Vol. 17: 37–40, 2017
https://doi.org/10.3354/esep00178

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Published September 1

Contribution to the Theme Section 'The ethics and practice of openness in life sciences data'

Big data and the emergence of new 'dissipative' structures

Daniel Pauly*

Institute for the Ocean and Fisheries, University of British Columbia, Vancouver, BC, Canada

ABSTRACT: This essay suggests that humanity has experienced several instances where lots of information ('big data') had to be accommodated, which led to new structures for channeling the subsequent data flows. These structures, such as articulated speech and writing, would be analogs to the 'dissipative structures' that emerge in physical systems characterized by strong energy (i.e. heat) gradients. Additional examples from oceanography, meteorology and ecology are given, with some emphasis on the prescient work of Alexander von Humboldt, whose identification of ecological communities was based on the occurrence records of multiple species. His lead was initially not followed up, but it can be now, as millions of occurrence records are available, along with the technology to manipulate them. The structures that will emerge in the process, however, are as unpredictable as dissipative structures in physical systems.

KEY WORDS: Information transfer · Language · Data sharing · Humboldt · Ecology · Occurrence records · Aquamaps

The term 'big data' was essentially unknown prior to the 1960s (Fig. 1), although several scientific disciplines were then already blossoming that produced and used huge quantities of data, and had addressed, or even resolved, the associated issue of data sharing.

It appears that the driver for the emergence of human speech was the need to keep track of social interactions in increasingly large groups of people (Dunbar 1998), and that the driver for the emergence of writing was the need to keep track of increasing numbers of commercial transactions (Lieberman 1980). Thus, it can be argued that language and writing were new structures created both for and by massive information transfers.

Similarly, the expansion of the European horizons in the Age of Discovery required a method to name the many animals and plants that were brought back from far away, beyond the 500 or so taxa that 'folk taxonomies' can usually handle (Berlin 1992), and thus the importance of the Linnaeus (1758) binomial and hierarchical system, which accommodated (and still does) an ever increasing terrestrial and oceanic biodiversity, in spite of various challenges (Boero 2010).

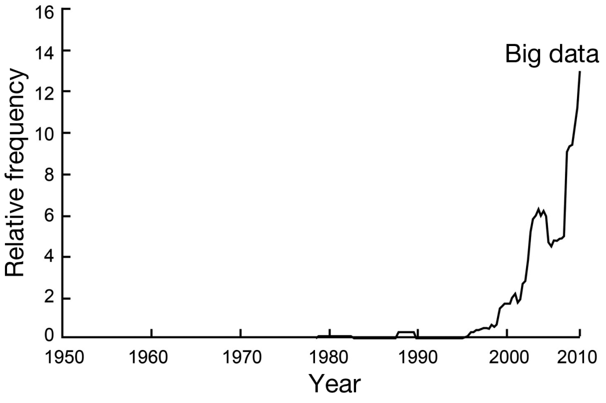

Fig. 1. Relative number of occurrences of the term 'big data' (including 'Big Data' and 'big data') in millions of books scanned by Google (see https://books.google.com/ngrams)

Thus, the pattern may be that more data or information lead to new structures to assimilate these data or express the information they contain.

This pattern may hold with texts incorporating lots of ideas, e.g. encyclopedias. One of these, the massive *Encyclopédie* of Diderot and D'Alembert

*Corresponding author: d.pauly@oceans.ubc.ca

Publisher: Inter-Research · www.int-res.com

(1751–1772), by summarizing the knowledge and aspiration of its time, can be safely credited with driving France toward new political structures expressing the aspiration of its people, rather than the presumptuous notions of its parasitic nobility (Roche 2006).

In biology, the advances enabled by the Linnean system allowed Charles Darwin to successfully complete the research program that he identified in 1838 ('Why do organism vary?') after his return from the voyage on the 'HMS Beagle', based also on the immense database of biological information he had acquired during his voyage and would expand in his lifetime. He won the day because biologists found a way to express their data through structures he had discovered, e.g. evolutionary trees.

On the other hand, Alexander von Humboldt — a hero to young Darwin — did not win the day, despite his fame in his lifetime. His brilliant understanding of ecology was too far ahead of his time and his major invention, ecological transects, was not followed up on.

