COMMENT · 15 OCTOBER 2019

Data — from objects to assets

How did data get so big? Through political, social and economic interests, shows Sabina Leonelli, in the fourth essay on how the past 150 years have shaped the science system, marking *Nature*'s anniversary.

Sabina Leonelli

Illustration by Señor Salme

Data. The confusingly plural cornerstone of research. The grounding for a scientific understanding of the world. Lightning rods for the negotiation of political, social and economic interests.

Over the past 150 years, ideas have shifted drastically as to what counts as data, which data are reliable and who owns them. Once regarded as stable objects whose significance was determined by a handful of professional interpreters, data are now reusable goods. Their mettle depends on the extent to which they are mobilized across contexts and aggregated with others. Growing in volume, variety and value, data have come to drive the very process of discovery.

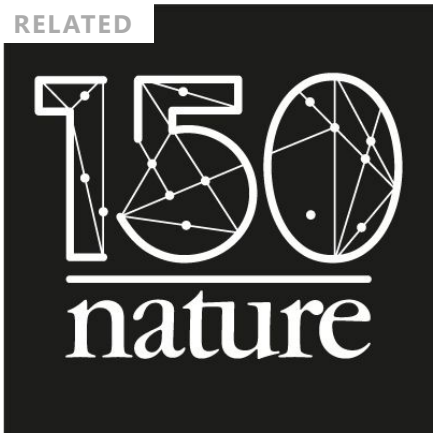
This explicit designation as assets has become possible only through a complex web of institutional, technological and economic developments. The history and consequences of how this web has been woven have repeatedly transformed research and its role in society.

Collecting commodities

Until the start of nineteenth century, efforts to collect facts and objects of study were spearheaded by visionary individuals, typically backed by wealthy patrons. Naturalists roamed the globe in search of biological specimens that were new to science. Court astronomers devised tools to observe new parts of the cosmos. The large quantities of data accumulated were systematized and analysed through simple and powerful models (think Kepler's laws) and classification systems (such as that developed by botanist Carl Linnaeus). Thus was born the myth of the heroic theoretician, mining order from the chaos of observations. This individualistic view was tied to an understanding of data as fundamentally private — their scientific value residing in conceptual interpretation.

The nineteenth century marked a shift. Data, as we now recognize them, became institutionalized as social commodities. Their intellectual, financial and political worth arose from investments, requiring regulation and oversight. The botanical wonder cabinet that was Paris's natural-history museum was reorganized as a world-leading, publicly accessible repository of objects of potential scientific value. By the 1850s, the natural-history museums of Berlin, London and New York City followed suit.

RELATED



The centralization of food markets spawned standardized approaches to the valuation and trade of organisms — such as the crop measures devised by the Chicago Board of Trade in Illinois. Cholera epidemics in Europe spurred large-scale collection of information on the spread and targets of disease. New methods of visualization and analysis emerged, such as physician John Snow's famous maps of how contaminated water spread cholera in central London.

National weather services started to build links between data collected regionally. The 1853 Brussels Convention on naval meteorology coordinated ships' logbooks into the first quasi-global data records for climate science. In Berlin, the first real bureau of standards, the Physikalische-Technische Reichsanstalt, was inaugurated in 1887 with physicist Hermann von Helmholtz as its founding director and a mandate to generate data needed for society as a whole. In the meantime, the US Army tasked the Library of the Surgeon-General's Office with collecting as many disease case reports as possible. Within 30 years, it had become the largest medical library in the world.

National treasures

By the turn of the twentieth century, the rise of nation states and the increasing demands of international trade drove initiatives to measure nature and society in a more systematic, objective way. National information infrastructures helped regions to share data, marking the start of a new informational globalism[1]. International entities, such as the League of Nations and the International Monetary Fund, yearned to globalize data collection and analysis for many purposes and across all scientific domains.

For example, the League of Nations Health Organization created the Permanent Commission on Biological Standardisation to monitor drug tests and biological assays from 1924. Well before the Second World War, there was increasing momentum to share information on employment, unemployment, wages and migration; from 1947, these data were amassed by the new International Statistical Commission. Such initiatives were fostered by an ever-expanding cadre of researchers, administrators, merchants and politicians.

All this fuelled the development of sophisticated approaches to quantification. Statistics emerged as a separate discipline – the main source of information for emerging insurance practices and public-health monitoring systems[2,3]. Techniques were developed to match the complexity of social exercises such as the census[4]. Population-level thinking gripped the life sciences, too – for good (genetics) and ill (eugenics). A new type of data collection focused on genetic mutants of a single model species[5,6], such as the fruit fly.

