Conflation of Crowdsourced and Authoritative Data to Enhance Geospatial Intelligence

Stefano Cavazzi, Brendan D. Mason, Neil Kirk, Gobe E. Hobona, and Roger C. Brackin
Envitia
UNITED KINGDOM

stefano.cavazzi@envitia.com brendan.mason@envitia.com neil.kirk@envitia.com
gobe.hobona@envitia.com roger.brackin@envitia.com

David Barber
Defence Science and Technology laboratory
UNITED KINGDOM

dbarber@dstl.gov.uk

ABSTRACT

The wealth and breadth of information available to the Intelligence, Surveillance and Reconnaissance (ISR) community has meant that features representing real-world phenomena may be duplicated or represented differently across several different datasets. Thus effective information fusion must address this significant challenge. Geospatial specialists are required to process and integrate an increasing number of sources of location-referenced information from authoritative data providers and the general public. It is accepted within the Defence community that crowdsourced data from open sources can add value to authoritative datasets when conflated appropriately. When such conflation is not properly managed, it can present the risks of confusing the user, or undermining the actual or perceived reliability of the data.

This paper is concerned with Data Conflation, which is commonly defined as the process of combining geographic information from overlapping sources so as to retain accurate data, minimise redundancy, and reconcile data conflicts. The work builds upon achievements and lessons learnt from previous research by the UK Ministry of Defence, particularly GI2RA (Geospatial Intelligence Integrated Reference Architecture), while also exploiting recent advances, such as crowdsourcing and the NATO Geospatial Information Framework (NGIF).

The paper describes a proof-of-concept involving conflation of an authoritative dataset from the Multinational Geospatial Co-production Program with a non-authoritative crowdsourced dataset from the OpenStreetMap initiative. The adoption of a harmonised pan-domain model is central in the proposed conflation process, as it enables the efficient harmonisation of multiple datasets from different domains. The research adopts the NATO Geospatial Information Model (NGIM), provided by NGIF, as the pan-domain model. Once the selected dataset has been transformed to the NGIM feature model, the matching process can be implemented with another previously discovered and transformed dataset. The resulting matched features are integrated according to predefined integration rules related to data content and geometric characteristics facilitating the creation of a new conflated dataset. The enriched dataset includes lineage and quality metadata ensuring traceability of the provenance of information. The developed concept, its implementation and evaluation with sample datasets are described in the paper.

1.0 SETTING THE STAGE

Data fusion has traditionally focused on processing physical or hard data while human observed or soft data has received less attention within data fusion processes. In the context of intelligence analysis, soft data can improve situational awareness where attributes, connections and interactions are difficult to observe with physical sensors (Gross *et al*., 2012). When data is generated by physical sensors as well as by crowdsourcing, its combination will likely require a hard/soft data fusion framework (Park *et al*., 2013).

The wealth and breadth of information available to the Intelligence, Surveillance and Reconnaissance (ISR) community has meant that features representing real-world phenomena may be duplicated or represented differently across several different geospatial datasets. However, the combined use of these datasets is made difficult as they may have been derived from different sources, using different acquisition methods and using differing data structures and attribution (Wiemann and Bernard, 2010). Nevertheless, it is accepted within the Defence community that crowdsourced data from open sources can add value to authoritative datasets when fused appropriately. Increasingly, geospatial specialists are being required to integrate an increasing number of sources of location-referenced information from authoritative data providers and the general public in the production of Geospatial Intelligence (GEOINT). Therefore, in order to create application specific high value information, hard/soft data fusion, in particular geospatial conflation is essential.

This research is motivated by the problem of determining an effective information fusion process to conflate authoritative and crowdsourced data. Geospatial data conflation is commonly defined as the process of combining geographic information from overlapping sources so as to retain accurate data, minimise redundancy, and reconcile data conflicts (Longley *et al*., 2005). A variation of this definition is offered by the Open Geospatial Consortium (OGC) which describes it as “the process of unifying two or more separate datasets, which share certain characteristics, into one integrated all-encompassing result” (OGC, 2010). The OGC definition recognises the fact that the datasets being combined may share any number of characteristics such as overlapping spatial extents, common thematic attributes and overlapping periods of time. Both of these definitions are considered to be relevant to the work described in this paper.

