The Evolution of Data Automation, and its Importance to the Australian Spatial Data Infrastructure

Michael DIXON, Australia

Key words: Data Automation, Data Management, LYNX

SUMMARY

The aim of this paper is to:

- Provide a common understanding of what Data Automation is;
- Discuss how PSMA Australia historically managed data, and how this impacted on business practices;
- Examine how PSMA Australia has embraced Data Automation and has implemented two innovative systems to manage the data processing cycle; and
- Review how the international spatial industry is addressing data management, and how Data Automation is influencing their business development.

Within Australia, spatial data management has historically been manually driven and managed by systems that are not centralised. Whilst progress has been made, manual activity is still a major feature and prone to error when conditions are not typical. PSMA Australia has recognised that significant automation is possible within the data management supply chain and that many manual steps can be replaced with carefully orchestrated processes. LYNX has already been internationally recognised and with the addition of new cutting-edge systems to support data management, PSMA Australia is able to demonstrate a number of outcomes not previously possible, for example:

- Decreases in processing time;
- Increased confidence & transparency in outputs;
- Reduction in the resource and maintenance costs for maintaining and developing datasets;
- Increased flexibility to quickly adapt to changing conditions; and
- Increased understanding and control of intellectual property;

This paper will discuss how the importance of Data Automation has been recognised globally, and how international initiatives like INSPIRE are endeavouring to achieve what PSMA Australia already has, but for the European continent.

The Evolution of Data Automation, and its Importance to the Australian Spatial Data Infrastructure

Michael DIXON, Australia

1. INTRODUCTION – WHAT IS AUTOMATION?

The term automation in IT systems is perhaps not well understood or, more precisely, its meaning varies for different users. Traditionally in data management, automation could be regarded as being an endpoint for a managed set of processes that has required less human intervention over subsequent iterations. This basic definition elaborates on roll of time in automation which is perhaps the hardest to comprehend for uneducated users. Some people expect that automation will be instantaneous, both in the time it takes for the task to be completed, and that, once the decision to automate has been made, little effort will be required for it to happen. The experience of PSMA Australia in automating its data management supply chain will be used to explain how it can be used to deliver real business benefit.

Automation occurs at a number of different levels in any process. In commercial off the shelf (COTS) software used in data management, at the lowest level individual tasks like selecting, reprojecting or exporting can all be completed with the definition of some parameters and the touch of a button. The same software may also be used to join multiple tasks together using simple scripting languages, or even through inbuilt interfaces to further automate some processes. Similarly in the online age of today, web services can be linked together with great flexibility to deliver increasingly sophisticated systems and, ultimately it is now possible to manage the execution of various COTS, database, web service and even human tasks through a single automated process. At each level, and in all cases, the automation is only achievable via knowledge of the inputs, outputs and each of the tasks used in the automated chain. Significant effort is required to understand the tools and develop the skill sets required to take processes from being human initiated to being controlled by a machine. Indeed it is likely that different skills will be required, changing from those relating to the individual tasks to those that concerned with how the tasks should be ordered together. In addition testing procedures and plans will need to be developed and executed to ensure that automation doesn't degrade the quality of outputs and reduce the benefits to the organisation.

PSMA Australia has long been on the path towards automation. However, it is a complex and challenging problem to take the final step required for its delivery. So why is it important to automate processes in data management? The most obvious reason is the generally excepted idea that an automated process will always be completed in less time than the same one that requires manual execution. The decision making that is required when moving from the end of one task to the start of the next will take time, even when the process for making the decision is trivial. The larger the number of decisions that need to be made, the larger the delay than can be expected; the process of understanding each of these decisions is one of the key concepts in building up an automated process. Whilst the reduction in time to complete the same process is

important, it isn't necessarily the most compelling reason for automating. Other factors such as the consistency of results and the ability to produce the same outputs at a lower cost may have more appeal. Taking the human element out of decision making will always lead to a consistency (which should not be confused with quality) in outputs. The concept of cost is not limited to purely monetary values either. The process of acquiring the tools and skills to automate will have a financial cost which organisations would typically expect to recoup at some point in the future. Having fewer people resources, either by less numbers or time spent on the same tasks, should allow the same resources to concentrate on producing additional quality measures, new datasets or some other innovation.

