Big data concepts for Natural Resource Management

Marsh Nick[1], Marsh Bonny[2]

1. Truii, Highgate Hill, Brisbane, 4101. Email: nick.marsh@truii.com

2. Yorb Pty Ltd, Highgate Hill, Brisbane, 4101. Email: bonny.marsh@yorb.com.au

Key Points

- There are tremendous opportunities for data driven Natural Resource Management decisions by combining local and public data resources.
- Online resources for data analysis and management can help NRM be a data focused industry.
- The constraint to better data-based NRM decisions are a lack of industry wide NRM performance data.

Abstract

The internet has become a valuable resource for data driven industries. Natural Resource Management (NRM) is a data driven industry, and is well placed to make the most of government initiatives which publish rich environmental data. These public data resources can now easily be combined with local data using online data analysis platforms. This paper illustrates opportunities gained through leveraging locally collected data with larger government datasets. To make the most of the rich online data resources, NRM agencies need to have their data easily accessible and combinable with the online resources. To become a more data driven industry, NRM agencies can embrace online team-based data sharing and management. This online approach allows a distributed solution to data management which empowers all team members to be data custodians and data analysts (data democratisation). A major limitation to the NRM industry being more data driven is a lack of industry wide performance data to act as benchmarks for determining both the likely cost and benefits of NRM activities.

Keywords

Data, analysis, visualization internet, teams, management

Introduction

Pushing public data online

We are in a time of ever shrinking national and state investment in natural resource management, but our NRM issues have never been more acute. In order to make a difference we have to do more with less. One key efficiency driver in the business world is 'data integration' through business intelligence (BI) systems. This rather boring jargon just means that we can gain efficiencies and competitive advantages in a business, by collating and analysing data across different business units, and by comparing that with broader industry benchmarks. Natural resource management organisations are not corporate giants, they are made up of enthusiastic custodians of the environment. However, this doesn't mean that we cannot learn something from the business world about how to use the data that we have to get the best environmental return for our investment and more importantly, to demonstrate the value of our work to our sponsor's.

Natural resource management is necessarily the domain of local decisions made using local data and knowledge. Despite the local nature of managing the environment, the problems faced by any given NRM agency are rarely unique and therefore, nor are the approaches to solving them. Recent data delivery and analysis initiatives can help firstly to inform these local decisions through the use of public data assets (that are now easily accessible), and secondly by providing online forums for the industry to share and leverage their data assets.

Moneyball the environment

It is time to "Moneyball" the environment. The term Moneyball comes from Lewis (2003) "The art of winning an unfair game" which was basis for the movie "MoneyBall" (2011). The moneyball story is about how in 2002 the Oaklands Athletics baseball team with a payroll budget of US$41M was able to compete successfully against much wealthier teams (such as the New York Yankees with US$125 payroll budget). The approach used by Oaklands Athletics was to move away from traditional metrics of a baseball player's value, and to use a data driven statistical approach to dig deep

into the data to buy great players that had been undervalued by the traditional metrics. Like cricket batting averages, baseball has long-established traditions of using statistics to summarise a player's performance. These traditional player statistics have played a major role in determining a player's value. However, when Oaklands analysed the data, the combination of traditional metrics of each player's value didn't match the team's overall performance. That is, a team of players that scored well in the traditional metrics didn't necessarily translate into a successful team on the park. Oakland Athletics developed their own more insightful metrics, bought undervalued players, and went on to make the playoffs in 2002 and 2003.

We are not suggesting that the methods by which we currently manage natural resources are inappropriate, but without quantifying the benefits of what we do, we will never know. It is time to moneyball the environment. To do this we need to quantify the benefits and compare that with the costs. To moneyball we need data and we need to be able to analyse it.

Fortunately one resource that is increasing for natural resource management is the availability of structured environmental data to help drive our decisions. To be more data driven in our decisions we need;

1) Easy access to relevant data,
2) Easy ways to manage and share data within the project team, and
3) Easy ways of combining, analysing and reporting our results.

These three topics are briefly discussed followed by some examples of where we could apply environmental moneyballing.

Access to relevant data

When faced with a natural resource management issue, the internet is an invaluable resource of archived reports, papers and opinions that help us understand our natural environment and see how others have approached the problem. However, almost all natural resource management issues require not just knowledge about principles and processes, but require interpreted data by which to apply that knowledge.

