class parlor, they are now being carefully sorted into stacks according to a precisely defined scheme. The housewife has hired a domestic maid to free herself up for this home-based piecework. Along with her two sisters, a brother-in-law, an unemployed trader, two widows and two unmarried young ladies from the neighborhood, she is earning good money helping to evaluate the results of the census. They work more than ten hours a day, seven days a week. For historian

Photo: akg-images / Imagno

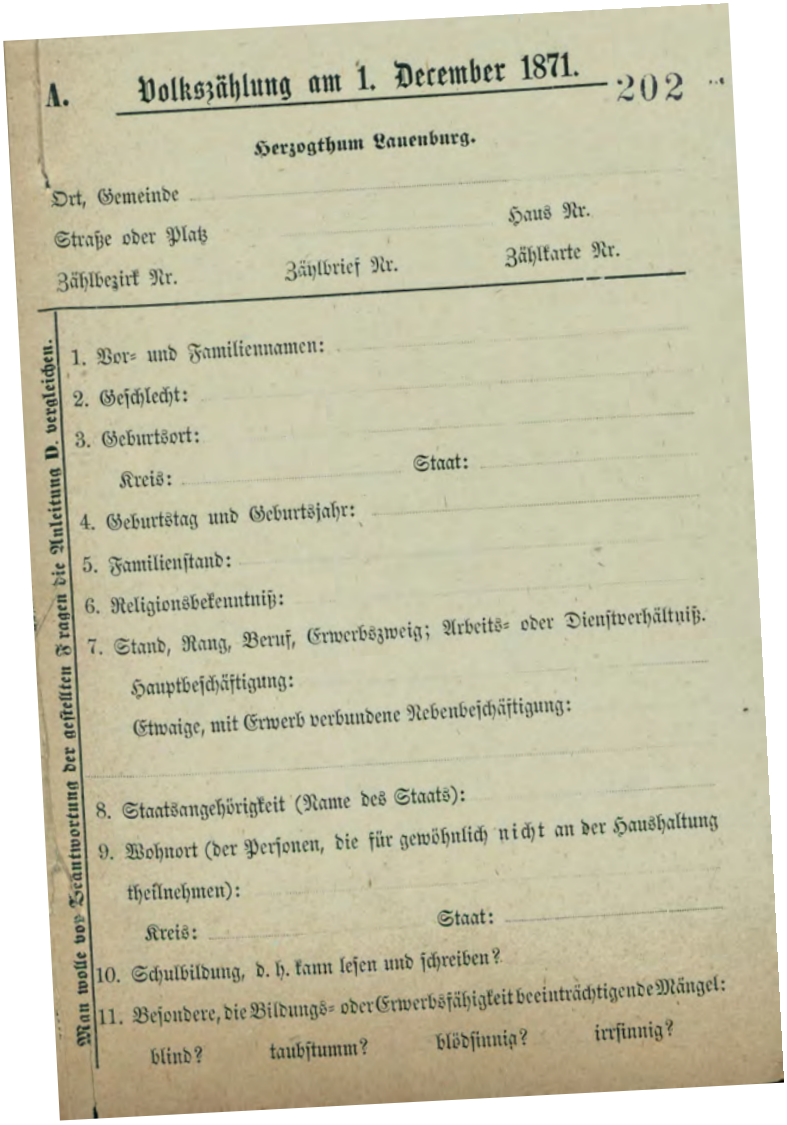
» 19th-century statisticians freed the data from rigid lists – they made data move. This was the beginning of modern data processing.

of science Christine von Oertzen, these two scenes reflect a crucial leap in the history of mass data processing.

"Then as now, the term data was used in very different ways. What is particularly interesting is that the Prussian authorities changed their method of conducting censuses in the 1860s. For the first time, they used a specific concept of data – it appears here in the sources," von Oertzen explains. "The authorities developed a definition of what they understood data to mean." It was Ernst Engel, appointed Director of the Royal Prussian Statistical Bureau in 1860, who established a vital conceptual distinction: he differentiated between the primary data collected in so-called enumeration lists and the processing of this data in tables. As the Director wrote, a table "contains a concentrated result, a summary and groupings of the information drawn from the lists." Engel was one of the leading figures behind the development of population statistics in Europe and, following Italy's example, introduced "counting slips" in Prussia in 1867. These slips made further processing of the gathered data in tables much simpler: the information collected from the enumeration lists was now transferred to the handy little cards, which were vaguely reminiscent of playing cards.