Automation in any sense is not particularly wise when inputs to the managed processes are not fixed. Subtle changes to complex processes upstream in data capture, transfer or delivery can lead to unanticipated issues arising which may require intervention and would ideally be managed as early as possible in the cycle. It would be virtually impossible to predict every scenario in terms of how incoming data could vary making it difficult to design any system to cater for it. This issue poses an ongoing challenge, and, without the inclusion of standards, reporting and issue management capability along the supply chain, it would be extremely difficult to manage.

The first step in the move towards automation is the defining of standards that data will be checked against. These standards, in a broad sense, attempt to cater for variations in the input data. When checking incoming data for conformance against a standard, only two results should be possible, pass or fail. The data that passes should be relatively simple to deal with and its onward flow should easily be possible. However, the non conforming data must be dealt with before any real degree of automation is possible. There are essentially three options available in dealing with the non conforming data;

1. Exclusion from output
2. Regard as an exception and include in output
3. Fix Issues to allow conformance and include in output

If non conforming data is excluded then it is likely that an incomplete product will be released. If a decision is made to include the non conforming data, an argument could be made that the standard that it is checked against may be redundant. "Fixing" the data to conform to the standards so that it can be included is by far the most challenging option and requires a paradigm shift in the philosophy of the people and organisations involved.

PSMA Australia, through its board of directors, has considered these options in significant detail and has firmly embraced option 3, as listed above. The challenge to find, fix and report data issues will be a fundamental paradigm shift aimed at providing higher quality data to the Value Added Reseller (VAR) network, and other users alike. This constitutes a significant body of work that is actively being pursued as part of the supply chain redevelopment.

With the decision to fix data having been made, the specific fixes need to be designed. They will be quite simple in the first instance, although levels of sophistication will likely be introduced at a later date. In developing these fixes it has provided an opportunity for users to go back to basics in terms of their understanding of the data, and in particular spatial geometries. The primary consideration in each instance will be to minimise actual changes to the data to make it conform to the particular standard. It is possible to automate this process; however, adequate reporting becomes even more crucial as data will inevitably display different levels of quality in early iterations of processes.

Each new process to check for standards conformance, and apply the associated fixes, needs to be somehow combined into the overall data management workflow. The sequence of when something should start, finish and what should occur subsequently requires an orchestration engine for management. This is a critical piece of "glue" that will go a long way to differentiating the new data management process from its predecessor. As suggested earlier, there is an expectation that it will be used to orchestrate web services, SQL scripts, human tasks and perhaps some other cots application. Business Process Management (BPM) software can be used as the orchestration engine, which typically has a comprehensive set of functionality. The BPM software allows for orchestration of varying levels of complexity to be easily specified and built, often using standards compliant notation. In addition, integration of "human tasks" (tasks that are directed to a human user to perform) allow human users to be included in automated workflows where it is more appropriate for tasks to be carried out by people. This is extremely important for completing tasks such as validation by visual inspection, but also for manually performing a task if previous attempts to perform it automatically have failed. Furthermore, included Business Activity Monitoring (BAM) features will allow data management users to obtain information regarding the status of orchestrations in real time, allowing them to respond more efficiently to issues that occur during execution.

2. BACKGROUND OF PSMA AUSTRALIA

PSMA Australia Limited is an unlisted public company. It was established under Australia's corporations law and is wholly owned by each of the state and territory governments and the Australian Government (known collectively as Jurisdictions). Chiefly PSMA Australia has a goal of coordinating the development and maintenance of fundamental national datasets using the information resources of government and increasingly, non-government entities.

PSMA Australia is the crucial link between the supply and demand sides of the national geospatial market. The organisation eliminates the difficulties of negotiating multiple licence agreements and the problems of integrating the data into a seamless consistent national dataset. Furthermore the existence of PSMA Australia minimises the duplication of effort within the market for organisations wishing to access national data (Figure 1).

The fundamental datasets that PSMA Australia currently manages are Administrative Boundaries (localities, local government areas, electoral and Australian Bureau of Statistics boundaries), Cadastral Boundaries (including property), Transport (roads, rail and air) and Topography, Post Code Boundaries, Points of Interest and the Geocoded National Address File (G-NAF). All of these data layers are managed through PSMA Australia developed software LYNX.

