Australian Government

Australian Public Service Better Practice Guide for Big Data

APRIL 2014

Joint work of the Data Analytics Centre of Excellence (chaired by the Australian Taxation Office) and the Big Data Working Group (chaired by the Department of Finance)

ISBN: 978-1-922096-31-9

Contents

Australian Public Service

Better Practice Guide for Big Data

Preface 1

Executive Summary 2

Introduction 3

- Big Data – a new paradigm? 4
- Scope and Audience 6

Establishing the business requirement 7

Implementing big data capability 10

- Infrastructure requirements 10
- Business Processes and Change Management 12
- Skills and Personnel 13
- Governance and Culture 14

Information management in the big data context 16

- Data Management 16
- Privacy 19
- Security 22

Big data project management 23

Conclusion 25

Responsible data analytics 26

Glossary 27

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. Developments in technology are adding new types of data and new methods of analysis to the advanced analytics capabilities already being used in government agencies today. Departments are now able to ask questions that were previously unanswerable, because the data wasn't available or the processing methods were not feasible. The application of big data and 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.

The Australian Public Service ICT Strategy 2012-2015[2] outlined the aims of improving service delivery, increasing efficiency of government operations and engaging openly. The Strategy identified the need to further develop government capability in Big Data to assist in achieving these aims. 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.

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).

As new technologies and tools are becoming available to make better use of the increasing volumes of structured and unstructured data this guide aims to provide advice to agencies on key considerations for adopting and using these tools, assisting 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.

---

[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

This Better Practice Guide aims to address the key considerations for government agencies when growing their capability in big data and big data analytics.

Big data represents an opportunity to address complex, high volume and high speed problems that have previously been beyond the capability of traditional methods. This includes finding new solutions for enhanced evidence based policy research, improved service delivery and increased efficiency.

Agencies need to identify the opportunities brought by big data and assess their alignment with strategic objectives and future business operating models. Agencies also need to articulate the pathway for developing a capability to take advantage of the opportunities.

Developing a capability in this area requires specific considerations of technology, business processes, governance, project management and skills.

Technology supporting big data represents a departure from today's information management technology. Organisations need to evaluate benefits and cost effectiveness of moving to newer technologies. In the early stages of developing capability agencies are advised to start small, consider bridging approaches with existing infrastructure where appropriate, be prepared to iterate solutions and plan for scalability.

To encourage successful projects agencies need to cultivate a culture of experimentation, adopting lean and agile methodologies to explore and deliver solutions. To realise the benefits from solutions developed agencies need to establish processes to adopt and transform service delivery, operating models and policy in response to solutions and insights.

The transformative power of big data insights, the changing technology, community attitudes, privacy and security considerations demand close governance of big data programs, this should include both internal and community engagement.

Government policy in relation to privacy and security continues to evolve in response to new technology, community attitudes and new risks. Industry standards and best practices are still forming and the profession is continuing to learn the possibilities and limits of the field. This better practice guide aims as far as it can to cover the big data territory generally relevant for government agencies.

Specific solutions and innovations will arise as agencies develop their capability and ongoing dialogue and communication of new approaches ought to continue through the Big Data Strategy Working Group and the Whole of Government Data Analytics Center of Excellence. Better practices that develop will be reflected in future iterations of this guide.

Further work is required to provide specific guidance and approaches for managing the responsible use of the data and data analytics to address vulnerabilities in relation to privacy, security, acquisition of data and the application of insights obtained from data. Work on this guidance is continuing and further consultation will be undertaken on the responsible use of data analytics.

Introduction

Big data technology and methods are still quite new and industry standards of better practice are still forming. To the extent that there are accepted standards and practices this Better Practice Guide aims to improve government agencies' competence in big data analytics by informing government agencies[4] about the adoption of big data[5] 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.
- incorporating guidance on the responsible use of data analytics[6]

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 (BPG) is intended initially as an introductory and educative resource for agencies looking to introduce a 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 the BPG 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 specific initiatives will be developed subsequent to this BPG 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

[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.

[6] To be developed by July 2014

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.

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" [7].

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[8].

