PRESIDENTIAL ADDRESS

The Challenges of Big Data in Expanding Geoscience: Embracing New Initiatives to Untangle our World

Dène Tarkyth

Anglo American Exploration (Canada) Ltd
800-700 West Pender Street, Vancouver
British Columbia, V6C 1G8, Canada
E-mail: dene.tarkyth@angloamerican.com

PREAMBLE

It was my pleasure to serve as the president of this organization through 2018 and part of 2019, and such an experience cannot help but remind me of the effort that comes from GAC staff and our many volunteers, but it also brought home the challenges that all of us face in organizing our time and activities in this so-called Information Age. We live in a world where both space and time are increasingly compressed, and all of us at times struggle to manage the demands of our work and our lives beyond the office walls. So I will start this address by asking you all to imagine that you had one extra day a week given to you - some time that you could spend on fun science and investigating exciting questions, or just catching up on work and life. Would we not all welcome such a gift? But then look back over the last few weeks, months or even years and think about how much time you spent searching for information, skimming papers to finding sample locations, compiling and cleaning up data, georeferencing maps….just some of the many basic things that need to get done before you can get to the fun part of your job as a geoscientist. There are estimates that geologists now spend 80% of their time searching for, formatting and organizing information and data, and I do not find these hard to believe.

A recent article highlighted the approach taken by Cameco, one of Canada's leading mining companies, to change how they manage data in order to save 20% of their geologists' time – one day a week – so that they would not have to spend countless hours looking for data and could do geology instead (Heffernan 2015). There are many efforts to amalgamate and process data in ways that make this process easier and more amenable to automation. A young student geologist at Princeton University, Julia Wilcots, undertook a summer project with a senior researcher at University of Wisconsin to examine the distribution of stromatolites through geological time by searching descriptive literature. Anyone who has worked in the Precambrian, or indeed in sedimentary rocks of any Eon or Era, can well imagine the immensity of that search. However, through the use of computer search techniques and the 'Geodeepdive' database, she was quickly able to identify over 10,000 papers that mentioned stromatolites (in the text, but not necessarily in the title) and extract the associated rock unit names from 10% of them. Then, by linking these results to the 'Macrostat' database, she was then able to come up with an estimate of the percentage of shallow marine rocks that contain stromatolites within different geological time periods. A more senior researcher at the University involved with the project estimated that doing this same search would have taken him sixteen months of tedium. The overall conclusions of the study – that the distribution of stromatolites is most closely linked to the abundance of dolomitic carbonate rocks (Peters et al. 2017) – are important, but the methodology demonstrates the ability of new techniques to unravel seemingly infinite tangles of data. What other questions could we address and what other problems could we solve as Earth Scientists if we were routinely able to query efficiently organized data with such rapidity?

As a science, geology continues to evolve towards a bigger view - from rocks alone, to facies, to entire sedimentary systems, to geodynamic environments, and to the Earth System as a whole. We increasingly recognize the interconnected nature of all geoscience data, and the need for a 'Big Context' to make sense of 'Big Data'. This address seeks to emphasize the great potential of the data explosion that confronts us but sometimes confounds us, and also to specifically highlight some of the new and exciting tools and techniques that can help us exploit it. I seek to provide but a glimpse of an ever-expanding branch of our science, which will feature more and more in our professional lives in the 21st century.

SNAPSHOTS OF SOME NEW INITIATIVES IN GEOSCIENCE DATA MANAGEMENT

In the 21st century, we hear so much about Big Data, Artificial Intelligence, Machine Learning, Data Analytics and their 'vast potential', but we are increasingly challenged to actually make use of that potential. Most of us are inundated with data and struggle to even keep up with the scientific literature. We think of the state of our own incomplete and fragmented datasets, and we find ourselves falling from the "Peak of Inflated Expectations" on the "Gartner Hype Cycle" into the aptly-named "Trough of Disillusionment" (Fig. 1;

Geoscience Canada, v. 46, https://doi.org/10.12789/geocanj.2019.46.152 pages 159–162

