A New G-NAF® Processing Environment

Brian MARWICK, Australia and Peter RICHARDS, Australia

Key words: G-NAF, Addressing, Street Name, Suburb.

SUMMARY

In 2004, PSMA Australia launched Australia's geocoded national address file called G-NAF®. Since the initial release, G-NAF has been updated every 3 months. G-NAF was built, and subsequently maintained, using addresses contributed every three months by the States and Territories of Australia, the Australian Electoral Commission and Australia Post. The methodology to build G-NAF utilises PSMA Australia's national Transport dataset (i.e. road centrelines), a component of the PSMA Australia's Administrative Boundary dataset (i.e. Localities) and the National Gazetteer as part of the validation of addresses provided by the contributors. This methodology, which systematically validates the various components of the Contributors addresses against the locality and transport reference datasets, has been shown to be most successful.

Over the six years since implementation, the quality of the addressing within G-NAF has steadily improved with each update. This is the result of a greater understanding of the issues associated with addressing in Australia being gained as well as efforts made to correct the addresses, both by PSMA Australia through the implementation of rules, and each of the Contributors in their source address datasets.

Notwithstanding the considerable take-up of G-NAF by organisations both within Government and the Private Sector, the redevelopment of the G-NAF maintenance processes is now required to allow G-NAF to take full advantage of PSMA Australia's LYNX environment. This new environment will pave the way to continually maintain G-NAF as new data is incrementally received from the various Contributors. Provision will also be made for an increase in the number of Contributors and the provision of geocodes from multiple sources. This will require changes to the management of metadata which accompanies G-NAF addresses such as Confidence levels. Potentially, the development and implementation of the new G-NAF processor will enable geocoded addresses to be available to users on a far more frequent basis in the future.

A New G-NAF® Processing Environment

Brian MARWICK, Australia and Peter RICHARDS, Australia

1. INTRODUCTION

It has been over five years since PSMA Australia released the initial build of Australia's G-NAF (Geocoded National Address File) in April 2004. During this time G-NAF has been updated on a three monthly cycle and distributed to an increasing number of users both the public and private sectors. Each update has seen an improvement in the quality of the address data through the efforts of the Contributors and PSMA Australia.

Over the twenty-two update cycles to date, there has been a considerable learning process associated with managing the many issues which have impacted the quality of addresses and identifying the steps necessary to improve the data. Notwithstanding the improvements made to date, further efforts are required to meet the expectations of the users in terms of data quality, timeliness and metadata.

This paper outlines the experiences of PSMA Australia in the maintenance of G-NAF and the anticipated future developments to ensure that G-NAF continues to meet the expectations of users. A brief background to the concepts on which G-NAF is built is also provided.

2. PSMA AUSTRALIA LIMITED

PSMA Australia Limited is an unlisted public company, established under Australia's Corporations Act, wholly owned by the State, Territory and Australian Governments.

Its primary role is to ensure that the substantial value inherently held within national spatial datasets can be readily accessed so as to deliver economic, social and environmental benefits to Australia. It accomplishes this by forming and managing a crucial supply chain between creators of fundamental spatial information and users of this information by aggregating, integrating and distributing national spatial datasets.

3. BACKGROUND TO G-NAF®

The concept of a National geocoded addresses file was developed by PSMA Australia in conjunction with a number of organisations in the mid to late 1990's. Following considerable investigations including an extensive pilot study in 1999, G-NAF was first released by PSMA Australia in April 2004 with the support of the respective State and Territory Governments, Australia Post (AUSP) and the Australia Electoral Commission (AEC) through their supply of address datasets. G-NAF was built by LogicaCMG (now Logica) in partnership with Geometry under contract to PSMA Australia following a tender process. At the completion of the initial build, PSMA Australia engaged Logica to undertake maintenance of G-NAF on a three monthly basis (Paull et al, 2005). As at 31 December 2009, twenty-two updates of G-NAF had been completed and made available to users.

4. OVERVIEW OF G-NAF® METHODOLOGY

The concept developed by PSMA Australia is based on the utilisation of addresses from multiple contributors with each Contributor dataset being subject to both spatial and aspatial validation of locality and street name components of each address against a set of reference files to validate the locality and street name components of each address. The five key components of the processing are outlined below.

