APLAP 2017–Conference Paper
The promises and pitfalls of big data: parliamentary perspectives

Dr Dianne Heriot
Parliamentary Librarian
Parliament of Australia

Introduction

As group of parliamentary library professionals, we are here to talk about what Big Data means for how we provide information, analysis and advice for our parliamentary clients.

Big Data is the big concept of the day. It seems to offer knowledge nirvana, information utopia: instant, seamless, searchable, complete, global, knowledge about everything.

I have had the opportunity to read the presentations that will be given over the next couple of days by my two Australian colleagues Catherine Lorimer and Dr Fiona Allen. I won't steal their thunder by covering the same material.

In essence, Fiona will talk about the inherent limitations of Big Data and questions the extent to which we in Parliamentary Libraries should be trying to use these techniques.

Catherine reflects on the way navigating the mass of parliamentary information can be strengthened by networks of collaboration between experienced human researchers.

I tackle this subject probably from more of a management perspective.

I have called this talk 'perils and pitfalls of big data' because I want to first explore what exactly big data offers us and our parliamentary clients; I then want to consider the various problems that we are likely to encounter when we are trying to capitalise on these new technologies.

Big data—and Big Information

Before we get into the detail, I think it's important to spend a few minutes discussing what we mean when we the use the term 'big data'—but also what it isn't.

The term, like many buzz words, is often used broadly and with other terms in a bit of a confused way: cloud computing, machine intelligence, the internet of things, data mining, data fusion and so on.

In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

The authors of a book helpfully named 'Big Data'[1] argue that Big Data is also characterised by several things:

- Complete data sets (n=everything) rather than data sampling, enabled by the automated storage of vast amounts of data relating to anything from mobile phone locational data, electricity network usage, or banking transactions

[1] V Mayer-Schonberger and K Cukier, *Big Data: a revolution that will transform how we live, work and think*, John Murray Publishers, London, 2013

- ‘Messy data’ which combines multiple and unstructured data sets, of varying quality and completeness and
- Analysis that is probabilistic rather than giving precise, accurate results: with large data sets, the interest lies in what it reveals in correlations between variables (sometimes quite unexpected) rather than in specific details. Importantly though, it may not reveal causation.

In reading this, it seems to me that when we in Libraries talk about Big Data, many of us actually mean something a bit different.

I think we are actually referring to something that I will call ‘Big Information’. A close cousin of Big Data, Big Information is more about the exponential growth in the amount of information available in digital, searchable format via the internet and delivered by search engines, social media, RSS feeds, email, podcasts, journals and so on.

At the risk of trivialising the issue, I think we can compare the two with fruit and vegetables.

Big data is like this large pile of mixed items. Big information more closely resembles the neat rows of ordered types. With Big data, we are probably interested in the relative mix of items: what might the types and ratios tell us about the size of the household, or its eating habits. With Big information, we want to be able to quickly find a nice peach.

When we talk about the challenges of Big Information, we are referring to a rather more human phenomena where people—and for our purposes in particular, politicians—are simply swamped by the volume of information, news, statistics, and opinion available.

Both Big Data and Big information are of course driven by many of the same technologies:

- the ever growing processing power of computers
- the digitisation of all forms of information, including text, voice, images and transactions
- the rapid fall of the costs of storing large volumes of data and the emergence of ‘cloud storage’ and
- near ubiquitous connectivity of devices and the internet of things, ranging from mobile phones to RFID tagged packages.

But there are still important differences in the characteristics of the two, which we should be aware of. I will talk about this in more detail later in the talk.

The promises of Big Data and Big Information

Let’s first talk about the promises of Big Data and Big Information.

Information to Parliaments

The first and most important is its potential to provide information to our parliamentarians informing their legislative, representational or oversight roles.

[Promise of Big Information]

The promise of Big Information for parliamentary libraries is fairly well understood.

In a basic sense, Big Information is crucial to our operations in familiar ways: the rapid growth of online databases for journals, media monitoring services, online government registries of—for example—company boards, as well as the digitisation of older parliamentary records.

We are finding that one of our fastest growth areas is the demand for detailed statistical information about electorates, especially where it can be graphically represented on maps. Doing this requires

access to a range of data sets both from government sources—such as the Australian Bureau of Statistics and the Australian Electoral Commission—as well as universities, research centres or industry groups.

The resulting analysis can provide insights into the distribution of factors such as educational levels, income, or age group on the ground. This in turn can guide planning for schools, hospitals, public or infrastructure; all issues of great interest to politicians.