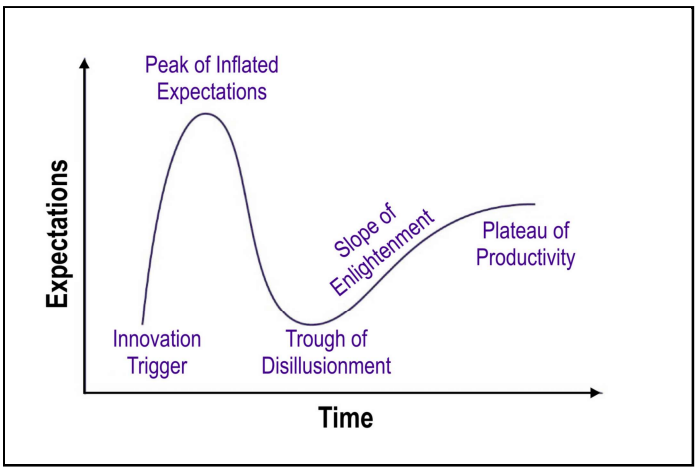

Figure 1. The "Gartner Hype Cycle", a concept that applies in many human endeavours. It was originally outlined by the Gartner Group, a global research and advisory firm in the United States, and has been widely applied to the development and marketing of new technologies, although it has also received some criticism from experts in the field. Figure source: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.

https://www.gartner.com). My hope is that we can eventually climb the "Slope of Enlightenment" and reach the "Plateau of Productivity", and that some of the new initiatives discussed below might measurably ease that ascent.

Several new 'nodes' have now sprung up to try to bring order to this overwhelming data chaos, and not just in academia. Mining companies are addressing the challenge, because they constantly need to revisit their historic archives, or those of predecessors, to try to find the next mine. The Prospectors and Developers Association of Canada (PDAC) recently initiated the Exploration Assessment Digital Data Formats (EADDF) project, which developed standard guidelines for all digital data submitted as part of mineral exploration company assessment reports. Such an effort is immensely valuable for the continuity of exploration as properties pass from company to company, and also to Government and academic researchers who tap into these valuable corporate data reservoirs.

Many analytical laboratories and related service providers now market data-management services along with the analyses or results that they provide to customers. Relatively high-priced subscription services such as SNL/IntierraRMG and Geofacets (available through Elsevier) recognize the additional value in providing their information in formats that are readily usable by the customer, including maps or other graphical products. The extra costs involved for the customer in acquiring such processed data are more than compensated by the elimination of the time and frustration involved in integrating and correlating raw results.

Australia is now a world leader in compiling and integrating mineral exploration industry data, and in bringing it together with information from Government geological surveys and other research sources. For example, it is now possible to view and download all the geological survey field sites for the country and all of the borehole locations drilled by industry projects via a single portal (https://portal.ga.gov.au). In Canada, most Provincial Geological Surveys provide geoscience data compilations via their websites, including advanced web application formats such as the excellent SIGÉOM system in Québec, and the Resource Atlas GIS maintained by the Geological Survey of Newfoundland and Labrador. These are just two examples of data delivery from Provincial geoscience, and most organizations now provide some system of this type. However, we have yet to achieve the nation-wide accessibility of federal, state and corporate data now provided to Australian geoscientists, or to reach a desirable level of consistency among the various delivery platforms.

However, many other data compilation initiatives have sprung up over the years across Canada. Bruce Eglington maintains the DateView and StratDB geochronological and lithostratigraphy databases, respectively, at the Saskatchewan Isotope Laboratory as part of some IGCP projects (most recently IGCP 648). These are key inputs to his global paleoenvironmental reconstruction modelling and other projects.

Elsewhere in the world, there are some excellent examples of similar initiatives that are linked to specific disciplines. For geochemistry, we have data repositories such as EarthChem (https://www.earthchem.org/), which is funded by the United States National Science Foundation (NSF), and GEOROC, a geochemical rock database in Germany. Other international and NSF-funded projects include the Paleobiology Database, Macrostrat for stratigraphy, and Neotoma for paleoecology. In Geophysics, there are several paleomagnetic databases such as the PALEOMAGIA Precambrian Database hosted at the University of Helsinki, and another version developed offline by Sergei Pisarevsky at Curtin University, Australia, based on an original compilation by Dr. M.W. McElhinney.

Many of these databases, and numerous others that cannot be referenced in this short article, are relatively isolated, standalone initiatives that are promoted and maintained by core groups of indefatigable academics, who most often maintain them as a side project alongside their main research and teaching roles. This is time-consuming work but of immense value to a much wider community in geoscience. Voluntary data submissions represent a very incomplete sample of the published data, as illustrated by the many obvious gaps in geochronological data compared to geochemistry data, which is clearly shown by maps produced from EarthChem (Figure 2).