What would have been needed for Humboldt's invention to win the day? Presumably, it would have required the naturalists who succeeded him to fully appreciate the idea that 'naked' occurrence records, each consisting of only (1) a species (scientific) name at (2) a given time (year, month, date) and (3) a location (as defined by a latitude and a longitude) are all that is needed to be able to perform a vast number of analyses on the biology and ecology of animals and plants[1]. Plotting such records had enabled Alexander von Humboldt to understand how altitude structured plant communities on the flanks of the Andes Mountains (von Humboldt & Bonpland 2010).

Following up on this insight would have enabled other biologists to quickly derive, with more records in more places, other structuring 'laws' of ecology, but they did not. Most naturalists at the time could not deal with quantitative data, as illustrated by Ernst Haeckel's polemic against Victor Hansen, who saw the patterns generated in the sea by (phyto)-'plankton' — which he named — where Haeckel saw patterns only in the shapes of individual planktonic cells (see Pauly 2004).

Indeed, the underappreciation of occurrence records as the raw data of ecology lasted all the way to contemporary research programs such as the Census of Marine Life (www.coml.org/), whose creation of an Ocean Biogeographic Information System (OBIS; www.iobis.org/), based on standardized occurrence records, was very much an afterthought (Pauly & Froese 2010).

The problem here, I think, arises from the silos in which the specialists of different taxa remained imprisoned, which prevented them from seeing occurrence records as a suitable currency for ecology. Arachnologists think a proper database of occurrence records should include the shapes of spider's webs, while ornithologists think that such a database should include bird songs, etc. Thus, taxonomists, who could (and should) have generated millions of occurrence records for ecology, and be connected to 'big data', failed to do so. Rather, taxonomy closed in on itself; as a result, it is now unjustly pushed out of many universities' life science departments (e.g. Blackmore 1996, Boero 2010), and replaced by genomics and related disciplines that are prolific producers and users of the current, digital form of 'big data'.

A discipline in which the transition to big data took place rather smoothly is physical oceanography. In the European Middle Ages and early modern period, individual mariners could accumulate the knowledge of the coasts and currents experienced in a lifetime, and various rulers had the information generated compiled into 'portolans' or other national maps, jealously kept from commercial competitors or potential enemies.

In the mid-19th century, however, a different mode of data exchange was found. Thus, Captain Matthew F. Maury of the nascent United States Navy offered much-improved maps to all mariners who contributed their personal observations on currents, depth soundings and other oceanographic variables (Williams 1963). The result of this intensive international data sharing was not only better maps and the discovery, among other things, of the mid-Atlantic ridge, but also the definition of oceanography as perhaps the first discipline shaped by big data sets and their exchange. Obviously, this advance was only possible because data on currents and depth soundings — soon to be accompanied by sea surface temperatures and salinity, the key quantities shaping the dynamics of water bodies — are readily standardized.

This development was closely followed (or paralleled) by the development of meteorology, which also could make sense of a growing body of data on the key features and movement of air masses (wind direction and intensity; air pressure, vapor content and temperature). Indeed, meteorology and oceano-

[1] A fourth element is usually required (and available) for biological specimens: the person who has done the sampling, which allows connecting the specimens in question to the scientific literature (Froese & Pauly 2013)

