Using Conflation for Keeping Data Harmonized and Up-to-date

Dan Lee
Esri
Email: dlee@esri.com

At the previous workshop, important questions were brought up on conflation (Breakout session notes, 2014). This paper is intended to follow up on that subject and extend the discussion in a few key areas. The paper gives an overview of the necessity of conflation in improving data harmonization, introduces the conflation tools and workflows in ArcGIS, and presents examples of potential uses of the conflation capabilities for solving multi-source and multi-scale data integration and updating problems.

1. Conflation for data harmonization

It is an increasing demand and practice for NMAs and GIS organizations to build and maintain comprehensive, consistent, and reliable databases that can support wide range of spatial analysis and mapping needs. One of the challenges they face is data harmonization among overlapping datasets and between adjacent datasets at borders. Data discrepancies may occur spatially, as shown in Figure 1, as well as in attributes, therefore, cause difficulties and errors in spatial analysis and result in poor quality maps. Without effective tools it is very costly to fix the problems.

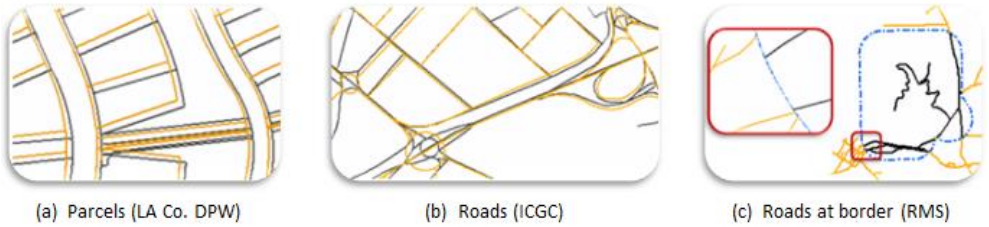

Figure 1. Spatial discrepancies in overlapping datasets (a) and (b) and in adjacent datasets along borders (c).

Conflation is the process of reconciliation of multi-source data for the best accuracy, completeness, and consistency. It involves matching of corresponding features and performing spatial adjustments and attribute transfer. Conflation is the solution for aligning legacy or outdated data with new or more accurate target data, unifying various data sources into one without redundancy, and resolving misconnections and other conflicts at borders. It can play an indispensable role in keeping data harmonized and up-to-date and optimizing data quality and usability.

Valuable efforts have been made to develop highly automated conflation tools, especially for popular linear features, such as hydrographic data (Stanislawski et al, 2002) and road data (Li and Liu 2012; Abdolmajidi et al. 2014).

2. Conflation tools and workflows

In ArcGIS Desktop 10.2.1, a new Conflation toolset was added under the Editing toolbox and contains five tools; another closely related new tool Detect Feature Changes was added to Data Comparison toolset under Data Management toolbox, as shown in Figure 2. Quick introductions of these tools are given below; some are designed for operating on overlapping datasets while others on adjacent datasets. More details can be found in the tool references in ArcGIS Desktop Help.

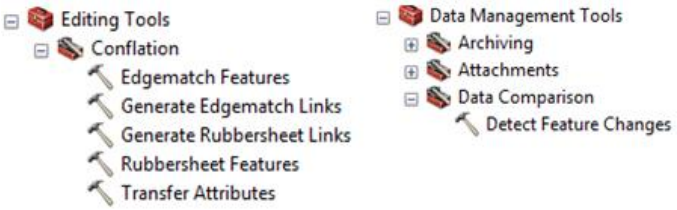

Figure 2. Conflation tools released in ArcGIS Desktop 10.2.1

2.1. Tools for overlapping datasets

Conflation of overlapping datasets is based on feature matching (FM). Some feature matching techniques were mentioned in a previous paper (Lee et al, 2014). The FM technique we choose to use is based on the fundamental analysis of the topological structures and feature pattern recognition (Yang et al, 2014). The three tools that are based on feature matching of overlapping datasets include: Generate Rubbersheet Links, Transfer Attributes, and Detect Feature Changes.