Big Data Analytics

Big data analytics[9] 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.
3. The complexity of the relationships with complex structures embedded in the data has reached a level that cannot be handled by the current tools and models of statistics and 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 relevant 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 a wider variety of available data relevant to the analytics problem. The data is messy because it consists of different types of

[7] Gartner, The Importance of 'Big Data: A Definition, http://www.gartner.com/id=2057415, Accessed 27 November 2013

[8] http://www.sbr.gov.au/

[9] Unless otherwise specified, reference to data, analytics, projects, technologies skills and personnel can be assumed to be in the context of big data

structured, semi-structured and unstructured content. There are complex coupling relationships in big data from syntactic, semantic, social, cultural, economic, organisational and other aspects. Rather than interrogating data, those analysing explore it to discover insights and understandings such as relevant data and relationships to explore further.

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 1.1: The Traditional and New Paradigm with Data

Traditional Paradigm	New Paradigm
Some of the data *For example: An online transaction records key data fields, a timestamp and IP address.*	**All of the data** *For example: Clickstream and path analysis of web based traffic, all data fields, timestamps, IP address, geospatial location where relevant, cross channel transaction monitoring from web, through to call centres.*
Clean Data *For example: Data sets are mostly relational, defined and delimited.*	**Messy Data** *For example: Data sets are not always relational or structured.*
Deterministic relationships *For example: In relational data stores, the data often has association, correlation, and dependency following classic mathematic or statistical principles, often designed and as a result of the data modelling process and the cleansing process. Including predictable statistical features such as independent and identically distributed variables.*	**Complex coupling relationships** *For example: Data can be coupled, duplicative, overlapping, incomplete, have multiple meanings all of which cannot be handled by classic relational learning theories and tools. Often the data does not lend itself to the standard types of statistical assumptions as relational data sets.*
Interrogation of Data to Test Hypotheses *For example: Defined data structures invite the generation and testing of hypotheses against known data fields and relationships.*	**Discovery of Insight** *For example: Undefined data structures invite exploration for the generation of insights and the discovery of relationships previously unknown.*
Lag-time Analysis of Data *For example: 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.*	**Real-time Analysis of Data** *For example: Data analysis occurs as the data is captured.*

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[10].

For those agencies currently not using data analytics, this document may assist strategic planners, business teams and data analysts to consider the value of this capability 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 analytics projects.

Technical APS personnel who manage data and/or do data analytics are invited to join the Data Analytics Centre of Excellence Community of Practice to share information of technical aspects of data management and data analytics, including achieving best practice with modelling and related requirements. To join the community, send an email to the Data Analytics Centre of Excellence (DataAnalyticsCentreofExcellence@ato.gov.au).

[10] 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 it is already being used by government to benefit agencies and the public[11]. 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 can benefit from data analytics and whether they need to build a capability to do so. Developing such a 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 the capability needs to take into account several factors:

1. Alignment of the capability with strategic objectives - consider the extent to which the agency's strategic objectives would be supported 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 the 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 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 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 data acquisition and analysis and the organisational environment to support the development of the technology, people and process capability required.

Once the strategic need for a capability has been identified, it is recommended that a vision for the capability is articulated and communicated to all stakeholders including the community and a program of big data projects established. Evaluation of such a program should be planned for at the program's commencement to determine outcomes, impacts and benefits.

[11] The Australian Public Service Big Data Strategy, Department of Finance and Deregulation, August 2013, http://www.finance.gov.au/big-data/, Accessed, 26 November 2013