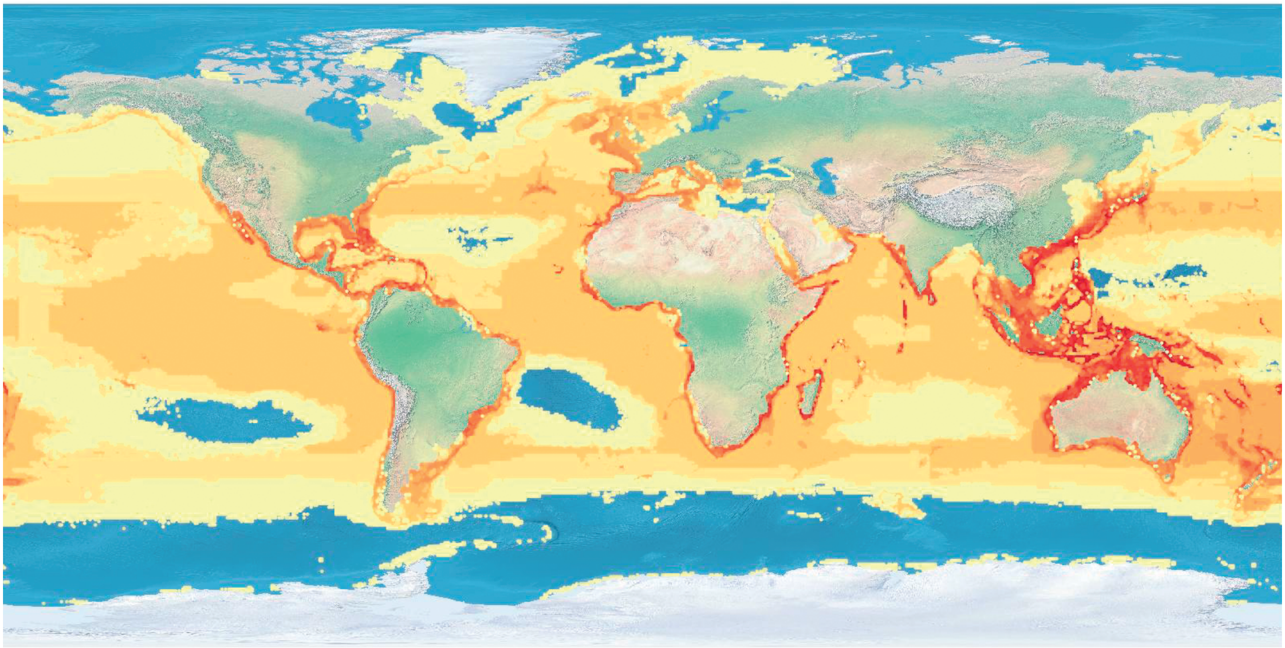

Fig. 2. Richness (red: high; yellow: low) of elasmobranch species (cartilaginous fish, consisting of sharks, rays and chimaeras) in the world oceans, as can be retrieved from Aquamaps (www.aquamaps.org), constructed from occurrence records and the environmental variable (temperature, depth, etc.) with which they could be associated

graphy are now converging. Their current and historic data sets (the latter enriched by rigorous data recovery programs, e.g. for atmospheric and oceanographic data gathered by the Axis powers during World War II) are jointly run for both short-term predictions of the weather and long-term predictions of the climate (Edwards 2010). Here, 'big data' not only created new patterns, but led to the emergence of programs of actions to undertake, or to ignore, at our own peril[2].

Big data may also help to overcome some of the divisions between the humanities and the sciences, e.g. through the introduction of quantitative approaches to study phenomena that have so far been approached phenomenologically. Examples are the study of 'Ngram' in millions of scanned books (see Fig. 1 and Michel et al. 2011, Stergiou 2017), or the construction of thousands of trees, and the selection of the most likely to depict the evolution of languages (Gray & Atkinson 2003), and even of creation and other myths (d'Huy 2016).

Until recently, ecology had no standard protocols for sharing data and no culture encouraging the practice, hence the frequent exhortations in leading scientific journals for more data recovering and sharing (e.g. Griffin 2017). However, ecology will eventually catch up with big data. Notably, it is likely that the hundreds of millions of occurrence records in the taxonomic literature and in museum collections will be retrieved by artificial intelligence programs. This would allow for better following up on biogeographical ideas such as those of Alfred Wallace (Barber et al. 2000). It would also allow for improving the extent and quality of the coverage of existing initiatives, such as OBIS, and thus for improving derived products, such as Aquamaps (Kaschner et al. 2008; www.aquamaps.org), which link these records with environmental parameters (temperature, depth) to generate probabilistic maps of the distribution of various marine and freshwater taxa (see e.g. Fig. 2).

The recovery of a massive number of occurrence records would also enable us to follow up, albeit belatedly, on Humboldt's ideas, and to track the effects of global warming on the distribution of communities of organisms. This topic is still in its infancy because many biologists persist in dealing with global warming one species at a time, despite concepts and approaches being available for dealing with ensembles of species (see e.g. Cheung et al. 2010, 2013).