Conflation consists of several sub-processes. The first step involves data discovery, analysis and comparison to ensure suitability to further processing steps. Then data needs to be adjusted to allow the conflation processing, including such operations as map alignment and spatial or thematic generalization. Only at this stage features can be matched using geometrical, topological and semantic attributes to achieve an unambiguous mapping. This is one of the biggest challenges in data conflation as there are a number of problems at this stage that need to be solved which include different coordinate reference systems, representations, resolutions or classifications. After the features have been matched it is possible to join or transfer the required attributes between the datasets to complete the data conflation process. When such conflation is not properly managed, there is the risk of undermining the actual or perceived reliability of the data. Hence/Thus Effective information fusion must address this significant challenge.

The benefits that data conflation of crowdsourced and authoritative information offers are:

- the ability to verify, at short notice, different vector datasets produced by different organisations;
- the ability to enrich authoritative datasets with information pulled-through from local knowledge;
- improved situational awareness through controlled information integration.

Previous UK MOD research, specifically the Geospatial Intelligence Integrated Reference Architecture (GI2RA) research project, proved the feasibility of using a pan-domain harmonised data model in the delivery of a software architecture to support the delivery of coherent and consistent Geospatial Intelligence (GEOINT). In the work described in this paper, the NATO Geospatial Information Framework (NGIF) was selected as the harmonised data model. NGIF is a suite of specifications for defining standardised geospatial products that will be used at all levels of command within NATO. The purpose of NGIF is to ensure interoperability between NATO, NATO Nations, and non-NATO Nations by defining a common standardized data model for the production, exchange and use of data and standard products. Version 1 of NGIF is primarily based on the US GEOINT Structure Implementation profile (GSIP), which was based on a series of domain models such as Additional Military Layers (AML), Aeronautical Information Exchange Model (AIXM), Electronic Navigational Charts (ENC) and others. The standardisation of a common data model and standardised portrayal rules for data and map products derived from one data model brings advantages in terms of semantic interoperability, but also provides some challenges in terms of coherence and maintenance.

2.0 LITERATURE REVIEW

Spatial data conflation is a specialized task within geoinformatics that is mainly used for detection of change, integration, enrichment and updating of geospatial datasets (Yuan and Tao, 1999). While the integration of geospatial data is widespread, conflation is still considered a difficult task due to the varying (differing) levels of accuracy or completeness of data collected in different ways and for different purposes (Stankute and Asche, 2011). Wiemann and Bernard (2010) identify two types of data conflation:

- Horizontal: referring to the merging of adjacent spatial data by edge-matching or zipping.
- Vertical: referring to datasets covering the same area including applications such as detection of changes, updating, enrichment and integration.

This paper focuses on vertical conflation of vector-to-vector data for the purposes of enriching spatial data through the augmentation of feature-level information from an assured dataset with other information from a crowdsourced dataset. Note that data conflation in this context focuses on vector data, the conflation of other information such as place names (Hastings, 2008) or points of interest (Song et al, 2014) is also of particular interest. Interest in conflation also remains high with the growing popularity and interest in the Open StreetMap (OSM) project. Pourabdollah *et al.* (2013) report on a quality assessment of OSM using conflation techniques.

Conflation as a process ranges from totally manual, to human assisted or fully automated depending on the nature and quality of the input data, the conflation requirements and the implementation system. The conflation process consists of several sub-processes; input data may require pre-processing to ensure compatibility including format translation, coordinate system conversion and other basic data preparation operations. Data checks can be incorporated in this first step to ensure internal consistency of the data with the conflation rules. Available metadata on geometric, thematic and structural properties can support this phase. A further step is also needed to align data models of input data; the role of harmonisation in this instance is to enable data conflation of data with different data models.

After these initial steps the real challenge of the conflation process is to identify and assign similar features to each other through some kind of similarity measure; geometric, attribution and/or topological (Tong *et al.*, 2014, Li and Goodchild, 2011). The matching operation has so far restricted existing approaches, limiting them to isolated operations (Zhang *et al.*, 2005); researchers have not yet discovered a generic way to conflate any type of data from various datasets. This central task in the conflation process may not be able to achieve unambiguous results requiring human intervention to disambiguate uncertainty (Freitas and Afonso, 2012). After the matching step, the features that have been successfully matched can be conflated according to predefined rules. Integration of matched data can involve either data attributes or geometry. Metadata can be updated with lineage, quality and other relevant information to ensure traceability and inform data content. Un-matched features can also be transferred to a new dataset for further investigation or added to the conflated data with a clear attribution from the input data to allow their identification.

