"Raw Data" Is an Oxymoron

Infrastructures Series
edited by Geoffrey Bowker and Paul N. Edwards

Lawrence M. Busch, *Standards: Recipes for Reality*

Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron*

"Raw Data" Is an Oxymoron

Edited by Lisa Gitelman

The MIT Press
Cambridge, Massachusetts
London, England

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Perpetua by Toppan Best-set Premedia Limited, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

••

10 9 8 7 6 5 4 3 2 1

Contents

8 *Data Bite Man: The Work of Sustaining a Long-Term Study*

David Ribes and Steven J. Jackson

Introduction

This chapter makes one basic point: the work of producing, preserving, and sharing data reshapes the organizational, technological, and cultural worlds around them. Data are ephemeral creatures that threaten to become corrupted, lost, or meaningless if not properly cared for. Long ago, data managers moved past speaking in narrow technical terminologies, such as "storage" and "transmission," and turned to a more nuanced vocabulary that included "data preservation," "curation," and "sharing." These terms are drawn from the language of library and archival practice; they speak to the arrangement of people and documents that sustains order and meaning within repositories.

In this chapter, we seek to push beyond the commodity fictions of data[1] that too often characterize and limit studies of data sharing. In particular, we tell stories of data production that reveal the complex assemblage of people, places, documents, and technologies that must be held in place to produce scientific data. The vehicle for our discussion is a distinctive case of long-term ecological research: stream chemistry data in the Baltimore region. We follow the practices of scientists and technicians from the field site to shared online repository as the natural world is translated, step by step, from flowing streams to ordered rows of well-described digital data, readily available for use in science.

The data, things, and people we care about here face particular temporal challenges. Data that stretch across years, decades, and ideally centuries are increasingly important within the ecological and climatological sciences that seek to generate "harder" evidence about longitudinal changes in the environment. Data must be comparable across time and sufficiently well described so as to facilitate integration with other data. Our cases demonstrate the staggering amount of work that goes into the production of information

for scientific purposes. But even more revealing are the mounting constraints scientists face in seeking to preserve data across time and to collect data that year after year continue to stand in for *the same phenomena*.

Conditions are continuously changing, whether environmental, human, or infrastructural. Sites where samples are collected are transformed over the years, becoming polluted by industrial growth and then purified as emission standards take effect. The academic cycle brings in new teams of graduate students (the laborers of scientific data production) and each such change threatens to tweak the delicate rituals of collection. New sensors promise automation and objectivity while subtly changing baseline readings and the accompanying human routines of collection and upkeep. What results is a complicated ontological choreography,[2] as scientists and technicians work to make data "the same" in a changing ecology of technologies, organizations, field sites, and institutional rearrangements.

In this context, data—long-term, comparable, and interoperable—become a sort of actor, shaping and reshaping the social worlds around them. Demanding and fickle, at the slightest change of condition they threaten to cease being useful for the scientific work they were born to accomplish. To bring us to ecological field sites in Baltimore, we first begin with three accounts of ecologies, scientific objects, and data archives that exemplify the ways phenomena shape the social orders that seek produce, manage, and preserve them. These accounts include (1) corn as a world builder, (2) flies that multiply data, and (3) data that threaten to overheat. We then turn to our detailed empirical analysis of production work for a long-term data stream within the ecosciences.

Corn Thrives in Industrial Ecology

As Michael Pollan describes, corn is an unlikely imperialist.[3] The species that has come to dominate global agriculture struggles to survive in the wild precisely for the reason that we humans find it so useful; with row after row of tightly packed kernels inside a thick protective husk, corn is more likely to rot than thrive in the absence of a creature with opposable thumbs to tear open the husk and individually plant the kernels. Even if an ear of corn somehow manages to lose its husk and fall to the soil, hundreds of seeds will sprout, crowd each other out, and die long before the reproductive cycle is complete. Corn, like more and more species then, has thrown its lot in with humans, adapting to the contemporary social world—and especially to industrial agribusiness—with such success that it has pushed nearly all other staple competitors out of business as a cornerstone of our food supply. Like Britain in its heyday, the sun never sets on the empire of corn.

Pollan offers a compelling picture of the trade routes of the corn empire, documenting the production of a raw ear of corn from a farm in Iowa, and then tracing all the steps it takes as it travels to the typical American consumer. We often think of this end product as the "raw" ears of corn that we purchase at the grocery store and imagine that it is shipped to our stores more or less directly from the farmer. But as Pollan describes, most corn enters our kitchen (and our bodies) through a much more circuitous route—losing its rustic form almost as soon as it is pulled from the ground. In our industrial economy, every portion of the plant is systematically stripped, collated, and processed to produce a standardized set of products, including the now-famous high-fructose corn syrup, cornstarch, MSG, maltodextrin, ethanol, and citric acid. Such derivatives are shipped to ranches and factories across the country where they serve as raw material for new products—constituting the basis of a full quarter of American processed food.

