Australian Public Service
Better Practice Guide for Big Data and
Big Data Analytics
Contents

Preface

Executive Summary

Introduction

Big Data – a new paradigm?
Scope and Audience

Establishing the Business Requirement

Implementing Big Data Capability

Infrastructure Requirement
Business Process and Change Management
Big data skills and Personnel
Governance and Culture

Information Management in the Big Data Context

Data Management
Privacy
Security

Big Data Project Management

Conclusion

Responsible Data Analytics

Glossary

Resources
Preface

The data held by Australian Government agencies has been recognised as a government and national asset\(^1\). The amount of data held by government is likely to grow as new technologies are adopted and an increasing amount of both structured and unstructured data become available from outside government. The application of big data analytics to this growing resource can increase the value of this asset to government and the Australian people.

Government policy development and service delivery will benefit from the effective and judicious use of big data analytics. Big data analytics can be used to streamline service delivery, create opportunities for innovation, and identify new service and policy approaches as well as support the effective delivery of existing programs across a broad range of government operations - from the maintenance of our national infrastructure, through the enhanced delivery of health services, to reduced response times for emergency personnel.

Technology and the internet are rapidly changing and changing the way government agencies access and use information. This includes opening up new communication channels, enriching the data within government and taking advantage of the growth and availability of data sources outside government.

The Australian Public Service ICT Strategy 2012-2015\(^2\) identified the need to further develop government capability in Big Data. The subsequent Australian Public Service Big Data Strategy\(^3\) outlined the potential of big data analytics to increase the value of the national information asset to government and the Australian people. The Government’s Policy for E-Government and the Digital Economy outlined that the Government will review the policy principles and actions in the Big Data Strategy and finalise a position by the end of 2014.

As new technologies and tools are becoming available to make better use of the increasing volumes of structured and unstructured data. This guide provides advice to agencies on key considerations for adopting and using these tools to assist agencies to make better use of their data assets whilst ensuring that the Government continues to protect the privacy rights of individuals and security of information.

This Better Practice Guide was developed with the assistance of the Big Data Working Group (a multi-agency working group established in February 2013) and the Data Analytics Centre of Excellence Leadership group (established August 2013).

---

\(^1\) FOI Act 1982 s3 (3) The Parliament also intends, by these objects, to increase recognition that information held by the Government is to be managed for public purposes, and is a national resource

\(^2\) Department of Finance, Australian Public Service Information and Communications Technology Strategy 2012-2015

\(^3\) Department of Finance, Australian Public Service Big Data Strategy
Executive Summary

{Summary to be developed}
Introduction

This Better Practice Guide aims to improve government agencies’ competence in big data analytics by informing government agencies about the adoption of big data including:

- identifying the business requirement for big data capability including advice to assist agencies identify where big data analytics might support improved service delivery and the development of better policy;
- developing the capability including infrastructure requirements and the role of cloud computing, skills, business processes and governance;
- considerations of information management in the big data context including
  - assisting agencies in identifying high value datasets,
  - advising on the government use of third party datasets, and the use of government data by third parties,
  - promoting privacy by design,
  - promoting Privacy Impact Assessments (PIA) and articulating peer review and quality assurance processes; and
- big data project management including necessary governance arrangements for big data analytics initiatives.

Government agencies have extensive experience in the application of information management principles that currently guide data management and data analytics practices, much of that experience will continue to apply in a big data context.

This better practice guide is intended initially as an introductory and educative resource for agencies looking to introduce a big data capability and the specific challenges and opportunities that accompany such an implementation. Often there will be elements of experience with implementing and using big data to a greater or lesser degree across government agencies. In this guide we aim to highlight some of the changes that are required to bring big data into the mainstream of agencies operations. More practical guidance on the management of big data initiatives will be developed subsequent to this better practice guide as part of a guide to responsible data analytics.

As outlined greater volumes and a wider variety of data enabled by new technologies presents some significant departures from conventional data management practice. To understand these further we outline the meaning of big data and big data analytics contained and explore how this is different from current practice.

---

4 This Guide does not aim to address the use of big data analytics by the intelligence and law enforcement communities.

5 This Guide does not aim to reproduce or restate better practice guidance for current data management practices. Additional resources on big data can be accessed at the Department of Finance.
Big Data – a new paradigm?

**Big Data**

As outlined in the Big Data Strategy, big data refers to the vast amount of data that is now generated and captured in a variety of formats and from a number of disparate sources.

Gartner’s widely accepted definition describes big data as “…high-volume, high velocity and/or high variety information assets that demand cost-effective innovative forms of information processing for enhanced insight, decision making and process optimization” \(^6\).

Big data exists in both structured and unstructured forms, including data generated by machines such as sensors, machine logs, mobile devices, GPS signals, transactional records and automated streams of information exchanged under initiative such as Standard Business Reporting\(^7\).