- Generate Rubbersheet Links (GRL) generates links between matched features and locations; these links are to be used by Rubbersheet Features (RF) tool to perform rubbersheeting spatial adjustment as illustrated in Figure 3.

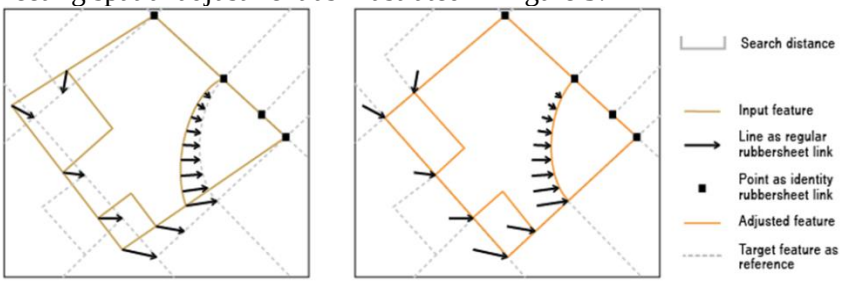

Figure 3: Illustrations of GRL (left) and RF (right) tools

- Transfer Attributes (TA) transfers specified fields from one dataset to another as illustrated in Figure 4.

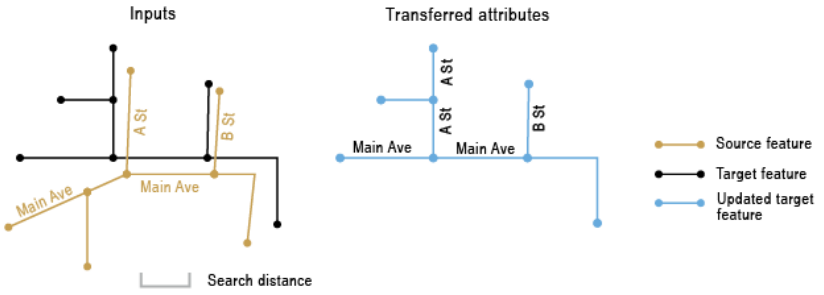

Figure 4. Illustration of TA tool

- Detect Feature Changes (DFC) identifies feature differences (spatial and attributes) between two datasets, typically the update features and the base features. It produces a feature class with a CHANGE_TYPE field, storing six possible values as illustrated in Figure 5:
  - Spatial (S) change: topology difference and shape deviation beyond the change tolerance
  - Attribute (A) change: attribute difference
  - Spatial and attribute (SA) change: spatial and attribute changes
  - No change (NC): 1:1 match without any spatial or attribute changes
  - New update feature (N): unmatched update features, usually new
  - To-be-Deleted base feature (D): unmatched base features, possibly to be deleted

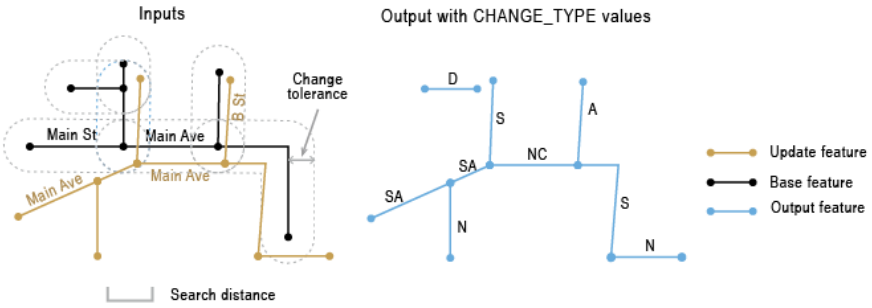

Figure 5: Illustration of DFC tool

Each of the three FM-based tools can be used independently to achieve a single goal, for example, you can run the TA tool to transfer some attributes from source to target features. Or they can be used in a workflow to accomplish more comprehensive goals, as explained later.