If we have domesticated corn, it has just as surely domesticated *us*. As Pollan argues, "It takes a certain kind of eater—an industrial eater—to consume these fractions of corn, and we are, or have evolved into, *that* supremely adapted creature: the eater of processed food."[4] And we are not the only species on the planet that has been so domesticated: in one memorable chapter, Pollan details the heroic efforts required to create the now ubiquitous corn-fed American steer, a particular challenge "since the cow is by nature not a corn eater."[5] Other chapters reveal how our financial system has been reconfigured to handle the deluge of industrialized corn, with new technologies like commodity markets and futures trading developed to support the ever-lengthening pathways between farmer and consumer.[6] Even the American farmer, an archetypal figure of autonomy and self-reliance, has been turned into a factory worker at the service of a commodity—corn—most varieties of which can now not even be *eaten* without substantial industrial processing.

As with corn so too with data . . .

Like corn on the cob that arrives to our grocery stores in conditions resembling its state in the field, we often think of raw data as following straight and commonsensical pathways from collection to database. Sometimes this is true (there are still farmer's markets, after all). However, the more common story—especially in today's "big science" projects—is an increasingly Pollan-esque one, with data moving through complex, multi-institutional networks, sharing more similarities with the production

of industrial corn than the traditional understandings of field or laboratory science. This is in some ways the *ambition* of contemporary "big science" investments: a more complex, dynamic, and commensurable world in which data really *can* flow freely like corn, leaving new systems, processes, and discoveries in their wake. To do this, we must domesticate data: establishing rituals and routines of collection, creating safe pathways for samples to travel, and setting metadata standards to render them comprehensible by others. And in doing so, data increasingly domesticate us.

Flies Dissatisfied with Information System

As historian of science Robert Kohler describes, the fruit fly *Drosophila* (and its most common lab species, D. Melanogaster) was not born as a laboratory animal per se.[7] Already "cosmopolitan," it has cohabited with us in cities for millennia; it is the fruit fly most likely to appear if you were to put a banana out on your window sill and then wait for the larvae to mature. Breeding ferociously in autumn, it is most plentiful at the beginning of the academic year—just in time for a fresh crop of undergraduate, graduate, and faculty experiments. As raw material for students' projects, it is readily available, cheap, easily maintained, and quickly restocked. Thus, there was always already an elective affinity between the labs of urban bioscience and what would become one of its most common objects. Melanogaster helped create a technology of research on which fly researchers came to depend for their professional livelihood. Once inside the lab, the fruit fly took on a new life of its own and came to drive research at paces never before seen in genetics—eventually demanding novel data management and classification strategies.

Scientists first began to use the fruit fly for genetic research in 1901 at Harvard and since then it has become a dominant species in this new ecosystem: the lab. While capable of sleepily surviving the outdoor winter, Drosophila took to the warmth and security of labs with perennial reproduction. Defining an entirely new criterion of fitness, its productivity in this new ecological niche pushed down the traditional species inhabiting the genetic lab: the rat and mouse, the pea and primrose.

One of the foremost early Drosophila scientists, Thomas Morgan, writing of the relentless reproductive productivity of Melanogaster, enthused: "It is wonderful material. They breed all the year round and give a new generation every sixteen days." As time passed, however, he became "overwhelmed with work": "who could have foreseen such a deluge. With various help I have passed one acute stage only I fear to pass on to another." Only months later Morgan declared himself, none too happily, to be "head

over ears in my flies."[8] The problem was not only the reproductive rate of fruit flies, but also their propensity to mutate in response to environmental change—the precise feature that made Drosophila so valuable to the geneticist interested in hereditary features and mutations across generations. Mendel's peas had been docile and well behaved by comparison: they were smooth or shriveled, and followed comparatively clear patterns of generational inheritance—a far cry from the seemingly endless variety of eye colors, wing shapes, and body sizes that emerged in the Drosophila "breeder reactor." In the face of this nineteenth-century data deluge, geneticists "had no choice but to adopt a fundamentally new system of naming and classifying factors."[9] In the lab, Drosophila became a new creature, one that could not exist outside that institution. But, it also reconfigured the lab itself, giving rise to new kinds of scientific places and persons, including "a new variety of experimental biologist, with distinctive repertoires of work and a distinctive culture of production"[10] In Kohler's striking language, experimental biologists became "lords of the fly," and the flies returned the favor.

Data Demand Care

Like Drosophila and Zea Mays, contemporary ecological data may be thought of as an awkward and improbable species that has nevertheless found its perfect ecological niche. Scientific data once fit on a few sheets of paper, which could last centuries if properly stored; now, we have cultivated strains of data so densely compacted they need us to take intricate care of them. As Cory Doctorow describes in a cover article for *Nature*, we have created immense industrial data centers to store and process all this scientific information.[11] In *Welcome to the Petacenter*, Doctorow stands in awe of the hundred-million-dollar computing centers that have been established to store the tens of thousands of terabytes (a terabyte being a thousand gigabytes) of data flowing from dozens of meteorological satellites, hundreds of genomic sequencers, thousands of ecological field sites, and the millions of sensors at the Large Hadron Collider. Just as the *Zea Mays* species of corn would die out in a couple seasons without our assistance, these computing centers would quickly overheat if not for the multistory cooling centers that control the massive quantity of heat they produce. If the primary, secondary, and tertiary cooling systems fail, it would only take ten minutes for the disk drives to bring their environment to a hazardous 42 °C (108 °F)—any hotter and they would begin to crack and break.