The counting slips provided a new way of accessing the information from the enumeration lists: it was now possible to handle the material in a literal sense. The slips could easily be counted, recounted or stacked and regrouped according to different criteria, so connections could be created between the various items of information from the survey lists. This had been precisely the problem with the marking process: another huge list had to be compiled for every new combination of criteria to be analyzed from the enumeration lists. The counting slips made it possible to correlate data. As Engel wrote in 1868: "The advantage of the counting slip method was that it allowed innumerable combinations of the individual data contained in the slips."

And Engel continued to optimize the method. A short while later he replaced the counting slips with individual counting cards that each respon-


A. Volkszählung am 1. December 1871. 202

Herzogthum Lauenburg.

Ort, Gemeinde

Straße oder Platz Haus Nr.

Zählbezirk Nr. Zählbrief Nr. Zählkarte Nr.

Man wolle vor Beantwortung der gestellten Fragen die Anleitung D. vergleichen.

1. Vor- und Familiennamen:
2. Geschlecht:
3. Geburtsort:
   Kreis: Staat:
4. Geburtstag und Geburtsjahr:
5. Familienstand:
6. Religionsbekenntniß:
7. Stand, Rang, Beruf, Erwerbszweig; Arbeits- oder Dienstverhältniß.
   Hauptbeschäftigung:
   Etwaige, mit Erwerb verbundene Nebenbeschäftigung:
8. Staatsangehörigkeit (Name des Staats):
9. Wohnort (der Personen, die für gewöhnlich nicht an der Haushaltung theilnehmen):
   Kreis: Staat:
10. Schulbildung, d. h. kann lesen und schreiben?
11. Besondere, die Bildungs- oder Erwerbsfähigkeit beeinträchtigende Mängel:
    blind? taubstumm? blödsinnig? irrsinnig?

Inconspicuous revolution: The Prussian counting card introduced in 1871 brought about a fundamental shift in processing census data.

Photo: Secret State Archives Prussian Cultural Heritage Foundation (GStA PK)

Desperately overcrowded: Population growth and mobility led to miserable living conditions, especially in big cities like Berlin. Improved census data analysis revealed such conditions in detail.

dent had to fill out themselves: they were approximately DIN A5 format in size, about four times larger than the counting slips, but equally manageable. On these, the residents of Prussia were required to provide numerous personal details, including age, place of birth, family and professional status, and reading ability. This saved Engel the trouble of having to use enumeration lists, and it did away with the interim stage of manually transferring data to counting slips.

SOCIAL INJUSTICES BECAME VISIBLE FOR THE FIRST TIME

"The Prussian statisticians were delighted at their new-found ability to combine different criteria," says historian Christine von Oertzen. They began analyzing the cards in three count runs, each focusing on several criteria. It was now possible to focus specifically on Catholic women in rural areas, for example, or unmarried Protestant workers in small towns. Being an ambitious statistician, this was precisely Engel's aim: he was in search of methods that not only improved the counting as such, but that also allowed more far-reaching insights to be gleaned from the material. "It's difficult for us today to grasp just what a major improvement this was," says von Oertzen. It was a breakthrough that allowed a previously unknown degree of differentiation in data analysis. "The Prussians wanted the census to provide a snapshot that captured the current situation." For the first time, the census material could be used to scrutinize social problems quantitatively, such as the high child mortality rate. Or else the information was broken down to see where large numbers of people who were not related to each other lived under one roof – another indicator of poverty.

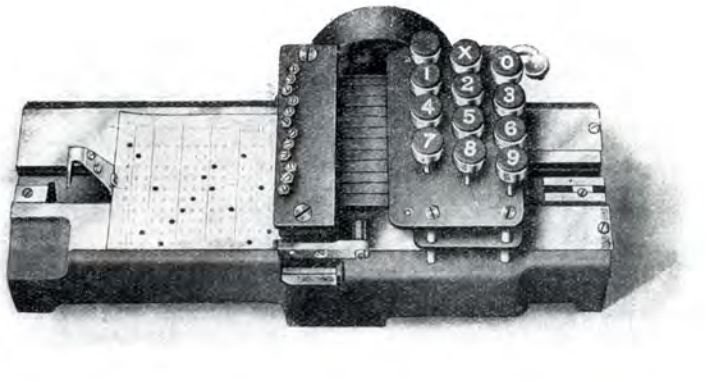
As a historian of science, Christine von Oertzen is particularly interested in the development of technologies and their concrete application. She regards the changeover from lists to maneuverable paper media such as counting slips and cards from 1860 onward as a data processing revolution that has thus far received little attention: "Statisticians freed the data from rigid lists – they made data move. This is what marks the beginning of modern data processing, not the introduction of Hollerith machines and mechanization." In von Oertzen's opinion, the significance of Hollerith's supposedly groundbreaking method is exaggerated.