**Big Data Analytics**

Big data analytics refers to:

1. Data analysis being undertaken that uses high volume of data from a variety of sources including structured, semi structured, unstructured or even incomplete data; and
2. The phenomenon whereby the size (volume) of the data sets within the data analysis and velocity with which they need to be analysed has outpaced the current abilities of standard business intelligence tools and methods of analysis.

To further clarify the distinction between big data and conventional data management we can consider the current practices:

- Traditional data analysis entails selecting a portion of the available data to analyse, such as taking a dataset from a data warehouse. The data is clean and complete with gaps filled and outliers removed. With this approach hypotheses are tested to see if the evidence supports them. Analysis is done after the data is collected and stored in a storage medium such as an enterprise data warehouse.

- In contrast, big data analysis uses all the available data. The data is messy because it consists of different types of structured, semi-structured and unstructured content. Rather than interrogating data, those analysing explore it to discover insights and understandings. In addition, analysis occurs as the data is captured.

In order to appreciate the shifts required it may be useful for agencies to consider big data as a new paradigm. Table 1 outlines the shifts required to move to the new paradigm.

---

Table 1. The Traditional and New Paradigm with Data

<table>
<thead>
<tr>
<th>Traditional Paradigm</th>
<th>New Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Some of the data</strong></td>
<td><strong>All of the data</strong></td>
</tr>
<tr>
<td><em>For example:</em> An online transaction records key data fields, a timestamp and IP address.</td>
<td><em>For example:</em> An online transaction records mouse movements, prior pages browsed and how long for, keystrokes, all data fields, timestamps, IP address, geospatial location, voice recording of subsequent related phone call as a result of “click to chat”.</td>
</tr>
<tr>
<td><strong>Clean Data</strong></td>
<td><strong>Messy Data</strong></td>
</tr>
<tr>
<td><em>For example:</em> Data sets are mostly relational, defined and delimited.</td>
<td><em>For example:</em> Data sets are not always relational or structured.</td>
</tr>
<tr>
<td><strong>Interrogation of Data to Test Hypotheses</strong></td>
<td><strong>Discovery of Insight</strong></td>
</tr>
<tr>
<td><em>For example:</em> Defined data structures invite the generation and testing of hypotheses against known data fields and relationships.</td>
<td><em>For example:</em> Undefined data structures invite exploration for the generation of insights and the discovery of relationships previously unknown.</td>
</tr>
<tr>
<td><strong>Lag-time Analysis of Data</strong></td>
<td><strong>Real-time Analysis of Data</strong></td>
</tr>
<tr>
<td><em>For example:</em> Data needs to be defined and structured prior to use, and then captured and collated. This duration of extracting data will vary but often involves a delay.</td>
<td><em>For example:</em> Data analysis occurs as the data is captured.</td>
</tr>
</tbody>
</table>

Data integration involving Commonwealth data for statistical and research purposes

Where a big data project involves the integration of Commonwealth data for statistical and research purposes, data users (researchers) and data custodians need to be aware of the Commonwealth arrangements in place that relate to this activity.

On 3 February 2010, the Portfolio Secretaries Meeting (now Secretaries Board) endorsed a set of high level principles for the integration of Commonwealth data for statistical and research purposes. Following the release of the high level principles a set of governance and institutional arrangements to support these principles was also endorsed by the Secretaries Board in October 2010.

Statistical data integration involves integrating unit record data from different administrative and/or survey sources to provide new datasets for statistical and research purposes. The approach leverages more information from the combination of individual datasets than is available from the individual datasets taken separately. Statistical integration aims to maximise the potential statistical value of existing and new datasets, to improve community health, as well as social and economic wellbeing by integrating data across multiple sources and by working with governments, the community and researchers to build a safe and effective environment for statistical data integration activities.
A complete description of the high level principles is available at [www.nss.gov.au](http://www.nss.gov.au) and is summarised below in Table 2:

Table 2: **High level principles for data integration**

<table>
<thead>
<tr>
<th>High level principles for data integration involving Commonwealth data for statistical and research purposes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strategic resource</strong></td>
<td><strong>Principle 1</strong>&lt;br&gt;Responsible agencies should treat data as a strategic resource and design and manage administrative data to support their wider statistical and research use.</td>
</tr>
<tr>
<td><strong>Custodian's accountability</strong></td>
<td><strong>Principle 2</strong>&lt;br&gt;Agencies responsible for source data used in statistical data integration remain individually accountable for their security and confidentiality.</td>
</tr>
<tr>
<td><strong>Integrator’s accountability</strong></td>
<td><strong>Principle 3</strong>&lt;br&gt;A responsible ‘integrating authority’ will be nominated for each statistical data integration proposal.</td>
</tr>
<tr>
<td><strong>Public benefit</strong></td>
<td><strong>Principle 4</strong>&lt;br&gt;Statistical integration should only occur where it provides significant overall benefit to the public.</td>
</tr>
<tr>
<td><strong>Statistical and research purposes</strong></td>
<td><strong>Principle 5</strong>&lt;br&gt;Statistical data integration must be used for statistical and research purposes only.</td>
</tr>
<tr>
<td><strong>Preserving privacy and confidentiality</strong></td>
<td><strong>Principle 6</strong>&lt;br&gt;Policies and procedures used in data integration must minimise any potential impact on privacy and confidentiality.</td>
</tr>
<tr>
<td><strong>Transparency</strong></td>
<td><strong>Principle 7</strong>&lt;br&gt;Statistical data integration will be conducted in an open and accountable way.</td>
</tr>
</tbody>
</table>
Scope and Audience