2.1 Data Fusion Research in Geospatial Intelligence Integrated Reference Architecture (GI2RA)

The GI2RA project began in 2007 to address the issues of geospatial information coherence, completeness and interoperability. The aim of GI2RA was to ensure that consistent information, at all levels of command, could be delivered to all potential users, for planning and execution (Envitia, 2010). Research within GI2RA included development of a harmonised data model. The harmonised data model was based on various defence formats such as Additional Military Layers (AML), Digital Aeronautical Flight Information File (DAFIF) and NGA Vector Map (VMAP). The feature types and attributes from the source data models were linked to the harmonised data model to support transformation of the source data into the target data model and fusion of the transformed data into a single dataset based on a single data model.

In order to practically evaluate and refine the information and harmonisation models, a deployment model and actual data encoding framework were used. The harmonised model was intentionally constrained to the needs of situational awareness users. The approach was proven to be practical and implementable through a series of trials and demonstrations. Since then however, there have been further developments within this area, most notably the release of a common pan-domain data model through NGIF.

2.2 Data Conflation Research in OGC testbeds

The OGC OWS-9 testbed (OGC, 2013) conducted research in the field of dataset conflation and geospatial web services. The research focused on the general service architecture, service interfaces and the development of workflows. The conflation architecture was mainly based on the chaining of web services, based on the Web Processing Service (WPS) standard, that consume data provided by other web services. The research provided processes for the following separate tasks:

- **Conflation based on geometric attributes**: Determining if a feature in the source data set is contained in the target data set; if it is contained, both features are linked, if it is not contained, the feature from the source data set is added to the target data set.
- **Conflation based on alphanumeric attributes**: Both, the source and the target dataset may possess different attributes; during the attribute conflation semantic reasoning is applied to determine equivalencies between attributes and if an attribute contained in the source data set is missing in the target data set, it is added to the target data set.

Within this work, an automated approach for dataset conflation was demonstrated using conflation rules that applied both geometry conflation (based on spatial similarity) as well as attribute conflation (based on semantic mediation). The realisation of this approach highlighted some disadvantages. First, every feature was returned with both source and target dataset attributes thereby potentially leading to uncertainty about the information presented. Second, the web services did not provide mechanisms to adjust conflation rules, as a result preventing the user from customising the conflation rules. The OWS-9 work however demonstrated that dataset conflation through web services was viable.

3.0 PROOF OF CONCEPT

This section demonstrates the possible use of NGIF to support conflation of data from different data sources. The adopted conflation workflow is presented in Figure 1. Within the workflow, the source data is discovered, its data model analysed and mapping rules created for the harmonisation into NGIF. Once the source data has been harmonised, the matching process can be implemented with another harmonised dataset. The matched features are integrated according to predefined integration rules related to data content and geometric characteristics resulting in the creation of a new conflated dataset. The enriched dataset includes lineage and quality metadata ensuring traceability and conformity to conflation requirements.

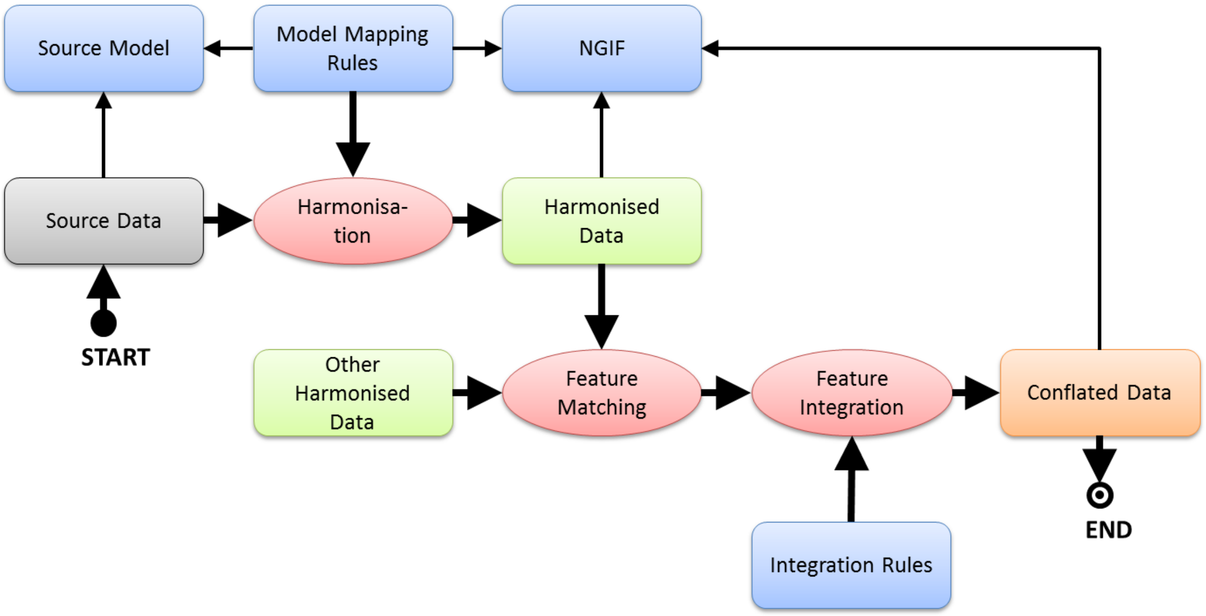

Figure 1: Information harmonisation and conflation workflow.