The collection of environmental data has not necessarily increased, but state and federal government initiatives to make data more easily accessible provides us with an opportunity to more rapidly access the data and apply it to the problem at hand. In times past, the availability of useful data sources was rarely known, and if a data source was known, then access to that data required identifying the data custodian, sending data requests, accepting data licencing agreements and eventually receiving the data. Once we had the data, it invariably had to be converted from some proprietary data structure into something readable by the software that we had at hand. Government initiatives to make publicly funded data sets more open has resulted in the development of some great data portals where the available data is easy to search, the data licencing is straightforward, (often adopting a creative commons licence) and much more uniform data structures are being used such as simple column based .csv files, or standard spatial data formats. Some useful resources for environmental data are:

www.bom.gov.au – climatological, satellite and river flow data

www.geosciences.gov.au – spatial land use and tenure layers

www.data.gov.au – federal government collation of data across all departments but dominated by environmental data (3,600 datasets)

www.data.qld.gov.au – Queensland government data across all departments with a high proportion of environmental data such as species observations (1,200 datasets). Equivalent data portals have been developed across most Australian states.

www.quandle.com – over 9 million indexed global datasets (dominated by world bank data)

www.abs.gov.au – Australian Bureau of Statistics data including census results but also specific industry research.

Whilst there has been a huge growth in the public availability of data collected in government sponsored programs, the same cannot be said for river management specific datasets collected by regional NRM bodies, or research groups. To

moneyball the environment we need to develop metrics of environmental performance and to do this we need quantitatively evaluated environmental programs. Surprisingly we have very little idea of the quantitative benefits of the stream management work, or even what type of work is carried out despite it being $100m/a industry in Australia (Rutherfurd *et. al.* 2004). The lack of quantitative data about environmental performance is well recognised; Price *et. al.* 2009 proposed a national stream rehabilitation data base to help at least quantify what rehabilitation work has been conducted and to be able to identify sites where monitoring work has been done. In addition to documenting the extent of work conducted and quantifying the environmental and social effectiveness of alternative NRM strategies we also need basic benchmarking measures to quantify the environmental condition at different points in time in order to gauge progress (Figure 1). The Wentworth Group of Concerned Scientists are in the process of working with regional NRM groups to develop and test an 'environmental accounts' approach, whereby the collective environmental condition of each NRM region is reported (Wentworth Group of Concerned Scientists 2008).

Whilst the ideal industry level data in Figure 1 is not yet directly available, most NRM agencies will have at least some local basis by which to be able to quantify what work has been done. In terms of quantifying the relative benefit of this work, in the absence of locally conducted experiments, we can refer to studies from elsewhere to help gain a sense of the quanta of benefit. A recent project conducted in South East Queensland (Thomson et al (2012)) set up a series of experiments in order to quantify just how much sediment was retained through various works and measures. The results are location specific, and monitoring conducted over a single wet season. However, in the absence of more appropriate local studies, Thomson's 2012 study could be used to quantify the relative sediment retention benefits of different works and measures.

The Wentworth Group of Concerned Scientists' environmental accounts project is still in development. However, previous levels of environmental accounting have been conducted at a national level and can be used as a surrogate. An example is the State of the Environment reporting which was conducted in 1996, 2001, 2006 and 2011 (see http://www.environment.gov.au/topics/science-and-research/state-environment-reporting /). Whilst the methods have varied between assessments and the data is not in an easily accessible format, there is none-the-less a baseline from which to consider the performance of our natural resource management activities.

The ideal collection of data for moneyballing the environment is not in place, however many useful components do exist and are publicly available.

