SPATIAL ARRANGEMENT AS A PART OF GEOSPATIAL FEATURE ONTOLOGIES

by

Jonathan R. Clark
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

__________________________________________ Dr. Anthony Stefanidis,
Dissertation Director
__________________________________________ Dr. Peggy Agouris,
Committee Member
__________________________________________ Dr. Kevin Curtin,
Committee Member
__________________________________________ Dr. Andrew Crooks,
Committee Member
__________________________________________ Dr. Peggy Agouris,
Department Chairperson
__________________________________________ Dr. Timothy L. Born,
Associate Dean for Student and Academic Affairs,
College of Science
__________________________________________ Dr. Vikas Chandhoke,
Dean, College of Science

Date: __________________________ Fall Semester 2012
George Mason University
Fairfax, Virginia
Spatial Arrangement as a Part of Geospatial Feature Ontologies

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Jonathan R. Clark
Master of Science
Mississippi State University, 1980
Bachelor of Science
Cornell University, 1973

Director: Dr. Anthony Stefanidis, Professor
Department of Geography and GeoInformation Science

Fall Semester 2012
George Mason University
Fairfax, Virginia
DEDICATION

This work is dedicated to Jonathan and Sarah. I hope that they have seen that learning is not simply a series of classes and tests, and that you are never too old to ask questions.

Never stop asking, never stop exploring. Learning is a lifelong pursuit.

For Carol, I reserve my deepest thanks. You listened when I needed to talk, ensured that I had the time and environment to do my studies, served as my sounding board when I needed to explore options, and helped me keep my eyes on the prize so I could finish this journey. This dissertation could never have been completed as a solo project and I consider this work to be as much your accomplishment as mine. Thank you.
ACKNOWLEDGEMENTS

As my dissertation director, Dr. Stefanidis, says: “...a dissertation is a team sport.” It is fitting, therefore, to recognize and thank each member of the team who helped in my work.

I’d like to thank Chuck Hayes and Rachana Ravi of Digital Globe, Inc. for their support in providing satellite imagery for the study. I would also like to recognize Dr. Caixia Wang, Ms. Xu Lu, and Mr. George Panteras for their efforts in data collection, data analysis, and for the many technical peer reviews.

My thanks go to Dr. Andrew Crooks for ensuring that my work was grounded in the experience and work of others, and helping me to see that geospatial analysis extends well beyond cartography into such fields as Agent-based Modeling. Thank you to Dr. Kevin Curtin for keeping me on-course in the scientific method, for reminding me at every turn that I needed to understand and apply the science of geography, and for pushing me to think about what my work will do to advance that science. To Dr. Peggy Agouris go my thanks for seeing places in my numerical analysis which would show me things in the data I had not thought of, and for establishing a professional and progressive atmosphere for research in the GGS Department.

My sincere gratitude to Dr. Anthony Stefanidis for being my chief mentor, an ever-patient listener, and an advisor and guide on each step of this road. He constantly reminded me that the goal of research was not necessarily to reject the null hypothesis, but to understand why you did, or did not reject it. This is how we learn.

