SUBMISSION 89 - SWIPEZY PTY LTD - DATA AVAILABILITY AND USE - PUBLIC INQUIRY

Abstract

The effective utilisation of unstructured big data documents presents an economic and societal benefit opportunity of comparable magnitude to that already being achieved from the utilization of structured big data.

Appropriate tools, infrastructure and expertise are now generally available such that Governments organisations are now able to realise these benefits.

Swipezy Pty Ltd
Level 2, 50 Bridge St
Sydney, NSW 2000
www.swipezy.com

Why effective management of unstructured big data documents can present economic benefits comparable to those achieved from the utilization of structured big data

The Australian Government is seeking to consider policies to increase the availability and use of data to boost innovation and competition in Australia. Furthermore, the Government recognises that the improved availability of data, combined with tools to use it, is creating economic opportunities.

Increased sharing of data across the public and private sectors could facilitate the improvement of interactions with Government and increase administrative efficiency.

Structured Data

Much of the current global discussion and economic activity in regard to enhanced utilisation of big data is focused on the "mining" of large structured data sets. In large part, this focus on the use of structured data sets has occurred as a result of the current availability of key and cost effective enablers including;

- **Compute efficiency**
  Cost effective, high throughput compute capacity
- **Data sets**
  Access to diverse large data sets with common formats
- **Skills**
  The broader availability of experienced and cost effective technical "mining" skills
- **Regulatory framework**
  Legal and societal regulatory context setting associated with data sharing provided as a result of Data Privacy regulations invoked by a range of Governments.

Much of the current global focus on structured big data is centred on the potential benefits arising from expanded re-purposing and reuse of structured data sets across multiple government agencies, organisations and individuals.

What about Documents (Big Information)?

Whilst the enhanced use of structured data sets continues to be a nationally strategic asset with opportunity to continue to yield additional significant economic and societal benefit Swipezy is of the view that the effective utilization and sharing of unstructured data in the form of documents ("Big Information") should also be considered in a similar way.

The effective "mining" of unstructured "Big Information" presents an economic and societal benefit opportunity of comparable magnitude to that already achieved from the utilization of structured big data.

Structured Big Data vs Unstructured Big Information

	Big Data	Big Information
Structure and format	• Structured (database) • Standardised	• Unstructured (documents and other knowledge base items) • Mainly non-standardised
Volumes	• Huge Volumes that are doubling every 2-3 years	• Huge Volumes – Doubling Every 2 Years • Trillions of existing Word & PDF documents • 500 Bn new Word Documents pa
Access and ability to be shared	• Largely accessible and shareable • BI tools well advanced	Until recently: • Not Accessible & not Shareable (largely) • 90% of documents frozen in pre-web formats • Mostly not searchable

The Three V's

The unstructured big data document (Big Information) environment including word processing files and PDF files has largely been overlooked, yet it has similar "three V" characteristics as observed within the structured data environment:

High volume	**Existing repositories of Trillions of existing Word and PDF documents.**
High velocity	It is estimated that in excess of 500 Billion new Word documents are created globally each year[1]
High variety	Every government agency, organisation and individual makes use of unstructured data. Whilst there are a wide range formats in which unstructured data is held most organisations and individuals globally utilise Microsoft Word as the basis for their word processing files and the originating environment for PDF documents.

Traditional Inhibitors

Historically, the delays in the utilisation of unstructured big data sources as an economic enabler have, in large part, occurred due to the lack of widely available enabling technologies such as;

- Distribution Platform
  The lack of a common, widely used information dissemination platform
- Standardised Format
  The lack of cost effective, scaleable tools to enable documents to be presented in a common format(s)
- Viewing Infrastructure
  The lack of widely available infrastructure for access to, and viewing of, documents
- Document Control
  The lack of consistent, simple master document version controls for appropriate "legal" date/time stamping

These technologies and tools are now available and in many cases already widely being utilised

[1] Chris Flores, Director Communications – Microsoft Windows Group (2014)

Distribution Platform

The World Wide Web ("the Web") has become a universal and standard tool for accessing unstructured data in all its forms. The internet has enabled:

- The ability to readily search, identify and access data, documents and information when content is made available in Hyper Text Markup Language (HTML) format
- The easy and cost effective distribution of data
- The ability to readily share data and information without the need for the intervention of highly skilled technical expertise.