These practice guidelines are for those who manage big-data and big-data analytics projects or are responsible for the use of data analytics solutions. They are also intended for business leaders and program leaders that are responsible for developing agency capability in the area of big data and big data analytics.\(^8\)

For those agencies currently not using big data or big data analytics, this document may assist strategic planners, business teams and data analysts to consider the value of big data to the current and future programs.

This document is also of relevance to those in industry, research and academia who can work as partners with government on big data analytics projects.

Technical APS personnel who manage big data and/or do big data analytics are invited to join the Data Analytics Centre of Excellence Community of Practice to share information of technical aspects of big data and big data analytics, including achieving best practice with modeling and related requirements. To join the community, send an email to the Data Analytics Centre of Excellence.

---

\(^8\) Technical APS personnel who manage big data and/or do big data analytics are invited to join the Data Analytics Centre of Excellence Community of Practice to share information of technical aspects of big data and big data analytics, including achieving best practice with modelling and related requirements.
Establishing the business requirement

The APS Big Data Strategy highlighted the opportunities and benefits of big data more generally and identified case studies where big data is already being used to benefit by government agencies\(^9\). The opportunities include the chance to improve and transform service delivery, enhance and inform policy development, supplement and enrich official statistics, provide business and economic opportunities, build skills in a key knowledge sector, and derive productivity benefits.

Big data is likely to have application in all government agencies now and into the future. Government agencies will need to consider the extent to which they for the potential to benefit from using big data and big data analytics and whether they need to build a capability to do so. Developing a big data capability requires significant commitment of resources and accompanying shifts in processes, culture and skills. Outside of the usual factors such as cost and return on investment, the decision to develop a Big Data capability needs to take into account several factors:

1. Alignment of a big data capability with strategic objectives - consider the extent to which the agency’s strategic objectives would be supported by a big data capability across the range of activities and over time.

2. The business model of the agency now and into the foreseeable future – consider the extent to which the current business model supports and would be supported by big data capability.

3. Current and future data availability – the extent and range of data sources available to the agency now, and the potential data sources, their cost, and barriers to access.

4. Maturity of the available technology and capability– consideration needs to be given to the extent to which current big data technology and capability can deliver the intended benefits, gather examples of what has been delivered and the practical experience of that implementation.

5. Likelihood of accruing benefits during the development of the capability – consideration needs to be given to whether there is an achievable pathway for developing a big data capability, the ability to take a stepwise approach and expand the solution across more aspects of agency activity as the technology is proven.

6. Availability of skilled personnel to manage big data acquisition and analysis and the organisational environment to support the development of the technology, people and process capability required to use big data.

Once the strategic need for a big data analytics capability has been identified, it is recommended that there is a program of big data projects that is prioritised and implemented.

Big data projects often fall into the domains of scientific, economic and social research, at an operational level big data analytics is applied to customer/client segmentation and marketing research, campaign management, behavioural economics initiatives, enhancing the service delivery experience and efficiency, intelligence discovery, fraud detection and risk scoring.

In particular the types of activities where big data is advantageous include:

- where there is a need to make rapid, high volume, informed decisions;
- where a broader variety of data sources is likely to reveal greater insights into business problems, this includes business problems where:
  - data is currently limited or not available
  - predictability of events is low
  - causal and correlated relationships are not well known; and
- unstructured data features as part of the business problem.

Example: Benefits of Big Data – A case study

(Under development)

Example: Applications of Big Data

(Under Development)
Implementing big data capability

Once the requirement for a big data capability is established there are several considerations that are specific to big data that agencies need to take into account in its implementation.

**Infrastructure requirements**

*Big data storage*

Structured data is traditionally stored in data warehouses and is extracted and manipulated using structured query language or SQL.

In contrast, all three forms of structured, semi structured and unstructured data are now being stored in large ‘big data appliances’. These devices use a parallel data-processing framework such as MapReduce\(^ {10}\) to manage the speeds and volumes required. Big data analysts use non-relational data-management languages which are designed to manage different types of data. In general these appliances need to be highly scalable and optimized for very fast data capture.

Considerations for estimating the type of appliance include\(^ {11}\) –

- Scalability of infrastructure is important to accommodate increasing data stores and uncertain data take on rates. To begin with agencies will need to have an understanding of the likely size of the data that will be captured and stored. Techniques for estimating big data storage requirements include;

- Extensibility – the ability to extend the architecture without introducing limitations;

- Performance criteria such as accessibility requirements, the number of users, the nature of the queries, the speed of the channels to and from the big data appliances, fault tolerance; and

- The compatibility requirements of the appliance. Does the appliance need to be tightly coupled with production systems, or is it a standalone development environment.