These hives of industrialized data storage are potent symbols and key infrastructures for the current era of "big" and "data-driven" science. But, the data center is also just

that, the *center* of a much larger and much more complex network that extends all the way from field sites and laboratories to desktop computers at universities in every corner of the world. Push beyond the chrome exterior of the data center and you will find a squeamish student taking spit samples and delivering them to a genomics lab; scratch the silicon surface and you'll uncover a frustrated field technician recalibrating a vandalized weather monitoring station for the third time that month, or a professor pleading with a county clerk for access to the latest tax assessment records. For, as we will demonstrate, data have domesticated science not only in the sanitized environments of the industrial data center, but also at every stage, moment, and site of scientific activity. In order to support our growing appetite for scientific knowledge, we have entered into a symbiotic relationship with data—remaking our material, technological, geographical, organizational, and social worlds into the kind of environments in which data can flourish.

Behind the Data Archive

This morning I'm working on a paper and I'm looking at data and I'm making graphs, writing this paper and the graphs are swell and the statistical analysis is coming up super well. I nearly went down the hall to thank the lab crew because whenever I do this. . . .You realize how many things have to go right in order to get that graph. I mean, so we had to design the study well but then the samples had to be collected right and then they had to be handled right and they had to be extracted right and then the chemical analysis and the incubation and like, so many . . . I always enjoy that process and I always enjoy realizing how much goes into it in order for it to come out right. So, I think this is an interesting topic.
—Ecoscientist

By turning our attention to the Petacenter we came closer to the invisible infrastructures of data. Technicians, robots, and cooling systems are increasingly hidden in the clouds of computing, laboring to preserve the data of the earth sciences and, agnostically, those of many others. However, the work of sustaining massive repositories reveals only a thin slice in the long chain of coordinated action that stretches back directly to a multitude of local sites and operations through which data in their "raw" form get mined, minted, and produced. What remain at repositories are the distilled products of these field sites; behind these centers lie an even more occluded set of activities that produce those data themselves.

For the remainder of this chapter we focus on a stream chemistry data set of the Baltimore region. Ecoscientists have been collecting these data for thirteen years. Each year their data sets grow. A further trickle adds one more column: 2011, 2012, 2013. . . . Each year these data must be made commensurate with those that came before. This is how such data accrue value for scientific research.

One ecologist draws an analogy between their research approach and the practice of urinalysis during a routine medical exam: "So, you go to the doctor and the doctor samples urine and they can tell something about how the patient, the body, is functioning based on the chemistry of the urine. And if you stress the patient, those stresses are going to be reflected in the chemistry of the urine and it's the same with a watershed." Changes in the environment are reflected in the chemistry of the stream flow. As our scientists like to say: seasons follow annual cycles, but ecological change occurs over decades and centuries. One way to follow such changes and disentangle such processes is through the patient work of building data sets that match those time frames. And one way of studying *that* is to follow a similarly patient approach in observing, studying, and collaborating with the people who do such work.

How to Measure the Same River Twice

For ecologists collecting data, the age-old maxim "you can never step in the same river twice" is not a philosophical reflection, but a practical problem. It is precisely "the same river" that from week to week ecologists wish to take temperature readings and collect water samples. Data only become longitudinal if they measure the same thing week to week and year to year. And yet it is also differences in those field sites over time that are of greatest interest to scientists. When are changes the right kind of changes? And, when are they no longer measuring the same river twice? In this section, we follow the work of scientists, students, and technicians as they each perform this delicate balancing act.

Over the last thirteen years, a lot has changed, much of it beyond the control of the research team. The conditions of possibility for production of data are continuously evolving. Each change threatens to compromise the comparability of data across the years, and thus, the very enterprise of a longitudinal data set. While today our stream-flow scientists take samples from sixteen different collection sites, that number has risen and dropped over the years. Occasionally, entire streambeds have ceased to be rivers at all—drying up as water consumption changes in Baltimore. Other sites have become

unviable because of neighborhood development projects or industrial activity. Sometimes sites are actively damaged or vandalized. Instruments, left behind from week to week, are stolen or left covered in graffiti. Economic conditions in Maryland contribute directly to this. Our scientists report that their instruments can be scavenged for parts or raw materials (such as copper cables). Posing additional challenges are the recent introduction of automated instruments to chemical water sampling and analysis in the Baltimore area: producing end-to-end changes in data routines and requiring months and years of painful and sometimes uncertain calibration work before the new results can be reliably matched to data produced by older techniques.

To understand the mutual construction of data and the everyday work of scientists, as well as the orientation to producing comparable longitudinal data, we cover these difficulties in three sections: (1) the weekly rituals and routines used to generate measurements that make up the database; (2) the field sites and instrumentation that both threaten and comprise the very purpose of the longitudinal study; and (3) those practices that carry data from field sites to the databases themselves. We developed these insights through ethnographic field research and from the accounts of data collectors who themselves characterize their difficulties and the lived work of data collection.

Routines and Ritual: "We Go Out on Wednesdays"

For the last sixteen years, teams of three or four ecoscientists, technicians, and graduate students have set out in a van once per week (most often on a Wednesday) to visit sixteen field sites in Baltimore county. The path is a circuit for the driver, repeated routinely. Sites are streambeds, located at driving distances of fifteen minutes apart to just over an hour. On a summer day in 2011, we join the on-site team for a day, acting as self-identified ethnographers and offering to participate in the manual labor.