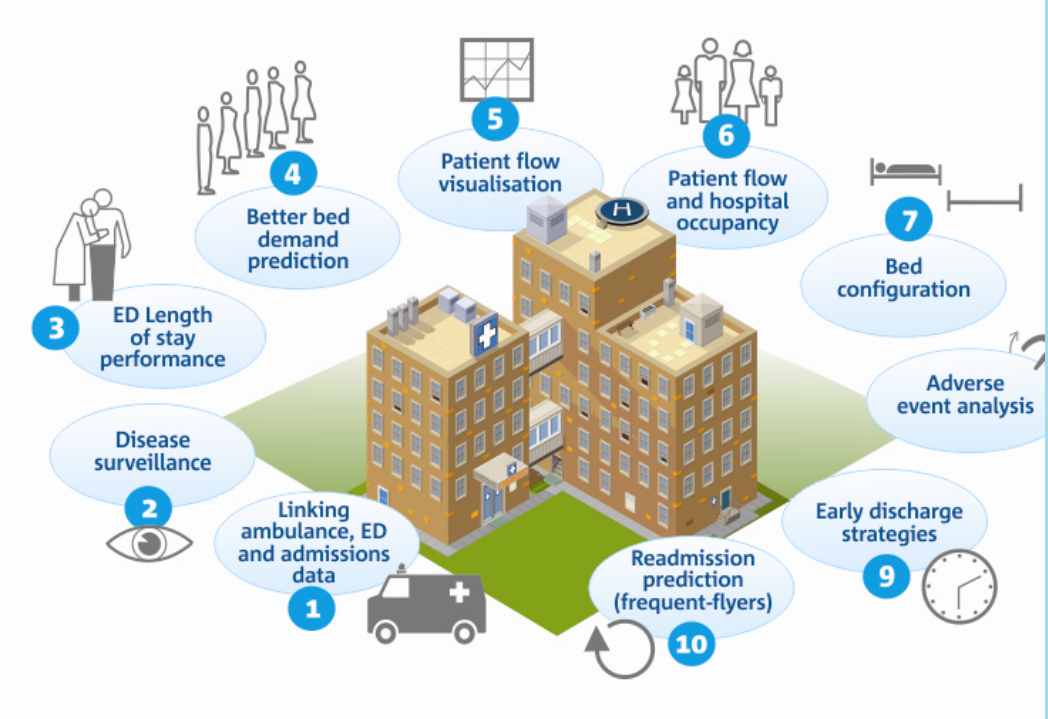
Case study - benefits of big data application - Patient Admissions Prediction Tool (PAPT)

The PAPT software predicts how many patients will arrive at emergency, their medical needs and how many will be admitted or discharged and allows on-the-ground staff to see what their patient load will be like in the next hour, the rest of the day, into next week, or even on holidays with varying dates, such as Easter and locally specific events.

The development of the PAPT software is a collaboration between the Australian e-Health Research Centre, Queensland Health, Griffith University and Queensland University of Technology. This partnership continues with new features and data sources are being incorporated including working with clinicians to build in process information, extending PAPT to predict admissions due to diseases from influenza through to chronic diseases.

The PAPT software has enabled the development of new responses to predicted large events For example during 'schoolies' week up to 20% of emergency presentations will be schoolies - having this information about presentations and admissions allows hospitals to plan the staff, medical supplies and beds needed to care for those schoolies and manage waiting times for other patients who are still arriving with other serious injuries. New responses have been developed such as work with Queensland Ambulance Service to establish a designated medic tent treating as many of these schoolies patients on the ground to free up beds in our hospital emergency department for more serious cases.

PAPT is used in 31 hospitals across Queensland and the forecasting provided by the tool assists with hospital bed management, staff resourcing, scheduling of elective surgery. For patients it should mean improved outcomes including timely delivery of emergency care, improved quality of care and less time spent in hospital. It is estimated that PAPT software has the potential to save $23 million a year in improved service efficiency for the health system if implemented in hospitals across Australia.



With Permission: Dr James Lind, Director of Access and Patient Flow Unit, Gold Coast University Hospital

<http://www.csiro.au/Outcomes/Health-and-Wellbeing/Technologies/PAPT.aspx#a1>, Accessed 30 January 2014

<http://www.health.qld.gov.au/news/stories/131115-schoolies.asp>, Accessed 30 January 2014