A match table can optionally be produced by each of the three FM-based tools. The match table stores full FM information including source and target FIDs, unique match group IDs, and match relationships, as shown in Figure 6. See the ArcGIS Help topic "About feature matching and the match table" for more explanations.

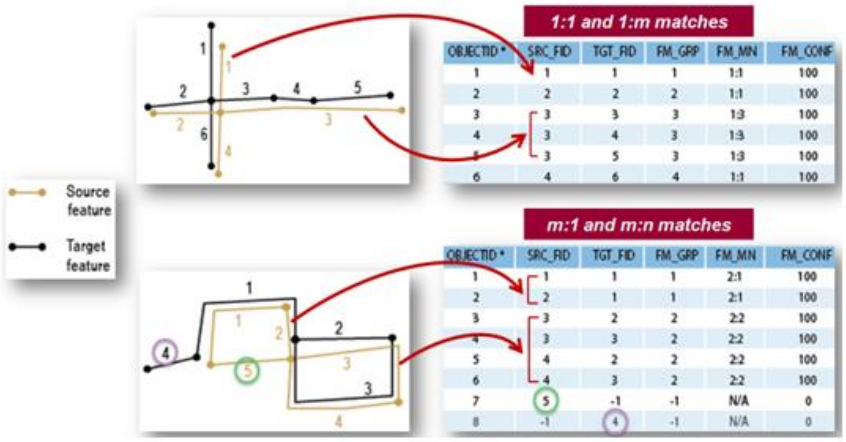

Figure 6: Match table – optional output by DFC, GRL, or TA tool

2.2. Tools for adjacent datasets

Edgematching (EM) involves generating links between corresponding features in adjacent datasets, followed by edgematching spatial adjustment to connect features.

- Generate Edgematch Links (GEL) tool finds corresponding features based on proximity, topology, continuity, and attributes (optional) analysis and generates links between matched features. The links carry source and target FIDs, as well as matching confidence values in the field EM_CONF. The more ambiguity the lower the EM_COMF value. The Edgematch Features (EF) tool performs edgematching spatial adjustment using the links, as shown in Figure 7. See the ArcGIS Help topic "About edgematching" for more information.

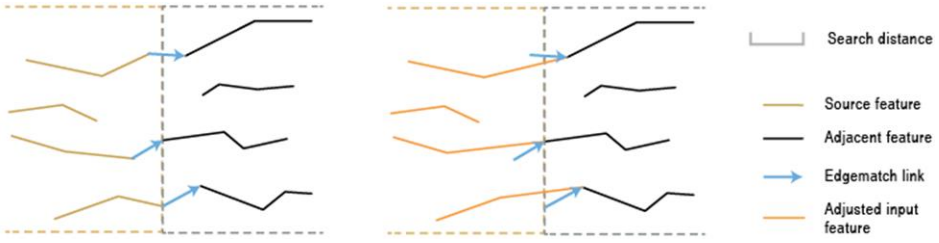

Figure 7: Illustrations of GEL (left) and EF (right) tools

2.3. Conflation workflows

In general there are three major components in conflation workflows: preprocessing, automated process and evaluation, and post-processing.

Preprocessing prepares the data to the best condition possible for analysis. The best practice includes validating geometry and topology, using consistent projection, selecting relevant features for processing, and so on. Details about preprocessing are not the focus of this paper.

Automated conflation and evaluations are done by a set of conflation workflow supplemental tools. In reality, the conflation tools may not produce perfect results due to the complexities of the data and the feature matching analysis. It is generally necessary to run a conflation tool, followed by evaluations to flag potential issues. This has led to the birth of a growing set of supplemental Conflation Workflow Tools, as shown in Figure 8.

