# Queensland Mining Journal analytics: new information from old data

**Cyril Gagnaire**
*DataCo*
*Level 1, 28 Ord Street*
*West Perth, WA 6005*
*Australia*
*cyril.gagnaire@dataco-australia.com*

**Neil Constantine**
*DataCo*
*Level 1, 28 Ord Street*
*West Perth, WA 6005*
*Australia*
*neil.constantine@dataco-australia.com*

**Michael Ball**
*DataCo*
*Level 1, 28 Ord Street*
*West Perth, WA 6005*
*Australia*
*michael.ball@dataco-australia.com*

## SUMMARY

The Queensland Mining Journal represents a wealth of information relating to mining activities in the state, from 1800 to the present day. This material has been scanned and made available in the public domain as high-quality image files. We have applied Google Vision optical character recognition, parsed the output JSON files through a domain-specific filter and indexed the content for presentation via an analytics dashboard based on mineralogy and which can be filtered by commodity age and location. The dashboard further incorporates mine site information from Queensland's Digital Exploration (QDEX) Data System allowing prospective miners to drill down into QDEX content based on mineral occurrence, mine status and deposit size. We plan to build on this activity using Elastic Search to improve our association of content from articles to spatial location.

This activity supports individuals and mining companies with an interest in Queensland to rapidly identify locations of interest. It offers a pragmatic approach using freely available information to support the Queensland authorities in attracting investment to their region through lowering the barrier in terms of effort level for companies looking to explore in the state. This comes at a time of heightened competition for investment and could ultimately lead to increased exploration activity and success, resulting in brownfield development of historic prospects and mine sites, minimising environmental impact of new exploration and mining yet generating income for the state through local activity and tax revenue on any mineral extraction.

**Key words:** mining, analytics, dashboard, Queensland

## INTRODUCTION

We have a background in oil and gas data management and combining our domain knowledge with emergent cloud-compute and analytics techniques. We have delivered multiple projects using analytics techniques to accelerate the legacy approach to data management, supporting companies with fast-tracked data discovery and extraction. These provide a competitive advantage as they allow for rapid review of multiple opportunities in terms of the available data and the likely extent to which the geological risk of an opportunity can be constrained by these data. This has shortened data review time frames traditionally measured in weeks down to days, whilst ensuring consistency and scalability of process. This information is typically presented to the end user in a simple dashboard allowing rapid data assessment and drill down to relevant data (Figure 1).

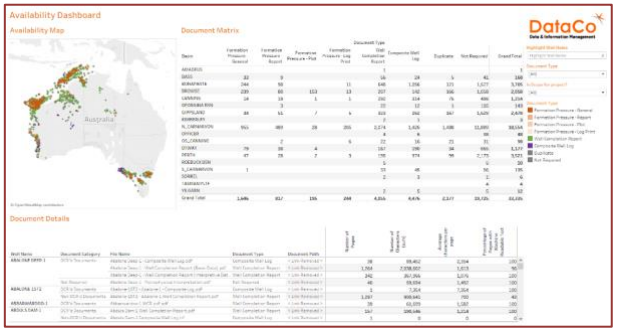


**Figure 1. Data Availability Dashboard, showing the accessibility of formation pressure data by document source across Australian open-file wells.**

Whilst looking to expand our footprint and test the approach on a domain outside of our traditional focus, we were directed to the Queensland Mining Journal (QMJ) in a scanned image format as a suitable test dataset, provided by the Department of Natural Resources, Mines and Energy (DNRME) in Queensland. These represent nearly 120 years of information covering multiple aspects of mining across Queensland and despite being scanned were held in an unindexed and relatively inaccessible format. The value proposition we identified was to surface this information in a workable format, allowing for identification of content relevant to the end user and presenting this at the user's fingertips. Further value would be delivered if we were able to combine this unstructured yet information-dense journal content with structured data from Queensland's Digital Exploration (QDEX) Data System.

This proposition stems from the common belief in the extractive industries that the best place to find a resource is in an area where this resource has already been found. Our idea was to surface this information from the locked-in QMJ image files and shortcut the laborious process of manually reading tens of thousands of pages content. This idea was unwittingly reinforced by the views of Mary Poulton, Co-Director at the Lowell Institute for Mineral Resources who was quoted in an article describing exploration for copper: "We have this saying in exploration that if you're hunting elephants, hunt in elephant country" (Chen, 2019). The same article asserts that "geologists will look at existing reports done by governments and universities and work with geophysics and geochemists to predict the probability of deposits". We had our proposition confirmed.

## METHODOLOGY

The first step in surfacing meaningful data from non-searchable file formats required the entire document set to undergo an Optical Character Recognition (OCR) process. For this we chose Google Cloud Platform (GCP) and its Vision API; an extremely powerful tool that provides a high-quality, analytics-ready dataset. Our input and output data were stored in Amazon Web Services (AWS) S3 for integration with our analytics tools.

Once the documents had undergone OCR and the outputs were sent back to our S3 data store, they underwent a transformation process to extract only the components required for subsequent analysis and to optimize the respective processes. The resultant files provided an understanding of not only what page a given word or phrase is on, but also in which paragraph and its exact location therein.