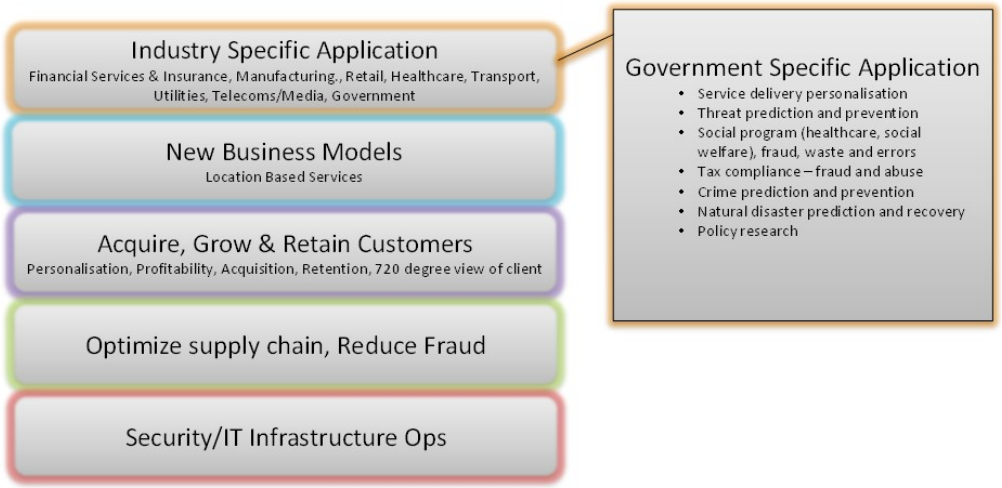
<http://www.csironewsblog.com> , Accessed 30 January 2014

Big data projects often fall into the domains of scientific, economic and social research, at an operational level 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.

Categories of business processes that can benefit from big data projects



Source: BigInsights Submission, www.biginsights.com

Implementing big data capability

Once the requirement for a 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

Storage

Structured data is traditionally stored in data warehouses and is extracted and manipulated using Structured Query Language or SQL. Using these technologies to pilot and explore the feasibility of projects may be worthwhile even where this requires different data schemas to be established.

While there is still a role for the relational databases, progressively all three forms of structured, semi structured and unstructured data are being stored in large 'big data appliances'[12] or clusters. These devices use a parallel data-processing framework such as MapReduce or Bulk Synchronous Parallel (BSP)[13] to manage the speeds and volumes required. In addition to processing frameworks, datastore structures are evolving to encompass the variety and multiplicity of relation types in the data. NoSQL[14] data stores span column storage, document stores, key-value stores and graph databases. Each of these types of data stores are optimised for kinds of data storage and retrieval uses. In general these appliances need to be highly scalable and optimised for very fast data capture and retrieval.

Considerations for estimating the type of appliance or cluster architecture include[15]

- 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 storage requirements include;

[12] The concept of an 'appliance' is that it is pre-configured and ready to run when data is added. Appliances generally offer specialised hardware, a management system and storage. Some vendors offer 'adaptors' to allow easier interface with traditional relational database management structures and, increasingly, versions of open source big data management software such as Hadoop and MapReduce. While appliances have their use, they are not required to work in the big data space.

[13] MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Bulk Synchronous Parallel (BSP) is another parallel processing algorithm – each have strengths and weakness for particular problem types. More generally these frameworks allow for the scaling up of existing analytics algorithms to large scales and fast speeds.

[14] NoSQL databases are Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable. Source: http://nosql-database.org/, Accessed 29 January 2014

[15] Adapted from http://data-informed.com/considerations-for-storage-appliances-and-nosql-systems-for-big-data-analytics-management/, Accessed 4 December 2013

- 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 and relevant storage configurations, the speed of the channels to and from the appliances, fault tolerance; and
- The compatibility requirements of the appliance. Does the appliance need to be tightly coupled with production systems, existing data stores, or is it a standalone development environment.

Processing Requirements

Suitable infrastructure options for performing analytics will depend on the computing problem and the business application. There are several options for performing 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.[16] An example is risk scoring individuals within a population.

Cloud Computing

Cloud technology[17] is likely to play a significant role for government agencies as an infrastructure platform for 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[18]. Importantly when considering cloud technology for applications, the confidentiality and security risks associated with cloud computing must be considered and managed alongside other risks.

Supercomputers/Clusters

Supercomputers are very tightly coupled computer clusters with lots of identical processors and an extremely fast, reliable network between the processors[19]. This infrastructure is suited for highly dependent calculations such as climate modelling or time series correlations.