The day trip has an easy feel to it, and the accompaniment of ethnographers raises no eyebrows; visitors are common. At the first signs of our expressed curiosity during an interview, the lead technician invited us to join on their trip. Our first scheduled visit to join was postponed because we competed with teachers for the two extra seats in the van. To have visiting scientists, elementary school teachers, or ecologically minded community members join is common and there is a commitment to outreach and education that piggybacks the collection ritual. When our turn comes we are warned only that we will have to be willing to get dirty and help carry a few things.

At 8 a.m. we congregate in the parking lot. Fitting ourselves from the plentiful supply of high rubber boots and spraying ourselves against mosquitoes, we are committed to a day's work. The team includes a lead technician in his mid-forties and two graduate students dressed in casually hip outdoor gear. We help load stacks of plastic bottles and a plentiful supply of fresh coffee into the trucks and head out from the University of Maryland's Baltimore County campus.

At 8:45 a.m. we arrive at our first sampling site. There is nothing in particular that visually distinguishes this first stop as a field site per se: nothing more than a road intersecting a small bridge, and a river swirling below. The driver pulls the van onto the gravelly side of the road and we disembark, bringing with us the necessary equipment: three small plastic bottles, three instruments affixed to the technician's belt, and an invaluable pen and field sheet in hand. The walk from road to river is a worn path through the underbrush and a hop into the streambed. Our boots protect our feet from the light flow of the water. We can still see the road, the passing traffic, and the neighborhood of single-unit homes around us.

The core of the collection ritual is as follows. First, we begin by inspecting a set of worn-looking gauge sticks that our scientists dug into the riverbed long ago. Each metered stick is partially immersed in the stream and a quick glance reveals the height of the water. Today, one stick is out of the water, no longer within the pathway of the river, thereby indicating a need for recalibration. On the field sheet our guides record the height of the water and note the misplaced stick (figure 8.1). Second comes sampling. Each of the three plastic bottles are dunked in the water, emptied, filled again, and then capped. One of the graduate student researchers records a series of matching numbers: one on each of the bottles, and one on the field sheet. The pen travels the short distance from bottle to field sheet, recording a matching number on each. Finally, each instrument is immersed in the stream, only to be emerged with readings for temperature, turbidity, oxygenation, and acidity. At 9:04 am we gather the bottles, check for debris, and climb back into the truck to head to the next site.

Half of the sites are located in the heart of Baltimore, nestled in residential neighborhoods and industrial zones. As the day wears on and we spiral out from the urban core of the city, the landscape and its residents change: from the dense interlocking residential neighborhoods and industrial zones of the downtown core to the lawns and open spaces of more affluent neighborhoods, and eventually to the more pastoral landscapes of state parks. Our collection sites track and reflect these variations. For instance, in the way we typically imagine these things two of the urban sites could not be considered

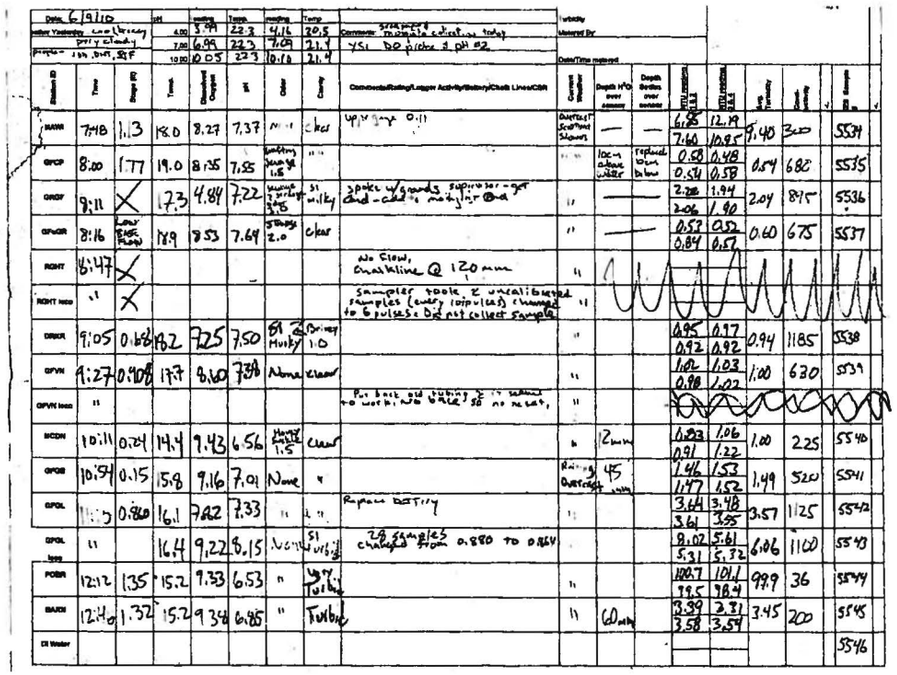

Figure 8.1 A completed field sheet

streams at all: to us they appear as sewage and drainage pipes. In the exurbs, our guides tell us, the water might be filtered down from private septic systems. Visually, very little sets the sampling sites apart from their surrounding urban landscape. Some of the sites are marked by discrete metal boxes containing automated sampling equipment, but to our untrained eyes these pass as electrical infrastructure.

