**What to do when your data doesn't tell the whole picture..?**

**Abstract**

Increasingly spatial data is being used in the decision making process and strategy development in natural resources. Can your data captured for project specific purposes be used to inform and determine the natural values and threats that exist? What would be the challenges of re-purposing data from one use to investigate another phenomenon?

Data can be captured through the use of ground based survey, aerial based or satellite platforms and come with a range of specific characteristics that can create large biases in any analysis. Having a key understanding of your data's limitations and original purpose is fundamental to the decision making process.

What about what to do when your organisational data doesn't show you what you would expect to see on the ground due to being out of date or lacking in coverage of your area of interest? One method to counteract the absence of complete data is expert elicitation or through aggregation and interpolation across larger areas. What are the pitfalls of this method, would it limit the suitability of the analysis to reflect what's occurring on the ground?

GHD has been working with clients in Victoria to overcome their data limitations to provide whole of catchment analysis in a range of natural resource management areas. The journey of this process has had many outcomes including the realisation that highly targeted on ground field survey is not always the most efficient way to understand and manage whole of catchment level issues.

**The Problem**

Have you ever tried to answer a question using data you knew your organisation had (or had been told existed), but couldn't find it? Was this data in a zip file, stored on a USB in someone's desk draw? All too often valuable data is captured in the field for projects or single use tasks and then never re-used as it's not easily locatable, inconsistent with other data, has no metadata or is simply forgotten about.

Data capture and fieldwork are relatively expensive tasks that organisations undertake themselves or have undertaken by other external parties to gain an understanding of a particular issue. However, this data which is used to gain insight is typically forgotten about and not treated as a tangible asset with future use. All too often the data captured is in a different data schema and different to what is captured by others, further hindering the opportunity to consolidate data into a single repository.

Attempting to carry out analysis using data that is unstructured and doesn't follow a common schema is a task fraught with challenges. The effort to piece together disparate datasets can be immense and if metadata isn't readily available the consistency of the consolidated data might be in question (Wigan, R., Clarke R. (2013). Big Data's Unintended Consequences).

When low emphasis is put onto the importance of the data that is created or collected, time is not set aside to ensure that data quality standards are met on some projects let alone fully implemented over time. Data quality metrics such as accuracy, consistency, completeness and accessibility[i] are the basic indicators that can be used to measure data quality. These indicators may be enhanced or made more appropriate to an organisation by thinking about how coherent the dataset is, what it was captured to achieve and how it should be used in the future. How often is this analysed, let

alone stored in accessible metadata within an organisation? The most common data quality issues prevalent are completeness and accessibility as the data captured has a single purpose to fulfil and often the data is a means to an end rather than seen as an important asset in its own right.

Highly detailed data is sometimes seen as a necessity of survey data, being that a higher resolution will give a more in depth understanding of a single phenomenon. However, this needs to be reviewed in conjunction with spatial extent. Trade-offs occasionally need to be made as a dataset of high resolution may answer one question and well but be less useful in answering ten more questions. Contrasting this, a dataset with a more course resolution may be more suited for answering other questions or may not be useful at all (Lindenmayer, D.F., Likens, G.E. (2018), Maintaining the culture of ecology).

With fieldwork and data capture occurring across many different parts of an organisation, opportunities for cooperative planning across departments and or timeframes are sometimes missed. Simple things that can be completed in the field such as taking geo-tagged photos may increase or even reduce the need for a future field trips as the required knowledge can be extracted from previously captured information.

**Solutions**

There are many solutions to the common data management problems that plague organisations. These solutions can be simple however, the task of who is responsible for managing and coordinating these solutions can be more complex. Some basic standards around data creation, processing, metadata standards, fieldwork planning and standardised workflow processes can achieve a high degree of data uniformity if correctly implemented and practised across the business. An understanding of an organisations overall objectives needs to include a key understanding of the where the data lifecycle falls within the objectives and how it is implemented.

Data standards for field and office based data capture are an integral part of intelligent data creation, and indeed efficient workflows. Data standards enable consistent data capture across staff, teams, offices and more importantly over time. As some projects run for years' data standards make future consolidation and integration of newly captured into the existing data model seamless, particularly for analysis of change over time. For example, a simple data capture standard form may include the types of fields to be captured, what is mandatory to be completed by the user and domain fields to limit the variety of free text entry. A more complex standard may conform to a common or standardised data schema specific for the organisation type or the common data use. However, it is valuable to capture data with a wider perspective of uses and later derive narrower subsets of data for particular uses.

Processes for data capture are key to consistent and reliable information being collected across different users and particularly over time. Collection of data that requires any identification, analysis of a condition or other potentially subjective attributes require robust frameworks for limiting biases or subjectivity at the time of data capture and categorisation.