In order to test the proposed conflation workflow, an experiment was set up including one authoritative dataset taken from the Multinational Geospatial Co-production Program (MGCP) and one non-authoritative from OSM, this is summarised in Figure 2. The selected data sources represent, as previously described, a typical reference data set with a well-structured data model (MGCP) and a crowdsourced semi-structured model but more up to date dataset (OSM). These were harmonised into the NATO Geospatial Information Model (NGIM), provided by NGIF, using mapping rules that convert them from their original data models into the information model provided by NGIF.

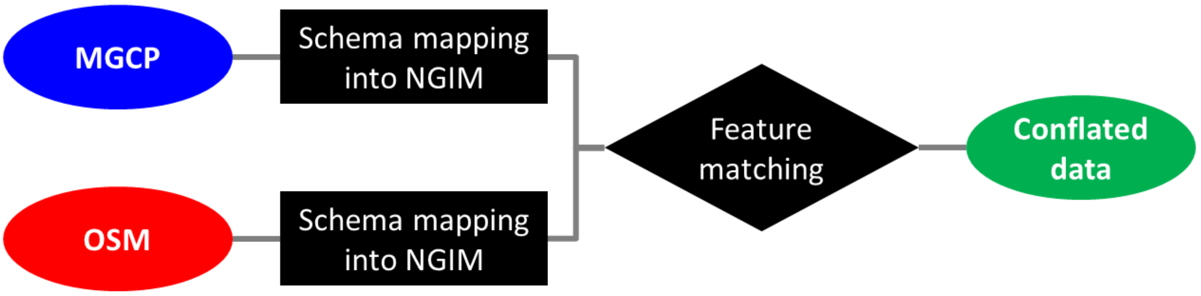

Figure 2: Conceptual representation of the workflow for conflation by alphanumeric attributes.

4.0 MATERIAL & METHODS

4.1 Study Area

The area selected for the experiment was Port-au-Prince in Haiti (an extract of 150 square kilometres covering the centre of the city). Port-au-Prince was extensively mapped during the Earthquake in 2010, with the OSM community becoming the default basemap for responding organisations such as Search and Rescue teams and NGOs. MGCP data was also published online to support relief efforts).

The MGCP data includes 836 road feature entities with all the major lines of communication. The OSM data includes 11559 road feature entities with a higher level of data density and detail than MGCP. Both datasets are shown in Figure 3**Error! Reference source not found.**.

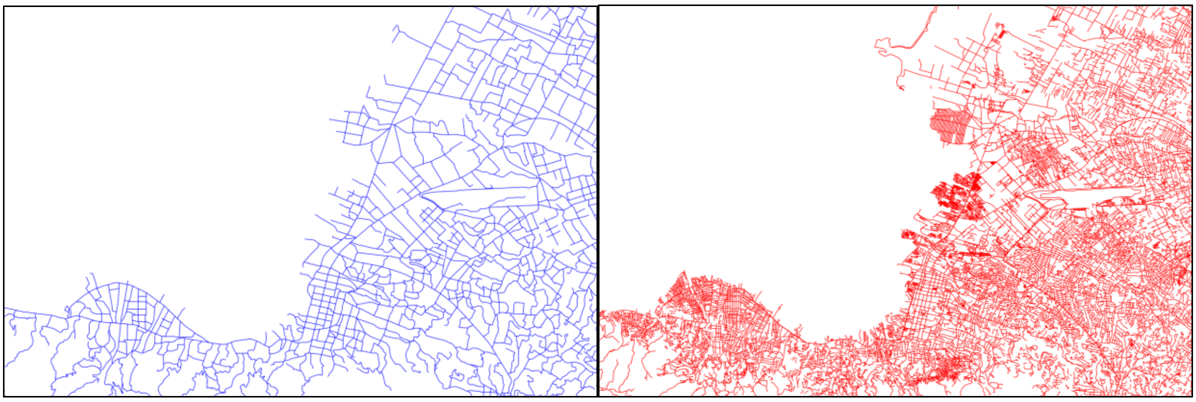

Figure 3: MGCP road data (left) and OSM data (right) for the selected study area.