In the final analysis, the true value of any database depends on the ease of searching for and retrieving data. Collections of data alone do not solve the problems that we face. Some really exciting work aimed at 'data harvesting' is happening at the University of Wisconsin-Madison where Shanan Peters leads the GeoDeepDive team (https://geodeepdive.org/). This initiative uses natural language processing (NLP) to identify contextual relationships and applies text and data mining (TDM) to harvest information from close to 300,000 newly published documents per month. This is where Julia Wilcot's summer project on stromatolites and geologic time was initiated. Shanan's team now has agreements with most major scientific publishers to access their publications via automated search methods. GeoDeepDive is one of several projects within the larger EarthCube initiative funded by the United States NSF.

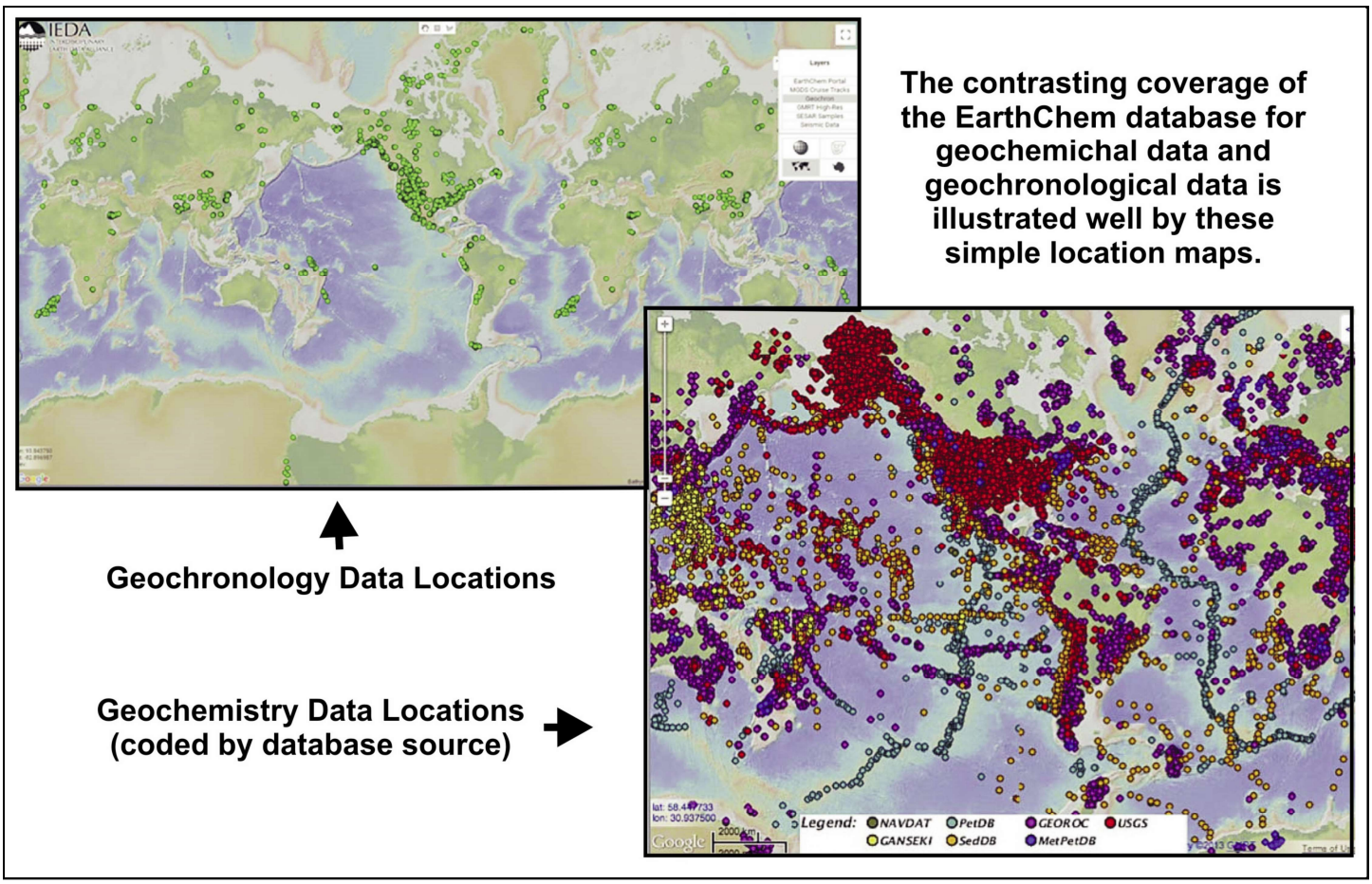

Figure 2. Maps produced from the EarthChem database showing the geographical restrictions of geochronological data compared to lithogeochemical data. IEDA = Interdisciplinary Earth Data Alliance. Figure sources: http://earthchem.org/portal (geochemistry sample image); http://app.iedadata.org/databrowser/ (geochronology image).