Furthermore, the knowledge and skills to enable Web searching are largely intuitive and extensively utilised by users of the Web.

Availability of this unstructured data in an on-line format also enables more effective management and structured/proceduralised retention of the data rather than, for example, risking the loss of data caused by the failures or loss of an individual's PC device.

Standardising Format

Recognizing that the Web is the access environment of choice for the majority of the World's unstructured data document users, there was a pressing requirement for the creation of cost-effective scaleable tools that would enable documents (typically created in desktop formats like Word and PDF) to be turned into structured HTML format.

Australia is a world leader in the availability of these tools. We now see the emergence of Australian Software as a Service (SaaS) platform providers with offerings that readily enable the automatic conversion, segmentation, classification and formatting of Word and PDF unstructured data documents into formatted searchable and enhanceable HTML Web Format.

Widely Available Viewing Infrastructure

Whilst business and government organisations have had ready desktop access to the Web for many years, it is only comparatively recently - with the advent of tablet and smartphone devices - that more extensive and deep "on the move" access to the web has become available for business and personal use.

Australia has one of the highest rates of smartphone use in the world with estimates that 83% of mobile phone users in Australia will have smartphones by 2017[2].

Similarly, high rates of internet connected mobile tablet device usage is indicative of a breadth of availability of suitable" viewing" device penetration in the Australian business and personal population.

Increasingly, web content is being automatically rendered for the individual device. This means that simple, consistent and usable access to the unstructured data documents under discussion is now readily and almost universally available.

Master Document Version Controls

For legal and comparative data integrity reasons it is vital that effective, consistent and accurate approaches are implemented to ensure master document and versioning control. Tools for the effective management of master document versioning are already widely

[2] http://www.statista.com/statistics/257041/smartphone-user-penetration-in-australia/

available, if not fully or universally implemented by Government agencies and business organisations.

Recognising that the majority of word processing and PDF documents will continue to be authored in Microsoft Word for the foreseeable future there are existing, robust approaches available for document and version control. These include:

- **Document Management Systems**
  Most agencies already use a formal Document Management System as the official repository for all master versions of documents according to predefined records storage procedures.
- **Web Content Management System**
  Most agencies already use a Web Content Management System (CMS) for version control of content and to ensure that only the authorised final version of the content is published on the website or intranet.
- **Collaboration Systems**
  Some agencies already use an organisation-wide collaboration system to share desktop-formatted documents. The new technologies discussed here enable these documents to be repurposed for the web.
- **Role management**
  CMS systems already enable organisations with web teams to define and implement processes that separate the roles of authoring, approval, publishing and curation of web content. However, as the tools referred to above now mean that anyone with Microsoft Word skills can make content available for the web, role management processes are now being extended to all involved staff.
- **Simple Classification tools**
  Classification and tagging of content is important if unstructured is to be filed in the right place and easily retrieved. The availability of consistent and easy-to-apply tools and processes mean that non-technical people can both classify and easily find required information.

Now that appropriate tools, infrastructure, security procedures and skills are cost effectively available to enable on-line availability of all unstructured data documents in an on-line format, Australia has a very significant opportunity to increase the availability and use of its unstructured data to boost innovation and competition in Australia.

Australian Governments, in particular, could realise a major operational efficiency dividend and provide substantially enhanced value to the Australian community with intelligent sharing and reuse of unstructured data – especially documents.

The Opportunity for Australian Government(s)

The harnessing of Big Information through standardizing unstructured documents offers dramatic potential efficiencies for Government.

Example 1 – Reduction in duplication

Every Government agency, Federal, State and Local, produces policies and procedures to manage their business operations. While naturally there are some procedures peculiar to particular agencies or circumstances, Swipezy estimates that approximately 70% of the content of these policies and procedures could be applied to other similar agencies with only minor variation.

For example, agency procedures relating to paper recycling, leave applications, superannuation change request and equal opportunity provisions would not have to be duplicated hundreds if not thousands of times if that information was standardised and made available to all.