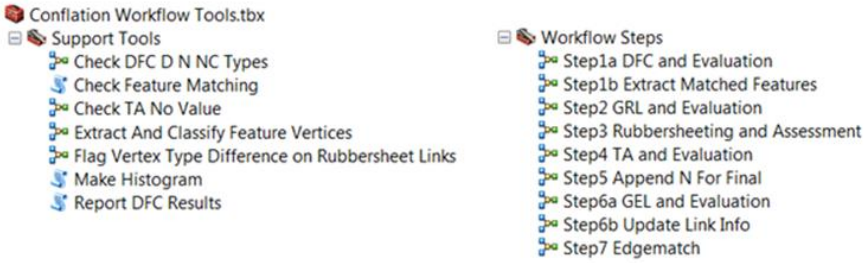

Figure 8. The supplemental Conflation Workflow tools.

- **Support Tools toolset** contains tools that analyze particular aspects of the data or the results of the conflation tools.
- **Workflow Steps toolset** contains tools that are more comprehensive and can be used in sequence (but not always necessary). These tools combine the conflation tools, the support tools, and additional geoprocessing tools to perform conflation and to derive information that facilitate post-processing.

Post-processing mainly involves interactively reviewing the potential issues flagged by the conflation and evaluation processes and making necessary edits and corrections. Features can be examined one by one or group by group according to unique flag values in a field and modified as needed. To make this process easier, a supplemental Conflation QA toolbar in the form of an Add-in is also provided to help reduce manual clicking, selecting, zooming to features, and updating field values so the work can be accomplished more efficiently.

3. Conflation scenarios

Conflation tools and the feature matching information they produce play essential roles in data harmonization. They can be used in a variety of scenarios involving multi-source base data, multi-theme data, and multi-scale data. The examples below show how conflation tools are used in: reconciliation of multi-source datasets, establishment of links for corresponding features in multi-scale databases, and change detection for updating.

3.1. Reconciliation of multi-source base data

The primary goal of conflation is to reconcile the discrepancies in multi-source datasets, overlapping datasets and adjacent datasets. Examples of both cases are given below.

Conflation of overlapping datasets

Conflation of overlapping datasets can be as simple as making spatial adjustment or attribute transfer from features of one dataset (source) to another (target) for better positional accuracy and attribute consistency. Or it can be as comprehensive as to unify features and information from multiple data sources for the best combined result.

Simple scenario: This use case requires spatial adjustment of one parcel dataset towards a more accurate parcel datasets, that is, the source (orange) and target (dark-grey) lines respectively (data from Dept. of Public Work, LA County, USA) in Figure 9, to improve positional accuracy and consistency. The Step1 through Step3 workflow tools were used along with necessary interactive quality improvement to address the flagged issues, see (b) in Figure 9. More than 4400 rubbersheet links were automatically generated, see (d) in Figure 9; the estimated matching accuracy was 93.84%. About 130 incorrect links were modified or deleted and 65 missed links were added. The source features were then adjusted, see (e) in Figure 9, using the improved links. See more details in the recent paper (Lee et al, 2014).

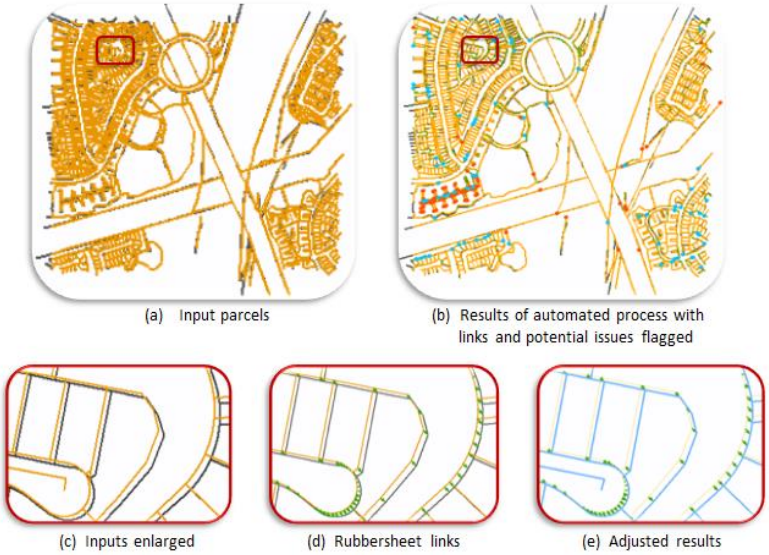

Figure 9. Spatial adjustment on parcel data.