The main characteristics of the two selected data sources are summarised in Table 1. MGCP is derived from high resolution, remotely sensed imagery where features are extracted to production tiles (1 degree by 1 degree) by highly trained personnel ensuring a high degree of quality assurance and conformity. The rapid update cycle and high spatio-temporal resolution make OSM a valuable source of information despite the uncertainties associated with its loosely defined data model and unrestricted collection techniques.

Table 1. Comparison of the selected data sources.

Multinational Geospatial Co-production Program	Open Street Map
• Administrative data • Quality assured • Normative status • Derived from imagery • 1:50k or 1:100k scales • Production units (1 degree by 1 degree)	• Crowdsourced data • Rapid update cycle • High spatio-temporal resolution • Data model based on nodes, ways and relations • Attributes stored in tags

4.2 Methodology

A conceptual representation of the proposed workflow for conflation based on alphanumeric attributes was presented in Figure 2 and a representation of the implementation (using Safe Software's FME) is presented in Figure 4. The two data sources (OSM and MGCP) were harmonised into NGIM using the *SchemaMapper* transformer in which customised mapping rules were imported from an external .csv file. The harmonised features were then matched via the attribute *RoadWayType*, after which only the resulting matched features were conflated. During conflation, the value of the attribute *RoadName* was transferred from OSM into the NGIM features and the accompanying metadata updated to reflect that the feature has been matched.

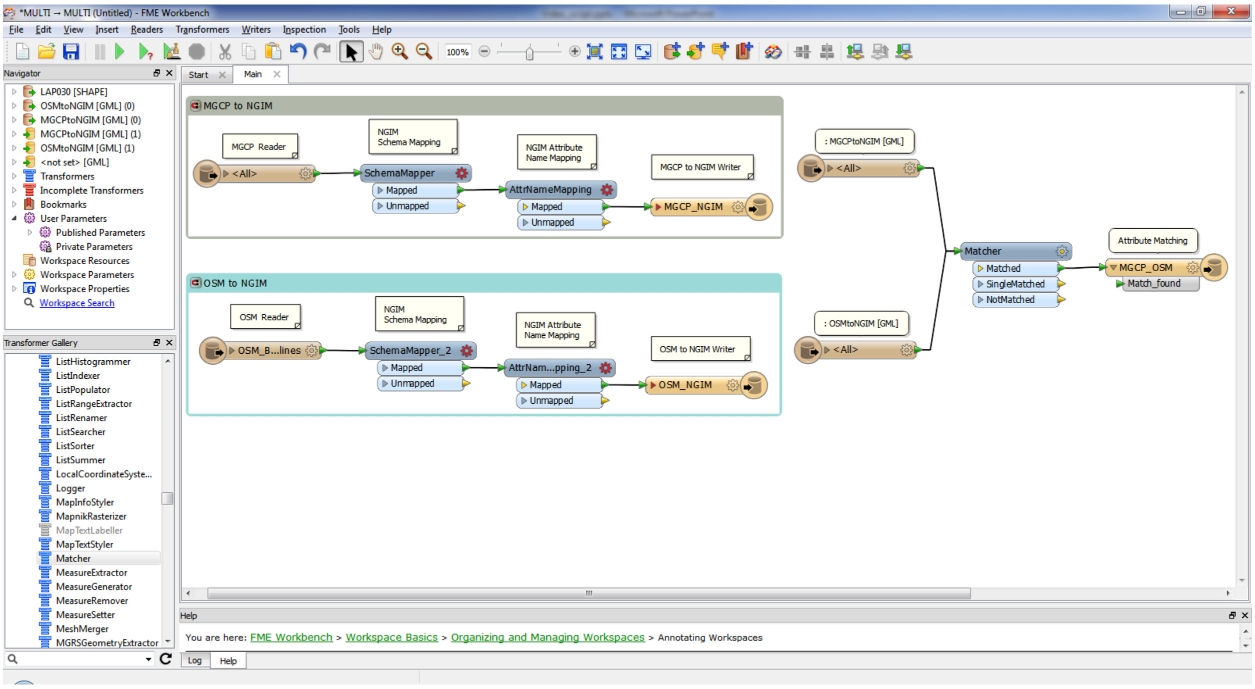

Figure 4: Implementation of the workflow for conflation by alphanumeric attributes.