Traveling in a dusty van piled with equipment, our ecoscientist team spends most of the day together—stopping occasionally to collect samples of water, take temperature measurements, and share a meal. The ritual has been repeated thousands of times, but no single practical or material element endures the years: students graduate to faculty, instruments become outdated or imprecise, even buckets wear out. We cannot even say that it is a routinized practice that persists, for that has been modified to fit novel instrumentation, changes in the sites of collection, or the execution of new subprojects in data collection.

We can only claim that the ritual is the same in the sense that we can claim that each time our scientists collect stream data they are stepping in the same river twice. What persists across each iteration of the ritual is the *comparability* of the data collected; it is the purpose and orientation of their activity. This comparability is the achievement of carefully coordinated effort that stretches out every week to the sixteen field sites and back to the labs at the University of Maryland.

In the language of anthropology, what differentiates routine from ritual is the meaning for participants. Routines are dry and mechanical. While routines are always adapting to local circumstance, changes make little difference to those involved. Rituals are lived. They may be enjoyable or tedious, but rituals are experienced as a feature of membership. Rituals tie activities to a past, and through enactment, reproduce that past into the present and future. The continuance of the data set is what sets the activities of these ecoscientists apart from routine.

An Orientation to Comparability

Arrival at each geographic site is initiated with a quick visual inspection for discrepancies: Is anything notably out of place? Are there higher or lower flows of stream water, residues of flooding, evidence of a disturbed sampling machine, graffiti on a bridge, or a strange smell? Such observations become the first raw data collected at each site, qualitatively recorded on the field sheet.

Four artifacts leave each field site: a field sheet and three bottled samples of stream water. The field sheet is a single-sided piece of paper; it begins each week as the same empty chart and ends each collection day with the qualitative and quantitative inscriptions recording observations for each of the sixteen sites. It is the documentary trace of the day's work. The samples of water only become data later; one of these bottles is processed in labs at the University of Maryland at the end of the day while the other two bottles travel to Milbrook, New York, for analysis weeks later. The top of the field sheet is analogous to the start of the collection day: it begins by documenting the date, the data collectors, and qualitative notes about the weather. The next step is calibration of the instruments, checking their accuracy against standardized acid and temperature metrics available in the labs at Maryland.

At the riverbed, the field site itself becomes data in front of our eyes through a practice of observation and a set of practical interventions. Smell is evaluated at each site and recorded on the field sheet: terse but florid descriptions accompany a number between 0 and 4: "pickles and propane, 3," "no smell, 0," "benzene (which is a

whisky-like smell), 4." Samples are collected in small plastic bottles that are first filled and then emptied into the stream to be sampled—thus, clearing any residue from last week's ritual. The practical activities of data and sample collection are technical, but not esoteric. By the fourth site, we visitors were invited to collect temperature samples or hold the field sheet, filling in the called-out measurements. To take a temperature reading, we had to wade into the middle of the water and hold out the thermometer upstream of our bodies.

The routine is simple and quickly learned, but experience teaches one how to manage outliers. If the smell of the field site is toxic ("methane," "chemical," "disgusting"), the site may be evacuated immediately, leaving only that trace recorded on the sheet. A single failed reading in the field sheet (what eventually becomes a blank entry in the database) does not threaten the comparability of the data; it is only over time that this failed reading becomes a concern.

During collection, participants are familiar with aligning a past set of data-capture activities with the circumstances presented to them at each field site; it is a form of standardization oriented to sustain alignment with past measurements. Observation and documentation at each site are focused on detecting changes relevant to the commensurability of past versions of the ritual. Such changes are meaningful in that they simultaneously threaten the data set and promise new developments in knowledge.

Each step in the activities of collection is routine and standardized. In this sense, the steps are mundane. Nevertheless, each activity is conducted with an orientation to comparability and managing situational differences. Differences are judged meaningful through activities of observation and made accountable through discussions between scientists conducted in situ at each site.[12] The database, the full archive of recordings stretching back sixteen years, is not physically present in the ritual. It is even likely that some of the technicians have never so much as glanced at the database. Yet, in the routinized activities of data collection, and in the perceptual orientation generated by the empty boxes of the field sheet, a concern with comparability (with that very database) is fostered.

Shifting Field Sites: Environment, Humans and Infrastructure over Time

For scientists, change in the field sites is the name of the game, but too great a change and these sites cease to be relevant at all. Change is both the source of new knowledge and an incipient threat to the comparability of longitudinal data. Determining whether a site is to be considered threatened is rarely a matter of on-site decision making—it does not occur as an instant in the data collection ritual. Rather, it occurs over a period of many site visits, as the scientists begin to notice a pattern week after week in their collection activities, and as the descriptive metadata on the field sheet pile up. Have environmental conditions systematically challenged data collection? Has new industry created a high point of pollution that cannot be considered representative of the environs? Chemicals are what interest our scientists most, but if a factory is built too close to a sampling site then the data are not generalizable. The environment itself is not considered static—transformations are expected. How do you know whether a change is revealing or compromising? In the section that follows, we step back from our ethnographic immersion to look upon past incidents in the oral history of the data set that have threatened its viability.

Environmental Changes: "There Is No Water to Measure"

In 2002, Baltimore was subject to a record drought. This drought caused visible transformations to the urban and forest ecologies within the county. Our researchers found many of their streambeds completely dried up. With no water to sample and no temperature to measure for months on end, little information was recorded in the field sheets.