---

[16] http://en.wikipedia.org/wiki/Grid_computing, Accessed 25 November 2013

[17] For further information and guidance regarding cloud computing see: http://agimo.gov.au/policy-guides-procurement/cloud/

[18] Derived from p7, A guide to implementing cloud services, Department of Finance, September 2012,

[19] http://www.gridcafe.org/EN/computational-problems.html, Accessed 25 November 2013

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 as well as research applications.

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[20]. 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 data capture, and to take on board the results of analysis.

In the early stages of a program there will be a tension between obtaining the capability to explore data, the practice of clearly articulating business objectives and expected return on investment. With more experience more educated opinions of the likely return will be available. Decision makers need to accommodate these early uncertainties in considering investment and approval of projects. 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, organisations must also commit to linking their 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. Analytics is often conducted away from those impacted by the outcomes. As far as is practicable projects would benefit from change management processes being planned from the commencement of the project, and for those processes to include impacted stakeholders.

It is important, when considering change management actions, to recognise that successful analytics borrows much more from scientific method than from engineering. The latter requires a great deal of certainty around tools, methodology, process and outcomes whereas analytics requires flexibility to enable analytics teams to test hypotheses and embark on discovery processes in their search for answers. Managing expectations of the process and the level of certainty should be emphasised.

[20] http://searchbusinessanalytics.techtarget.com/definition/in-memory-analytics, Accessed 25 November 2013

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."[21]

A team implementing 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 will be deployed operationally skilled 'change' managers or 'campaign managers' are also beneficial as part of the project team[22]. This cross disciplinary approach is recommended whether these teams exist within an agency, across agencies or with external service providers and research partners.

Research Partnerships, Analytics Services and Industry Collaboration

It is recognized globally that analytics and big data skills are in short supply. There is a high level of mathematical, statistical and computer science skills required. There is an increasing number of courses undergraduate and postgraduate becoming available to train new data scientists. Over time, other elements of managing data analytics projects are being incorporated into these courses.

To cultivate a capable and public sector ready analytics profession agencies should look to the growing number of opportunities to partner with academic, scientific and research institutions. While this guide does not aim to undertake a complete survey of these organisations, examples include CSIRO, NICTA, Advanced Analytics Institute at University Technology of Sydney, Deakin School of Information and Business Analytics among others.

Consideration should also be given to partnering with ventures such as Cooperative Research Centres, and other industry partners to encourage collaboration and innovation in approaches, these approaches can yield benefits beyond those that could be achieved by Government alone.

Commercial entities such as IT companies and consulting firms are also developing offerings. One such example is Hewlett Packard which operates HP Labs as a research facility for Big Data and analytics.

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

[21] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011

[22] Tackling the big data talent challenge, Hudson, Industry Leaders Series, 2013

different business problems will be the key to maintaining a workforce that will evolve as technologies evolve.

Because the field is growing rapidly 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 technologies by reducing some of the barriers to adoption such as inexperience with the legal, ICT, security and privacy considerations. Encouraging an active community of analytics professionals across government agencies will increase learning opportunities and the identification of collaboration opportunities.

The diversity of situations and applications to which the technology can be applied will also require agencies to consider the need to branch out into data mining specialisations such as text mining, optimisation and operations research.

More generally, tools are increasingly becoming available that make insights and information accessible to a wider range of personnel outside of the technical expertise. The challenge for government departments is to make the data and ability to generate insights from the data available to a broader audience, reducing reliance on the scarce technical resources.

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.

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[23].

As with many other large analysis projects that may identify scope for change, 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, capability development will need to be governed and expectations will need to be managed at the outset of big data projects; this includes stakeholders, technical personnel and the community.

Because projects can be complex and lead to uncertain results their success can often hinge on discovery and the exploitation of opportunities that emerge, it is important that agencies cultivate a dynamic culture of discovery, experimentation,

[23] Responsible Data Analytics: {guide to be developed by July 2014}

evaluation and feedback and explore the use of methodologies such as 'Lean Startup' and 'Agile'[24] that may be more suited to analytics projects over traditional 'waterfall' methodologies. It is this culture of enquiry and managing the expectation of data analysis 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 capability on the senior management team is an important consideration to ensure that projects traverse functional boundaries. It is also important to model a culture of discovery and experimentation, and for the management team to develop a fluency in the identification of potential value from big data.