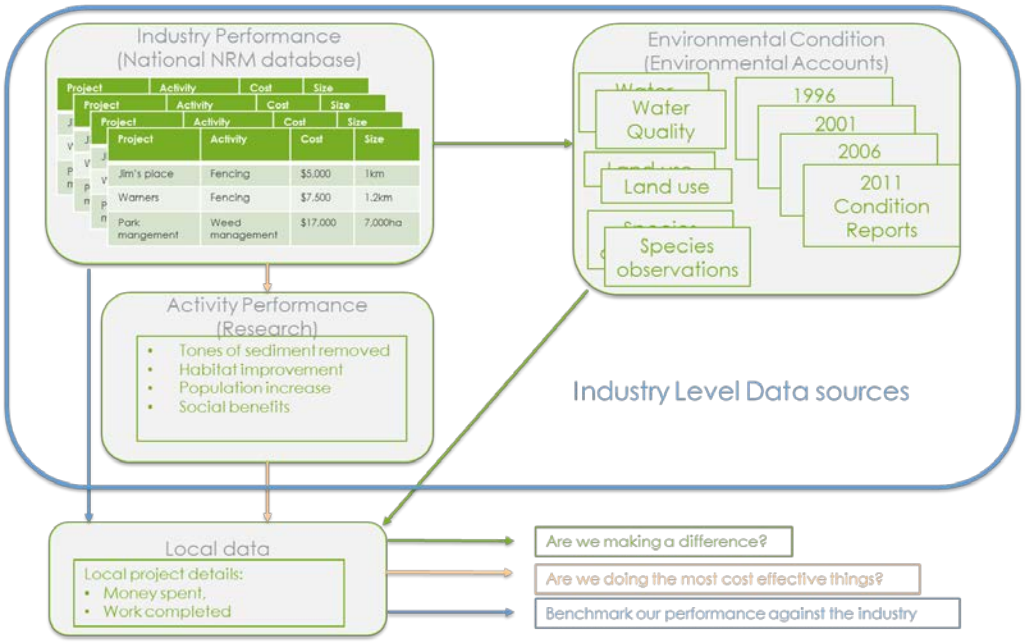

Figure 1. To Moneyball the environment we require industry level data about program delivery, environmental condition and activity level data about the benefits of different NRM activities.

Share and manage data

The key to being data driven is to democratise the data management and analysis process. Natural Resource Management groups are often eclectic associations of earnest environmental custodians who are rarely housed in the same corporate structure. Data handling and management is often a one-off, short-term, project specific operation, with the solution devised by the local practitioner. This approach is at odds with our traditional view of data managing where the preferred approach is an integrated corporate data management system with tight data management protocols and institutionalised data custodian protocols. Whilst, corporate databases are entirely appropriate for large scale ongoing data collection and management systems, they simply become a burdensome overhead for small project based operations. This often results in practitioners circumventing the corporate database system in favour of a local quick solution. The result of the local quick solution, is that datasets are disconnected and have no point of truth, and have no lineage of history which prevents them being updated or added to. The trick for data management is to ensure data lineage, whilst embracing the distributed nature of NRM organisations.

This lack of strategy for data management leads to the ubiquitous problem of data wrangling where we spend a lot of time handling and rehandling data. Wickham (in press) suggests that up to 80% of data analysis is spent on data handling. Through the use of cloud based data storage and management functionality we can have a compromise solution which doesn't require burdensome database overheads, yet allows data to be kept in an accessible place for all team members so a point of truth is maintained and the 80% data handling overhead can be dramatically reduced. These tools can be as simple as online file sharing tools such as dropbox.com, box.com or online spreadsheets provided by Google.com and Microsoft that allow access from anywhere to shared files. Moving up in sophistication are online project management tools such as basecamp.com that allow a more robust level of version control and other project management activities such as scheduling and meetings. There are more data specific applications such as Truii.com that are specifically tailored toward allowing a project team to keep on top of their data by providing versioning, data wrangling tools and data visualisation and analysis. If one chooses to create more traditional relational databases of structured information, ZOHO.com provides tools for non-database engineers to build databases which are housed online and accessible by the entire project team. These collective tools reduce the need for dedicated in-house database designers and database administrators, and effectively allow us to operate on a project by project basis.

Limitations to internet based data tools

There are several reasons why we don't use online resources for better data management;

They are new: the internet has been a massive archive. It is where we place content of things past, search others content and download it to our local computers. The concept of doing work online, such as using a live spreadsheet shared over the internet is a relatively new concept and these new ways of working take some time to be adopted.

Security: Users are reluctant to rely on a third party to keep their information secure. Security is the most critical issue for online data warehousing providers, as such they tend to spend a lot of energy protecting the security of their systems. No system is foolproof, however most will encrypt your data as it is transferred over the internet, and add a second level of encryption when it is stored in their databases and provide backups of different versions of your data in case a file is lost or corrupted. This is in contrast to your personal computer, where most users are happy to email attached files back and forth unencrypted with no real backup plan. When your hard drive fails it doesn't make the news, but if a hacker accesses an online data repository it does, hence we have a rather distorted view of how safe locally vs online data storage systems are.

Bandwidth: Some online solutions create duplicate copies of files on each users computer which is great when working offline. However this can be burdensome for large data collections which are replicated on many local machines. Most providers (e.g. dropbox.com, googledocs) will now allow the duplication settings to be controlled by the users to limit the duplication of files.

Data custodianship: If the collective data files are shared, then there is no clear line of authority and data custodianship. Our traditional approach of having a single gatekeeper doesn't apply and if we don't empower the group with data responsibilities, the result is that nobody takes care to keep the data in order. Most online solutions facilitate varied levels of responsibility by allowing the appointment of different levels of access to files. Additionally many online providers will provide a versioning system so that edited files can be rolled back, or at a minimum old versions can be retrieved.