We now had the text for 1,509 journals, consisting of 73,673 pages and over 52 million words. Our approach to making sense of this large dataset was to first apply domain- and geographic-specific lexicon searches to build context around each journal and its content. This was achieved via python regular-expression searching and the generation of a further analytics dataset, specific to the metadata being extracted. The primary searches were based on commodity (major and minor) and place name, both sourced from Geoscience Australia (GA). We also surfaced publicly available Mine Site data from QDEX.

We then applied standard Business Intelligence (BI) techniques to shape the final data set into a format that allows for a BI tool to create various visualisations. The data has then been exposed through two Microsoft Power BI dashboards, allowing users to sort, filter and highlight the data in various ways depending on their use case.

## RESULTS

Without modifying the underlying repositories or data sources which have been established with many decades of effort, in a two-week sprint we were able to deliver a user-friendly dashboard, answering both the initial value proposition and identifying ideas for future activity.

The dashboard surfaces QMJ content as spatialised and time-bound references to mineralogy, filtered at the highest level as major or minor commodity types (Figure 2). A map display shows the spatialised locations of these references whilst a bump chart shows the frequency of occurrence of each mineral over time. Unsurprisingly, gold dominates but with copper as the second most commonly mentioned mineral. The left-hand panel provides a reference and a link to the article mentioning the mineral.

We built the dashboard to offer the ability to reduce the time range and select one or more minerals to provide a clearer view. This provides a fast means for the user to identify the journals and the pages mentioning the minerals of interest, vastly reducing the time to access and then consume this information.

A second page (Figure 3) presents the QDEX content, linked to a map display based upon the mine site location. This is filterable by commodity, mine status and deposit size, supporting queries such as: *show me abandoned copper mines or occurrences of copper where the deposit size is classed as medium or larger* (Figure 4).

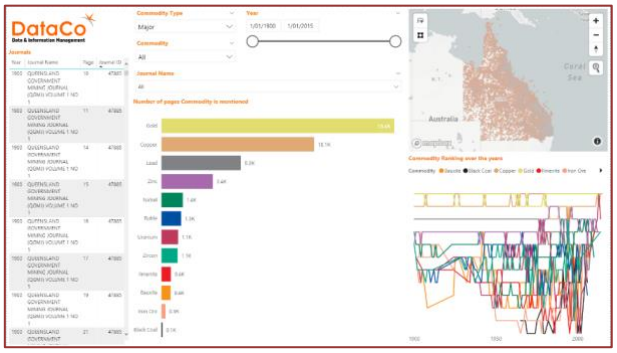


**Figure 2. Dashboard 'landing page' for QMJ content.**

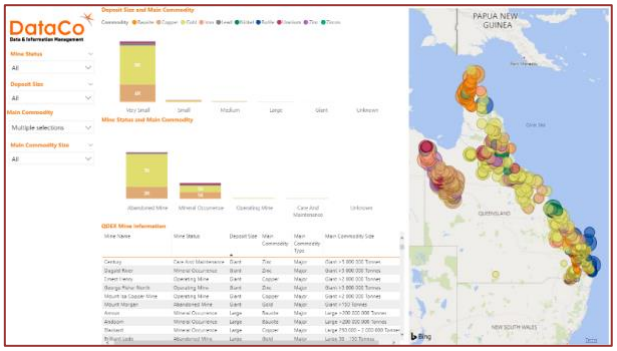


**Figure 3. Dashboard 'landing page' for QDEX data**

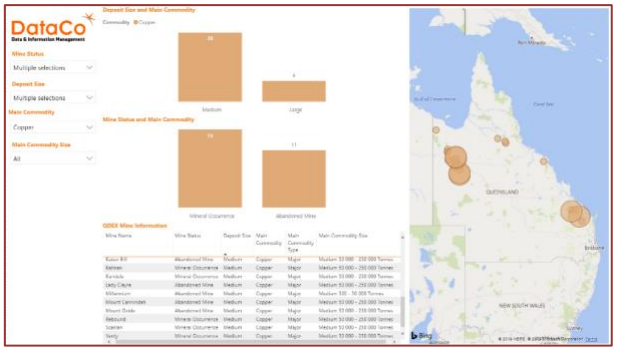


**Figure 4. Filtered for abandoned copper mines or mineral occurrence of copper for medium or larger deposit sizes.**

## CONCLUSION

Whilst there remains significant work to be done to further enrich this content and support additional use cases, we believe we have demonstrated the ability to rapidly apply scalable techniques to an untested data source and allow the domain expert to extract data and information, leading to new ideas and associations which otherwise would not have been achieved.

The Google Vision approach should be modified to allow for better identification of articles rather than just page location, allowing improved specificity of mineral references.

Elastic Search or similar tools capable of Natural Language Processing should be applied to allow sentiment analysis to

rank returns by favourable mention and improve the ability to spatialise references, which in turn will improve the link to QDEX content which has mine site as the primary key (albeit a non-unique primary key!).

In summary, the ability to apply what are fast becoming standard techniques to a range of data sources across different domains has the promise of simultaneously simplifying and enriching the data mining process, leading to better constrained exploration activity and ultimately commercial successful discoveries.

## ACKNOWLEDGEMENTS

We wish to acknowledge Simon Atkinson for bringing the QMJ content to our awareness and all DataCo staff globally who have had input into the process and dashboard design.

## REFERENCES

Chen, Angela, 2019, Where will the materials for our clean energy future come from: www.theverge.com, Feb 15, 2019.