Forward looking fieldwork planning within organisations can increase efficiencies in data capture and data post processing. Planning, travel and safety considerations all add to the cost of fieldwork, hence combining fieldwork in areas across teams or departments could enable economic efficiencies and or reduce the total amount of fieldwork required.

Field data capture can, in some cases, be enhanced or potentially replaced by different methods such as sensor networks, remote sensing of image or video footage from different platforms. A wide range of earth observation satellite platforms now exist providing a for a range of imagery, video or radar, other platforms such as aerial and drone provide a wide range of data products also such as imagery or LiDAR. Sensor networks are not a new, however advances in technology have increased their popularity as the technology gets more advanced, capability increases and cost reduces. These have the ability to provide long term monitoring data that may reduce fieldwork requirements and dramatically increase the temporal and spatial scale of the data captured. Remote sensing of imagery or video enables the use of different sensors to capture data such as hyperspectral imagery or LiDAR in the field that cater for a wide range of applications. More significantly, these methods enable data to be stored, used, reviewed for multiple purposes over time.

Post capture quality assurance (QA) is a fundamental step prior to any data being used for analysis or shared with others. This sole purpose of this post capture QA process is to ensure the most complete and accurate dataset is created, and limitations understood. The QA would capture and fix gaps, illogical entries and flag outliers while also ensure the data is consistent and complete for the purpose it was captured. This may ideally be an automated (or semi-automated) process that would validate the data against existing data already captured or validate against a published standard.

Consolidation and publishing to enable easy access of organisation data is the key to getting the most value out of existing data and guiding future decisions regarding where fieldwork is required or if existing data can be repurposed. The consolidation of existing data into a single repository and publishing to a platform that enables sharing across an organisation will provide a single source of known, verified data useable by more than just the original user. This will feed into the first step of any decision making process when the question is asked, "*what do we already know about area x or phenomena y from our organisational data?*", which should be asked prior to any data capture or fieldwork.

The idea of data as an asset is not new, and a consolidated repository of organisational data should be treated as just that- a valuable asset. The effort that goes into the overall process from capture to consolidation of data therefore requires recognition that while it may be for a specific project or undertaking, the data captured will be used in the future.

Lack of complete spatial coverage data can lead to the decision that more data is required, and the fieldwork planning process to start. However, sometimes existing data that covers part of the study area of a catchment for example may be a close approximation of the wider area that is data deficient. Data extrapolation across areas without data is a method that needs a good understanding of the phenomena on the ground to be able to accurately validate whether any interpolation is representative of the truth. Interpolation of data across wide areas is prone to create erroneous or invalid data, and therefore the process is hazard prone and requires careful checking of any extrapolated data before use.

Expert elicitation is a method that could be used to back fill or create meaningful data where it doesn't exist from the knowledge of people inside or outside an organisation. This elicitation can be undertaken in different methods, whether it be through surveys, workshops or via interactive web maps to gather data. This can sometimes replace the need for fieldwork, and if used in conjunction with data extrapolation result in a robust dataset if multiple experts are involved giving higher confidence to the data.

**The real world challenges**

GHD has completed a project with a water organisation to consolidate and standardise spatial data captured across 51 separate survey projects along watercourses they manage. The organisation developed a target data schema they wished for all the data to align to, which, once combined would ensure the disparate information could be queried and treated as a single dataset. Many challenges were encountered in this process; firstly, being that limited to no metadata was supplied with the GIS data, only lengthy reports detailing the entire project scope, method and outcome. This is useful background, however a lot of effort is required to distil this into a concise metadata statement of roughly 2 pages, while keeping required information yet without the length of an 80-page report.

The next challenge became apparent when comparing the list of attributes of the target schema and the data provided from the varying projects. Some projects were highly targeted on one particular phenomena such as weed coverage, and included no attributes other than the weed species and an identifier for the point or polygon. This increased the effort to combine the data as mandatory attributes need to be manually collated and created for example date recorded from other sources such as the report.

The end result of this work was a single consolidated dataset. This dataset combined and standardised information from 263 separate files, many with unique data schemas and value mapping. Aligning the datasets into the target schema was a laborious task, however the transformation of this data into the new uniform data schema now functions as a base that can be added to without difficulty. This data schema is also the data standard that all fieldwork must be captured to and delivered as.

**Conclusion**

Forward planning can make data more useful asset for the organisation, however careful thought and planning needs to be done to understand the suitability of organisational data for decision making. There are distinct limitations for using targeted data for extrapolation to catchment scale that need to be taken into account. A consolidated data repository of organisational data is a key asset that can be used to share and increase the knowledge held by an organisation.

---

[i] Xia, J (2012) Metrics to Measure Open Geospatial Data Quality, Indianapolis University
Wigan, R., Clarke R. (2013). Big Data's Unintended Consequences, IEEE Computer 46
Lindenmayer, D.F., Likens, G.E. (2018), Maintaining the culture of ecology, Frontiers in Ecology and the Environment