3. WHAT IS LYNX?[1]

PSMA delivered its first dataset in 1995; it was the result of more than two years work and involved tens of private sector organisations. The effort was so large that an update was not considered until the lead up to the next scheduled census in 2001. Following the 2001 Census, PSMA Australia set about putting arrangements in place so that it could provide annual updates to its national dataset. This involved documenting the structure and content of the input data that would be provided by mapping agencies and developing a specification defining the final product that was required. This then formed the basis for a request for tender and ultimately the appointment of a private sector firm to manage the supply of updates to the dataset. Internally, these organisations were referred to as Data Managers. Over the next three years more datasets were added and update cycles moved from an annual update to being quarterly. The increase in datasets and updates constituted a 100 fold increase in the number of dataset updates performed annually compared to the situation in 2000.

To manage this increase several initiatives were instigated including the implementation of a storage environment that would hold all the PSMA Australia data products in a single harmonised and highly normalised data schema. This was not a trivial task and took some 18 months to design and implement after considerable consultation with data contributors and Value Adding Resellers (VARs). The importance of this decision cannot be understated. It has proven to be a vital step, crucial in many of the activities that were to follow. The implementation of this storage environment then enabled the automation of quality assurance tasks. This alone had a significant impact on the quality of data structure and consistency.

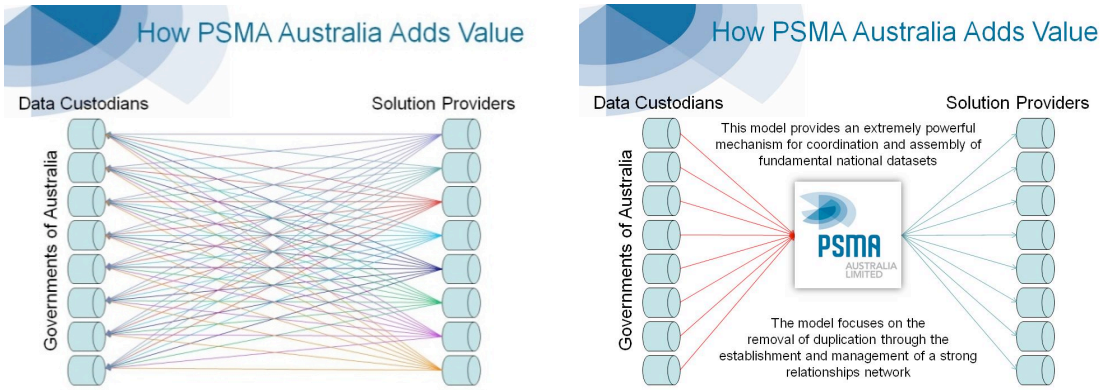

Figure 1: PSMA Australia as the link between upstream suppliers and downstream users of spatial information.

[1] Largely taken from Paull 2009

The new environment also enabled the establishment of a secure web portal for provision of reports and product information, data manager deliverable acceptance and VAR data requests that could be fulfilled by automated DVD burning and mailing, email or FTP delivery. The system was coined LYNX (by virtue of the connections that it assisted in streamlining) and was launched by the then Special Minister of State, The Honourable Gary Nairn in 2006. During the following year, the system was recognised by URISA[2] amongst an internationally field as an exemplary enterprise system within government.

The technical processes required to deliver the national datasets are still generally completed by data managers. Typically these data managers have been external to PSMA and have been contracted to integrate the datasets using their expertise in a particular set of technologies not usually mandated by PSMA. As these organisations are typically highly skilled the processes for conflation have been refined over time to the point where they are now relatively efficient. Importantly the contracts governing the relationship between PSMA and data managers are oriented around producing outputs rather than defining any processes that should be followed.

4. IMPROVING DATA MANAGEMENT

Perhaps the primary reason behind the review of current data management processes is the time it takes for data to be processed. The current release cycle is quarterly. This means that, for any release, the data has not actively been maintained for three months prior. There are good reasons why this is currently the case, but primarily it is because the cycle must cater for the dependency between some of the data sets. This means that processing must occur in a particular sequence, with only some tasks occurring in parallel where possible. A shift to smaller, more frequent updates, coupled with ad hoc data supplies, will produce a requirement for processing to occur in a non-sequential way. This change will enable significant reduction in the time taken for information to be delivered through the supply chain to the VAR network, and beyond. The term non-sequential should not be confused with the natural hierarchy that exists between datasets (i.e., an address contains a road and a locality). Hence, an authoritative locality and roads dataset will still underpin any G-NAF update.