4.1 Mapping the Contributor Addresses to the G-NAF Address Model

The initial component of the process requires the mapping of all addresses supplied to PSMA Australia on a three monthly basis into the G-NAF address format. This format is consistent with the Australian Standard AS4819. As part of this process any address which does not have the mandatory components, namely Number_First, street name and locality will be rejected. A number of addresses in rural areas will sometimes have a building or property name identifying a caravan park or a farming property, however as they have no number_first under the current process these will also be rejected. Work is currently underway to identify these properties and create rules to allow them into G-NAF.

4.2 Identification of Possible Errors in Contributor Addresses

The G-NAF process utilises an extensive rules base to ”correct” Contributor addresses where locality names, street names or street types can be clearly identified as being incorrect. The rules have been developed through both programmatic means and manual interrogation of the addresses over the five years of updates. Where errors are identified and corrections made, alias locality and alias street locality tables will be populated as part of the correction process.

Changes to the Contributor data are made available to the Contributors to allow them to correct their address data in future supplies. When the Contributor rectifies their address the rule will no longer be applied however it will remain in the database for future use if required. As a result of the extensive improvement process which has occurred over the past five years, there are some 200,000 rules available to support the validation of addresses.

4.3 Validation against Reference datasets

The G-NAF process ensures all addresses have a locality and a street name and that these fields have been validated spatially and aspatially against PSMA Australia’s gazetted locality boundary dataset and the national road dataset. These datasets are also maintained by PSMA Australia on a three monthly basis using data supplied by each State and Territory Government prior to the commencement of any G-NAF processing.

Where a locality name does not match a gazetted locality name, it will be checked against the Commonwealth Gazetteer, which contains all place names in Australia, in an effort to identify in which locality the address exists. Failing this, the address will be rejected from G-NAF until the correct locality can be determined. The only exception to this rule is South Australia, where un-gazetted topographic localities are used in lieu of full coverage of gazetted localities across the State.

Where a street name does not match the PSMA Australia transport reference dataset, Transport & Topography, within a specific locality, the address will also be rejected unless there are at least five addresses using the street name within that locality or multiple contributors have supplied addresses using the street name. Rejected addresses will be reviewed to identify if any resolution is possible.

This variation was adopted during the initial building of G-NAF where a considerable number of addresses were being rejected due to the absence of street names or the use of alternative street names in the Transport Datasets supplied by the respective State and Territory Governments. The omission of street names could occur as a result of street names not being gazetted either through a failure of Local Government to follow the gazettal process at some point in time, or perhaps the road being associated with relatively new land development and the gazettal is still working its way through the system.

Since the initial build of G-NAF, considerable improvements in the alignment of State / Territory Government address and transport datasets has occurred which has reduced the impact of this variation. The use of this approach for the acceptance of addresses, street names not in the Transport Reference dataset will be reviewed in the near future based on the improvements made by the respective State and Territory Governments.

4.4 Merging Addresses from Multiple Contributors

G-NAF currently utilises national address data from three Contributors (i.e. the State and Territory Governments, AUSP and AEC) to build a singular dataset where associated metadata (i.e. Confidence level) provides information as to the level of congruency achieved through a merging (i.e. matching) process based on eleven (11) key fields. Through this process each address will be allocated a 'Confidence level" to indicate the extent of the match achieved. Where a Contributor address fails to achieve a match on all key fields it will remain as an address in its own right within G-NAF. The Confidence levels used are:

- 2 All three contributors have submitted this address.
- 1 Only two Contributors have submitted this address.
- 0 Only a single contributor has submitted this address.
- -1 The address is no longer supported by any contributor (i.e. has been retired).

The fields on which the addresses must match are:

- Locality
- Street Name
- Street Type
- Street suffix
- Primary postcode
- Flat number
- Level Number
- Number_First
- Number_First prefix
- Number_First suffix
- Number_Last

4.5 Allocation of Geocodes

The G-NAF process geocodes each address to an accuracy dependant on the level of validation achieved against the Locality and Transport reference dataset. For example, an address which only passes the locality validation will only be allocated a locality geocode (i.e. the centre of the locality). An address which passes the street locality validation will also obtain a geocode which is the midpoint along the street centreline within a locality. As all addresses supplied by the State and Territory Governments include geocodes accurate to the parcel level, any address which has been contributed to by the State and Territory Governments will also have a parcel level geocode. Given the match levels now being achieved, in excess of 90% of the addresses in G-NAF now have geocodes to the parcel level. Metadata is provided indicating the level of geocoding achieved.

In some situations, sufficient parcel level geocoded addresses may exist along a street to use "gap" geocoding thus improving the accuracy of geocodes to a number of addresses previously with geocodes only at the street locality level.

