Feature

Big Data Knowledge Discovery project launched

The University of Sydney School of Geosciences EarthByte Group has teamed up with National ICT Australia (NICTA), Securities Industry Research Centre of Asia-Pacific Limited (SIRCA: a non-profit organisation enabling big data research) and Macquarie University to unearth 'big data' insights for the natural sciences. The $12 m, three-year research and innovation project will advance fundamental mathematics and statistics to provide a framework, and methodologies and tools for data-enabled scientific insight and discovery. It is supported by $4 m from the Science and Industry Endowment Fund (SIEF) and $8 m from the research collaborators over the life of a three-year research project. It will combine NICTA's world-class machine learning capabilities and SIRCA's expertise in big data software engineering with three natural science groups from the University of Sydney and Macquarie University:

1. Geosciences and Earth Dynamics and Tectonics, led by Dietmar Müller at the University of Sydney
2. Terrestrial Ecology, led by Mark Westoby at Macquarie University
3. Physics and Mathematics of Complex Laser Systems, led by Deb Kane at Macquarie University.

The essence of the Big Data Knowledge Discovery project is to bring some of the brightest people in the world in computer science from NICTA (in machine learning and analytics) and SIRCA (in software and big data) together with three of Australia's most distinguished natural scientists in physics, plant science and geosciences to tackle grand scientific challenges in completely new ways. The EarthByte team has previously had success in applying big data mining for constructing long-term earthquake hazard maps along subduction zones (Müller & Landgrebe 2012) and with generating Australia's first opal prospectivity map (Merdith *et al. 2013).*

How do we distinguish underlying trends in datasets from random variations – or 'noise' – and extract meaningful information? Can we use geological and geophysical data more quantitatively to find out what Australia looked like between 2 and 1 billion years ago, when it consisted of several continents surrounded by subduction zones, generating some of the world's richest metal deposits? These are the sorts of questions the project will address, drawing on the skills of the multidisciplinary team involved. NICTA is Australia's Information and Communications Technology Research Centre of Excellence and SIRCA's technology is used by over 400

Big Data Knowledge Discovery project members. Image courtesy SIRCA.

leading institutions in the financial services industry worldwide. SIRCA technology, in partnership with Thompson Reuters, powers the largest and most comprehensive research database of historical financial markets data in the world, and many of the methodologies they have developed can equally be applied to spatio-temporal data analysis in geology and geophysics. Working together, the group plans on developing a new space-time data-mining approach, exploiting similarities with other research fields such as finance and taking a new look at large geo-datasets to unravel the structure and evolution of Australia in a global plate tectonic context.

The project plans to assimilate combined geological and geophysical data into a database and software framework, which will facilitate interpretations of geological processes that have previously been very difficult or impossible to achieve. A prototype for this approach was outlined by Wright *et al.* (2013) who established a methodology for integrating open-access paleogeographic (Langford *et al.* 1996) and paleobiology (Alroy 2003) data with plate tectonics. Established data-mining methods will be connected to the *GPlates* open-source software (**http://www.gplates.org**), which is being developed by the University of Sydney with international partners. *GPlates* is a tool for visualising and analysing the interactions of moving tectonic plates embedded in an evolving network of plate boundaries through time (Williams *et al.* 2012).

The premise of the geo-portion of the SIEF Big Data Knowledge Discovery project is to develop an understanding of Australia's geology beyond the present-day setting, trying to reconstruct the original tectonic setting in which particular geological provinces were formed. Spatio-temporal data mining has never been applied to Australian mineral exploration because most mineral deposits that we know of are reasonably close to the surface: resorting to spatio-temporal data mining was not necessary to discover them and we didn't necessarily need to understand the original tectonic setting of ore deposits. But that is changing, as the UNCOVER initiative of the Australian Academy of Science recently pointed out. UNCOVER has developed a strategic framework for better understanding Australia's metallogenic evolution (**http://www.science.org.au/policy/uncover.html**).

The argument is that if we combine all the data we have in a spatio-temporal context, we should be able to gather enough intelligence to narrow down the regions where large mineral deposits may be hidden underneath the weathered surface layer. The EarthByte Group has recently shown that this approach works for opal exploration, by creating the first opal prospectivity map for the Great Artesian Basin based on spatio-temporal data mining.

Unlike gold exploration, for example, there are no accepted concepts or methodologies available to guide opal miners to where new fields may be found. There are as many theories as there are opal miners. The EarthByte Group used its *GPlates* software in conjunction with other GIS software to analyse huge datasets on the geology, geophysics and age of more than 1 000 sites, mostly in the Great Artesian Basin, where gem-quality opal has been found. The group analysed past landscapes stretching back to the early Cretaceous. This work formed the basis for identifying characteristics common to the sites. The group found that geological conditions under which opal formed resulted from a very particular sequence of surface environments over geological time (Landgrebe *et al.* 2013). These conditions involved alternating shallow seas and river systems followed by uplift and erosion.

The resulting prospectivity map (Merdith *et al.* 2013), which is shown with this article, indicates targets for exploration. Targets are in the southwestern reaches of the Great Artesian Basin in South Australia, in a northwest–southeast corridor in central Queensland and near the opal centre of Lightning Ridge, NSW. Without *GPlates*, such seamless analysis of huge datasets through geological time would have been impossible.

Interestingly, a new opal field was recently discovered 75 km southwest of Lightning Ridge, but before the new prospectivity map was published. This is precisely an area where the new map suggests high opal prospectivity, giving credence to the team's approach to mining data through geological time and space.

At the SIEF project launch, Ian Chubb, Chief Scientist of Australia, said, "If this project succeeds in its admirably ambitious aims, Australia could one day be home to a new generation of big data analytics [*sic*] tools that could be used by all manner of scientists around the world to advance knowledge discovery".

DIETMAR MÜLLER

Opal is Australia's national gemstone, but no new significant opal discoveries have been made since the early 1900s. Image courtesy Dietmar Müller.

A Google Earth mop showing regions prospective for opal (light) vs non-prospective regions (dark) bosed on data mining. Image courtesy Dietmar Müller.

REFERENCES

Alroy J 2003. Global databases will yield reliable measures of global biodiversity. *Paleobiology* 29, 26–29.

Landgrebe TCW, Merdith A, Dutkiewicz A & Müller RD 2013. Relationships between palaeogeography and opal occurrence in Australia: a data-mining approach. *Computers and Geosciences* 56, 76–82.

Langford RP, Wilford GE, Truswell EM, Totterdell JM, Yeung M, Isem AR, Yeates AN, Bradshaw M, Brakel AT, Olissoff S, Cook PJ & Strusz DL 1996. *Palaeogeographic atlas of Australia: time dependent summarisation of sedimentological data based on several datasets, between 550 Ma to present day.* Australian Government, Canberra.

Merdith AS, Landgrebe TCW, Dutkiewicz A & Müller RD 2013. Towards a predictive model for opal exploration using a spatio-temporal data mining approach. *Australian Journal of Earth Sciences* 60, 217–229.

Müller RD & Landgrebe TCW 2012. The link between great earthquakes and the subduction of oceanic fracture zones. *Solid Earth* 3, 447–465.

Williams SE, Muller RD, Landgrebe TCW & Whittaker JM 2012. An open-source software environment for visualizing and refining plate tectonic reconstructions using high-resolution geological and geophysical data sets. *GSA Today* 22, 4–9.

Wright N, Zahirovic S, Müller RD & Seton M 2013. Towards community-driven paleogeographic reconstructions: integrating open-access paleogeographic and paleobiology data with plate tectonics. *Biogeosciences* 10, 1529–1541.

Scan for the latest GSA news: