Semantic Method of Conflation

James Ressler[1a], Eric Freese[1b], Veleria Boaten[1c]

[1] Northrop Grumman Corporation
1010 Market Street, Suite 1740, St. Louis, Missouri, 63101

[a]james.ressler@ngc.com, [b]eric.freese@ngc.com, [c]veleria.boaten@ngc.com

Abstract. The fusion of geospatial entities, commonly called feature conflation, requires aligning the type and characteristics of the entities. Similar to the problem of entity disambiguation in natural language parsing and metadata recognition, feature conflation involves identification of matching elements and properties in order to determine similarity among entities in terms of location and description. This paper presents a semi-automated semantic process for feature conflation that solves the type-matching problem using ontologies to determine similar feature types, and then uses business rules to automate the merge of geospatial features.

Keywords: ontology, semantic, conflation, geospatial, feature, business rule

1 Executive Summary

Once a cartographer's specialty, the task of computationally conflating various geospatial sources has been attempted for decades with limited success. As digital geospatial information became more prevalent than traditional cartographic means, the need for conflation likewise increased to reduce complexity, account for varying specialization of products, and enhance information content. While merging various geometric levels of abstraction and precision continues to be a challenge for cartographic research, the use of semantic technologies shows promise for automatically determining common feature types among complementary sets of data. The method developed in this paper illustrates the feasibility of creating feature type business rules for conflation that are inferred using an authoritative ontology. The application of semantic feature conflation creates opportunities to automate conflation and take advantage of an expanding web of geospatial information.

2 Introduction

Conflation is the process of merging similar geospatial representations of the same physical entity into a single form. This paper addresses the problem of conflation with a method to infer similar entities by using a feature type ontology, a method called

Semantic Feature Matching (SFM). The paper presents the topic in the following organization:

- Section 3 describes the problem of feature conflation.
- Section 4 describes the creation of a feature type vocabulary with an ontology
- Section 5 defines the application of semantic inference to develop business rules for conflation
- Section 6 illustrates of the feature conflation process using two maritime sources.
- Sections 7 & 8 describe related work and conclusions

3 Feature Conflation

3.1 Matching problem

A “feature” in geospatial information systems is an entity with reference to a location on the Earth that has both geometric representation and information attributes. For example, a line consisting of connected latitude and longitude points represents a road with additional information such as the name of the road, direction, speed limit, surface, etc. Features may represent real-world entities such as roads, rivers, lakes, and buildings, or man-made artifacts such as boundaries, counties, and waypoints. Features serve many purposes, from making maps, to overlaying information on images, to guiding electronic navigation systems (such as locating street addresses on a road), to depicting property ownership and tax evaluation. The goal of feature conflation is to merge geospatial information from multiple sources into a combined set of features that is superior in positional accuracy and attribution than the original sources. The origin of feature data is varied by the sources from which it is derived, such as aerial/satellite photography (generally called imagery), manually transcribed from maps, automatically digitized from various mapping products, or converted from other digital feature representations. These different representations of features drive the need for conflation in order to combine information about the same feature into a single representation. Other types of spatial data exist for specific purposes (raster imagery, surfaces, and coverage areas of elevation or temporal information). These types of geospatial data pose different challenges for conflation, but are not in the scope of feature conflation.

3.2 Types of features and representations

Features are described in terms of geometry and attribution. The geometry of a feature is a mathematical representation of the entity’s spatial extent, which is frequently a simplification of the feature’s location. The most basic representation of geometry is point, line and polygon. Variations of these basic geometry types are point elevations, curved lines, arcs, rings, rectangles, which are commonly called “vector” representations of feature data [8]. Figure 1 illustrates the three basic types of feature geometry: points (well), lines (rivers) and lake (polygon). In most cases, points and lines are a generalization of a more complex feature that is in reality a polygon. For example, a point generalizes a city region and a line generalizes a road. At different mapping levels of detail (scales), the appropriate level of generalization is adjusted. When conflation occurs between different levels of representation, the more accurate geometric feature may be retained while transferring attributes from the more general features. Hence, conflation also involves matching features between levels of generalization, such as matching the point for a building with a polygon.

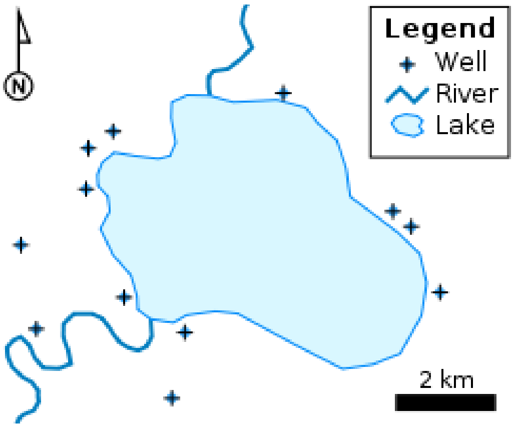


Figure 1: Feature Geometry[1]

Besides the geometric properties of a feature, a feature has a “type” metadata property to indicate what the feature represents (e.g. boundary, road, building). The taxonomy of feature types is a distinguishing characteristic of a feature dataset that is driven by the domain and purpose of the data. Just as the features on a state highway map differ from those on an agricultural chart of crop yields, the description of feature types varies from one domain to the next. A city map would only distinguish between a residential and industrial buildings while a municipal property and zoning map would distinguish between residential buildings (single-family dwelling, condominium, multi-floor apartment) type and businesses based upon the type of business (restaurant, grocery, general retail, specialty retail, automotive, etc.). Stemming from the variety in sources, domains and purposes of geospatial features, a wide variety of names exist for the same types of feature. A “bridge” in one source may be named an “overpass” or “viaduct” in another source. The variation in feature types is one of the principal challenges of conflation. For conflation to be automated, feature types must be correctly matched between different sources. Because many different types of features exist in close proximity to each other, a feature cannot be merged solely based upon its location or type of geometry. Without clear type classification, ridiculous combinations of features are possible such as conflating the shoreline of a lake with a roadway or boardwalk adjacent to the lake. The variation in type descriptions is a characteristic that has inhibited the automation of feature conflation.

Besides geometry and type, the information about a feature, called the attributes, are the other properties used in conflation. Feature attributes are either physical properties or metadata. Property attributes describe the characteristics of the feature, such as road surface material, communications tower height, and bridge load capacity.

---

[1] [1] http://commons.wikimedia.org/wiki/File:Simple_vector_map.svg (courtesy Open Geospatial Consortium)

Metadata attributes provide information about the abstract representation of the feature, such as the date it was last modified, the organization or person that derived the feature information, and the positional accuracy of the geometry. Combining the feature attributes in conflation also requires recognition of different names for the same attribute.

3.3 Conflation Process

Given sets of features, the process of feature conflation involves three overall stages: similarity, matching, and selection [2]. Additional pre- and post-processing of the feature data exist in various conflation algorithms for the purpose of transforming the data into a common format and spatial representation. The similarity stage of conflation sorts out the types, attributes, and geometry of the sources into like sets that are candidates to be matched. The matching stage pairs similar features that are within the geometric and attribute tolerance of conflation into candidates to be merged. The selection stage takes the matched feature and changes the geometry and attributes of the matched features to conform to the desired output. The logic to govern this process overall are referred to a business rules for conflation. The similarity stages uses business rules to pair like feature types and attributes. These rules are used in the matching and selection conflation processing.

3.4 Examples of conflation algorithms and software

The matching and selection stages of conflation are implemented in several commercially available conflation applications; the most prevalent are described here. The semantic feature matching complements these applications by providing a set of feature type business rules to govern the matching stage.

- ACS™ conflation with ArcGIS. ACS™, developed by Northrop Grumman, is a plug-in module to the ArcMap interface of ESRI ArcGIS software. ACS™ conflation is based upon matching feature codes and conflation of features based upon geometry and attribute values. There are many predefined mappings and priority attribute selections in ACS™ based upon the feature code, product type and NFDD/FACC attribute names. The ACS™ conflation process consists of several steps in completing multiple forms of information to encapsulate the business rules and settings for conflation.
- ConfleX – Citygate. An extension to ArcGIS, Conflex provides conflation of the various geometry types in a straightforward, highly automated workflow. “ConfleX is a MapObjects software-based conflation tool that uses artificial intelligent (AI) technology to automatically transfer attributes from a source to a target map.”[2]
- ESEA MapMerger another ESRI ArcGIS extension is favorable for simplicity of control for distance, attribution and merged results.

---

[2] http://gis.esri.com/partners/partners-user/index.cfm?fuseaction=product&BP_ID=257&PID=4986

- GeoMedia Fusion. Fusion is a component of the GeoMedia Professional software developed by Intergraph. Business rules in Fusion consist of multiple forms of information that work collectively to narrow the matching of features that are input to the selection of geometry and attributes to the output feature set. GeoMedia uses feature "connections" as input to the conflation process, each of which comes from an input layer (file, database, or feature service). When two features are matched through conflation, they are considered "linked." Fusion does not contain pre-defined business rules for common feature datasets, feature types and attributes. Fusion is flexible to define rules for any feature types and attributes that are loaded into GeoMedia.

In all of these products, the a-priori identification of feature and attributes to be matched are essential to determining matched results. Each conflation software uses different means of defining business rules. Until standards are developed for representing conflation business rules, a translation is required for each conflation software application to specify rules in a product-specific format.

4 Feature Type Vocabulary / Ontology

4.1 Need for a feature type vocabulary

As mentioned previously, a method is needed that allows features to be classified in such as way as to allow conflation between datasets from different sources. This classification scheme should, at a minimum, define a base set of features that can be identified. It can also define relationship between features as well as attributes of the features. These will allow for more precise conflation of feature data while preventing cases of improper conflation. For example, a ramp can be considered a part of a road. If the ramp and the road features were to be conflated, the risk of creating bad data is minimal. However, an entrance/exit can be considered a part of a building. If these two feature types were conflated, the risk of bad data is much higher. Therefore, the classifications of and the relationships between features become very important when considering options for conflation. These are represented and maintained by the concept of ontology initially described by Gruber [6] and promoted by recent semantic web initiatives.

4.2 Building a feature type ontology

Several candidate ontology methods/schemas were evaluated to determine which would be best suited for modeling GIS feature data. In [10], a survey of spatial ontology evaluated several upper and domain ontology that could be reused. For rule representations, SWRL, ruleML, and OWL taxonomies were evaluated for a method that would extend existing ontology and be supported by commonly available reasoners, such as Pellet. It was determined that the W3C Simple Knowledge Organization System (SKOS) [13] would provide the best set of semantics with which to model a feature ontology. While SKOS is typically used to model the broader and narrower relationships between concepts in thesaurus applications, these relationships also apply to geographic features. We incorporated the SKOS extensions which further refine the broader, narrower and related relationships by specifying partitive and instantitive relationships between features. SKOS also works well with the W3C Web Ontology Language (OWL) [4] which allows us to more accurately model the more specialized constraints between classes. The use of OWL and SKOS provides an industry standard method of modeling and communicating information contained with the ontology.

The National System for Geospatial-Intelligence (NSG) Application Schema (NAS)[3] was selected as the feature model upon which the master feature ontology is based. The NAS defines a set of over 500 features (or information entities) and classifies them based on utility and location. Each feature has a single label, a definition, a set of constraints and descriptive attributes, and relationships with other feature types. Application of the feature type ontology is described further in Section 5.

One of the weaknesses encountered in the NAS model is the definition of a single label for each feature. As mentioned previously, one of the stages of conflation is determining similarity. When data comes from different sources, one of the methods employed for determining similarity is the examination of labels and other textual information to provide hints as to the type of feature being examined. A single label is insufficient for determining the possible similarities between several features. What was needed was a database of synonyms from which similarities could be discovered based on the text within the feature data. We developed an ontology based on WordNet [8] which provides definitions, preferred and alternate labels for synonym sets (rather than individual words), as well as broader and narrower terms. This ontology was also encoded using SKOS.

The NAS ontology and the WordNet ontology were then connected using the *skos:related* property. This property is used to denote a relationship between two concepts without specifying any type of hierarchy between them. In some cases, more than one WordNet concept was associated with a single NAS feature. The relationship allows the conflation process to locate additional terms that might occur within the descriptions or labels of a feature, but might not be defined in the NAS. For example, the NAS defines a feature called "road." By connecting this feature to the appropriate "road" synonym set in WordNet, we now have additional labels such as "route" available. If we choose to allow the use of the broader and narrower relationships within WordNet, the number of options expanded greatly. For example, from "road" we can follow the hierarchy to narrower terms such as "highway" and on to "expressway" (aka freeway, motorway, throughway, etc.) and even "turnpike". International versions of WordNet could also be added to the ontology. This would provide the ability to conflate datasets that use different languages within the datasets. The availability of these extra terms increases the likelihood that some similarity will be found between features that have textual labels or descriptions within the data.

---

[3] The NAS model and schema are available from the Geospatial Intelligence Standards Working Group (GWG) at http://www.gwg.nga.mil

Additional processes have been developed which take partial words and compare them to known feature labels to determine the likelihood that they are describing a specific feature type.

The NSG defines Feature Data Dictionary (NFDD) classification codes, also called the Feature and Attribute Coding Catalog (FACC). These codes are another set of synonyms for the features. For example, "AP030" is the NFDD code for the feature type "road." Many older GIS systems used the NFDD codes to identify the types of features being described. The inclusion of the NFDD codes in this ontology provides a bridge from older data to newer data, allowing conflation to take place on an even wider set of data.

5 Semantic Feature Type Matching

The algorithm used for semantic feature type matching follows a three step processes depicted in Figure 2. In summary, the steps of the process are: 1) feature ingest; 2) concept mapping; and 3) feature rule inference. These steps are elaborated in the following sections.

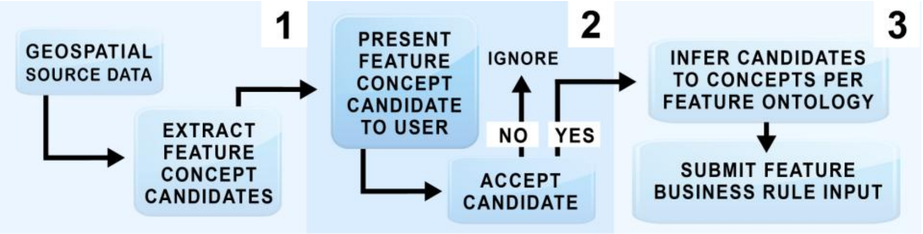


Figure 2 – The Semantic Type Matching process flow is divided into three sub-processes.

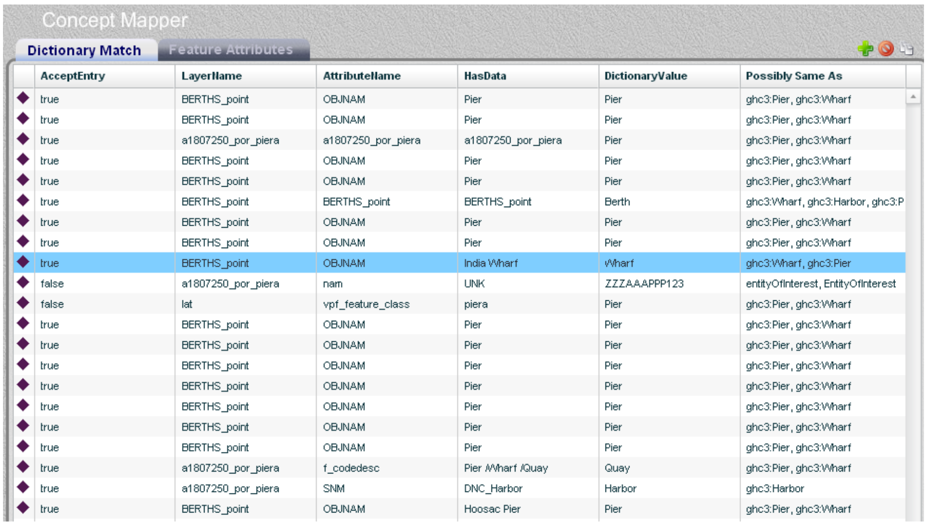
5.1 Feature Ingest

The process flow is started by ingesting various geospatial data sources to extract feature concept candidates based solely on content. Within Sub-process 1 of Figure 2, this is achieved through ingesting source data into a common ontology and filtering concepts through SPARQL Protocol and RDF Query Language (SPARQL) [14] queries. Through the use of SKOS, the team developed a Resource Description Framework (RDF) vocabulary around various geographic feature themes. The vocabulary is designed assist in determining feature information for geographic source data. Through the use of SKOS, the source data elements which are conflation candidates can be totally independent for the conflation workflow. The process asserts the *skos:narrower* relationship between the feature instances within the vocabulary and conflation candidates extracted from data files that were converted to RDF. This helps save time for the subject matter experts in examining unknown data source formats for candidate feature types.

To start the matching process, the user submits feature data to identify conflation candidates. Through a user interface, a user ingests various feature sources, such as XML files using a Geographic Markup Language (GML) schema, Microsoft Excel spreadsheets, ESRI Shape DBF files, and tab-delimited text files. Third party products convert other geospatial sources to one of these formats. Once the source file is uploaded, the source data collection goes through a series of transformations to extract source data, layer information, file name information, and any other textual information. The process then uses the extracted information and compares content against the established vocabulary, as described in the next section. The conflation candidates are passed in a Notation3 (N3) [11] file that contains data that may match the feature vocabulary as well as other unknown feature attributes that may be relevant, but did not match any feature types within the feature ontology.

5.2 Concept Mapping

The NAS-based ontology is the basis for the taxonomy used to assign feature type identification of conflation candidates. In Sub-process 2 of Figure 2, the feature type candidates are asserted as the NAS concepts are set to the entity types extracted from Sub-process 1 based upon the feature concept associated with a dictionary value. The matches are presented as candidate feature concepts to the user as shown in the "HasData" column of Figure 3.



	AcceptEntry	LayerName	AttributeName	HasData	DictionaryValue	Possibly Same As
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	a1807250_por_piera	a1807250_por_piera	a1807250_por_piera	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	BERTHS_point	BERTHS_point	Berth	ghc3:Wharf, ghc3:Harbor, ghc3:P
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	India Wharf	Wharf	ghc3:Wharf, ghc3:Pier
◆	false	a1807250_por_piera	nam	UNK	ZZZAAAPPP123	entityOfInterest, EntityOfInterest
◆	false	lat	vpf_feature_class	piera	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	BERTHS_point	OBJNAM	Pier	Pier	ghc3:Pier, ghc3:Wharf
◆	true	a1807250_por_piera	f_codedesc	Pier /Wharf /Quay	Quay	ghc3:Pier, ghc3:Wharf
◆	true	a1807250_por_piera	SNM	DNC_Harbor	Harbor	ghc3:Harbor
◆	true	BERTHS_point	OBJNAM	Hoosac Pier	Pier	ghc3:Pier, ghc3:Wharf

Figure 3: Semantic Feature Matching Client – mapped concepts

In the line highlighted, the attributed OBJNAM with value "IndiaWharf" matched the feature dictionary item "Wharf" as a candidate feature concept. Valid feature concepts are given along with "unknown" values (tagged as "ZZZAAAPPP123"). Unrecognizable values such as named places, feature identifiers, dates, organizations, codes, and abbreviations, are omitted from the candidates.

A subject matter expert (SME) on geospatial features is able to select the feature types recommended from the conflation candidates. When concepts are inferred by the ontology, the proposed concepts require minimum oversight by a SME. This step provides the SME the ability to examine the semantic identification of conflation candidates instead of manually examining each data source. The typical process to conflate two different sources requires a SME to review datasets in native format to find common features. By using semantic inference and data transformations, candidates are identified according to a common vocabulary for the SME using the method described in section 5.3. When the vocabulary is well established, a less-than-expert conflation user will be able to re-use business rules based upon the accepted ontology, which increases the degree of automation.

The user can accept or reject the feature concept candidates that were extracted from the source data, as indicated by "AcceptEntry" column of Figure 3 where "False" ignores the candidate. After the initial selection of feature type, the user is given the opportunity to accept or reject identified candidates. For the unknown terms, the user may also manually map these terms to concepts within the feature ontology.

5.3 Feature Rule Inference

Once the candidates are evaluated, the accepted types are passed along to Sub-process 3 of Figure 2. The feature candidates are asserted by a *skos:narrower* relationship to a broader feature concept. The inference of common concepts based on source information and common feature type vocabulary relies upon the source-specific knowledge to develop conflation business rules. Using concepts and definitions defined by NAS that incorporate the semantic relationships from SKOS, the conflation business rules are proposed in the form of a feature concept list. The inferred feature types are indicated in the "Possibly Same As" column of Figure 3.

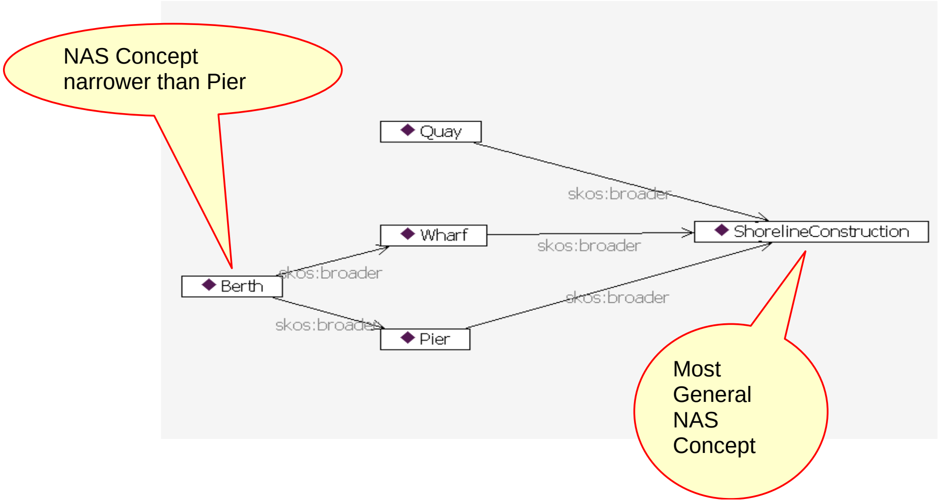


Figure 4: SKOS Relations of Maritime Features

Populated with feature concepts, a SPARQL Query is executed with the knowledge base to generate the feature type similarity rule result set. The SPARQL query is as follows:

```
SELECT DISTINCT ?ClassName ?DataInstance ?FeatSimName
?GeneralConcept ?DerivedFrom
WHERE {
   ?DataInstance :dictionaryValue ?DerivedFrom .
   ?DataInstance :hasData ?FeatSimName .
   ?conceptParent rdfs:label ?GeneralConcept .
   ?DataInstance skos:broader ?conceptParent .
   ?DataInstance rdf:type ?ClassName .
   ?DataInstance rdf:type :dictionaryEntry .
   ?DataInstance rdf:type skos:Concept .
FILTER (?ClassName = :dictionaryEntry) }
```

A subset of the vocabulary for maritime features is shown in Figure 4. Based upon this ontology, the resultant XML from this query for a maritime feature data set is used for illustration. In the SPARQL query, the FeatSimName is the common dictionary value used to match the feature data. In the example, the names “Pier”, “Berth”, “a1807250_por_piera” and “Pier/Wharf/Quay” have the common derived concept of Pier. Other similar concepts are associated with the same query. A sample of the SPARQL query results for four feature types is listed below (the feature names are underlined for emphasis). The SPARQL results are transformed into a feature similarity business rule format using XSLT to create business rules containing groups of FeatSimName from <result> that have common DerivedFrom concepts as illustrated in section 6.

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="ClassName"/>
    <variable name="DataInstance"/>
    <variable name="FeatSimName"/>
    <variable name="GeneralConcept"/>
    <variable name="DerivedFrom"/>
  </head>

 <results>
   <result>
       <binding name="ClassName">
<uri>http://www.ngc.com/FoundFeatures#dictionaryEntry
</uri></binding>
       <binding name="DataInstance">
<uri>http://www.ngc.com/FoundFeatures#dictionaryEntry_T
est21</uri></binding>
       <binding name="FeatSimName">
```

```
<literal>Pier</literal></binding>
       <binding name="GeneralConcept">
<literal>Pier</literal></binding>
      <binding name="DerivedFrom">
<literal>Pier</literal></binding>
    </result>
    <result>
…
       <binding name="FeatSimName">
<literal>BERTHS point</literal> </binding>
…      <binding name="DerivedFrom">
<literal>Pier</literal></binding>
   </result>
    <result>
…
       <binding name="FeatSimName">
<literal>a1807250 por piera</literal> </binding>
…     <binding name="DerivedFrom">
<literal>Pier</literal></binding>
    </result>
</sparql>
```

6 Automated Conflation Illustration

One example of conflating different representations of the same feature types used the Electronic Nautical Chart (ENC) and Digital Nautical Chart (DNC™) to merge attribution from the ENC to same features in the DNC™. As described previously, Figure 3 depicts the feature concepts mapped for the Berth and Pier feature types of an ENC and DNC™ product using the processes described in Section 5.

One resultant business rule for a Berth and Pier illustration taken from the XML excerpt in Section 5.3 is listed below.

```
<featGrp id="G-Pier">
    <group-name>Pier</group-name>
    <about/>
    <feats>
        <featRef ref="F-a1807250_por_piera"/>
        <featRef ref="F-BERTHS_point"/>
        <featRef ref="F-Hoosac_Pier"/>
        <featRef ref="F-Pier_"/>
        <featRef ref="F-Pier_Wharf_Quay"/>
        <featRef ref="F-piera"/>
    </feats>
</featGrp>
```

The DNC™ and ENC (S-57) feature data shown in Figure 5 was loaded into ArcGIS. from an ESRI geodatabase or a shapefile format; future enhancements will acquire feature data using web service interfaces.

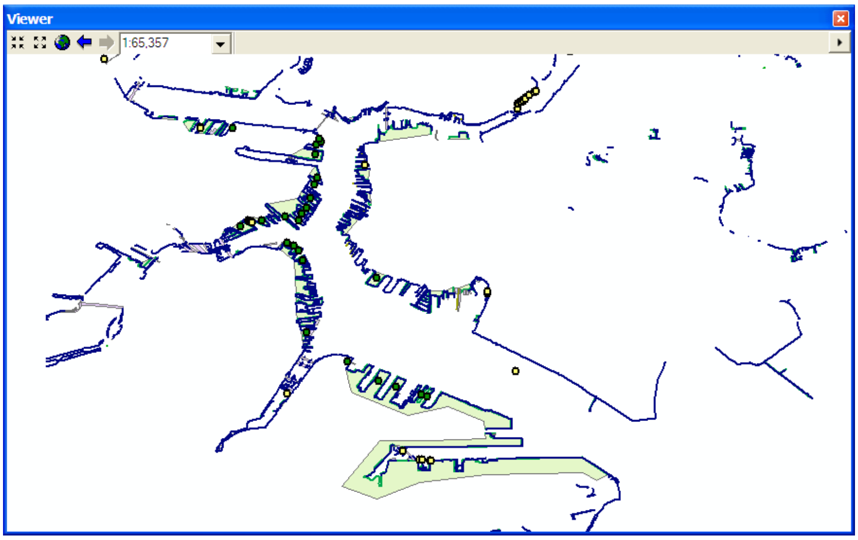
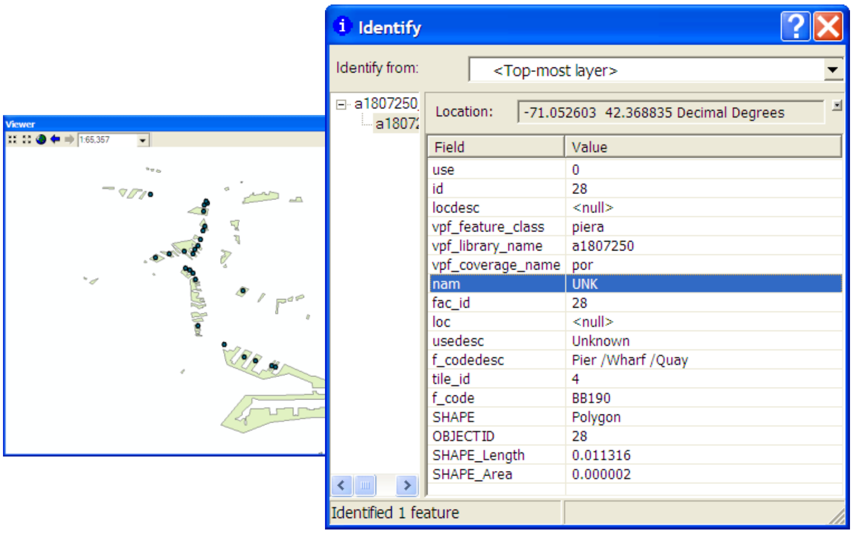


Figure 5: DNC™ and ENC Feature Data in ArcGIS

Using the generated business rules, the conflation execution was performed by the ACS™ conflation application. The conflation algorithm matches the features using the generated business rules and matches the attributes according to additional business rules to produce a new layer containing the results. The merged ENC berth and DNC™ pier features are shown in Figure 6.



Identify from: <Top-most layer>

Location: -71.052603 42.368835 Decimal Degrees

Field	Value
use	0
id	28
locdesc	<null>
vpf_feature_class	piera
vpf_library_name	a1807250
vpf_coverage_name	por
nam	UNK
fac_id	28
loc	<null>
usedesc	Unknown
f_codedesc	Pier /Wharf /Quay
tile_id	4
f_code	BB190
SHAPE	Polygon
OBJECTID	28
SHAPE_Length	0.011316
SHAPE_Area	0.000002

Identified 1 feature

Figure 6: DNC™ and ENC Conflation Results (one feature selected)

In this case, several ENC point berths had the attribution merged with the DNC™ area pier features. The resultant set retained the DNC™ Pier geometry, but included the ENC attribute values, including the name (NAM) of all berths located on each matched pier. As dictated by the user's needs, the merged data set could be saved as another layer or published as an updated product.

7 Related Work

Volz [11] published a paper on matching geospatial schemas based upon feature data and location. Volz's method creates Multi-representational (MRep) Relations between matched features with an associated value. The paper mentions creating ontology mapping to semantically transform related features in a related work, but doesn't explicitly use the ontology in deriving the MRep relations. In Volz's approach, the schema correlations are implied or determined by human decision. In the SFM method, the correlation between feature types is based upon ontologies that have relations between classes.

A preceding work by Fonseca [5] establishes the utility of ontological representation of real-world entities in a GIS in an Ontology-Driven GIS (ODGIS). This work states "our proposal is to use ontologies to match the features found in the images to classes in the ontologies." The application of ontologies in this work used a community accepted high level ontology that was related to ontologies for a particular domain, task and application. These inter-related ontologies were used to classify geospatial features derived from imagery into common objects. The goal was to integrate the geospatial entities into a hierarchical representation where entities are related to common classes. While the approach is similar to SFM in concept, the end result of ODGIS does not create a conflated set of features, but a classification of features into a predefined hierarchy.

Cobb's paper in 1998 [3] provided a foundation for developing conflation applications. Her work established principles of conflation of various geometric combinations of nodes and lines. The feature type matching in this work relied upon the definitive vocabulary at the time, FACC, and similarity of FACC codes. Products that did not adhere to FACC were incompatible for conflation without an explicit type conversion. Having definitive feature type ontology, such as based upon the NAS, Wordnet and FACC, overcomes this limitation while enabling use of different conflation applications' geometric feature matching and selection algorithms.

The approach of semantic feature mapping presented here has implications in a broader field of entity disambiguation [7, 1] and multi-source identity resolution [15] that applies to more than geospatial conflation -- to create associations between other complementary data sources for knowledge discovery and fusion. Zong [16] developed a minimal rule-based algorithm to disambiguate geographic place names on web pages with a graph-based gazetteer of named locations. Named locations are one type of feature that allows a specific matching algorithm, though it is unsuitable for conflation of all feature types. The development of semantic conflation for geospatial data can be assisted by entity disambiguation of natural language and possibly contribute new methods.

8 Future Work and Conclusions

The problem of feature type matching is one aspect of conflation, a process that requires additional business rules for automating the matching and selection step of conflation. Once feature type matching is established, other rules are necessary to determine the common attributes between sources. For all common attributes between two sources to be conflated, a business rule is required to reconcile differences between the attribute name, measurement, language, enumeration or other characteristic. For example, if a tower feature in one dataset is to match a smokestack in another set, the height of the tower and smokestack is an important comparison to determine a feature match (in addition to location). If the tower source has a HEIGHT attribute while the smokestack has an ELEVATION attribute, a business rule to match these attributes is needed. Further, knowing the units of each attribute and the measurement of height (relative to ground surface versus distance above sea level) are needed for a valid comparison of tower and smokestack heights. In general, the determination of attribute matching business rules is a further area of research that can utilize the semantics of features.

This method can benefit conflation research that utilizes a wider variety of complementary sources beyond geospatial products. When geo-located by address, street intersection, or named location, non-structured sources contain relevant information that can be used for conflation. The Terra World[4] project at USC has developed frameworks that integrate traditional structured data (such as relational databases), semi-structured data (such as XML files), and spatio-temporal data (such as bus schedules and road networks) so that it can be queried to produce results that would be otherwise difficult or impossible to acquire through other means. Even as a multitude of unstructured geospatial data is available from public sources, structured web services are becoming more common (such as KML feeds for GoogleEarth, Web Feature Services and ArcGIS services) within open and closed geospatial systems

Discovery and ingest of an expanded world of geospatial content is a useful application of conflation that takes advantage of intelligent automation in a semantic web. Rather than maintaining mappings between every conceivable feature type and code, applying semantic information contained within the feature data with an established base of knowledge enables algorithms to infer the relationships between geospatial entities. These relationships are queried to build feature type rules that fit the data to be conflated. As demonstrated by this approach, these rules aid the

---

[4] http://infolab.usc.edu/projects/terraworld/index.jsp

automation of feature conflation and contribute towards developing other applications of data fusion.

References

1. Brauer, F., Loser, A. Do, H., “ Mapping Enterprise Entities to Text Segments,” *PIKM’08*, October 30, 2008, ACM 978-1-60558-257-3/08/10.
2. Brennan, P., and Evans, C., OGC® OWS-5 Conflation Engineering Report, Open Geospatial Consortium report OGC 07-160 r1, 2008-04-18.
3. Cobb, M. et. al., “A Rule-based Approach for the Conflation of Attributed Vector Data,” *GeoInformatica* 2:1, 7-35 (1998).
4. Dean M. and Schreiber G., eds. “OWL Web Ontology Language Reference,” W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-ref-20040210/
5. Fonseca, F., Egenhofer, M. and Aguoris, P., “Using Ontologies for Integrated Geographic Information Systems,” *Transactions in GIS*, 6(3), 2002.
6. Gruber, T., “Toward Principles for the Design of Ontologies Used for Knowledge Sharing,” International Journal Human-Computer Studies 43, p.907-928, 1993.
7. Hassel, J., Aleman-Meza, B., and Arpinar, B., “Ontology-Driven Automatic Entity Disambiguation in Unstructured Text,” Computer Science Department, University of Georgia, 2006.
8. Miller, G., “WordNet – a lexical database for the English language,” Princeton University Cognitive Science Laboratory, http://wordnet.princeton.edu/.
9. Open Geospatial Consortium (OGC©), Geographic Markup Language (GML) Encoding Standard, version 3.2.1, OGC 07-036, 2007-08-27.
10. Ressler, J., Dean, M. and Kolas, D., “Geospatial Ontology Trade Study”, paper for The National Center for Ontological Research (NCOR) “Ontology for the Intelligence Community" conference, 2007, Columbia, Maryland.
11. Volz, S, “Data-Driven Matching of Geospatial Schemas,” A.G. Cohn and D.M. Mark (Eds.): COSIT 2005, LNCS 3693, pp. 115 – 132, 2005.
12. W3C, “Notation3 (N3): A readable RDF syntax” W3C Team Submission 14 January 2008, http://www.w3.org/TeamSubmission/n3/
13. W3C, ”Simple Knowledge Organization System (SKOS)“, http://www.w3.org/2004/02/skos/
14. W3C, “SPARQL Query Language for RDF,” W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/.
15. Yankova, M., Saggion, H., and Cunningham, H., “Adopting Ontologies for Multisource Identity Resolution,” *OBI ’08*, October 2008, ACM 978-1-60558-219-1/10/08.
16. Zong, W., Wu, D., Sun, A., Lim, E., “On Assigning Place Names to Geography Related Web Pages,” *JCDL’05*, June 7–11, 2005, ACM 1581138768/05/0006.