In physics, there is an analog to the above contention that massive data or information lead to new

[2] Indeed, we are the first civilization that will be able to predict its own demise (see Oreskes & Conway 2014)

structures to assimilate these data or express the information they contain. This analog relates to the 'dissipative structures' (Nicolis & Prigogine 1977) that emerge when energy gradients become so strong that energy is not transferred by a linear increase of the mechanism used when the gradients are weak. Such dissipative structures emerge spontaneously in pots of boiling water, or as the Hadley (wind) cells that transfer heat from the tropics to the poles. Indeed, life itself may be a dissipative structure, as well.

Here, I simply contend that massive data create the structures through which they are processed and will flow, and that the shapes and dynamics of these new structures are not predictable from the shapes and dynamics of the structures that accommodated the smaller data flows.

There is a long tradition of old men predicting a future that they will not experience and that mostly does not turn out the way they predicted. However, this author — also an old man — predicts that we cannot predict what big data will create in the longer term, in any scientific discipline and in society at large, good or bad.

LITERATURE CITED

Barber PH, Palumbi SR, Erdmann MV, Moosa MK (2000) Biogeography: a marine Wallace's line? Nature 406: 692–693

Berlin B (1992) Ethnobiological classification: principles of categorization of plants and animals in traditional societies. Princeton University Press, Princeton, NJ

Blackmore S (1996) Knowing the Earth's biodiversity: challenges for the infrastructure of systematic biology. Science 274:63–64

Boero F (2010) The study of species in the era of biodiversity: a tale of stupidity. Diversity 2:115–126

Cheung WWL, Lam VWY, Sarmiento JL, Kearney K, Watson R, Zeller D, Pauly D (2010) Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. Glob Change Biol 16:24–35

Cheung WWL, Watson R, Pauly D (2013) Signature of ocean warming in global fisheries catch. Nature 497:365–368

d'Huy J (2016) The evolution of myths. Sci Am 315:62–69

Dunbar R (1998) Grooming, gossip, and the evolution of language. Harvard University Press, Cambridge, MA

Edwards PN (2010) A vast machine: computer models, climate data and the politics of climate warming. MIT Press, Cambridge, MA

Froese R, Pauly D (2013) Fish stocks, Vol 3. In: Levin S (ed) Encyclopedia of biodiversity, 2nd edn. Academic Press/ Elsevier, Waltham, MA, p 477–487

Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435–439

Griffin E (2017) Rescue old data before it's too late. Nature 545:267

Kaschner K, Ready JS, Agbayani E, Rius J and others (eds) (2008) AquaMaps Environmental Dataset: Half-Degree Cells Authority File (HCAF). Available at www.aquamaps.org

Lieberman SJ (1980) Of clay pebbles, hollow clay balls, and writing: a Sumerian view. Am J Archaeol 84:339–358

Linnaeus C (1758) Systema naturae per regna tria naturae, secundum classes, ordinus, genera, species, cum characteribus, differentiis, synonymis, locis. Tomus I. Editio decima, reformata. Impensis Direct. Laurentii Salvii, Holmiae

Michel JB, Shen YK, Aiden AP, Veres A and others (2011) Quantitative analysis of culture using millions of digitized books. Science 331:176–182

Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations. Wiley, New York, NY

Oreskes N, Conway EM (2014) The collapse of western civilization: a view from the future. Columbia University Press, New York, NY

Pauly D (2004) Darwin's fishes: an encyclopedia of ichthyology, ecology and evolution. Cambridge University Press, Cambridge

Pauly D, Froese R (2010) Account in the dark. Nat Geosci 3: 662–663

Roche D (2006) Encyclopedias and the diffusion of knowledge. In: Goldie M, Wokler R (eds) The Cambridge history of eighteenth-century political thought. Cambridge University Press, Cambridge, p 172–194

Stergiou KI (2017) The most famous fish: human relationships with fish as inferred from the corpus of online English books (1800–2000). Ethics Sci Environ Polit 17:9–18

von Humboldt A, Bonpland A (2010) Essay on the geography of plants. University of Chicago Press, Chicago, IL

Williams FL (1963) Matthew Fontaine Maury, scientist of the sea. Rutgers University Press, New Brunswick, NJ

Editorial responsibility: Konstantinos Stergiou, Thessaloniki, Greece

Submitted: April 25, 2017; Accepted: June 25, 2017
Proofs received from author(s): August 15, 2017