5. CURRENT STATUS OF G-NAF®

5.1 G-NAF User Base

Since its implementation in 2004, the use of G-NAF has grown steadily through both the private and public sectors. Address data from G-NAF is embedded in numerous online systems where address searching is required. The telecommunications, banking, finance and insurance industries and government have been found to be the primary users of the G-NAF dataset in recent years.[1] A number of State Emergency Services are now using G-NAF to support their dispatch services. It is now an integral component of the National Address Management Framework (NAMF) and has been mandated as the address reference dataset to be used for a number of Federal, State and Territory Government Agencies.

5.2 Data Quality Improvement

As previously mentioned, considerable improvement of address quality and geocoding has been achieved over the past five years. The following table provides an overview of the geocoding accuracy and the matching levels between Contributors (i.e. Confidence) currently being achieved on a National level and shows the overall improvement that has taken place since the initial update in August 2004. The rate of improvement achieved has been relatively consistent over the period.

[1] Finding from G-NAF Product Lifecycle Recommendations Report, v1.0, 2009.

Summary of Geocode Reliability of Principal Addresses in G-NAF							
Geocode Reliability		Aug 04 (Update 1)		Nov 09 (Update 22)		Change in No. of Addresses	% change
		No. of Addresses	%	No. of Addresses	%		
1	GPS Derived level	0	0.0%	0	0.00%	0	0.0%
2	Within Site Boundary	9,889,867	82.2%	11,459,442	91.0%	1,569,575	8.8%
3	Gap Geocoded	226,159	1.9%	173,527	1.4%	-52,632	-0.5%
4	Street Level	1,369,532	11.4%	902,970	7.1%	-466,562	-4.3%
5	Locality Level	515,479	4.3%	62,475	0.5%	-453,004	-3.7%
6	Topo Level	32,680	0.3%	566	0.0%	-32,114	-0.3%

Table 1: Summary of Geocode Reliability in G-NAF

Summary of Confidence of Principal Addresses							
Confidence Level		Aug 04 (Update 1)		Nov 09 (Update 22)		Change in No. of Addresses	% change
		No. of Addresses	%	No. of Addresses	%		
2	Three Contributors	5,286,958	43.9%	7,321,578	58.1%	2,034,620	14.2%
1	Two Contributors	2,458,817	20.5%	2,232,670	17.7%	-226,147	-2.8%
0	One contributor	4,287,943	35.6%	3,044,727	24.2%	-1,243,216	-11.4%
Total Principal Addresses		12,033,718	100.0%	12,598,975	100.0%	565,257	

Table 2: Summary of Confidence Levels in G-NAF

6. SIGNIFICANT DATA RELATED ISSUES IMPACTING DATA QUALITY

Over the past five years a number of significant issues have restricted the efforts of PSMA Australia to improve the quality of the addresses as reflected by the level of matching achieved between the three Contributors and the number of geocodes at a parcel level. As an interim step, address anomalies caused by several of these issues have been resolved through the implementation of processes to identify the faulty addresses and then correct them through the use of rules. Where the issues can be clearly identified the Contributors have been notified. Some of the major issues encountered to date are outlined below.

6.1 Incorrect Road (Street) Name or Road (Street) Type

Initially the major reason for mismatching between addresses was the use of incorrect road names either as a result of alias names, incorrect spelling or incorrect road types (eg. "avenue" instead of "crescent"). The majority of these have now been rectified by the various Contributors as a result of a progressive improvement program. This is reflected in the jump of in excess of two million in Confidence Two addresses (i.e. all three Contributors match) in the table above.

Of those remaining to be resolved, many will be road names in rural areas where the use of alias names is more common. For example, often the State Government will allocate road names, such Princes Highway, for major roads and highways form a network across the State. Given major highways of this nature go through many communities they may have multiple names at a community level. It is often these local names that are used for addressing. These are being progressively resolved as alias road names are identified and linked to the correct road name in G-NAF through the rule creation process.

6.2 Street Addressing within Private Developments

The absence of street addresses within private developments, such as gated communities, shopping centres, caravan parks and retirement villages in the respective State Governments' address datasets remains the single largest issue with regards to both the matching of addresses, and improvements in the level of geocoding. Many of the 900,000 addresses that are only geocoded to a street locality level (i.e. mid-point along the street centreline) will be the result of this issue.

This issue has arisen as each of these developments is generally viewed by Local Government as a singular property rather than the multiple properties that form these often large development complexes. The road network and the individual servicing of the units within these developments are not the responsibility of Local Government and as a result properties within the complexes are not assigned individual addresses. Given State Government address datasets are sourced from Local Governments as a general rule, these addresses do not exist in the State and territory Government contributions in G-NAF.