Herman Hollerith, an engineer, presented his invention at the Paris Exposition in 1889: a method using punch

Photo: akg-images

THE NEW CENSUS OF THE UNITED STATES—THE ELECTRICAL ENUMERATING MECHANISM.—[See page 132.]

Photo: akg-images



Left Hollerith machines – tabulating machines based on punch cards – were first used in the US census of 1890. At the time, this required laborious preparatory work, as all data had to be punched into the cards by hand.

Right It wasn't until later that punch keyboards were sophisticated enough to be operated quickly – as in the case of this model from the 1920s.

Photo: akg-images

cards with machines for sorting and tabulating. It was first used for the American census in 1890. The idea occurred to Hollerith when he saw ticket inspectors in America punching railway tickets as a way of storing information: the ticket was punched in a different place depending on who presented the ticket (to denote the place of boarding, destination, travel class or fare, for example). The advantage of the Hollerith cards was that they could be read by machines, so the counting and sorting process was considerably faster. For the 1890 census, though, the information still had to be punched into the 63 million cards by hand.

DATA ANALYSIS WAS LIKE BRINGING IN THE HARVEST

Some European states, such as Austria-Hungary and the Russian Empire, introduced Hollerith's system right away, and it is generally regarded as a crucial step forward in the history of modern data processing.

At the turn of the century, however, the Prussians believed their own manual method was at least equally effective. As moveable data carriers, the Prussian counting cards were based on the same principle as the Hollerith cards. According to Christine von Oertzen, European statisticians such as Engel established a key cornerstone of the information technology era 20 years prior by introducing the principle of card counting.

The use of slips and cards also allowed the Prussian authorities to outsource the job of data analysis quite literally: it became a task that was typically carried out by women in their homes. The state delegated this work to the wives of its census workers and officials, and held the latter responsible. Hefty wage deductions were put in place to punish slipshod work and thus keep revisions to a minimum. It was during her archival research that von Oertzen came across the unusually large Prenzlauer Berg tabulating team described above. "The data had to be analyzed quickly – it was seasonal labor, like bringing in the harvest," she explains. "We often tend to regard data as not being physical." But when von Oertzen started to dig deeper into the history of Prussian data processing before 1900, the data started to grow "hands and feet," as she puts it – it took on the concrete, tangible form of millions of cards sent back and forth between the census bureau and many private dwellings in Berlin.

Incidentally, the Prussian statisticians were quick to draw attention to the fact that Hollerith machines threatened to take people's jobs away. Ernst Engel's successor Emil Blenck insisted that his agency had a mandate to provide work first and foremost for war veterans – though he conveniently failed to mention that it was no longer impoverished veterans doing most of the work but in fact middle-class housewives.

IMPOSING ORDER ON AN AMBIGUOUS REALITY

While they were busy sorting, stacking and counting census data in their parlors, the women were faced with the fundamental dilemma underlying all data processing: forcing a complex and often ambiguous reality into the supposedly distinct statistical categories provided. In the Prussian census of December 1, 1890, for example, respondents were required to indicate "Kinship or other relationship with the head of the household." The answers not only came in millions of different scripts – some barely legible – but they also covered an enormously diverse range of terms, since people were expected to enter the information in their own words. The women had to

Women often performed the work of transferring data onto punch cards, as here in the US census office in 1908. Piano players were given preference because of their ability to operate the punch keyboard quickly and without errors.

classify the responses into seven categories. The census bureau wanted foster children and pensioners counted in one category, for example, but soldiers, subtenants, and day lodgers – night workers who rented a bed that was unused during the daytime – to be subsumed under different rubrics. "The women were required to sort the cards before counting them. This was an essential operation – anything but mechanical or mindless," says Christine von Oertzen. "It required considerable interpretation and analysis. Diligence and reliability weren't sufficient: the women had to be relatively well educated to be able to classify the information correctly.

BIG DATA DEPENDS ON HUMAN WORK, AS WELL

The census bureau included a sample sheet with model answers, and this shows just how difficult it was to fit the data into the given categories. When it came to the respondent's relationship with the head of household, for example, the statistics were supposed to reflect two separate categories: "Category 2: Servants to the head of household" and "Category 3: Helpmates to the head of household." The examples provided in the instruction sheet stipulated that Category 2 should include rural maidservants, governesses, lady's companions, "household helpers," housekeepers, household support staff and maids, as well as menials and coach drivers, while Category 3 was to include "Workers, house tutors, apprentices and head housekeepers" as well as those fitting such a general description as "in work." Why did those who described themselves as a "housekeeper" fall into Category 2 while others who stated their position as "head housekeeper" fell into Category 3?