\(^{10}\) MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster.

Processing Requirements

Suitable infrastructure options for performing big data analytics will depend on the computing problem and the business application. There are several options for performing big data analytics, agencies will need to consider the mix that is right for their purpose:

Grid computing

Grid Computing uses available processing capacity across a grid of processors. Workload balancing allows high availability and parallel processing of the analytics algorithms. This arrangement is well-suited to applications in which multiple parallel computations can take place independently, without the need to communicate intermediate results between processors. An example is risk scoring individuals within a population.

Cloud Computing

Cloud technology is likely to play a significant role for government agencies as an infrastructure platform for big data applications. This is because big data applications in many circumstances will have stable functional requirements of the underlying infrastructure, their demand for processing and storage is relatively uncertain, and in many instances big data processes are not yet tightly integrated with in-house applications or processes or can be performed as discrete components of end to end business processes. Importantly when considering cloud technology for big data applications, the confidentiality and security risks must also be considered and managed.

Supercomputers/Clusters

Supercomputers are very tightly coupled computer clusters with lots of identical processors and an extremely fast, reliable network between the processors. This infrastructure is suited for highly dependent calculations such as climate modelling or time series correlations.

In-database processing

This technology executes analytics within a data storage appliance. Movement of data is reduced as the processing is performed where the data resides and provides faster run times. In-database processing is more suited to data discovery and exploration, research applications.

---

14 Derived from p7, A guide to implementing cloud services, Department of Finance, September 2012,
In-memory processing.

As processing power increases technology is being increasingly made available that performs the analytics within the memory of an analytics engine. Caching data and using RAM to process analytics functions vastly reduces the query response time, enabling real-time analytics processing. In-memory processing is being used to bring analytics to fast transaction style events such as online interactions informed by analytics.

Business Processes and Change Management

As with any technology that results in new insights and discovery it is important for agencies to consider how they can change business processes to support big data capture, and to take on board the results of big data analysis.

Deriving benefits from data is an iterative process. Improving business processes can lead to richer data, which in turn can drive further business innovations and efficiencies. To benefit from big data, organizations must also commit to linking their big data program to the continuous improvement of their business processes.

Change management practice is an important consideration for agencies that use big data to alter business processes, case assignment and workloads. Big data analytics is often conducted away from those impacted by the outcomes. As far as is practicable big data projects would benefit from change management processes being planned from the commencement of the project.

Big Data Skills and Personnel

“A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data.”

A team implementing big data projects needs a blend of skills to be effective. Depending on the nature of the application these will consist of the leadership, the data management experts, the ‘domain’ expertise or business expertise, the data scientists/miners and project managers and communicators. Where insights from big data will be deployed operationally skilled ‘change’ managers or ‘campaign managers’ are also beneficial as part of the project team. This cross disciplinary approach is recommended whether these teams exist within an agency, across agencies or with external service providers and research partners.

17 Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011
18 Tackling the big data talent challenge, Hudson, Industry Leaders Series, 2013
The technical personnel and skills are the core capability in a big data team. Personnel will need to be skilled in big data technologies in so far as they differ from traditional data warehousing and analysis. They will need this in addition to their traditional skill sets of statistics, machine learning and the model execution and deployment.

Technical personnel should also be trained in and familiar with the general privacy and security obligations as well as any agency specific legislative obligations. Such training would benefit from being grounded in the types of scenarios data scientists are likely to encounter.

Agencies should also consider the demonstrated capacity of their data scientists to innovate, learn new skills and adopt new technologies. This flexibility and capacity to continue to learn and successfully apply new technologies and techniques to different business problems will be the key to maintaining a workforce that will evolve as big data technologies evolve.

Because big data is a growing capability and its applications are broad, there is a significant opportunity to actively share lessons learned and experiences across government agencies. This accelerates the adoption of big data technologies by reducing some of the barriers to adopting big data such as inexperience with the legal, ICT, security and privacy considerations relevant to big data. Encouraging an active community of big data analytics professionals across government agencies will increase learning opportunities and the identification of collaboration opportunities.

The diversity of big data will also require agencies to consider the need to branch out into data mining specialisations such as text mining, optimisation and operations research.

**Whole of Government Data Analytics Centre of Excellence**

The Whole of Government Data Analytics Centre of Excellence (DACoE) shares information, and develops tools and platforms that make better use of data analytics to support better analysis of data trends, inform policy development and the design and delivery of tailored services and enhance understanding and competency across agencies.

The DACoE operates a Community of Practice open to all APS personnel with an interest in data analytics. Two seminar sessions have been run to date around the topics of implementing an analytics capability, and text analytics.

Personnel interested in joining the community of practice are invited to email the DACoE.
Governance and Culture

As with traditional data management and analytics, governance of programs and projects is critical. A strong governance framework will include sensible risk management and a focus on information security, privacy management and agency specific legislative obligations as they relate to data use, acquisition and secrecy.