A similar situation also exists where units under single land tenure are leased or rented and as this is not reflected in the Land Registration system, no unit numbers will be available within the State Government address supply.

The addresses allocated by the Body Corporate / Owner responsible for the management of the complexes are used however by all the residents and as such are included in G-NAF through the contribution of AEC and AUSP. In some States this situation is currently under review particularly given the increase in the number of these communities over the past decade and the impact of the emergency services attending these developments.

A similar situation also exists where facilities owned by the Commonwealth Government, such as Army, Air force and Naval Bases, exist. Many of these bases include significant residential developments; however, as these are not under the control of Local Government no addresses are recorded in the State Government address files.

6.3 Historical or "Retired" addresses

An issue which has been difficult to resolve is the inclusion of "historical" addresses in Contributor address datasets. In some situations they have been associated with whole communities such as disused mining towns, however the addresses have not been retired from the Contributor datasets. To date, no rules have been implemented to prevent these addresses entering G-NAF; however, this may be required in the near future in the absence of any action by the Contributors. Whilst it may be possible to exclude the addresses associated with abandoned communities, the identification of individual addresses that are no longer used as a result of renumbering or a new development superseding the existing address is more difficult. The most likely approach will be to seek assistance from the Contributors and see if they are able to ascertain the validity of the address. An alternative approach for PSMA Australia may be to remove any address that has been supported by only one Contributor for more than five years. Such an approach could only be adopted after all the checks had been undertaken to ensure the address had not matched as a result of road name issues

6.4 Ranged Addresses

An ongoing impediment to the matching of addresses is the use of ranged addresses. Whilst Local Government may allocate ranged addresses to properties in some areas to take into account the potential for redevelopment of the property in the future, the use by the owners of the ranged address is problematic. It would appear most property owners tend to use only the first number as the address when notifying agencies of their address. As a result, G-NAF will have both addresses as principal addresses. A rule has been implemented in G-NAF to use the ranged address as the principal address where two contributors support the ranged address and make the single numbered addresses within the range as the alias addresses. The requirement for two Contributors supporting the ranged address was required as difficulties were encountered with variable ranges. This approach is currently under review as its success in assisting the matching of addresses has been minimal.

7. A NEW G-NAF® PROCESSOR

7.1 Overview

Whilst address data improvement has been the major focus for the past five years, it is apparent that significant improvements in currency of the data will be required within the next two years, given feedback now being received from users. To support a change from a three monthly update cycle together with other improvements such as full integration into PSMA Australia's LYNX[2] environment, the entire G-NAF processing environment will be redeveloped. This will require the G-NAF Processor to be moved from its existing Data Manager's environment (i.e. Logica) into the LYNX environment. In this environment, the new G-NAF processor will exist as a hosted service container. Web services will be used by permitted users to access the new G-NAF processor's functionality. The diagram below provides an overview of the various interfaces between G-NAF Processor, PSMA Australia's data management processes and the LYNX Services Framework.

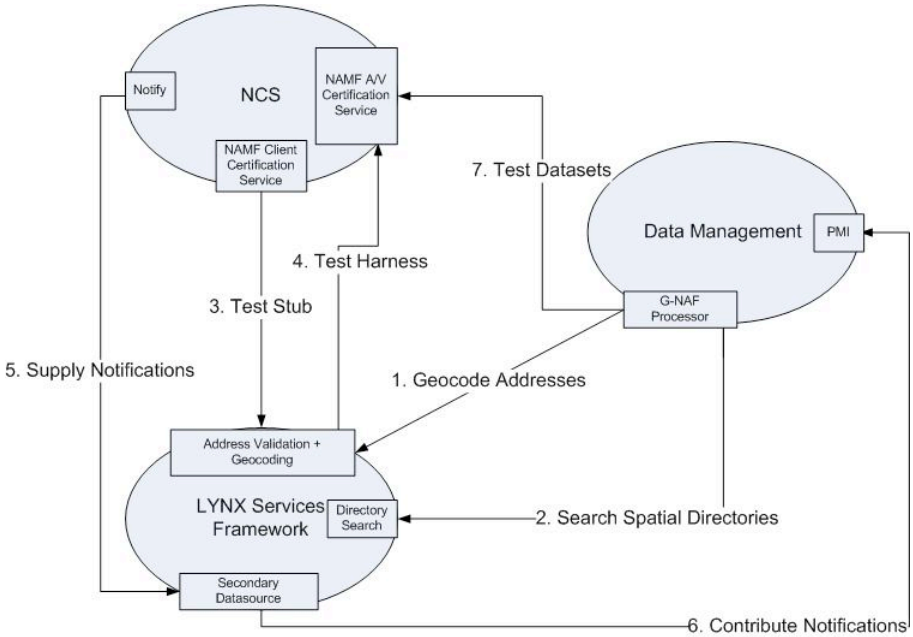

Figure 1: PSMA Australia's Data Management Processes

A key component of the new G-NAF Processor will be the implementation of the G-NAF Knowledge database which will store:

- All current addresses;
- All retired address;
- All the valid rules that have been created;
- All the aliases that have arisen as a result of these rules;
- All the addresses that have failed validation and have not as yet been resolved; and
- The metadata related to the processing of all addresses.

[2] LYNX is the database infrastructure developed by PSMA Australia. LYNX has been developed with SOA architecture, and will be providing services to both government and private sectors.

The storage of all this information will assist in overcoming the current deficiencies of the existing processor where both non validated addresses, and the rules used to identify possible errors in the Contributor addresses, are not stored within the G-NAF database.

7.2 Outcomes and Benefits of the New G-NAF Processor

It is expected that the new G-NAF processing environment will have the following outcomes and benefits:

- Continued provision of high quality datasets at an increased frequency;
- Integration with the 1Spatial's Radius Studio being used by PSMA Australia for the maintenance of the Reference datasets ;
- Integration with the LYNX Services Framework;
- Inclusion of the National Address Management Framework (NAMF) notification services;
- PSMA Australia to have a greater control of the processes behind the data creation;
- A more adaptive processing environment as the new processor will be modular and make greater use of web services;
- Increased capability for accepting data from non traditional sources;
- Closer relationship with Jurisdictions through improved feedback processes available via LYNX and NAMF;
- Reduced labour intensive data processing; and
- Reduced costs through the optimisation of the data maintenance cost structure.

7.3 New Functional Requirements

The following sections provide an overview of some elements of the new functional requirements for the proposed new G-NAF processing environment.

7.3.1 Non Sequential Maintenance

The requirement for increased frequency to at least weekly, and possibly daily, updates fundamentally changes the manner in which G-NAF is processed. Currently G-NAF is processed following the updating of the two primary reference datasets, namely the Locality (Suburbs) and Transport (Road Centrelines) datasets. Furthermore, in the new LYNX environment, it is envisaged that updating cycles will eventually be continual as updates are received from the Contributors. Given G-NAF is aligned or referenced to the Locality and Transport datasets, the G-NAF processing environment must ensure any updates of these reference datasets are immediately reflected in G-NAF notwithstanding that no address updates have been received. Alternatively, address updates from Contributors will be processed on receipt rather than waiting for the updating of the Reference dataset. The new processing environment must ensure G-NAF maintains synchronicity with the Reference Datasets at all times. This will be achieved through the use of rules in a similar manner to that currently used.

7.3.2 Support for both Full File and Incremental Contributions

Currently all Contributors provide full address supplies each update. The G-NAF process identifies the changes and updates G-NAF accordingly. It is anticipated that the introduction of more frequent updates, incremental supplies will be provided by some, if not all

Contributors. Given this, the new G-NAF processing environment will need to be capable of accepting and processing both full and incremental supplies. This includes the option of receiving updates via a web service possibly as they are "created" by the contributors.

7.3.3 Improved Management of the Rules

As previously indicated, significant use has been made to date of "rules" to support the "correction" of Contributor addresses where errors can be clearly identified. The new G-NAF environment will require the integration of the rules, currently approx 200,000 into the database (i.e. G-NAF Knowledge Base) thus allowing the generation of metadata to enable users to identify which addresses have been impacted by any rules during the processing. This will be increasingly important given that the rules will be an integral part of the non sequential updating process. Rigorous management of these rules will be required.

7.3.4 Integration with the LYNX Services Framework and the National Address Management Framework (NAMF)

Through integration with PSMA Australia's LYNX Services Framework (LSF), the new G-NAF Processor will have the capacity to utilise more sophisticated address parsing and matching algorithms through a standardised interface. Furthermore, through integration with the NAMF, it is possible for PSMA Australia to receive notifications concerning addresses that are either not currently in G-NAF, or are incorrectly stored in G-NAF, allowing the currency and correctness of the dataset to improve. PSMA Australia is looking at how it might utilise these notification services to provide an interim solution for more frequent updates of the G-NAF, given the new G-NAF processor is unlikely to become available until early 2011. This might be achieved through the creation of a supplementary dataset, for use between updates of G-NAF.