The metonymy of "urinalysis" breaks down when the body of the environment cannot be read from its fluids. Stream flow can be reported as a zero, a finding in itself, but with no accompanying samples there is no chemistry to analyze in labs. As such, nothing can be reported at all in those fields of the database. However, the situation reversed radically in 2003 and 2004 due to reports of record moisture and renewed flow in Baltimore streams. The term for this reversal is a "climate pulse." These "pulses" are precisely the kinds of changes our scientists hope to examine in a longitudinal study. A short-term study, months to years, could be ruined by the inability to collect samples, but in a long-term project such pulses became data in analysis that stretched across decades.

Human Changes: A New Sewage System

In 1999, the City of Baltimore Department of Public Works (DPW) entered into a consent decree with the Environmental Protection Agency (EPA) to address sanitary and combined sewer overflows across the entire city. In short, Baltimore has experienced a population collapse over the last few decades. From the perspective of sewage,

this decline presents unique "scaling down" challenges for the city's infrastructure: a system designed for a citizenry of over a million people quickly came to support less than six hundred thousand.

Our stream chemistry researchers have mixed feelings regarding this transformation. On the one hand, urban ecologies are of great interest to them: the effects of "coupled systems" (natural and human changes) is precisely what they seek to study. On the other hand, such large-scale interventions present a virtually uncountable number of variables to manage in their studies: population, demographics for that population, policy and legal interventions leading to engineering overhauls, and of course innumerable changes to the sewage system itself. For some, entire trajectories of investigation had to be scrapped. For others this presented a natural experimental condition: "This, for us, was a great experimental opportunity because we had seven years of background data off these twelve streams and a few of them were very strongly affected by the sewage improvements and a few of them were not."

Considered as an "experiment," the new sewage system provided a unique occasion for a novel study that no other researcher has had the ability to enact. Long-term data stretching before and after a change will open a window of understanding on urban renewal. Many cities in America and around the world are going through a similar process. But, how are these new data to be reconciled as a single longitudinal arc? Scores of variables that were well understood are thrown into a complex flux—making environmental claims difficult for those scientists to assert.

Instruments: Breakdown and Automation

In a longitudinal study instruments come to be part of the field sites themselves. At each of the sixteen sites, meter sticks are strategically placed in the streams. These sticks are dug into the ground on metal poles or affixed to the walls of overpasses. These allow for quick and standardized gauges of the height of the water flow, on each occasion measured from the same location. However, water flows are not static—by their very nature they continuously dig away at their own streambeds. As one scientist noted: "Sometimes our poles stop being in the water at all. That's when we realize that our readings might have been off for some time. That's a pain: we'll have to adjust recent data, and figure out where to put the gauge meters next."

Local residents are sources of consternation, as they interfere with instruments, sometimes in ways that make it difficult to know this even happened. Each site has a rain meter—a small open pipe that fills with rainwater—providing a measurement of rainfall; these "pipes seem to be an irresistible temptation for kids to pee in." There is always the urge to find designs that avoid such human interference or to increase the instruments' precision, thereby reducing labor through automation. But, each such improvement in infrastructure inevitably presents a challenge to the record and, thus, to the sustainability of the long-term endeavor. While a gauge meter is crude and occasionally needs calibration (e.g., reaffixing the stick), it is also very reliable, easily available, and requires very little user expertise. Our scientists are conservative toward their instruments, protocols, and objects of study. They do not add a new chemical test to the repertoire as it becomes available without an assurance that the test will remain available, affordable, and able to keep measuring "the same thing" across the years.

These ecologists fight a three-front war with their closest allies. In order to sustain a comparable archive, data demand the taming of unruly field sites, humans, and infrastructure. A dance of stability and change emerges in an ongoing effort to isolate environmental transformations that can stand in for something broader than a streambed.

Cascading Rituals: From Field to Lab

We dived from the warm petacenter directly to the flowing field site, revealing the background work of creating comparable long-term data. Thus far, we have focused on how participants are oriented to aligning a past data archive with a present practice of data collection. Differences found at any given site—whether environmental, human, or infrastructural—are subjected to a test of relevance: is the ritual of collection threatened, and in turn, will changes present difficulties to comparability with past data? This is an accurate description of the orientation of participants *as they go about the task of collecting data and samples on site*; nonetheless, it leaves all the space between the archive and the field site unaccounted for.

Only a small portion of what comes to be considered raw data is actually generated in the field. Specialized tools, such as thermometers and meter sticks generate quantitative results, such as temperature and gauge, while, moderate training of the senses produces qualitative perceptions, such as smells. With only these mediators, our ecoscientists have for over a decade transformed the natural world into data on site. Bruno Latour has described these as the moments whereby matter becomes form, where some of the materiality of streams is sloughed off in favor of greater mobility of data.[13] Transformed into numbers and writing on field sheets, these mobile facts can be easily ported

back for data entry. But, the greater part of *becoming raw data* only comes to pass further down a chain of mediations: in the laboratories of Maryland or Milbrook.

Producing those data means isolating and transporting little bits of streams back to labs in ways that preserve meaningful relationships to those streams. These bits are called "samples": a straightforward term that belies the work that meaningfully sustains them as representing a stream at a particular point in time. How is the relationship between a field site and its sample preserved? This is a mundane question, as are the methods used to transport and coordinate those samples. On the right side of the field sheet is a number that facilitates the movement of the sample from field site to lab: the "sample #." The same number is placed on the sample bottles: once on the bottle itself and once on the lid. It is this number that holds together the relationship among a date, a field site, and the bits of water that are trucked away.

Many of these bits of river continue on for years as samples rather than data, preserved in massive cold rooms in the basements of Cary Institute of Ecosystem Studies in Milbrook. Every month, the samples for several weeks of ritualized work are loaded onto a chilled truck that travels from Baltimore to Milbrook. The samples are then carefully ported to these cold rooms. This physical archive preserves samples of Baltimore streams, stretching back almost two decades, along with bits of other bodies of water that go back even further. In the face of a new laboratory technique or scientific question, these samples could be used to regenerate an entire new data stream stretching back as long as the numerical archive. It is this simple alignment, a number on a field sheet matched to two on a bottle, that make accountable the representativeness of each sample for decades, and possibly centuries. If this simple numbering ritual fails, so too does the chain that connects the field to the lab: "That's one long-term study we haven't done. The life of the label glue over time. I dread to think that one day we'll walk in to the cold storage room and hundreds of labels will be lying scattered beside the bottles. But I think the extra label tape we put on the lids will hold up."

These sample numbers are a kind of data that never make it into the final archive. They are used to coordinate the movement of samples across physical and temporal distances, after which they are discarded. We could call these numbers procedural metadata. Metadata, as the meme goes, are data that describe data. Usually, we think of these as contextual information: the date, time, and location of a measurement, who took the reading, when was the instrument last calibrated. These kinds of metadata can be used to understand and evaluate data at later points in time, or used by those unfamiliar with the collection rituals. But, procedural metadata serve only in the interim

periods, as samples are transported. A routinized check comparing samples to field sheets occurs at each end of each trip: from field site to Baltimore labs, or from Maryland to Milbrook.

There are others bits of the river that are shed along the way, never making it to the main database. For instance, conductivity is another measure taken right at the field site. This measure is recorded on the field sheets in the second to last column, but it travels no further along the chain. Conductivity accompanies our scientists back to the lab, but there it is forgotten, or rather, buried in a mound of archived field sheets: "It is possible to go back to these data sheets if necessary, but they would have to dig." When we asked why these data points made it no further, the glib response was simply that no entry existed for conductivity in the database. Meanwhile, well-worn columns of the database were filled with qualitative observations about smells and random field events.

This labeling ritual, the notations on samples, the checks at each point of transport, are the cascades of rituals that tie together field sites to samples to databases. We have only scratched the surface of these events. Our scientists described how samples are placed in the car just to prevent overturning. Bottles, whether filled or empty, are transported with sealed lids to prevent cross-contamination. Shifting the contents of water from one bottle to another or to an instrument involves isolation from other samples to ensure none are confused. We observed cascades of rituals, from the moment of a sample's collection in the river to its placement in the lab refrigerator.

At each of these tiny transitions data again threaten to become unruly masses. A misinscribed sample number, a confusion of two bottles, or the spilling of a sample during filtration can all threaten the chain that links a date and a field site to a sample and its eventual transformation into data. A myriad of ritualized activities seek to solidify this chain, but small mistakes and accidents still occur. At best, a mistake or accident is caught and a data point is lost. While a single data point is a loss, in the grand scheme of a longitudinal database, it is a fairly small one. Scientists who use these data expect such things: anomalies and outliers that must be thrown out, missing data points that can be interpolated or extrapolated. At worst, a mistake becomes systematic (as with a misplaced gauge stick), whereby entire sections of a data stream must be reconstructed or altogether thrown out.

The metaphor of a chain is revealing: it helps us understand the heterogeneous work of custodianship stretching from field site to lab and from lab to databases. Only at the end of these mediations can we meaningfully speak of "raw data." Nevertheless, the

chain is also obfuscating—implying a beginning and end for data. The archive for this chemistry stream flow is not singular; rather, it is quite literally distributed across databases, field sheets, and a physical archive of samples. When asked "Where is the stream data archive?," a researcher will first insistently direct us to a public online page with an embedded web service. But thereafter, with only a little further prodding from the interviewer, the database becomes multimedia: it is digital; it is paper and pen; it is water. The digital database is its public face, accessible around the world. It signals the presence of an archive to scientists. In the field sheets, salinity data remains silently annotated, awaiting their analyst, and in the water archive, a promise of future discoveries. All three of these, and multiple other components, make up the archive of raw data.

Conclusion

We tell ourselves that we live in an era of aggregation and automation. From this perspective, raw data patiently await assembly: potable water, environmental damage, or climate change? Click. Shuttled from data storage to a computing center, the analytical engines of the twenty-first century assemble statistics, graphs, and ever more clever visualizations in response to these and many other questions we have not yet thought to ask.

But there are stories *behind* these stories. What we have offered here is another narrative, one of temperamental and delicate creatures, whose existence and fraternity with one another depend on a complex assemblage of people, instruments, and practices dedicated to their production, management, and care. Like corn and flies before them, data demand and build the human, organizational, and infrastructural worlds around them—enforcing a burden of care and work that disappears beneath (but ultimately, constitutes) the futuristic possibilities of the petacenter. Where then does raw data begin and end? If such a clear and objective dividing line exists, we have not yet found it.