GeoDeepDive currently provides access to some 10 million documents and represents a very valuable tool for researchers in all sectors of geoscience.

To make it easier to get data from published research, the main scientific publishers are phasing in new standards for supplementary data over the next two years. One of these is a requirement for the use of the System for Earth Sample Registration (SESAR; http://www.geosamples.org/) in which researchers must register samples with a unique sample identifier termed an alphanumeric International Geo Sample Number (IGSN). This works like the more familiar doi or ISBN reference used for publications (see http://www.geosamples.org for more discussion). It will ensure that basic data such as sample location are accurately captured, and allow unambiguous linking of information in cases where different analyses such as lithogeochemistry, U–Pb dating or isotopic analyses are completed on material from the same sample, even if they are from different laboratories and published in different papers over the years. This is in its own right a valuable initiative that can bring order to data and avoid unnecessary duplication of effort.

FUNDING AND SUPPORT

Of course, initiatives in data management and compilation require funding, just like any other scientific endeavour. There are several funding models for making geoscience data more accessible and useable. The EarthCube project is a joint effort funded by the United States NSF Directorate for Geosciences and the Division of Advanced Cyberinfrastructure. Its wider objective is to develop the cyber-infrastructure of technology and systems that will allow sharing and accessing all types of geoscience data and related resources. It is probably the most robust data system presently used by the North American geoscience community. Other funding models include industry-academia consortia, of which the Mineral Deposits Research Unit at the University of British Columbia is probably one of the best-known examples in Canada. Direct government support for data integration initiatives is present through many geological survey data portals around the world, and through International consortia such as OneGeology (http://www.onegeology.org/). This includes geological surveys from around the world, representing 118 countries, and also UNESCO, the International Union of Geological Sci-

ences (IUGS) and the Commission for the Geological Map of the World. Other services are available as subscription services aimed more at corporate clients, including SNL/IntierreRMG and Geofacets.

China recently announced $75 million in funding over ten years for a Deep Time Digital Earth (DDE) database initiative developed as an International Union of Geosciences (IUGS) project, although there is currently no funding outside of China for this. Non-experts will do much of the data harvesting, which will be verified by experts. The DDE grew out of the Geobiodiversity Database, which was started by paleontologist Fan Junxuan of Nanjing University in 2006 and became the official database of the International Commission on Stratigraphy in 2012. DDE was officially supported by the IUGS in February 2019 as one of its projects.

Extensive collaboration between EarthCube and DDE is unlikely at this stage due to political issues, although technical discussions between scientists continue. The co-chairs of some of the working groups within the IUGS-DDE initiative include North American researchers such as Bruce Eglington in Saskatchewan, Kirsten Lehnert at the Lamont-Doherty Earth Observatory in New York (on geochronology, geochemistry and petrology) and Isabella Montanez at University of California, Davis (on sedimentology).

CLOSING REMARKS AND SUGGESTIONS

In an age of highly collaborative research and systems thinking, efficient access to good data provides a competitive edge to companies, and promotes scientific insight and discovery. As members of the geoscience community, we all need to be more aware of how we can use these resources in our work and also make efforts to enhance them to improve the activities of our community. I urge GAC members to get actively involved in these North American and International Geoscience Data initiatives, by harnessing the power of 'Big Data' to investigate your own scientific problems and, above all, to further the growth of this potential by contributing data and information. Ask what queries to a system like GeoDeepDive could do for your own research questions, and also think about how the research of others might be aided through access to information that you could contribute. I also strongly encourage the promotion of skills development in these areas amongst employees, students and professional colleagues. The challenge of unravelling 'Big Data' may at times seem overwhelming, but it provides methods to visualize forests as well as identify individual trees. Canadian geoscientists have much to contribute to these valuable initiatives!

ACKNOWLEDGEMENTS

I would like to acknowledge my employer for supporting my participation in the Geological Association of Canada, and Bruce Eglington for many stimulating discussions about data initiatives.

REFERENCES

Heffernan, V., 2015, Freeing up time for mine-finding: the Cameco solution: Earth Explorer, Jan 21—15, p. 8–10, https://www.geosoft.com/media/uploads/resources/reports/ee-data-management-report-jan21-15-A4-web.pdf.

Peters, S.E., Husson, J.M., and Wilcots, J., 2017, The rise and fall of stromatolites in shallow marine environments: Geology, v. 45, p. 487–490, https://doi.org/10.1130/G38931.1.