In a big data environment agencies are required to respect privacy and be responsible for the safe and proper use of data, particularly when the data being used is sensitive. This includes the requirement for agencies to have clear and transparent privacy policies and provide ethical leadership on their use of big data\(^{19}\).

As with many other large analysis projects that may identify scope for change, big data projects can be risky – they require experimentation and discovery, they can deliver unexpected results and sometimes no significant results, in extremely rare cases they can deliver false results. As such, governance of the development of a big data capability should also include managing expectations at the outset of big data projects; this includes stakeholders, technical personnel and the community.

Because projects using big data can be complex and lead to uncertain results, it is important that agencies cultivate a dynamic culture of discovery, experimentation, evaluation and feedback. It is this culture of enquiry and managing the expectation of data yielding a result that will enable agencies to realise the full potential of big data and avoid the risks of overinvesting.

A sponsor or champion of the big data capability on the senior management team is an important consideration to ensure that big data projects traverse functional boundaries, a culture of discovery and experimentation is modelled, and the management team develop a fluency in the identification of potential value from big data.

\(^{19}\) Responsible Data Analytics: {guide to be developed by July 2014}
Cultivating a dynamic culture of discovery at Department of Immigration and Border Protection

The Department of immigration and Border Protection (DIBP) hosts a research week for its data scientists. The research week concept exemplifies the culture of innovation. It recognises that successful innovation requires a cultural setting that accepts some risk and relies on staff who are actively working to support corporate goals and processes.

DIBP promotes a culture of innovation by encouraging staff to think and try out their ideas. The laboratory environment allows staff to test their ideas in a safe and affordable environment. Active, solutions-focussed thinking is encouraged, reinforced by a strong partnership between analysts and business areas.

While a few of the projects undertaken have a direct path into mainstream processes, projects that prove unsuccessful are not considered a failure since the lessons learnt are of value in themselves.

All participants are given the opportunity to learn from their own and their colleagues’ research, including those where the conclusion of the project is that something cannot be made to work.

In addition to the benefits of projects that benefit mainstream processes, and findings that shortcut future development pathways, there are indirect benefits to cultivating a culture of innovation. This approach encourages and rewards specialist staff to stay in touch with developments, test ideas in a real-world environment and facilitates better collaboration for other work. This is both challenging and attractive to analytics professionals and presents DIBP as an exciting place to work.
Information management in the big data context

Data Management

Commonwealth data, including data used by big data analytics projects, needs to be authentic, accurate and reliable if it is to be used to support accountability and decision making in agencies. The creation, collection, management, use and disposal of agency data are governed by a number of legislative and regulatory requirements and government policies and plans. These regulatory requirements also apply to big data sources and the management of big data.

Big data is like other data with respect to existing policy, regulation and legislation. Agencies need to ensure that, when considering their acquisition and management of big data, they are complying with their obligations under these policies.

Agencies are already experienced in traditional management of data and the issues of the information management lifecycle. Traditional information management practices will need to continue for the foreseeable future, and be supplemented increasingly with the consideration of big data as agencies develop their capability. This section aims to outline some of the specific shifts agencies may experience in their adoption of big data management.

Acquiring big data

To realise the benefits of big data it is important to identify the variety of diverse data sources that may be applicable, including those potentially tangentially related to the business problem (for example: accelerometer readings from mobile devices on roads can be used to detect infrastructure problems such as potholes). Traditionally agencies have built, acquired and publicly shared structured data sets. Agencies also have access to their own semi-structured and unstructured data stored in various media such as document repositories and other large storage devices.

To realise the potential of big data it is important for agencies to develop systematic intelligence on data available, identify high-value data sources and produce and maintain information asset registers.

Potential data sources include:

- Administration activities and agency transaction systems;
- Unstructured and structured information available from querying the internet;
- Businesses who are using data they own and making it available as part of a new data trading business model;
- Government data sets that are made available to citizens and businesses via data.gov.au and other data portals;

---


21 Where Commonwealth administrative data is integrated with one or more data sources at the unit record level for statistical or research purposes, then the arrangements for data integration involving Commonwealth data for statistical and research purposes apply.
- Publicly available data from non-government sources;
- Data from sensors, mobile devices and other data collection sources; and
- Research data sets.

Agencies looking to develop intelligence on data sources need to produce and maintain details such as the value of the source, the metadata that has been captured, whether data is publicly available or is restricted in its access, cost of the data, the intellectual property rights of the data (e.g. creative commons) and other relevant information required by users to make decisions on use of the data.

When designing and managing administrative datasets, the responsible agency should consider the potential statistical value of the datasets for public good, both in terms of use by their own agency, and use more broadly. In particular the interoperability of common data elements should be considered and specific elements such as spatial and socioeconomic data elements should be considered in the context of the Statistical Spatial Framework. For example an agency could collect information such as health or wealth statistics that includes relevant spatial information such as the residential address of the patient, the operating addresses of healthcare providers including hospitals and transport nodes. Where these are geo-referenced it is important that the relevant statistical and spatial metadata is accessible so that it can be understood and useful across the statistical and spatial data-user communities.