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 actively work 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 secure environment. Active and solutions-focussed thinking is encouraged 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 research and the research of their colleagues, including those projects where the result is proven to be unfeasible.

In addition to the benefits towards mainstream processes, there may also be findings that shortcut future development pathways. These are some of the indirect benefits associated with supporting a culture of innovation. This approach encourages and rewards specialist staff to stay in touch with developments, test ideas in real-world environments and facilitate better collaboration in all areas of work. This is both challenging and attractive to analytics professionals and presents DIBP as an exciting place to work.

[24] http://theleanstartup.com/principles & http://agilemethodology.org/

Information management in the big data context

Data Management

Commonwealth data, including data used by 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 is governed by a number of legislative and regulatory requirements[25], government policies and plans. These regulatory requirements also apply to the management of all data sources.

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 new practices 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.

[25] The Australian Public Service Big Data Strategy, Department of Finance and Deregulation, August 2013, http://www.finance.gov.au/big-data/, Accessed, 26 November 2013

Potential data sources include:

- Administration activities and agency transaction systems[26];
- 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;
- 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, any associated metadata, whether the data is publicly available or is restricted in its access, the cost, the intellectual property rights (e.g. creative commons) and any other relevant information required by users to make decisions about the 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.[27] 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[28]. 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.

[26] 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.

[27] High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes, National Statistical Service

[28] National Statistical Service, Statistical Spatial Framework. http://www.nss.gov.au/nss/home.NSF/pages/Statistical%20Spatial%20Framework

Data integration involving Commonwealth data for statistical and research purposes

Acquiring data from government agencies for operational purposes is undertaken under the auspices of the relevant agency legislation and managed under memoranda of understanding where applicable. Under these arrangements data normally has identifiers to enable matching of records for administrative purposes. The data may only be used for the purpose for which it was collected under these circumstances.

To obtain value from Australia's national data assets held by government agencies the opportunity exists to integrate and de-identify data for statistical and research purposes.

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.

Where a big data project is identified that 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.

A complete description of the high level principles is available at www.nss.gov.au and is summarised below in Table 2:

Table 2: High level principles for data integration involving Commonwealth data for statistical and research purposes

1	Strategic resource	Responsible agencies should treat data as a strategic resource and design and manage administrative data to support their wider statistical and research use.
2	Custodian's accountability	Agencies responsible for source data used in statistical data integration remain individually accountable for their security and confidentiality.
3	Integrator's accountability	A responsible 'integrating authority' will be nominated for each statistical data integration proposal.
4	Public benefit	Statistical integration should only occur where it provides significant overall benefit to the public.
5	Statistical and research purposes	Statistical data integration must be used for statistical and research purposes only.
6	Preserving privacy and confidentiality	Policies and procedures used in data integration must minimise any potential impact on privacy and confidentiality.
7	Transparency	Statistical data integration will be conducted in an open and accountable way.

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 Commissioner[29]

'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 Commissioner[30]

Where data sources include personal information (i.e., 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 regulates the handling of personal information throughout the information lifecycle, including collection, storage and security, use, disclosure, and destruction.

The Australian Privacy Principles 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 include:

- The collection of personal information from sources other than the individual.
- The creation of new data through 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 (an example of the mosaic effect[31]).
- The potential for unstructured information sources to hold personal information not known by the individual.

Agencies engaging 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. These are:

- Robust de-identification capabilities. Identifiers can be removed via a number of methods including deletion (safe harbour[32]), masking, aggregation and other statistical techniques (collectively known as expert determination[33]). As well as

---

[29] http://www.oaic.gov.au/news-and-events/media-releases/privacy-media-releases/privacy-is-a-priority-for-the-australian-community, Accessed 25 November 2013

[30] OAIC community attitudes survey: http://www.oaic.gov.au/privacy/privacy-resources/privacy-reports/oaic-community-attitudes-to-privacy-survey-research-report-2013

[31] 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.

[32] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html, Accessed 26 November 2013

[33] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html, Accessed 26 November 2013

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[34].