Comprehensive scenario: This use case requires a unification of two road datasets A and B (data from Ohio Dept. of Transportation, USA) into a new dataset C. The process involves taking the more accurate features and attributes from dataset A, transferring attributes from matched features of B to A, and merging in features that only exist in dataset B. The resulting dataset C contains matched and unmatched features from both inputs with the positional accuracy of the better dataset and all the attributes. The Step1 through Step5 workflow tools were used along with necessary interactive quality improvement. Since the input data quality and similarity is quite good, the estimated FM accuracy was 98.74%. See more details in the recent paper (Lee, et al, 2014).

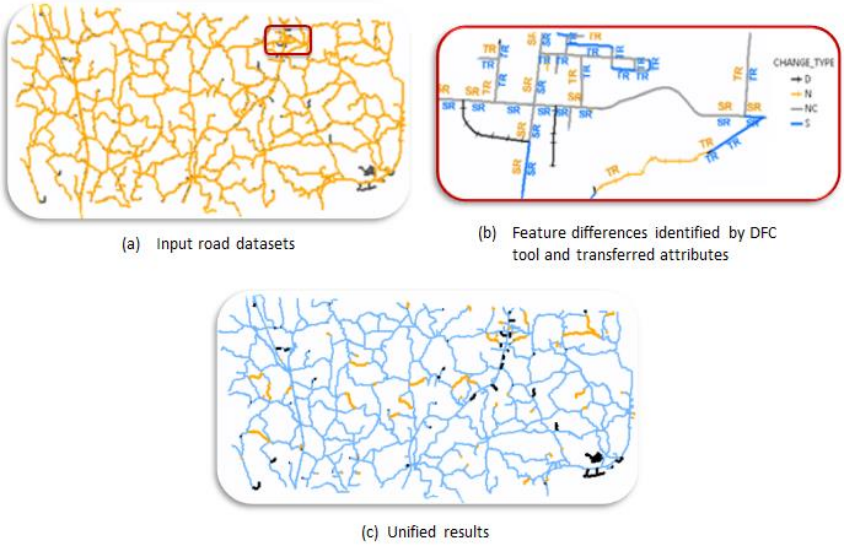

Figure 10. Unification of road data

Edgematching of adjacent datasets

Edgematching is the solution to cross-border data issues, as mentioned early. Edge refers to where adjacent data areas meet with or without an explicit boundary. Edgematching is the process of ensuring clean and correct continuation of adjacent datasets at their meeting edges. One of the use cases to be presented at the ICC 2015 (Lee et al, 2015) is the edgematching of European Location Framework (ELF) hydrographic data, between Norway and Sweden. The Step6a through Step7 workflow tools were used along with necessary interactive quality improvement. The top row of Figure 11 shows the inputs, the automatically generated edgematch links with their midpoints symbolized, and the only one flagged location of intersecting links. The bottom row shows two examples of edgematch adjustment result. The adjusted features are connected at border locations, marked by black circles, derived from the links. Hydrographic data tend to be less congested than road data; less ambiguity results in higher matching accuracy, estimated above 95% in this case, and mostly correct links.

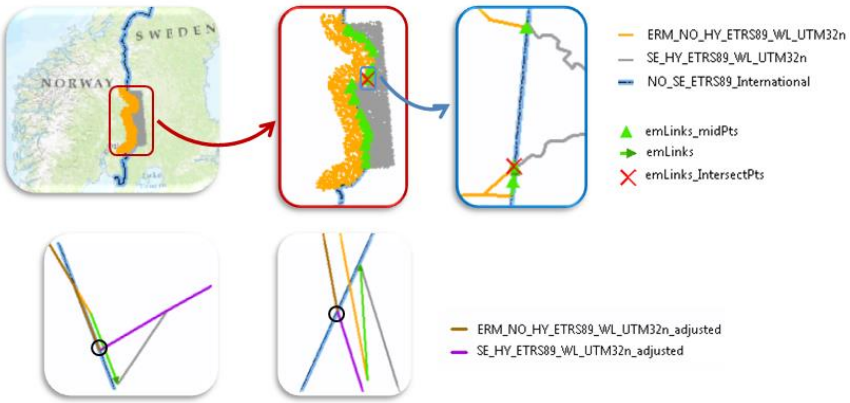

Figure 11. Edgematching of ELF hydrographic data.