**Indexing Data**

The costs of data storage is continually falling, and the insights to be gained from more data is creating an incentive to hold more data. Indexing data and ensuring that it has adequate metadata is increasingly important in the big data context. This allows it to be easily retrieved and used for business purposes.

As big data tools and technologies will be able to analyse the content of unstructured and semi-structured data files such as audio, video, images, tweets, wikis and blogs there will be a growing requirement to ensure that the metadata for these types of files is also adequate. At this stage the most effective way to do this is to create a semantic data model layer that helps users to make sense of the stored data.

---

22 High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes, National Statistical Service


24 http://www.gartner.com/it-glossary/semantic-data-model - A method of organizing data that reflects the basic meaning of data items and the relationships among them. This organization makes it easier to develop application programs and to maintain the consistency of data when it is updated.
Harvesting versus Storing Data

With an increasing rate of data storage there is a related growth in the range of business risks including security and privacy breaches, multiple versions of the truth, reputational risks, and excess costs. Where appropriate, agencies need to look at options such as ‘harvesting’ data, using it immediately and then discarding it. This can limit the risks associated with acquisition and storage of large amounts of data.

Harvesting data from various sources to meet specific business requirements is selective in that a just-in-time approach is used to acquire only the data that is required. It is also a repeatable process that can be done as often as necessary. It therefore can negate the need to store the extracted data unless there is an operational, legal or administrative reason that compels its retention.

Data retention, accessibility and disposal

Australian Government data is considered a Commonwealth record and this is no different for big data. Relevant archiving, data management and disposal obligations apply. In the era of big data it is likely that new information and new data can be created from existing data and records held by government. The same rules of archiving, data management and disposal will apply to any created data items; these created data items will need to be assessed against relevant criteria to determine their status and treatment as a Commonwealth record.

Semantic Data Models

In the big data context data collections are likely to be highly fragmented, they are likely to have different metadata schemes, overlapping classifications, and incongruent definitions. A semantic model allows for a congruence to be achieved across disparate data sets allowing for analysis with a richer set of information, and allowing for additional sources of knowledge to be incorporated into those data sets.

* Adapted from, New Directions in Data Analytics in The ABS, Presented by Ric Clarke, Assistant Statistician, Australian Bureau of Statistics, 27 November 2013

---

Privacy

‘In the last 5 years we have seen a significant change in how people communicate and interact online. People’s attitude to the importance of personal privacy protection is changing at the same time,’
Professor McMillan, Australian Information Commissioner26

‘The OAIC’s 2013 Community Attitudes to Privacy survey results show that 96% of Australian expect to be informed how their information is handled, and if it is lost. It is clear that the Australian public continues to insist that their personal information in handled with the highest possible standards’
Timothy Pilgrim, Australian Privacy Commissioner27

Where data sources for big data activities include personal information (ie, information about individual who are identifiable, or can be identified from the information) the application of the Privacy Act 1988 (Privacy Act) must also be considered. The Privacy Act which regulates the handling of personal information throughout the information lifecycle, including collection, storage and security, use, disclosure, and destruction.

From March 2014, the Australian Privacy Principles will replace the National Privacy Principles and Information Privacy Principles and will apply to organisations, businesses (other than some types of small businesses), and Australian, ACT and Norfolk Island Government agencies.

Privacy principles apply in the current context of data and information management, and they continue to apply in the context of big data. Considerations specific to big data relate to some specific features of big data and big data projects. This includes:

- The collection of personal information from sources other than the individual.
- The creation of new data through big data analytics that generates enhanced information about a person.
- The capacity to compromise anonymity through the collation of a range of data that reveals identity (for example the mosaic effect28).
- The potential for unstructured information sources to hold personal information not known by the individual.

In order to protect privacy considering and complying with requirements set out in the Australian Privacy Principles will ensure adequate consideration for the protection of privacy. Agencies engaged in big data projects will need to ensure that they give adequate consideration to protecting privacy paying attention to three areas that are particularly relevant for big data. These are:

- Robust de-identification capabilities. Identifiers can be removed via a number of methods including deletion (safe harbour29), masking, aggregation and other statistical techniques (collectively known as expert


28 The concept whereby data elements that is isolation appear anonymous can amount to a privacy breach when combined. This is increasing as a possibility as data analysts become more adept at joining disparate data sets that can result in revealing identity.

determination\(^\text{30}\)). As well as these techniques consideration should also be given to appropriate separation of duties and knowledge of the personnel working with the data. Existing draft guidance on de-identification is available from the Office of the Australian Information Commissioner\(^\text{31}\).

- **Privacy by design\(^\text{32}\)** - agencies will need to ensure that privacy risks are adequately managed at all stages of a project, and in all facets of a project. This includes technology, people and process considerations. There are some considerations that are likely to impact on big data projects specifically:
  - Where big data projects incorporate the use of online and mobile applications or other sensors, it is important to consider the privacy safeguards from the point of collection.
  - The potential generation of new personal information through the bringing together of data sets.
  - The potential discovery of new personal information in unstructured data sources.