A selection of histological slides of lung biopsies

Microscopy slides used in the first detailed UK report of a link between lung cancer and asbestos. Credit: SSPL/Getty

The two world wars severely disrupted data collection and sharing in the short term. But from the 1940s, the huge military investment in intelligence and information technologies kick-started the drive towards mechanized computing. The space race was perhaps the most notable cold-war contribution to globalized data systems and practices, particularly satellite technology. This produced the first global view of the planet and spurred the inauguration of the Intelsat system for worldwide civil-communications networks in the 1960s.

The World Meteorological Organization was founded in 1950 to oversee the international linkage of regional weather services, for instance in the Global Atmospheric Research Program. The International Geophysical Year of 1957–58 marked a step change in the commitment of Earth sciences to global data exchange, and was a diplomatic achievement in the middle of the cold war[7].

ESSAY SERIES: LESSONS FROM THE PAST FOR THE FUTURE OF RESEARCH

Read more of this collection published to mark *Nature*'s 150th anniversary, in which leading historians explore how the past century and a half has forged some of the defining features of today's scientific system.

We ignore the past at our peril

Government: Discovery is always political

China: How science made a superpower

Identity: How advances have repeatedly changed who we think we are

Data: From objects to assets

Can marketplace science be trusted?

Ethical research: the long and bumpy road from shirked to shared

Science must move with the times

Global goods

From the 1970s, almost every scientific field was building global, digitalized infrastructures for data sharing. The United Nations consolidated its global environmental monitoring system just

as the World Health Organization systematized its efforts to map the spread of infectious diseases. The holy grail became the development of tools, such as computer models, that could crunch numbers at a previously unimaginable scale.

Increasingly, data were seen as sharable assets for repurposing, the value of which could change depending on their use. This view owed much to the cybernetics movement, with its emphasis on modularity and complexity[8]. Once again, the shifting role of data was also informed by the growth of international trade and the rising recognition of research as an engine of economic growth, military power and international relations.

Also in the 1970s, big science such as studies of particle collisions at Los Alamos National Laboratory in New Mexico and at CERN, Europe's particle-physics lab near Geneva, Switzerland, took centre stage. Here, the production and trade of data were no longer the responsibility of individual researchers. Rather, they were the output of large investment and collective efforts performed in centralized experimental facilities. Such centralization was unfeasible in many fields, for instance in environmental, biological and climate sciences, which work with observational rather than experimental data. Yet even those disciplines were focused on building networks for sharing information so it could be fed into new computational tools.

Hollerith data machine in an office with employees, 1963

A Hollerith data machine at a steel works in Sheffield, UK, in 1963. The electromechanical device helped workers to tabulate statistics stored on punch cards. Credit: Paul Walters Worldwide Photography Ltd/Heritage Images/Getty

Since the 1980s, portable computers, modelling and simulations have shaped data collection, manipulation and archiving. Climate scientists have developed ways to use legacy records to reconstruct a history of the atmosphere at the global level. This effort drove the pooling of international data, culminating in 1992 in the Global Climate Observing System.

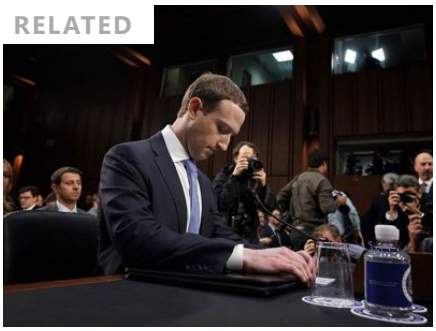
In biology, the quest to map moved to the molecular level with big genetic sequencing projects, first in model organisms such as the nematode worm *Caenorhabditis elegans*, then through the Human Genome Project[9]. Sequencing databases were reimagined as playgrounds for discovery to facilitate immediate sharing, visualization and analysis online at a low cost, transforming the massive investment in genomic data production into useful knowledge.

Open season

As global data infrastructures and related institutions burgeoned, the resources needed to maintain them have mushroomed, and in ways that do not fit contemporary regimes of funding, credit and communication. For example, the curators of biological databases do essential work. But they do not routinely publish in top-ranking journals and might not be recognized or rewarded as high-level researchers. Similarly, keeping digital platforms robust and fit for purpose requires serious investment. The more data move around and are repurposed, the more vulnerable they are to unwarranted and even misleading forms of manipulation.

Over the past few decades, the Open Science movement has called for widespread data sharing as fundamental to better research. This has prompted several changes. One is the birth of journals devoted largely to the publication of data sets. Another is ambitious investment in data infrastructures, exemplified by the European Open Science Cloud. And the FAIR guidelines were crafted for how data should be labelled and managed to make them reusable[10]. There have also been calls to improve rewards for data stewards (such as technicians, archivists and curators), to raise their professional status from support workers to knowledge creators[11].