In reviewing the results of the conflation, it was observed that only some of the features were correctly matched by alphanumeric attributes. The factors affecting conflation through alphanumeric attributes were identified as being:

- OSM data is unstructured, in terms of the attributes used to describe features
- Most of the attributes in the OSM data had NULL values

To overcome these issues, a conflation workflow based on geometric attributes was developed using the principle of intersection. The workflow intersects the buffers of nodes along the MGCP roads with OSM roads as illustrated in Figure 5.

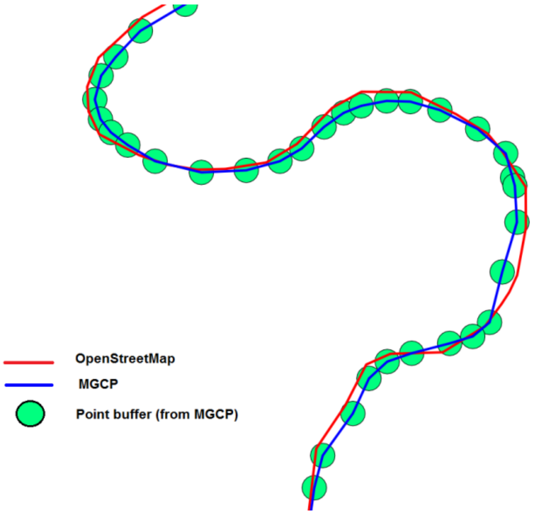

Figure 5: An illustration of the concept behind conflation based on geometric attributes.

The approach adopted for conflation of linear datasets through geometric attributes is a multi-step process. A prerequisite for the datasets being conflated is to have unique identifiers for each feature. The unique identifiers do not need to be universally unique but are required to be unique within the datasets being conflated.

The first step involves generation of points along linear MGCP features. The GRASS GIS function v.to.points was used for this purpose, from within the Open Source QGIS Software. The main variable in this step is the maximum distance between the generated points.

The next step involves creating a buffer around the points generated in the previous step. The main variable in this step is the buffer size. Smaller buffers have fewer false positives but may have more false negatives. Therefore it is advisable to use a buffer size that accounts for the typical width of a road without extending unnecessarily beyond. Such a buffer partially compensates for collection errors or errors introduced by GPS multipath.

The next step involves intersecting the generated MGCP point buffers with the OSM linear features in order to produce a spatial join (with the MGCP buffers being the target of the spatial join). A match is found when multiple MGCP point buffers with the same identifier (taken from the same road) intersect with the same OSM road. An example match is shown in Figure 5 and an overview of several roads is shown in Figure 6.