We have cast corn, flies, and petacenters as the surprising conquerors of their environments, demanding suitable treatment from their human coinhabitants. In comparison, the practical collection of water samples may seem local and mundane, but it is at this level of granularity that data exert a continuous force, on a weekly basis, requiring a new round of collection and care. It is these local collection and data practices, combined with similar practices around the world that make seeing large-scale and long-term phenomena possible. They are the very stuff of global knowledge. With only a little push from the interviewer our informants agree. A scientist worth his or her mettle never stops their investigation at publically available data. This is not where the archives of raw data reside. Rather, the archive's borders stretch to a receding horizon that include the pen and paper field sheets backfiled for years, a cold room of samples, and the uncaptured experience of scientists and technicians entrusted with the production of the archive. The field sheets are not paper relics, but rather, a source of data awaiting their user. The samples are a promise of a renovated archive, in line with the newest analytical techniques. The field techs standing knee-deep in the pickle-smelling waters of exurban Baltimore are not the opposite of global knowledge—they are *participants* in its assembly.

Acknowledgments

We would like to thank our collaborators in field research and conceptualization of this chapter: Stuart Geiger and Mathew Burton. We also thank Jessica Beth Polk for her careful and observant eyes.

Notes

1. This is a term we develop by analogy to Foucault's "commodity fiction of power," in which power is modeled as an undifferentiated force, reserve, or entity separable from specific contexts of action. The commodity fiction of data performs the same trick on data, assuming or projecting a world where data floats free of its origins, shedding form, substance, and history, and is thereby rendered free to travel the world as an undifferentiated and universal currency; but as the chapters in this volume make clear, data is stickier than that (Michel Foucault, "Two Lectures," in *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*, ed. Colin Gordon [New York: Pantheon Books, 1980]).

2. C. Thompson, *Making Parents: The Ontological Choreography of Reproductive Technologies* (Cambridge, MA: MIT Press, 2005).

3. M. Pollan, *The Omnivore's Dilemma* (New York: Penguin Books, 2006).

4. Ibid., 90.

5. Ibid., 64.

6. W. Cronon, *Nature's Metropolis: Chicago and the Great West* (New York, London: W. W. Norton & Company, 1991).

7. R. E. Kohler, *Lords of the Fly: Drosophila Genetics and the Experimental Life* (Chicago: University of Chicago Press, 1994).

8. Ibid., 47–48.

9. Ibid., 61.

10. Ibid., 49.

11. C. Doctorow, "Big Data: Welcome to the Petacentre," *Nature* 455, no. 7209 (September 4, 2008): 16–21.

12. C. Goodwin, "Professional Vision," *American Anthropologist* 96, no. 3 (1994): 606–633.

13. B. Latour, *Pandora's Hope: Essays on the Reality of Science Studies* (Cambridge, MA: Harvard University Press, 1999), see chap. 2.

Data Flakes: An Afterword to "Raw Data" Is an Oxymoron

Geoffrey C. Bowker

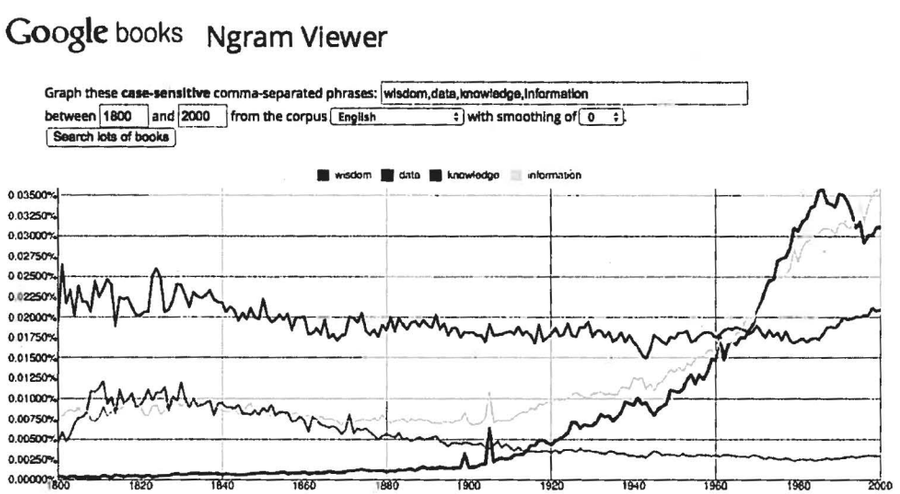

Figure 9.1 Google Ngram (http://books.google.com/ngrams) showing data and information on historic rise and knowledge and wisdom in historic decline. *Note:* Information and data peak in the late twentieth century (data is the darker line); wisdom and knowledge are in gradual decline.

Let me be hyperbolic and assert that we are entering into the dataverse. "Entering" is a key word here—it is through the labors of millions of sensors, click-workers and of course our collective selves that we are being entered.

It has been a longer-term process than most would have thought, before they read this marvelous volume. It has also been ineluctable. Harry Harrison imagined "the stainless steel rat" who could continue to swarm in worlds of concrete, glass, and cameras