"There's this idea that handling data is straightforward because the data itself is self-explanatory – that counting is all that's required, which is a simple task. I believe that's an illusion," says Christine von Oertzen. Her study shows vividly just how much the data collected had to be analyzed and evaluated more than 100 years ago. And today, in the much-vaunted age of big data? "Of course we're interested in continuity and ruptures," says the historian. Despite digitization, there is still a lot of human work involved, she says – even for big data today, at the beginning of the 21st century: the mass of data must be made compatible, and it must be updated and maintained for ongoing use. "These are things we are only too inclined to overlook," says von Oertzen.

IN THE PAST, TOO, QUANTITY WAS DEEMED TO MATTER MOST

And what about the assumption that digital data is a new kind of scientific object, and that computerized data processing represents a new scientific method? "Some people believe that scientific research will be exclusively data-driven in the future," says the researcher. The claim is that science will become a simple matter of using automated algorithms to process huge datasets rather than putting forward hypotheses and testing them. Von Oertzen's study of mass data gathering in the past has tended to make her skeptical of this idea.

The dream of achieving completeness in scientific data gathering – an increasingly widespread vision in the age of big data – is something Christine von Oertzen has also seen before. "In the 19th century there was an enormously enthusiastic belief that data could be used to create a comprehensive record of reality," she says. Scientists in a broad range of disciplines attempted to amass particular data in search of an

Photo: Library of Congress, Prints and Photographs Collection, Washington, D.C.

» There's this idea that handling data is straightforward because the data itself is self-explanatory. That's an illusion.

overall picture – whether in astronomy, linguistics, evolutionary biology or taxonomy. The motto for many research projects back then was: quantity matters most.

Yet this was precisely what caused problems, too. Libraries and scholars used card indexes in their attempt to get a handle on the vastly increasing flood of information. David Sepkoski, co-organizer of the working group, traces the origins of data-driven research in taxonomy and paleontology. He examines how the study of paleontology, which originated in the 19th century, involved the development of classification systems for fossils over a long period of time, and how scientists classified and archived information on extinct species of a bygone era using paper tools to create databases – long before the advent of computers. Paleontologist Heinrich Georg Bronn (1800–1862) drew on existing catalogs and compendia, for example, but reorganized the mass of data they contained. He subjected the data to quantitative analysis by restructuring it according to his own scientific hypotheses, compiling charts and diagrams to illustrate at a glance the emergence, proliferation, diversification and extinction of species. The system he used to reorganize the material on paper was later used as a model for electronic and digital paleontological databases.

In observational disciplines such as astronomy, which had always been oriented toward data collection, the quantity of data exploded in the wake of new technological capabilities such as photographing the night sky or using electronic and ultimately digital super telescopes. As a result, astronomers' work shifted more and more from observing the sky toward merging different data formats to analyze and correlate the collected data in a meaningful way. Sharing and circulating data thus became the core activity of astronomy, transforming the culture of the entire discipline.

DATA TODAY CAN BE DETACHED FROM ITS CONTEXT

Large-scale geophysical data became a veritable exchange currency during the Cold War, as Elena Aronova, co-organizer of the Berlin-based working group, discovered. American and Soviet data centers collected and archived vast masses of data in analog form, but the vision of making this material freely available to scientists in both the East and the West was only partly put into practice, hampered not only by political constraints but also by technological limitations of analog storage media.

What is new in the digital age, according to the group of historians of science in Berlin, is the ability to detach data entirely from its original context. Once collected and digitized, information is no longer limited to a specific location as it was with the data centers in the Cold War: it can be freed entirely from its original context and used in other ways. This is what happened, for example, with medical data collected in the 1990s from the Pima indigenous people of the Gila River Indian Community Reservation in Arizona: the medical data of the members of this tribe of American Indians was originally collected with the consent of the individuals involved, the aim being to study excess weight and diabetic tendencies within the group. This particular collection of data has since become freely available online and is now used mainly to optimize computer-based machine learning. The movement of this data – without the consent of the original subjects – highlights the complicated politics of data mobility in the digital age.

Mapping the world through data raises new issues and has now reached new dimensions as a result of modern-day digitization. However, if we look back at the data practices of the past, we quickly realize just how old the foundations are that shape datification as we know it today. ◂

TO THE POINT

- **Scientists were already collecting large quantities of data in the 18th and 19th centuries in the hope of being able to create a snapshot of reality. Scientific work shifted increasingly toward data analysis.**
- **The Prussian Statistical Bureau revolutionized data processing in the mid-19th century by using counting cards. This enabled data to be combined according to different criteria, thereby revealing unfamiliar interconnections.**