The use of outsourced data managers to undertake the technical processes behind the release cycle has been extremely successful in allowing PSMA to produce regularly updated datasets. Whilst no doubt providing benefits, it has meant that PSMA Australia progressively moved further away from understanding the finer detail of the conflation processes. Over time, data manager processes evolved to become more sophisticated and efficient, but only within the confines of the data manager's own specific task, rather than PSMA Australia's data management processes as a whole. This observation should not be regarded as a fault of the data manager. On the contrary, it demonstrates their ability to continuously improve. However, without access to the "bigger picture" of the supply chain, the data mangers have not been encouraged to innovate beyond their original mandate of producing a defined set of outputs. The redevelopment of processes will provide the opportunity to acquire the intimate level of detail required to

[2] The Urban and Regional Information Systems Association, http://www.urisa.org/

have a comprehensive understanding of the issues that need to be solved in order to streamline processing. This redevelopment should not be seen as a shift away from using the expertise of industry to help with this process. In fact, once the new processes are developed, opportunities will still exist for their management and execution, empowered by increased input from PSMA.

The lack of complete transparency of the processing methodology has made it difficult for PSMA Australia, and downstream users, to be confident about both the repeatability of the outputs, and the accountability for any issues that may have occurred. Again this should not be taken as a criticism of data managers' work. It is simply realisation that more visibility of processes would provide a greater level of confidence in what is being delivered. Typically organizations have a "champion" responsible for a particular process and, whilst they complete their work in a normal manner the output would typically be "normal". It is when a new situation arises, or a delegate is required to complete the same task, that issues often arise. The closer organisations can move toward "full" automation (see automation section below), the less likely that a change in inputs or personnel will cause issues. It should be noted that different inputs and personnel do not always cause a negative impact. Introduction of new personnel can be positive, due to the injection of a new perspective(s). Furthermore, changes to inputs caused by data quality improvements in contributor data may render some downstream processes obsolete, and allowing for the obsolete processes to be retired.

Any user of data will invariably have some questions about it at some point in time. These questions could be the result of issues with processing, delivery or around the data itself. Historically metadata statements, coupled with personal relationships, have been used as the basis for answering questions about data/products. This approach makes risk and issue management more problematic, and leaves any organisation exposed to losing corporate knowledge through staff rotation. Currently much of this knowledge is kept in desktop tools, such as spreadsheets and emails, without any significant coordination. The ability to centralise, organise and share this information would provide a vital business tool, and is one of the key initiatives within the LYNX redevelopment.

5. DECREASING TIME TO MARKET

Infrastructure capable of supporting time critical access to spatial data is now becoming more widespread. Examples, such as GNSS corrections data and traffic incident reports, show that processes behind data collection, processing and dissemination can be streamlined to significantly reduce time to market. Whilst not all users of data require such immediate access, there are applications where this will be required, both now and into the future.

As already discussed, the current PSMA Australia data processing cycle takes approximately ninety days from the acceptance of data to its release. If downstream VARs require additional processing to repackage the data for their own products, an additional delay will be incurred. The cumulative impact of these delays may mean data is upwards of six months old before it makes it into end user applications. Data users, such as those from emergency services, rightly find this delay unacceptable for their

own use. However, it should be noted that not all users have indicated that the current update frequency is a problem. The reduction in the time required for all PSMA Australia datasets to be processed is a key driver for the entire LYNX redevelopment. The current processing environments, whilst being fit for the requirements as they exist now, are fragmented in the sense that they have all been developed in isolation, and are not closely tied together with any real level of sophistication.
The redevelopment of LYNX will be in the form of a Services Oriented Architecture (SOA) providing ability to alter the IT functionality and respond with agility and flexibility to rapidly changing market demands and new technological developments. Kumar, Dakshinamoorthy and Krishnan (2007) studied the impact of SOA adoption on the performance of electronic supply chains for a cross section of businesses. They found that adoption of SOA leads to better performance of the supply chain. This result provides confidence that PSMA Australia's adoption of SOA will enable a significant reduction in the time it takes the data to get to downstream users.