- Privacy by design[35] - 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. Specific considerations that are likely to impact on big data projects include:
  - Where 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[36] –. 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 help to shape and evolve with the project's development. PIAs are one way of assisting implementation of Privacy by Design.

A number of agencies have additional legislated requirements to comply with. In some cases (e.g. the Migration Act) impose additional (and typically more stringent) requirements.

[34] Office of the Australian Information Commissioner, De-identification resources May 2013, http://www.oaic.gov.au/privacy/privacy-engaging-with-you/previous-privacy-consultations/de-identification-resources-may-2013/

[35] Further information on privacy by design http://www.privacybydesign.ca/index.php/about-pbd/

[36] Privacy Impact Assessment Guide, Reviewed May 2010, Office of Australian Information Commissioner

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.

Source: Information Privacy Commissioner, Ontario, Canada, http://www.privacybydesign.ca/index.php/about-pbd/

Security

In the context of big data, security policies apply as they do for traditional data. The Protective Security Policy Framework[37] and supporting guidelines[38] 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[39].

Considerations of security specific to big data include:

- 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; and
- Increased distribution of physical and virtual locations of data storage.

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

[37] Attorney General's Department; Protective Security Policy Framework

[38] http://protectivesecurity.gov.au/informationsecurity/Pages/Supporting-guidelines-to-information-security-(including-the-classification-system).aspx

[39] OAIC Information Security guide: 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 readily understood by those outside the data science profession.
- There is 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 projects including-
  - The possibility that the data acquired will not provide the insights sought.
  - Findings and results from projects can be uncertain and counterintuitive.
  - False discoveries are possible –large amounts of data can produce statistically significant results that arise purely by chance.
  - During the exploratory process, data may present unexpected, positive, outcomes that may or may not be related directly to the initial issue being addressed.

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

Earlier better practice guidance on managing scientific projects is available from the ANAO[40]. Much of the advice in this guidance is still relevant and used today. It is also worth considering that 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[41].

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

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

[40] Management of Scientific Research and Development Projects in Commonwealth Agencies Better Practice Guide For Senior Management, ANAO 2003

[41] http://agilemethodology.org/

- An adequate risk adjusted expectation of return on investment to account for uncertainty.
- A commitment to an incremental and iterative[42] 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 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 project.
- Privacy by design principles should be applied at all stages of the project.
- Stakeholder engagement plans for projects should include transparent review processes of the data ownership, data acquisition, data management, results, and the application of the findings where appropriate.

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

[42] http://scrumreferencecard.com/scrum-reference-card/, Accessed 5 December 2013

Conclusion

This guide aims to identify developments in current information management and analytics practices. 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 projects, strengthen and articulate existing processes for the responsible use of data analytics, monitor technical advances, develop appropriate skills to extract value from information and adequately protect privacy and security of information.

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

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.56
> The guide will incorporate the NSS produced *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*57, 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 and responsible management of analytics projects 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.[43]

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).[44]

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.[45]

[43] IBM, What is data scientist, http://www-01.ibm.com/software/data/infosphere/data-scientist/

[44] Office of the Australian Information Commissioner, Information Policy Agency Resource 1, http://www.oaic.gov.au/privacy/privacy-engaging-with-you/previous-privacy-consultations/de-identification-resources/information-policy-agency-resource-1-de-identification-of-data-and-information-consultation-draft-april-201

[45] Office of the Australian Information Commissioner, FOI guidelines - archive, http://www.oaic.gov.au/freedom-of-information/freedom-of-information-archive/foi-guidelines-archive/part-5-exemptions-version-1-1-oct-2011#_Toc286409227

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.[46]
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.[47]
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.[48]
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.

[46] http://www.privacybydesign.ca/

[47] Office of the Australian Information Commissioner, Privacy Impact Assessment Guide, http://www.oaic.gov.au/privacy/privacy-resources/privacy-guides/privacy-impact-assessment-guide

[48] Office of the Australian Information Commissioner, Open public sector information: from principles to practice, http://www.oaic.gov.au/information-policy/information-policy-resources/information-policy-reports/open-public-sector-information-from-principles-to-practice#_Toc348534327