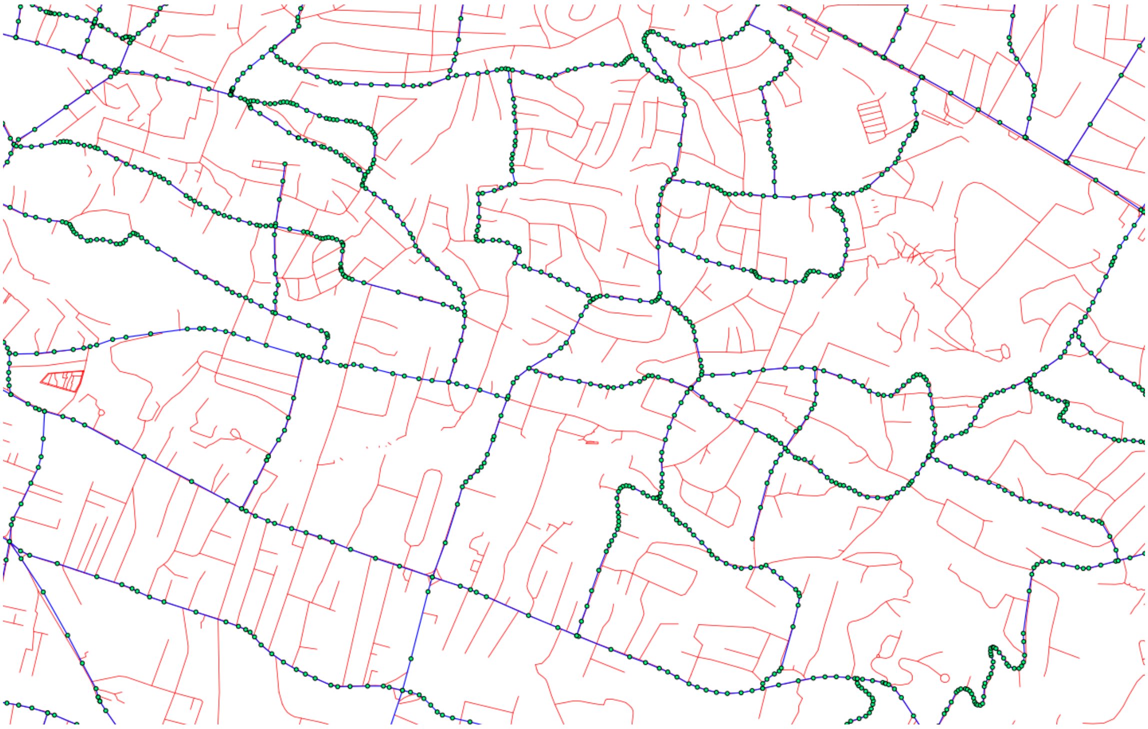

Figure 6: An overview of the MGCP roads (blue), point buffers (green) and OSM roads (red).

The final step involves a frequency count of the number of unique MGCP identifiers that match OSM features with a common identifier. The confidence of match between each MGCP road and OSM road increases with the frequency count. Any matches with frequency count of 1 are assumed to be false positives. Once identified, the matches assumed to be false positives are removed from the frequency count.

5.0 RESULTS AND DISCUSSION

A review of the outputs shows that out of 836 roads in the MGCP dataset, 670 MGCP roads can be matched to OSM roads if we consider at least 3 intersecting point buffers along the roads (i.e. just over 80% of the MGCP roads can be matched). If we consider at least 2 intersecting point buffers, then we match 727 MGCP roads, which is 87% of the MGCP roads.

The presented proof of concept demonstrates the feasibility of integrating authoritative and non-authoritative data sources. The selected datasets have very different data models, MGCP has a well-structured data schema with well-populated tables while OSM has a less-structured model but more up-to-date local information. By matching OSM data to the MGCP data it has been possible to add street names (where they exist) to the MGCP data and other information not available from the current collection process of MGCP. The initial harmonisation at the data model level, has proved essential in the workflow as it enables the filtering of feature types and matching of key individual features such as major roads. The inclusion of both alphanumeric and geometric attributes within the conflation process improves the performance of the process. Application of the GI2RA approach was found to be essential to conflation workflows involving both the alphanumeric and geometric attributes.

There were several fields that could not be mapped due to the absence of any semantically equivalent attributes in the source or target feature types. Most of the MGCP attributes that could not be mapped were due to the MGCP feature type being split into two separate feature types by NGIF. Most of the OSM features that could not be matched using alphanumeric attributes suffer from the variable structure (or lack of uniform structure) of the OSM dataset. OSM lacks topological consistency or spatial accuracy in some areas. This might be the result of the collection procedures carried out by less experienced volunteers omitting basic quality management procedures or the inaccuracies resulting from basic GPS devices. A number of pre-processing steps might be required to address the previously described issue. More advanced similarity measurements (including semantic and topological analysis) might also need to be implemented to optimise the feature matching.

A review of the results suggests that there were few, if any, false positives. We credit the configurability of the approach for this performance, as the distance of points and the size of the buffers allows the sensitivity of the process to be adjusted according to the local road network characteristics. For example, in urban areas smaller segments and low buffer radiuses can be used to account for high feature density. While in rural areas where feature densities are much lower, larger search radiuses can be used to overcome the limited data available. As OSM is typically derived from data collected using low cost GPS devices measurement precision can be low. The orientation of some road segments can, therefore, be different between authoritative and crowdsourced datasets. The use of customisable buffer sizes was found to address this issue.

Involving both alphanumeric and geometric attributes in the conflation process appears to address the issue that OSM roads are only segments of MGCP roads due to scale differences and the volunteered nature of OSM (i.e. one person may contribute part of a road and another person may contribute another part). The approach adopted also appears to address the issue of differences in position between feature entities.

The potential for automation of the conflation proof-of-concept, at scale, was considered. The key reason being that to exploit much of the crowdsourced data that is available on the World Wide Web, there is a need to have access to a significant amount of computing resources. Smaller datasets such as a national road network can be conflated using typical desktop computers. However, as the need to exploit even more crowdsourced data increases (as is currently the case), the need for more computation is likely to increase with it. Related research examining the application of Grid Computing (Hobona *et al*. 2010) and Cloud Computing techniques in geospatial analysis (Hobona *et al*. 2011) could provide insight into how so-called ‘Big Data’ technologies could be used to provide data conflation at scale.