6. ONGOING IMPROVEMENT

As indicated previously a fundamental reason for moving towards automation is to utilise human operators more effectively. Perhaps the single biggest enabler for this to occur is the provision of high quality consistent reports about the strengths and weaknesses in existing processes. In terms of data processing, reports are written for two scenarios:

1. What was routinely updated?
2. What had to be fixed (and how) so that a non-routine update could be completed?

It is the reports for the second scenario that are the most critical. If these reports are provided back upstream to the original data providers, and they are able to adequately convey the information about the weaknesses in the data, then reasonably these issues may be fixed, removing the requirement for repeated processing during subsequent iterations. Furthermore, if the modified data was provided back to the contributor, they may be able to replace the original data, extending the process of continuous improvement along the entire length of the supply chain. If, however, the returned data is not accepted, or adequate resources are not in place to make sure that the identified weaknesses in the data are corrected, there is a strong likelihood that the data maintenance cycle will become more cumbersome with the burden of updating data each time it is provided.

The coordinated management of these issues is a fundamental consideration in the redevelopment of LYNX. The transparent flow of information back up the supply chain to the authoritative maintainer should provide transparency and an additional degree of confidence regarding the entire supply chain. The potential inclusion of feature level metadata detailing which standards (business rules) have been applied to it, and what changes have been made to it, will enable better decision making for data users. In addition, end users, including the general public (with appropriate limitations), should have the ability to lodge and track issues using the same system.

7. FUTURE DATA MAINTENANCE

The primary philosophies behind the changes to PSMA Australia data management processes have already been described. The concepts of automation, standards, reporting and orchestration have been discussed and how they are critical in terms of improving on the current data management processes. The following section provides an overview of some of the key experiences to date in terms of redeveloping the data maintenance processes.

Based on a broad examination and testing, the primary tools that are expected to be utilised in the revised processes are ActiveVOS (Active Endpoints), FME (Safe Software), Oracle (Locator/Spatial) and Radius Studio (1Spatial). The term 'expected to' is included here because the new method for the delivery of data, pre-processing and some parts of the data management processes have not been yet been finalised. However, all components have been successfully tested in various development projects.

With the addition of ActiveVOS, the role of the traditional technical expert is changing to someone that not only understands the individual data management tasks, but also comprehends how they all should be coupled together to produce the required outputs. Experience has shown that whilst this sounds simple, it requires significant skill including an understanding of Business Process Execution Language (BPEL) and other standards for the implementation to be successful. At this stage the ActiveVOS implementation is limited to a development environment. The development of the automated processes using Radius Studio has largely been completed for polygon datasets, and is well advanced for line datasets. As such, this topic will be discussed in more detail below.

The most important step in the development of an automated environment is the creation of the business rules or standards that data will be checked against. These business rules should be simple, unambiguous and ideally mirror the constraints applied by the people responsible for the creation of the real world feature. In terms of administrative boundaries, simple examples might be that each area has a name, and adjoining areas should not overlap nor have gaps between them. The documenting of these business rules, and associated technical detail, has had the added benefit of enabling PSMA Australia to learn even more about its core business. Importantly, it became clear early into the redevelopment process that many of these business rules were uniform across different data themes and, in fact, could be reused during processing. This now provides a powerful tool set that may be updated only once to alter processes across all data. The various rules can be applied to each theme merely by mapping the required spatial and aspatial attributes to a common schema as the data is input to the relevant processes. The prospect of reusing particular aspects of existing data management processes should have the additional benefit of easing the burden and cost of introducing new datasets as they materialise.

The volume of work required to redevelop processes that have previously been outsourced is quite significant and should not be underestimated. The experience has been akin to training new staff. The nature of using external assistance to deliver

outputs means not every minute level of detail about all processes will be documented. These oversights are typically only discovered at the most inappropriate time. Whilst this may seem like a negative, it should be regarded as the opposite, as it requires a fundamental comprehension of even the most basic components to reproduce any outputs. These steps have added some time to the redevelopment process as the number and complexity of these issues have been difficult to predict.

