The Uses and Benefits of Big Data for Geological Surveys

Michael H. STEPHENSON

British Geological Survey, Nottingham, NG12 5GG, UK


‘Long tail’ data is the difficult-to-get-at data that sits in libraries, institutes and on the computers of individual scientists. Informatics specialists like to contrast it with the smaller number of large, more accessible data sets (e.g. Sinha et al., 2013). The name ‘long tail’ derives from graphs drawn of the size of data sets against their number: there are relatively few large datasets and a lot of smaller ones. Geological science has more long tail data than sciences like physics or meteorology, probably because historically it has been less associated with big science infrastructure and sensors. Much of this ‘long tail’ data resides in geological surveys – institutes created by nations to survey and ‘inventorise’ geological resources.

The fact that ‘long tail’ data is difficult to get at probably holds back progress in geological science and the impactful use of data. The low discoverability of long tail geoscience data, and its heterogeneity make it difficult to bring it together for geological synthesis but more importantly impedes machine learning and artificial intelligence. Informaticians describe ‘data islands’ in geological libraries, in the records of geological surveys and on desktop computers.

For many large and long-established geological surveys, improving the discoverability of these data islands and the ease of compilation ( interoperability) of geoscience data involves making historical paper data in islands available to cyberspace. A unique collaboration between the British Geological Survey (BGS) and computer scientists of the GeoBiodiversity Database (GBDB) is opening up some of this data. The BGS has an abundance of biostratigraphical data collected over almost two centuries associated with about 3 million fossils and thousands of localities and stratigraphic sections, to exacting and consistent standards. The data has great potential for science, but much of it is contained within paper documents or simple document scans and so is inaccessible to big data tools. It needs lifting from the page and into cyberspace. The BGS began working recently with GBDB (the official database of the International Commission on Stratigraphy (ICS)) which is almost unique in being the only large database to hold sequences of fossils tied to sections, rather than just spot collections. To date GBDB and BGS scientists have placed live manipulable data from more than 6000 UK stratigraphic sections on a public access website. The project is also using machine-learning methods to get at biostratigraphical information directly from text.

When the ‘long tail’ data is brought into cyberspace alongside the more accessible, larger data sets gathered from sensors (e.g. gravmag, seismic and groundwater) what benefits do geological surveys expect?

For many geological surveys, a strategic trend has been away from the making of static 2D maps, to 3D maps, and finally to 4D representation of the subsurface that enables prediction of processes on a timescale that matters to human lives and livelihoods. This might mean being able to predict earthquakes, landslides, groundwater flooding, groundwater drought, subsidence, shrink-swell and sinkhole formation - being able to assimilate and make judgements from data using automatic systems. Often this means distinguishing between real ‘change points’ in time series data and anomalies. Change points are true harbingers of systematic change–i.e. a reflection of the real system rather than an outlier. Geological surveys would like to develop statistical methods to identify change points and anomalies in geoscience 4D data. This would not only help in predicting change, but also in communication with the public to help geoscientists to advise monitoring and remediation/response for regulatory bodies.

Finally bringing together big data that is interoperable and that is georeferenced both in a palaeogeographical and modern geographical sense can help us understand the long term development of resources for example metalogenesis and sedimentary basin formation. In the future largescale models will be needed to understand, for example, the distribution of clusters of porphyry copper deposits (PCDs) by linking georeferenced plate motion and geometric properties of subducted slab data. Linked databases, including groundwater and aquifer data, recharge data, meteorological data, sediment flux, subcrop geology, basin subsidence, sequence stratigraphy, compaction, geomechanics and tectonics data could lead to more accurate models for groundwater storage, underpinning sustainable development in poor countries vulnerable to climate change. Similarly, linked databases including sediment flux, subcrop geology, basin subsidence, sequence stratigraphy, compaction, geomechanics and tectonics data could be useful in mapping, for example, salt deposits across large basins to understand salt’s capacity for storage for gases like hydrogen or compressed air for the low carbon transition.

* Corresponding author. E-mail: mhste@bgs.ac.uk

© 2019 Geological Society of China
https://onlinelibrary.wiley.com/journal/17556724
Key words: long tail data, informatics, biostratigraphy, geological survey

Acknowledgments: This abstract is submitted with the permission of the CEO of the British Geological Survey

References

About the author
Michael H. STEPHENSON, male, born in 1961 Derbyshire, UK; PhD; graduated from Imperial College and Sheffield University; Director of Science and Technology at the British Geological Survey. He is interested in stratigraphy and palynology in deep time and resource geology, including shale gas. Email: mhste @ bgs.ac.uk; phone: +44(0) 115936 3577.