Finally while this paper has focused on the conflation of foundation geospatial data, mainly roads, it is possible that the approach could be made more generic allowing the conflation of any datasets with geographical attributes. This is significant as intelligence production and analysis starts to rely on an ever increasing number of heterogeneous datasets.

6.0 CONCLUSIONS

This report has investigated the present state of data conflation within the MOD defence community. The developed concept and its implementation demonstrate the feasibility of data conflation of datasets from different sources to achieve enriched information. The report concludes that a harmonised model for geospatial information such as the one provided by NGIF can, indeed, enable data conflation. This conclusion addresses the first of the research questions presented in this report. The following conclusions can also be drawn from the report:

- Differences in vocabularies present minor complications in feature level mapping of MGCP into NGIM, and severe difficulties for OSM into NGIM.
- OSM has high spatial resolution information but poor attribution.
- A combination of Alphanumeric and Geometric attributes conflation has the potential to improve the matching process.
- The cartographic scale of the datasets should be as close as possible.

7.0 ACKNOWLEDGEMENTS

This research work was supported by the DSTL Advanced Geospatial Information and intelligence Services (AGIS) research project under the terms of contract number DSTLX-1000091287.

8.0 REFERENCES

1. Gross, G.A., Nagi, R., Sambhoos, K., Schlegel, D.R., Shapiro, S.C., Tauer, G. (2012). Towards hard and soft data fusion: Processing architecture and implementation for the joint fusion and analysis of hard and soft intelligence data. FUSION 2012, 955-962.
2. Park, B., Johannson, A., Nicholson, D., (2013). Soft Data Fusion of Categorical Crowdsourced Data and its Application to Urban Situation Assessment. 3rd IMA Conference on Mathematics in Defence Proceedings papers, 6.
3. Wiemann, S., Bernard, L. (2010). Conflation Services within Spatial Data Infrastructures. In: 13th Agile International Conference on Geographic Information Science 2010. Guimarães, Portugal.
4. Longley, P.A, Goodchild, M.F., Maguire, D.J., Rhind, D.W., (2005). Geographic Information Systems and Science (Second edition). Chichester: Wiley, 560.
5. OGC, (2010). Fusion Standards Study, Phase 2 Engineering Report. OGC 10-184, 67.
6. Yuan, S., Tao, C., (1999). Development of Conflation Components. Proceedings of Geoinformatics, Ann Arbor. 1-13.
7. Stankutė, S., Asche, H., (2011). Improvement of Spatial Data Quality Using the Data Conflation. Computational Science and Its Applications - ICCSA 2011.
8. Hastings, J. T., (2008). Automated conflation of digital gazetteer data. International Journal of Geographical Infrmation Science 22(10): 1109-1127.
9. Song, G., Linna, L., Wenwen, L., Krzysztof J., Yue Z., (2014). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. Computers, Environment and Urban Systems.
10. Pourabdollah, A., Morley, J., Feldman, S., Jackson, M., (2013). Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network. ISPRS International Journal of Geo-Information (3): 704 - 728.

11. Tong, X., Liang, D., Jin, Y., (2014). A linear road object matching method for conflation based on optimization and logistic regression. International Journal of Geographical Infrmation Science (4): 824-846.

12. Li, L., Goodchild. M. F., (2011). An optimisation model for linear feature matching in geographical data conflation. International Journal of Image & Data Fusion 2(4): 309-328.

13. Zhang, M., Shi, W., Meng, L., (2005). A Generic Matching Algorithm for Line Networks of Different Resolutions. In: 8th ICA Workshop on Generalisation and Multiple Representation. A Coruna, Spain.

14. Freitas, S., Afonso, A.P., (2012). Distributed Vector based Spatial Data Conflation Services. In: XIII Brazilian Symposium on GeoInformatics. São Paulo, Brazil.

15. Envitia, (2010). GI2RA: Technical Summary and Recommendations. Envitia GI2RA-SRS006-03, 32.

16. OGC, (2013). OWS-9 CCI Conflation with Provenance Engineering Report. OGC 12-159, 43.

17. Hobona, G., Fairbairn, D., Hiden, H., James, P., (2010). Orchestration of Grid-enabled Geospatial Web Services in Geoscientific Workflows, IEEE Transactions on Automation Science and Engineering, 7 (2), 407 – 411.

18. Hobona, G., Jackson, M., Anand, S., (2011). Implementing Geospatial Web Services for Cloud Computing. In: Zhao P. and Di L., (ed.), Geospatial Web Services: Advances in Information Interoperability Information Science Reference, Hershey, New York. 287-308.