The avoidance of duplication within agencies and between agencies represent a significant efficiency dividend opportunity for Government.

Example 2 – Contextualising content

Many agencies require their staff and the Public to navigate to several places on their website to retrieve information – a practice that usually requires multiple downloads. For instance, a person interested in an Agency's recycling policy may have to download long PDFs including:

- The Carbon Management Plan
- The Waste and Resource Management Strategy
- The Sustainable Procurement Policy
- The Biodiversity Action Plan
- The Environmental Awareness Strategy

Each of these documents may contain "nuggets" of desired content, but often significant amounts of the material within each is not needed.

This problem is exacerbated when the desired information resides across different agencies and when the search is invoked on a mobile device where downloads are not only problematic, but where it is difficult to read the material.

Modern tools solve these problems:

- Focused downloading of data and information rather than the need to download complete documents. This is both quicker to obtain the specific data or information sought and less resource intensive
- It is simple to tag information at the page or paragraph level.
- The tags make it simple to build contextual links to related data.
- In short, related information is aggregated in a fraction of the time of traditional methods.

Example 3 – Enhanced efficiency of accessing data and effectiveness of decision making

Within the current operational processes utilized by Government agencies it is often difficult for staff to:

- Identify appropriate pools of data and information
- Efficiently identify specific, relevant content within the pools of data and information
- Search for content within attached documents – e.g. PDF files
- Ensure that a sufficiently comprehensive sourcing of existing internal data and information has occurred

Modern tools can federate search across multiple systems and document types. This ability to readily conduct an on-line search of an extended range of documents provides the opportunity for Government agencies to:

- Significantly enhance the speed with which suitable sources of data and information are identified (efficiency)

- Gain greater visibility of a wider range of relevant sources (many of which are potentially already resident within the Agency but may not be visible to the requestor as they are not readily visible); and
- Enhance the effectiveness of decision making.

Example 4 – Opportunity for enhancement of the Productivity Commission's own processes

The Productivity Commission's has a significant opportunity for the enhancement of their own current process in seeking submissions.

- ### Format not optimised for Web

  The Commission is seeking submissions that will be soon be published online for public viewing. Unfortunately, each of the submissions received by the Commission will have been created in desktop formats – Microsoft Word or PDF - but then uploaded, most likely in PDF format.

  Modern tools could have automated the conversion of each submission into web format and then allowed any web visitor to easily consume any submission on any device without the need to download anything.

- ### Reading and searching submissions is problematic

  As these submissions do not utilise an HTML based format approach the Productivity Commission does not have a ready ability to simply search and navigate within and between documents.

  This introduces internal inefficiencies within the Commission in the interpretation and evaluation of submissions.

  Furthermore, when submissions are made available publically interested parties will be forced to download every required submission in entirety with downstream limitations on the presentation of the material (not rendered for specific device viewing) and limitations in the ability to efficiently navigate within and between documents.

  This is inefficient and not user friendly for interested parties – particularly on mobile devices such as tablet devices.

- ### Linking and contextuality

  Each of these submissions will have little or no metadata, tagging or cross-linking capabilities meaning that there is no way to navigate to related submissions or topics.

Recommendations

1. Mandated Web format

All final form Government documents (unstructured data) should be mandated as a standard operational procedure to be available on-line in Web format to enable ready searching, identification, accessing and potential enhancement of the documents.

This will enhance the efficiency and effectiveness of both internal & external communications, use of aggregated data sources and enable innovative applications of data.

2. Appropriate Security

Access to each and every Government document available via the Internet must continue to be assessed in terms of appropriate audiences and controlled via an appropriate security mechanism.

3. Simple Tagging of documents

All new Govt. information should be tagged with appropriate keyword descriptors. These should be easy to insert by non-technical people – very likely from a point and click interface. This will enhance searchability and contextuality of related information and result in substantial time savings across Government for publishers and consumers alike.

4. Thinking Mobile

When it comes to information publishing, Government must think about how it is to be consumed - Information must be made web and mobile friendly. It is neither acceptable, nor pragmatic to force the 60% of consumers who choose to access the web on a mobile device, to consume information in analogue format.

The good news is that the tools are here now to make the above a reality.