GROWING PAINS IN THE ERA OF BIG DATA

In the five seconds it takes you to read this sentence our global community will have generated another 800 trillion bytes of data. This year alone, the world will add 5 trillion gigabytes—that's 5 zettabytes or 5,000,000,000,000,000,000,000 bytes—to the ongoing, frenzy of data accumulation (VASEM 2015). No wonder everyone is talking about "big data" and the expanding toolbox of "big data analytics."

The popular press gives the impression that the big data explosion is borne of industries from the medical, social science, communications, and financial sectors. After all, recent increases in data volume can be attributed to the exponential production and use of electronic devices, from phones and wearable sensors to surveillance cameras, that can monitor, collect, and store virtually every type of information from our daily lives.

So where are the Earth sciences in all of this? An Internet search of "big data in Earth sciences" suggests that there has been little discussion of the topic. Indeed, a recent article by Kuo et al. (2014) asks, "How will big data technologies benefit Earth science research?" Does this imply that big data sets, data mining, and processing analytics are foreign to the Earth sciences? Nothing could be further from the truth!

Many decades before the term "big data" was coined, the seismology, geodetics, and planetary sciences were already using approaches that fall into the realm of "big data analytics." These fields mined existing data sets for physics-based patterns and employed their respective data sets as virtual testing grounds to develop new methods for data processing, modeling, and simulation. For example, many of the historical drivers of the supercomputer industry actually emerged from the Earth sciences. Thus, our field is a "been there, done that" older brother of big data analytics.

Looking ahead, all groups working with big data, including the Earth sciences, are facing two intertwined challenges—capacity and preservation. Capacity is critical because the growth of data sets is outstripping our ability to keep up. Researchers in academia are concerned because many of the decisions regarding who pays for this infrastructure are not yet reconciled. Should universities or funding agencies build and maintain these resources? Given that the growth in data volume is exponential, with no end in sight, this is an urgent challenge requiring constant vigilance.

The problem of capacity was recently highlighted from a business perspective by Vinton Cerf, the Vice President and Chief Internet Evangelist of Google, at the Virginia Academy of Science Engineering and Medicine Big Data Summit (VASEM 2015). He acknowledged that everyone has enjoyed dramatic increases in storage density while storage costs have been declining. This has allowed companies like Google to keep up so far. But he notes, "If we have to start throwing data away, we will need protocols and decision tools for doing that. This will be a process fraught with risk as what we think is important today may not be important later." Who will decide what is important and should be kept and who will decide what is thrown away to make room for the new?

This brings us to the other challenge: data preservation (UNESCO 2012). Not unlike the archiving of physical samples, such as ice cores or marine sediments, there are obvious arguments for preserving digital data sets.


Patricia Dove

Digital data sets have a unique usefulness for testing new theories or reanalyzing the basis for past interpretations. But also, new advances in data analytics and data mining may uncover previously unrecognized patterns or may reveal new ways to probe into archived data. Given the irreproducible nature of these data sets, particularly those collected in the past, it would seem the value of many is truly incalculable.

The Earth science community needs to address the problems of capacity and preservation in the face of exponential growth in data. The problems cannot be ignored. From the images of grandparents on your phone to the torrents of data coming from supercomputers, information requires software that can interpret the bits and bytes using a current operating system. Many of us are already experiencing "brownouts" when lapses in backward compatibility prevent us from opening simple files that we created only a decade ago.

Now let's fast forward 50, 100, or even 1000 years into the future. Archiving data sets for future generations, whether big or small, is not trivial. We need to preserve the original application software, the original operating systems, and the original computers that ran the operating systems in order to read and use old digital data (VASEM 2015). If we don't, we risk losing records of the 21st century and leaving a "digital Dark Age" for future generations (Ghosh 2015).

The Earth science community cannot wait for other groups to sort out these issues. A new frontier is upon us, and our discipline needs top people at all levels to plan for the best use of current resources. Furthermore, both early career and established scientists need to plan for the future—short and long term—by providing decision makers with informed recommendations.

There is no time to waste! And this reminds me to check the small herd of space-eating *Ceratotherium annoyius*, umm… filing cabinets that are living in my office. After thinking about these issues, maybe I won't turn them out to pasture quite yet.

REFERENCES

Ghosh P (2015) Google's Vint Cerf warns of 'digital dark age'. BBC News, 13 February 2015. www.bbc.com/news/science-environment-31450389

Kuo K-S, Clune TL, Ramachandran R (2014) Earth science data analysis in the era of big data. Earthzine, 4 April 2014. www.earthzine.org/2014/04/04/Earth-science-data-analysis-in-the-era-of-big-data

UNESCO (2012) The Memory of the World in the Digital age: Digitization and Preservation. 26–8 September 2012, Vancouver BC. Conference Proceedings. Duranti L, Shaffer E (eds) United Nations Educational, Scientific, and Cultural Organization, Paris. 1458 pp

VASEM (2015) Virginia Academy of Science Engineering and Medicine: Big Data Summit. Report of December 4–5, 2015. National Academy of Sciences, Washington, DC. 19 pp

Patricia M. Dove, Virginia Tech
Principal Editor