- **Privacy Impact Assessments\(^\text{33}\)** — A privacy impact assessment (PIA) is a tool used to describe how personal information flows in a project and is used to help analyse the possible privacy impacts on individuals and identify recommended options for managing, minimising or eradicating these impacts. Ideally, a PIA should be commenced in the planning stages of a project; PIAs work most effectively when they evolve with and help to shape the project's development. PIAs are one way of assisting implementation of Privacy by Design.

---


7 principles of privacy by design

1. **Proactive not Reactive; Preventative not Remedial**
   The Privacy by Design (PbD) approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy-invasive events before they happen. PbD does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred – it aims to prevent them from occurring. In short, Privacy by Design comes before-the-fact, not after.

2. **Privacy as the Default Setting**
   We can all be certain of one thing – the default rules! Privacy by Design seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy – it is built into the system, by default.

3. **Privacy Embedded into Design**
   Privacy is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. The result is that it becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

4. **Full Functionality – Positive-Sum, not Zero-Sum**
   Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. Privacy by Design avoids the pretence of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

5. **End-to-End Security – Full Lifecycle Protection**
   Privacy by Design, having been embedded into the system prior to the first element of information being collected, extends throughout the entire lifecycle of the data involved, from start to finish. This ensures that at the end of the process, all data are securely destroyed, in a timely fashion. Thus, Privacy by Design ensures cradle to grave, lifecycle management of information, end-to-end.

6. **Visibility and Transparency – Keep it Open**
   Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember, trust but verify.

7. **Respect for User Privacy – Keep it User-Centric**
   Above all, Privacy by Design requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric.

Security

In the context of big data, security policies apply as they do for traditional data. The Protective Security Policy Framework\textsuperscript{34} and supporting guidelines\textsuperscript{35} outline the requirements for government agencies in the management of the security of data including managing cross border flows of information and outsourced arrangements. Where data contains personal information the Office of the Australian Information Commissioner also provides guidance on information security\textsuperscript{36}.

Considerations of security specific to big data include, among others:

- Increased value of the information asset as government agencies enrich their data, its aggregation and the insights derived from it.
- Perceptions of the increased value of information as an asset, as private sector business models seek to profit from providing data sources for research and insight become more commonplace.
- The increasing range of data acquisition channels and their potential vulnerability.
- The unknowability of the content of unstructured data sources upon acquisition.
- Increased distribution of physical and virtual locations of data storage.

These considerations and the extent to which they apply to a big data project or agency should modify the assessment of the risks and the existing controls from the classification of data and its subsequent protection.

\textsuperscript{34} Attorney General’s Department; \url{Protective Security Policy Framework}

\textsuperscript{35} \url{http://protectivesecurity.gov.au/informationsecurity/Pages/Supporting-guidelines-to-information-security-(including-the-classification-system).aspx}

\textsuperscript{36} OAIC Information Security guide; \url{http://www.oaic.gov.au/privacy/privacy-resources/privacy-guides/guide-to-information-security}
Big data project management

There are some key characteristics of big data projects that need to be considered when applying project management methods:

- Big data projects combine software development, business process and scientific research.
- Information security and privacy are key considerations in all big data projects.
- Often the methods used in big data projects are highly technical and not understood by those outside the data science profession.
- There are a wider range of stakeholders in big data projects than is often apparent. Often their data is used as part of the project, or they are impacted by the findings of the project.
- When applied to service delivery type applications, big data findings realise their full benefit when applied to improve business processes – this often involves a change in business processes or the type of work done and is often neglected in the project scope.
- There are some considerations with respect to project outcomes and managing project risks that are relevant for big data projects including-
  - The possibility that the data acquired will not provide the insights sought.
  - Findings and results from big data projects can be uncertain and counterintuitive.
  - False discoveries are possible – in the big data context large amounts of data can produce statistically significant results that arise purely by chance.

These characteristics introduce some core differences in the way big data projects need to be managed.

Earlier better practice guidance on managing scientific projects is available from the ANAO\textsuperscript{37}. Much of the advice in this guidance is still relevant and used today. It is also worth considering that big data projects can often share characteristics of software development projects, and the nature of the uncertainty associated with obtaining insight from data lends itself to Agile development methodologies. Agile approaches are typically used in software development to help businesses respond to unpredictability\textsuperscript{38}.

Some features of project management that become important in managing big data projects are:

- As with all projects, a clear and accountable owner of the business objectives and project outcomes.

\textsuperscript{37} Management of Scientific Research and Development Projects in Commonwealth Agencies Better Practice Guide For Senior Management, ANAO 2003

\textsuperscript{38} \url{http://agilemethodology.org/}
• An adequate risk adjusted expectation of return on investment to account for uncertainty.

• A commitment to an incremental and iterative\textsuperscript{39} development methodology, including the discipline to stop projects after the feasibility stage if they do not demonstrate potential.