3.2. Establishment of links for multi-scale databases

At the previous workshop the Institut Cartogràfic i Geològic de Catalunya (ICGC) and Esri jointly presented the first steps in the implementation of the ICGC MRDB for topographic data (Baella et al. 2014). The conflation tools were tested for performing the matching between features in two different scales: BT-5M and BT-25M. Among the three tested scenarios was the establishment of links between road centerlines of the BT-5M and the BT-25M and attribute transfer from BT-25M to BT-5M road centerlines. Without repeating the processing details, the transferred attributes (unique feature IDs from BT-25M) and the established link IDs (matched feature group IDs) are shown in Figure 12. The estimated matching accuracy was 95.6%.

Figure 12. Transferred feature IDs and the established link IDs.

3.3. Change detection and updating

Change detection is the first step in data updating, at the base data level or for multi-scale databases. Knowing where and what the changes are helps you assess how significant they are

and whether or not you need to proceed with updating. The example in Figure 13 illustrates street changes. The update features (orange) and base features (black) have spatial and attribute differences. Using the DFC tool introduced early the six types of changes are detected and written out, as shown in Figure 13.

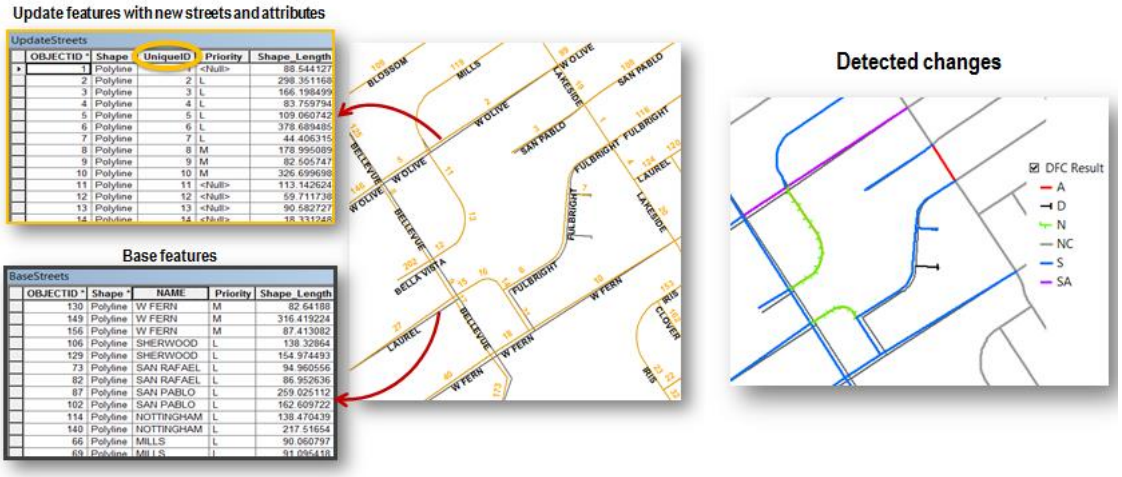

Figure 13. Example of change detection.

If the changes are at the base data level, the workflows for overlapping datasets presented above can be used to unify the changes into the base data. If the changes needed to be propagated to multi-scale databases, the established links presented above can be used to identify affected features across scales and the appropriate actions, such as generalization, should take place to modify or regenerate features for smaller scale databases.

4. Work and research in progress

There are still challenges to overcome in order to improve the conflation solutions. The main focuses of our planned future work are highlighted below; some questions are raised for discussion.