RELATED



We ignore the past at our peril

These reforms are temporary solutions to a large-scale crisis of the contemporary research system, rooted in the inability to reconcile the diverse social and scientific aspects of data. The crisis recalls how the twentieth century reconfigured research data as political and economic assets. Their ownership can confer and signal power, and their release can constitute a security threat — as in the cold-war efforts to contain geological data that could have signalled nuclear testing. Now, new technologies are intersecting with emerging regimes of data ownership and trade. Starting from the 2000s, a handful of corporations has created — and wielded control over — new kinds of data left by billions of people as they meet, work, play, shop and interact online. (Think Amazon and Google.)

As algorithms become ever more opaque, the transparency and accountability of techniques and tools used to interpret data are declining. Whereas data curators remain the Cinderellas of academia, those who understand and control data management have climbed company ranks. And concerns are growing around data property rights, especially in the wake of misuses of personal data by the likes of Facebook and the UK company Cambridge Analytica.

Such tensions between data as public goods and private commodities have long shaped practices and technologies. Consider, for instance, the acrimonious debate over the ownership

and dissemination of genomic data in the 1990s. On that occasion, free sharing won out through the establishment of the Bermuda Rules — an agreement among publicly funded researchers to deposit their sequences in public databases as soon as possible[12]. Wildly successful, this paved the way for open-data practices in other fields. Yet it also emphasized the financial advantages of owning genomic data[13,14] — a lesson swiftly learnt by companies that sequence and claim to interpret clients' genomes, which typically retain and use such data. Another example is the vast number of patents being filed for synthetic organisms by the chemical industries.

Denise Harwood diagnoses an overheated CPU at the Google Data Center

Banks of servers at one of Google's US data centres. Credit: Connie Zhou/Google/ZUMA Press

Value added

The use of big data as input for artificial-intelligence systems relies on the promise of global, comprehensive, easily available data riches. In principle, the marriage of powerful analytical tools with big biological data can support personalized medicine and precision agriculture. Similarly, social data hoovered up from Internet platforms and social-media services can inform evidence-based policy, business strategies and education. Yet history shows that moving research data around is not so simple. Underpinning technical questions around integration and use are thorny social, ethical and semantic issues.

How can different research cultures be encouraged to communicate effectively? What is the best way to collect, share and interpret data generated by the state, industry or social media? Which experts and stakeholders should have a say in data management and analysis? Who should have access to what, when and how? Addressing these issues requires effective administration and monitoring, and a long-term vision of the research domain at hand[15,16]. It also demands a repertoire of skills, methods and institutions geared to the study of specific research objects[17].

In summary, data generation, processing and analysis are unavoidably value-laden. The scientific legitimacy of these activities depends on the extent to which such values are held up for public scrutiny. Indeed, the best examples of data-intensive research to this day include strategies and methods to explicitly account for the choices made during data collection, storage, dissemination and analysis.

Model-organism databases such as PomBase (for the fission yeast *Schizosaccharomyces pombe*) and FlyBase (for *Drosophila*), for instance, clearly signal the provenance of what they store, including information about who created the data, for what purpose and under which experimental circumstances. Users can then assess the quality and significance of data[18]. Similarly, the Catalogue of Somatic Mutations in Cancer (COSMIC) captures the provenance of its holdings and the interpretive decisions taken by its curators while processing them. This helps clinicians to reassess the value of the information[19].

The more such assumptions and judgement are filtered by large digital infrastructures, the easier it becomes to hide or lose them, making it impossible for future generations to situate the data adequately. Data are cultural artefacts whose significance is clear only once their provenance — and subsequent processing — is known.

Technological development, particularly digitization, has revolutionized the production, methods, dissemination, aims, players and role of science. Just as important, however, are the broad shifts in the processes, rules and institutions that have determined who does what, under which conditions and why. Governance, in a word. Data emerge from this reading of history as relational objects, the very identity of which as sources of evidence — let alone their significance and interpretation — depends on the interests, goals and motives of the people involved, and their institutional and financial context. Extracting knowledge from data is not a neutral act.

Building robust records of the judgements baked into data systems, supplemented by explicit reflections on whom they represent, include or exclude will enhance the accountability of future uses of data. It also helps to bring questions of value to the heart of research, rather than pretending that they are external to the scientific process, as has arguably happened in bioethics[20]. This is a crucial step towards making big-data sciences into reliable allies for tackling the grave social and environmental challenges of the twenty-first century.

Nature **574**, 317-320 (2019)

doi: 10.1038/d41586-019-03062-w

References

1. Hewson, M. in *Approaches to Global Governance Theory* (eds Hewson, M. & Sinclair, T. J.) Ch. 5 (State Univ. New York Press, 1999)

show more

Nature ISSN 1476-4687 (online)