Privacy: Maintaining tight controls over the privacy of stakeholders, such as ensuring landholder contact details remains confidential is critical for any community based project. When using internet based tools, it is therefore important to ensure that only de-identified data is made public. Any data that has sensitive information simply should not pass beyond the project team.

Combining analysing and reporting

Data analysis

To moneyball the environment we need data, we need data democratisation to empower the team to take collective responsibility for the data, and lastly we require tools for data analysis. This role has traditionally been the domain of the data analyst or statistician. However, in many instances complicated statistics and difficult to use software packages are not required for pulling out key relationships in data. Two very powerful techniques are filtering and bivariate visualisations. Filtering is simply subsampling a large data set to allow you to focus on the important data. For example, from a water quality data set you may want to create a subset of only those samples that were not taken during a storm event, so you simply extract the water quality data from the main file on days of low stream flow. The second powerful technique is simple visualisation on one variable against another in a standard scatter plot. Environmental systems are complicated and there are many interacting elements, but as a first cast, being able to plot the dry weather water quality against the number of upstream kilometre of intact riparian vegetation will give you an indication of the strength and importance of the relationship before getting too carried away with tests of significance. These combinations sound simple, but their implementation requires both spatial and temporal data analysis. Until recently, spatial and temporal analysis were very separate domains, with spatial analysis being an expensive and specialist skill due largely to the complicated software used to create and analyse spatial data. There has been a growth of online spatial analysis tools, and even the Google Earth application provides an inexpensive and powerful GIS application.

There are an increasing number of online tools aimed at the non-specialist user to allow rapid data visualisation and analysis. Tableau.com has a powerful collection of data visualisation tools allowing the creation of graphs and even maps, Datahero.com has online charting facilities, and Statwing.com is an online statistical package for non-statisticians. Truii.com also has data analysis and visualisation tools. One intellectual hurdle to overcome with data analysis is to be able to answer the question: 'When is the analysis complete?' In order to satisfy scientific curiosity, we always want more data and more sophisticated statistical analysis, however for business decisions, a simple graph relating one variable to another may be more than adequate. In order to know when the analysis is complete the focus should be on the audience that you are trying to convince. A scientific audience may not require the same analysis as your CEO, community member or funding sponsor (Rutherfurd *et al.* 2000).

Conclusions

The software functionality required for storing, sharing and analysing data is not new but applying each bit of functionality was time consuming and often required a new piece of software to be purchased and installed. We are entering a new period of data democratisation where all this data management and analysis functionality is available on the web along with many of the data sources that we are analysing. The vast collections of government environmental data being made easily accessible presents a great opportunity for the NRM industry to leverage their own local data sets against public data assets. The key limitation that hinders the NRM industry from fully embracing this data driven decision making is a lack of industry wide reporting to allow the establishment of benchmarks and consistent quantification of environmental condition. It is always difficult to get started when there is no agreed data standards in place about what metrics should be reported and how. Price *et al.* (2009) have suggested a starting data schema for just such an industry wide collation of river restoration projects. The fields suggested by Price *et al.* will inevitably require review and update to allow them to be tailored for specific locations or project situations. The focus need not be on perfect and complete data from the start. Rather on the process and discipline of becoming a data driven industry where we can leverage our local project data against broader public data to demonstrate the value of investing in the environment.

References

Lewis (2003) The art of winning and Unfair Game. W.W. Norton and Company Inc. PP288.

State of the Environment 2011 Committee (2011). Australia state of the environment 2011. Independent report to the Australian Government Minister for Sustainability, Environment, Water, Population and Communities. Canberra: DSEWPaC, 2011.

Rutherfurd ID, Ladson, AR, and Stewardson MJ (2004). Evaluating stream rehabilitation projects: reasons not to, and approaches if you have to. *Australian Journal of Water Resources* 18(1):57–68.

Rutherfurd, I., Jerie, K., Marsh, N. (2000) *A Rehabilitation Manual for Australian Streams*. Land and water Resources Research and Development Corporation, Canberra, and Cooperative Research Centre For Catchment Hydrology, Monash University, Melbourne.

Price P, Lovett S and Davies P, A (2009). National synthesis of river restoration projects. Waterlines report, National Water Commission, Canberra

Wentworth Group of Concerned Scientists (2008). Accounting for Nature: A Model for Building the National Environmental Accounts of Australia. Wentworth Group of Concerned Scientists, www.wentworthgroup.org.

Thomson, B., Hardy, J., Parker, N., Rogers, B. (2012). Monitoring of targeted works to reduce sediment export to waterways entering Moreton Bay. SEQ Catchments Ltd, Brisbane.

Wickham, H (in Press) Tidy Data, Journal of Statistical Software.