However, there is also a cautionary element to this. The promise of Big Data is changing our clients' expectations: they increasingly expect there to be complete information informing issues before parliament.

They increasingly expect information and analysis that crosses international jurisdictional (and often language) borders. They expect the full history of their issue of interest. They sometimes expect that can answer nuanced qualitative questions that the quantitative data cannot support.

And more than anything, they want it quickly.

[Promise of Big Data]
But these points are not really about Big Data.

As I noted earlier, Big Information is about finding a single piece of information in a large sea of data. Big Data is analysis of ALL the data set to detect trends and patterns that are not visible when looking at specific bits of information.

These may hold less direct promise for Parliamentary services. Making use of Big Data ourselves requires the capacity to collect, store and analyse the huge data streams involved. My colleague Fiona explores in more detail whether this is something we can or perhaps should be doing. And for technical reasons, this type of work is likely to be the domain of large companies, universities and think tanks.

However, Big Data is still relevant to Libraries in several ways.

[Purchasing or commissioning]
There are of course opportunities for Parliamentary Libraries to purchase Big Data products. For example, we have been experimenting with commercial products that provide 'sentiment metrics'—analysis of reactions on social media to various issues or people. These products are based on analysis of the information generated across social media platforms such as Facebook, Twitter, or Instagram. These are gradually becoming popular with clients, but there is a learning curve on their part, as well as ours, in understanding how to use these products.

Equally, as the use of Big Data techniques becomes ever more widespread, our clients may begin to want us to commission universities or data companies to conduct modelling on behalf of the Parliament; particularly in support of parliamentary committee inquiries.

[Analysing and advising]
This brings me onto a second role: the power and application of data analysis and modelling for analysing public policy issues is likely to see players in the legislative and political process deploying these tools more often in support of their positions. In these 'battles of the models', Libraries may be increasingly be asked by clients to analyse and assess their validity.

[Legislating data]
Finally, the availability of data and the proliferating ways in which it can be used in business, advertising, finance, law enforcement, political marketing, product delivery etc, means that Parliaments will increasingly have to grapple with the complexities of data access, use and privacy. And if parliament is grappling with these issues, we need to be able to advise them.

I don't think I'd be the first to observe that privacy and the data use are likely to be defining issues of our age.

In all these respects, there is a need for a combination of quite sophisticated skills across disciplinary areas that many Libraries do not currently possess.

Understanding our clients

Big data actually offers a different range of uses if we apply the techniques to the narrower data sets of parliament itself. With everything digitised and logged, there are enormous opportunities to analyse patterns in how parliamentarians use our products, which ones, on what subjects, when they do so and how they access them. In the Library environments where we can record every book borrowed, every article downloaded, every online magazine accessed, and by who; we can get remarkably nuanced information about which products are used, by who, and when.

Similarly, monitoring what issues are getting the most attention in parliamentary proceedings can give a valuable insight into what issues are of most interest to our clients, thereby guiding what subjects we might publish on.

These techniques therefore offer a way to go beyond the often anecdotal 'feel' of researchers and give some useful quantifiable information to aid decision making.

And its pitfalls

So it's clear that big data and big information has much to offer parliamentary libraries and our clients. And I'm sure that many of you have sat through presentations showing a slideshow of various products that promise information utopia at the click of a mouse.

But there are pitfalls on this road, which it pays to be aware of.

As a general point, my own experience has been that implementing these systems is often harder than it looks. It takes longer and costs more than you expect. And it is far from easy to deliver on its promises.

I will just focus on a just a few issues here.

Paradigm shift

I'll start with what I think is a fundamental consequence of the technology changes.

With gigantic volumes of information and data pouring into the Parliament, Parliamentary Libraries do not have the same job they used to. In an age of information scarcity, we collected information. In an age of Big Information, our clients don't have any difficulty in finding things. Quite the opposite: so our challenge is to synthesize information and help clients to assess the veracity and reliability.

This observation is almost a truism in Library discussions. But there is a danger in perceiving the change but not understanding its profound implications. In 1903 the Wright Brothers flew their primitive first aircraft: a little over 100 years later, the first Airbus A380 flew. The differences in

capability, scale, function—and the skills needed to operate and maintain them—are barely recognisable.

The information revolution has brought about similar changes in our environment—they are perhaps less obvious but profound.

We have a similar need to rethink what we do and the skills needed to do it. New technologies that offer great promises also offer irrelevance to those unable to take advantage of them. And that counts as a definite pitfall.

Data quality

Moving onto more specific matters: data quality can be a big issue.