• Where new, more valuable but unexpected insights might arise from big data, a process to consider the current project scope a new and maybe additional course of action.

• An engaged business owner able to commit to considering and implementing results and findings.

• Where technical methods are used to transform data into insights, peer review of methodology used, its reproducibility and statistical soundness are encouraged.

• Similarly quality assurance processes need to apply to all aspects of the project including the data sources, data management and data analytics.

• Privacy and security assessments should form specific components of a broader risk assessment of the big data project.

• Privacy by design principles should be applied at all stages of the big data project.

• Stakeholder engagement plans for big data projects should include transparent review processes of the data acquisition, data management, results, and the application of the findings where appropriate.

Big data projects will benefit from the allowance and adjustment for these requirements in a traditional project management methodology.

\textsuperscript{39} \url{http://scrumreferencecard.com/scrum-reference-card/}, Accessed 5 December 2013
Conclusion

This guide aims to identify developments in current information management and analytics practices as they relate to big data. Adopting these practices to take advantage of big data will allow agencies to deliver enhanced services, improved policy development and identify new services and opportunities to make use of the national information asset that is Australian Government data.

The Big Data Strategy Working Group and Data Analytics Centre of Excellence will continue to work with government agencies to identify opportunities for big data projects, strengthen and articulate existing processes for the responsible use of big data, monitor technical advances, develop appropriate skills to extract value from big data and adequately protect privacy and security of information. In addition, one of the Government’s election commitments is to seek proposals from agencies, researchers and the private sector for joint projects using big data that have promising efficiency or service quality payoffs.

The field of information management and analytics is rapidly evolving, and this guide, in as far as it is able, aims to describe the technology, skills and processes required. This guide will be updated and reviewed every two years to reflect developments in technology and policy.
Responsible data analytics

{ Under Development }

The APS Big Data Strategy outlined an action to develop a guide to responsible data analytics.

**Action 4: Develop a guide to responsible data analytics [by July 2014]**

The Big Data Working Group will work in conjunction with the DACoE to develop a guide to responsible data analytics. This guide will focus on the governance of big data projects and will incorporate the recommendations and guidance of the OAIC in regards to privacy.

The guide will also include information for agencies on the role of the National Statistical Service (NSS) and the Cross Portfolio Data Integration Oversight Board and its secretariat.

The guide will incorporate the NSS produced *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*, this includes how and when agencies should interact with the secretariat as they develop big data projects that involve the integration of data held by Commonwealth agencies. The guide will also investigate the potential for a transparent review process to support these projects.

This guide is under development and will focus in more detail on practical management of big data and is expected to be developed by July 2014.
Glossary

Cloud computing
Cloud computing is an ICT sourcing and delivery model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
This cloud model promotes availability and is composed of five essential characteristics: on demand self service, broad network access, resource pooling, rapid elasticity and measured service.

Data scientists
A data scientist has strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems; they will pick the right problems that have the most value to the organization.
Whereas a traditional data analyst may look only at data from a single source a data scientist will most likely explore and examine data from multiple disparate sources. The data scientist will sift through incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem. A data scientist does not simply collect and report on data, but also looks at it from many angles, determines what it means, then recommends ways to apply the data.40

De-identification
De-identification is a process by which a collection of data or information (for example, a dataset) is altered to remove or obscure personal identifiers and personal information (that is, information that would allow the identification of individuals who are the source or subject of the data or information).41

Information assets
Information in the form of a core strategic asset required to meet organisational outcomes and relevant legislative and administrative requirements.

Information assets register
In accordance with Principle 5 of the Open PSI principles, an information asset register is a central, publicly available list of an agency’s information assets intended to increase the discoverability and reusability of agency information assets by both internal and external users.

Mosaic effect
The concept whereby data elements that in isolation appear anonymous can lead to a privacy breach when combined.42

Privacy-by-design
Privacy-by-design refers to privacy protections being built into everyday agency/business practices. Privacy and data protection are considered throughout the entire life cycle of a big data project. Privacy-by-design helps ensure the effective implementation of privacy protections.43

---

43 http://www.privacybydesign.ca/
Privacy impact assessment (PIA)
A privacy impact assessment (PIA) is a tool used to describe how personal information flows in a project. PIAs are also used to help analyse the possible privacy impacts on individuals and identify recommended options for managing, minimising or eradicating these impacts.44

Public sector information (PSI)
Data, information or content that is generated, created, collected, processed, preserved, maintained, disseminated or funded by (or for) the government or public institutions.45

Semi-structured data
Semi-structured data is data that does not conform to a formal structure based on standardised data models. However semi-structured data may contain tags or other meta-data to organise it.

Structured data
The term structured data refers to data that is identifiable and organized in a structured way. The most common form of structured data is a database where specific information is stored based on a methodology of columns and rows.
Structured data is machine readable and also efficiently organised for human readers.

Unstructured data
The term unstructured data refers to any data that has little identifiable structure. Images, videos, email, documents and text fall into the category of unstructured data.

Resources

{ Under Development }

{ Under Development }