Arguably the most import rules that have been written are those that define what changes will be made to data to make it conform to the base business rules where they are not initially met. The underlying philosophy is always to produce the most accurate dataset possible. However, if issues exist in the source data, they will be rectified henceforth. It is this point that provides one of the keys to the automation of the data processing cycle. The automatic application of a fix to modify data so that it will conform to business rules means that the dataset can be updated in a single process. However, it could be said there has been a degree of concern about this move towards fixing data. The alternative is to continue to produce a dataset that is potentially being cleaned by many downstream users, resulting in significant duplication of effort. To date evidence suggests that most fixes are extremely small (sub centimetre), but are nonetheless required to create a topologically consistent product within each data theme.

One challenge already described when automatically manipulating data is that one solution will invariably not cope with all possible scenarios. Over time, it has been possible to introduce higher levels of sophistication in subsequent iterations of processes. Thresholds can be added to avoid automating a change which would produce negative results. In addition a manual step (e.g. visual inspection), could be requested as a final supporting check. These thresholds could be based on spatial or aspatial content including content derived during any earlier process. A pertinent example has been the treatment of voids or holes in the localities theme. The business rule states that there should not be any gaps between adjoining localities. The accompanying fix is to create a new polygon in the gap, and add that into the adjoining polygon sharing its longest boundary. On the surface, this would seem to be straight forward, but it is not until features such as water bodies are considered that there is the realization that a locality could easily be incorrectly extended to cover these features. There are potentially numerous methods to manage this issue however the choice made in this instance was to maintain a point dataset which would be a list of the exceptions to this rule. Whilst a polygon would be created to fill in the gap, it would not be added into the dataset because it contains one exception point.

Overall, the experience to date has been extremely positive. Users have made comment about the increased quality in the automated data products produced using the new technologies. The time for production has been reduced by around 80%, with more gains expected to be made as additional components of the LYNX redevelopment come on board. In addition, a single resource is capable of overseeing the update for all of the datasets developed to date. Whilst this number will grow, it is still a reduction with respect to what has been required previously. Finally, this project has been recognised

within the industry, receiving the Innovation and Commercialisation award at the recent 2009 Asia Pacific Spatial Excellence Awards[3].

8. CONCLUSION

PSMA Australia has a history of producing quality national datasets. The data management processes are managed via LYNX, with technical processes typically being delivered by external industry partners. As part of the future program of work for the redevelopment of LYNX, PSMA is reviewing their data management processes, which is helping to foster an increased understanding of its core business. A number of enhancements to the current processes have been identified, with continuous improvement, automation and a paradigm shift to find, fix and report data issues being key concepts for development. These new strategies for data management will be capable of supporting the continued development of the industry, and continuing PSMA Australia's role in spatially enabling the information economy.

REFERENCES

Kumar, K., Dakshinamoorthy, V. & Krishnan, M. S., 2007, *Does SOA Improve the Supply Chain? An Empirical Analysis of the Impact of SOA Adoption on Electronic Supply Chain Performance,* System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on Jan. 2007 pages: 171b-181b.

Paull, D. L., 2009*, LYNX: PSMA Australia's information infrastructure facilitating collaboration and delivery capability across the governments of Australia,* GSDI 11 World Conference, Rotterdam, June15-19, 2009.

BIOGRAPHICAL NOTES

Michael Dixon is the Business Manager for the product management group at PSMA Australia. He has previously filled the role of a product manager for several national datasets, including Administrative Boundaries, Transport and Topography and CadLite. In his current role, he oversees operational activities including the current data maintenance cycle, technical infrastructure and is also the leader of technical development for new data management processes and new data products. He has significant industry experience, with a long history in local government coupled with a masters degree with spatial specialisation from UNSW.

CONTACTS

Mr Michael DIXON
PSMA Australia Limited
Level 1, 115 Canberra Ave
GRIFFITH
ACT 2603
AUSTRALIA
T. +612 6295 7033
F. +612 6295 7756
E. michael.dixon@psma.com.au
Web: http://www.psma.com.au

[3] APSEA 2009 Awards: http://www.walis.wa.gov.au/forum/general_info/apsea-gala-dinner