As I mentioned earlier, by definition, big data systems collect large amounts of information from numerous places, via automatic data feeds or through 'scraping' up data sets from the internet. The accuracy and completeness of this data may be variable and once it's been sucked up, it can be unclear where it came from. Of course minor errors may not matter in the context of very large datasets.

But this goes back to my earlier point about the difference between Big Data and Big Information. Issues of accuracy can be ok when the point is to detect general correlations. However, both the technique and the mindset that goes with it may be unsuitable for some applications. In many cases, our products are required not to see the broad outlines of any issues but rather to find a very specific piece of information. Both our clients and the public will expect our products to be accurate and trustworthy. So it's important for everyone concerned in a project to understand what type of information is needed.

Data formats

Data formats can also be a problem. We have had experience of importing data sets into our mapping software that is structured in ways that led to misleading results. But importantly, this was not immediately obvious to the developers or project staff; it was only when the statisticians were doing detailed testing of the prototype system that they noticed it. Had we released it publicly, it could have caused us—and worse, our clients—real embarrassment.

In the meantime, fixing the system involves a time consuming and thereby expensive process of restructuring the data.

Procurement risk

A second issue, probably applicable mostly to Big Information, is the increased procurement risks associated our increasing use of online reference and databases. This arises through exposure to currency variations meaning that maintaining a subscription to a key service can unpredictably jump its forecast budget by substantial amounts – in addition to the more predictable but ongoing price increases.

It also increases our exposure to 'supply chain' risks: the reliability of our services—including time critical services like media monitoring—are the responsibility of vendors rather than our own colleagues. But often in practice, you may be able to contract out a service, but you can't always contract out the responsibility.

And where parliamentary data goes public

Another interesting pitfall of Big Data that arose related to a civil society group dong a 'GovHack': a process where an organisation invited participants to get together to explore ways to utilise public parliamentary data—relating to, for example voting patterns—to build apps or a website showing

citizens how their politicians acted in parliament. We agreed to send one of our researchers along to observe and assist.

For us, the interesting result was that the 'hackers', while expert in coding data applications, had limited knowledge of parliament and its procedures. In short, the data they were accessing did not mean what they thought it did. For example, many decisions in the chambers do not require a formal division, so basing analysis of voting on the records of formal divisions can be very misleading.

I won't go into the detail. But our researcher did end up giving some impromptu tutorials on the workings of parliament and how to interpret the resulting data.

My point in relation to both these points is not to criticise the efforts of the developers. Rather, that making use of data requires technical expertise in the realms of both technology and subject matter.

Navigating promises and pitfalls

Having now traversed some of the promises and pitfalls of the information revolution, I'll conclude with some thoughts on how to manage these issues in practice.

These observations are deliberately high level. They may perhaps seem obvious. But I've also observed that they are areas where teams are very prone to failure.

Project management

The first lesson is a classic project management one: beware of scope creep and mixed objectives. IT projects in general generally involve people from different areas—IT, the business area etc. The team members often see the project as a way to meet a variety of their own business needs. The problem of course may be that several people, all trying to 'kill two birds with one stone', try to take the project in different and possibly conflicting directions.

The basic tools of project management are therefore critical, even with small projects.

Skills development

My second lesson learned is about building skills.

In terms of specific projects, Library research specialists, IT systems specialists, and data analysts all have their own areas of knowledge. Each group tend to see a project through the lens of their own profession, and communicate in their own professional language. This seems obvious. But in practice, it can be quite exceptionally difficult to get them to actually understand what the other means.

Solving this problem depends on building the skills of all staff involved. Everyone appreciates that Library staff may need skills in data systems and IT. But equally, the IT crew may need training in what we do so they understand our key issues.

This leads to the wider observation that for Parliamentary Libraries as a whole, we need to recognise that the overall mix of skills and knowledge is changing. We have always needed to be able to put together teams that have advanced research skills, broad knowledge of parliament and politics, and a sophisticated understanding of public policy. We now need to add diverse capabilities in coding algorithms, data analysis, IT app development, statistics, and design of visualisation tools.

Conclusion

I'll conclude with the thought that the forces that are driving the information revolution—massive computing power, cheap storage, and ubiquitous connectivity—will also deliver the raw material that will drive the phenomena of Big Data and Big Information.

Most libraries have come a long way in making use of Big Information but we are still in the early stages of understanding Big Data.

Both offer many opportunities to deliver information, analysis and advice to our parliamentary clients. And they are opportunities that we don't have the option of declining. To do so risks irrelevance and extinction.

So like it or not, we're on a 'Big' Adventure. So we should enjoy the ride and use the opportunities to revolutionise what we do but also step around the pitfalls.

[END]