8. CONCLUSION

The methodology adopted by PSMA Australia to build and subsequently maintain G-NAF has proved to be most successful over the past six years. By bringing together the three National datasets, it has ensured the inclusion of virtually all of addresses in use within Australia. At the same time it has been able to build on the considerable improvements the respective Contributors have made to their individual datasets over a five year period. The use of processing rules by PSMA Australia as part of the maintenance has allowed the identification and correction of errors in locality and street names of Contributor addresses. It has also allowed reporting back to the Contributors the issues involved. This steady improvement in G-NAF has undoubtedly assisted in the acceptance it is now receiving in the marketplace.

A number of significant issues still exist in terms of further improving data quality and, most importantly, reducing the time between updates. The proposal by PSMA Australia to redevelop the G-NAF Processing environment such that it is tightly integrated with its new LYNX environment, including the web based services, will allow the much needed reduction in the update period from three monthly to be progressively reduced to possibly daily over the next few years. This will however require changes in the manner in which address contributions are supplied to PSMA Australia possibly via web services.

A further significant improvement in the data quality from a geocoding perspective will be dependent to some degree on the success of the respective States and Territories in ensuring their address datasets include addresses in "private" developments, such as gated communities, caravan parks and Commonwealth Government facilities unless alternative sources of addresses with this data can be found. The new G-NAF processor will be capable of supporting such an approach.

In summary, PSMA Australia's G-NAF has made a valuable contribution to the Australian Spatial Data infrastructure over the past five years and work is in progress to ensure it continues to meet the user needs into the future.

REFERENCES

Paull, D. and Marwick, B., 2005, **Maintaining Australia's Geocoded National Address File (G-NAF)**, Spatial Science Institute Conference Proceedings 2005, Melbourne.

Coslett, D., 2009, **Requirements for G-NAF Processor**, Internal PSMA Australia document.

MacDonald, S., 2009, **G-NAF Product Lifecycle Recommendations Report**, Internal PSMA Australia document.

BIOGRAPHICAL NOTES

Brian Marwick is currently employed on a part time basis by PSMA Australia following his move to semi retirement in 2008. He is also currently undertaking part time a Masters of Geomatics Engineering at the University of Melbourne. Brian's experience in the spatial information industry extends over some forty years. During this time he has moved from a project surveying role on major construction projects to senior management roles in both the public and private sectors.

Since 1980, Brian has been actively involved in both technical and management capacities in the implementation of the spatial information technology that has evolved during this period. Throughout his time in the industry Brian has served on various Government Advisory Committees associated with the Spatial Industry. He was President of the Institution of Surveyors, Australia in 1998 and served on its Council for some nine years. Brian has also been a member of the University of Melbourne Geomatics Course Advisory Committee since 1986 and has chaired this Committee for the past ten years. He was also an External Member of the University of Melbourne Engineering Faculty for some fifteen years.

Brian is a Licensed Surveyor in Victoria and a Fellow of the Surveying and Spatial Institute of Australia.

Peter Richards is currently a Senior Project Manager with PSMA Australia. He is jointly responsible for the day-to-day management of the Dataset Maintenance Program and ensures the timely delivery of well maintained and quality-assured national datasets and metadata to clients. Peter was instrumental in managing the integration of the topographic dataset, as well as building and the ongoing maintenance of the Geocoded National Address File (G-NAF).

Peter has a Bachelor of Information Technology (Geospatial Information Systems) and a Diploma of Surveying and Land Engineering. He is a committee member on the Australian Standards for Street Addressing (AS:4819), and is also a committee member on the Interchange of Client Information – with a focus on Address Interchange (AS:4590-2006).

CONTACTS

Mr Brian MARWICK
PSMA Australia Limited
Level 1, 115 Canberra Avenue
Griffith, ACT
AUSTRALIA
Tel. +612 62957033
Fax + 612 62957756
Email: brian.marwick@psma.com.au
Web: www.psma.com.au
Affiliations: Fellow of SSSI

Mr Peter RICHARDS
PSMA Australia Limited
Level 1, 115 Canberra Avenue
Griffith, ACT
AUSTRALIA
Tel. +612 62957033
Fax + 612 62957756
Email: peter.richards@psma.com.au
Web: www.psma.com.au