- **Enhancement of feature matching:** the key areas in feature matching are pattern recognition and similarity analysis. Incorrect matches often occur where feature situations are complicated and ambiguities are high. The examples include highway interchanges, roundabout with multiple ramps, complex parcel or structure shapes, and winding roads or rivers. Better recognition of these objects and conditions are needed to improve match accuracy and confidence level. From data structure side, the question to be raised for discussion is: What can be done to make data structure and feature classifications better facilitate feature matching?
- **Improvement of the workflows**: more work can be done mainly in two areas: a) further evaluate of conflation results to catch any missed errors and minimize false alarms; b) more efficient tools and integrated environment for the interactive quality improvement processes.
- **Tools and workflows for multi-theme data**: the current tools work on linear features; new tools and options are to be developed for point, polygon, or different feature types in context. Workflows for syncronizing spatially related feature in other themes after some of them are adjusted need to be laid out as well. The question to be raised for discussion is: What are the effective ways to describe and store the interrelations among theme features so the impact of conflation can be evaluated?
- **Research on hybrid approaches**: High resolution and highly accurate data sources such as GPS, imagery, and lidar data can provide the most up-to-date ground-truth information. Using these data sources may help guide conflation work, especially in the determination of accurate feature positions and in change detection. Feature extractions from imagery and lidar data still remain challenging and may require a

combination of automated and interactive processes to accomplish. We need to explore the potential of using the hybrid approaches in conflation workflows.

5. Conclusions

As multi-source geographic data are being frequently produced and easily obtained, it is important to keep data consistent and up-to-date within GIS and mapping organizations and beyond their ownerships and borders. The conflation tools and workflows presented in this paper combine highly automated feature matching, spatial adjustment, attribute transfer, and change detection processes with manageable interactive processes. The set of test results proves that the automatic matching accuracy of 85 – 95% or better is achievable; the final results are further improved through the interactive inspections and corrections.

It is inevitable that conflation can play a critical role in multi-source data integration, quality improvement, and harmonization, therefore make geospatial data more reliable and valuable. Working closely with NMAs, government agencies, and other interested users has allowed us to receive feedbacks and requirements. Further research and development are underway to bring more automation and efficiency to conflation solutions and to better support overlapping and adjacent database integration, maintenance, collaboration, and multi-scale mapping and analysis.

Reference

Abdolmajidi E, Will J, Harrie L, Mansourian A (2014) Comparison of matching methods of user generated and authoritative geographic data, The 17th ICA Generalization Workshop, 2014, Vienna, Austria

Baella B, Lee D, Lleopart A, Pla M (2014) ICGC MRDB for topographic data: first steps in the implementation, The 17th ICA Generalization Workshop, 2014, Vienna, Austria

Breakout session notes (2014) Conflation & Matching, The 17th ICA Generalization Workshop, 2014, Vienna, Austria, http://generalisation.icaci.org/images/files/workshop/workshop2014/slides/Conflation_Matching.pdf. Accessed on April 15, 2015

Lee D, Yang W, Ahmed N (2014) Conflation in Geoprocessing Framework - Case Studies, GEOProcessing, 2014, Barcelona, Spain

Lee D, Yang W, Ahmed N (2015) Improving Cross-border Data Reliability Through Edgematching, to be presented at The 27th International Cartographic Conference**,** 2015, Rio de Janeiro, Brazil

Li Y and Liu C (2012) Spatial approaches for conflating GIS roadway datasets, Sustainable Transportation Systems, 2012, pp. 290-298, http://dx.doi.org/10.1061/9780784412299.0035. Accessed on April 15, 2015

Stanislawski L, Nelson C, and Hamann M (2002) Automated Conflation of Reach Data for the National Hydrography Dataset", http://proceedings.esri.com/library/userconf/proc02/pap1207/p1207.htm. Accessed on April 15, 2015

Yang W, Lee D, and Ahmed N, "Pattern Based Feature Matching for Geospatial Data Conflation", GEOProcessing, 2014, Barcelona, Spain