Thanks go to Ms. Teri Fede for guiding me through the graduate processes within the College of Science, and helping to ensure that my program stayed on track and on schedule.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Abstract</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2: Literature Review</td>
<td>13</td>
</tr>
<tr>
<td>2.1. Overview</td>
<td>13</td>
</tr>
<tr>
<td>2.2. Semantics</td>
<td>13</td>
</tr>
<tr>
<td>2.3. Knowledge Representation</td>
<td>18</td>
</tr>
<tr>
<td>2.4. Ontologies</td>
<td>19</td>
</tr>
<tr>
<td>2.5. Space and Context</td>
<td>36</td>
</tr>
<tr>
<td>2.6. Similarity</td>
<td>46</td>
</tr>
<tr>
<td>2.7. What Gap Areas Will This Research Address and Fill?</td>
<td>58</td>
</tr>
<tr>
<td>CHAPTER 3: Methods</td>
<td>59</td>
</tr>
<tr>
<td>3.1. The Approach</td>
<td>59</td>
</tr>
<tr>
<td>3.2. A Workflow</td>
<td>60</td>
</tr>
<tr>
<td>3.3. Geospatial Information</td>
<td>61</td>
</tr>
<tr>
<td>3.4. Feature Ontology</td>
<td>73</td>
</tr>
<tr>
<td>3.5. Numerical Analysis</td>
<td>83</td>
</tr>
<tr>
<td>3.6. Similarity</td>
<td>87</td>
</tr>
<tr>
<td>CHAPTER 4: Results</td>
<td>95</td>
</tr>
<tr>
<td>4.1. The Metrics and Ontology</td>
<td>95</td>
</tr>
<tr>
<td>4.2. Importance of Each Metric</td>
<td>103</td>
</tr>
<tr>
<td>4.3. Statistical Examination of the Metrics</td>
<td>104</td>
</tr>
<tr>
<td>4.4. The Final Ontology and Metrics</td>
<td>110</td>
</tr>
<tr>
<td>4.5. Issues with the Ontology</td>
<td>111</td>
</tr>
<tr>
<td>4.6. Assessing Similarity</td>
<td>117</td>
</tr>
<tr>
<td>CHAPTER 5. Conclusions and Future Research</td>
<td>130</td>
</tr>
<tr>
<td>Appendices</td>
<td>136</td>
</tr>
<tr>
<td>Appendix A: Ontology of Container Terminals</td>
<td>137</td>
</tr>
<tr>
<td>Appendix B: Study Sites</td>
<td>150</td>
</tr>
<tr>
<td>Appendix C: Metrics Definitions</td>
<td>168</td>
</tr>
<tr>
<td>Appendix D: Glossary of Terms</td>
<td>173</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>176</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>188</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1: List of training and test sites for the study.</td>
<td>66</td>
</tr>
<tr>
<td>Table 2: List of initial metrics.</td>
<td>80</td>
</tr>
<tr>
<td>Table 3: Metrics collected from the 22 sampled training sites.</td>
<td>97</td>
</tr>
<tr>
<td>Table 4: Descriptive statistics for the metrics from the 22 sampled training sites.</td>
<td>98</td>
</tr>
<tr>
<td>Table 5: Descriptive statistics for the original 22 metrics.</td>
<td>107</td>
</tr>
<tr>
<td>Table 6: List of training and test sites.</td>
<td>152</td>
</tr>
<tr>
<td>Table 7: Metrics used in the study, along with their functional basis.</td>
<td>169</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1: An ontology of the topics in the Literature Review chapter. ........................................14
Figure 2: The “Semiotic Triangle”. ..................................................................................16
Figure 3: Ontologies by strength. .......................................................................................21
Figure 4: Evolution in how features are conceptualized ...................................................35
Figure 5: Progression of select research efforts ...................................................................40
Figure 6: Facial recognition using simple graphs .................................................................45
Figure 7: Detection of human forms in imagery .................................................................45
Figure 8: Types of graph matching .....................................................................................53
Figure 9: The study effort shown as a general workflow ...................................................61
Figure 10: Commercial container terminal at Port Constanta, Romania. .........................63
Figure 11: An airport as an example of a complex feature type .........................................64
Figure 12: Study site locations .........................................................................................67
Figure 13: Container terminal ontology shown as a graph .................................................82
Figure 14: Results of sensitivity analysis ..........................................................................84
Figure 15: Two container terminals with different physical layouts .................................89
Figure 16: Approach used for comparing the ontologies .....................................................92
Figure 17: Similarity is expressed as accumulated Z values .............................................93
Figure 18: Tornado charts showing the effect of normalization .......................................100
Figure 19: Effect of weighting each metric .......................................................................102
Figure 20: Results of initial sensitivity analysis ...............................................................103
Figure 21: Probability distribution plots for 2 example metrics .......................................106
Figure 22: A portion of the initial correlation results for all 22 metrics .........................109
Figure 23: Multiple Regression results .............................................................................109
Figure 24: Final ontology shown in tabular and graph forms ...........................................111
Figure 25: Container Stuffing Facility (CSF) ..................................................................112
Figure 26: Geometry for the components of the Pasir Panjag terminal ..........................116
Figure 27: Final similarity scores for the four test sites ....................................................118
Figure 28: Vehicle in/out processing building at the Los Angeles test site ......................120
Figure 29: Graphic plots and statistics for synthesized test sites ......................................123
Figure 30: Probability plots for similarity scores .............................................................123
Figure 31: Similarity of test sites under different scoring methods ..................................127
Figure 32: Aerial view of a portion of a large container terminal .....................................140
Figure 33: Container Terminal at the Port of Long Beach, California ............................141
Figure 34: The Shingamman and Gamann container terminals ........................................ 142
Figure 35: A portion of a storage yard ............................................................. 144
Figure 36: View of the loading apron at the Colombo, Sri Lanka port ...................... 145
Figure 37: Cargo transfer Area (CTA) in Hamburg, Germany .................................. 146
Figure 38: A Container Stuffing Facility (CSF) in Gdansk, Poland .......................... 147
Figure 39: Example of a container terminal ...................................................... 148
Figure 40: Pier-based container operations at Puerto Barrios, Guatemala ................. 149
Figure 41: Map of study site locations ............................................................... 150
Figure 42: Training site at Busan, South Korea .................................................. 153
Figure 43: Training site at Cape Town, South Africa .......................................... 154
Figure 44: Training site at Durban, South Africa ............................................. 154
Figure 45: Training site at Fremantle, Australia .............................................. 155
Figure 46: Training site at Hamburg, Germany .................................................. 155
Figure 47: Training site at Jawaharlal, India .................................................... 156
Figure 48: Training site at Long Beach, USA .................................................. 156
Figure 49: Training site at Los Angeles, USA .................................................... 157
Figure 50: Training site at Luanda, Angola ....................................................... 157
Figure 51: Training site at Napoli, Italy ............................................................ 158
Figure 52: Training site at New York, USA ...................................................... 158
Figure 53: Training site at Newark, USA .......................................................... 159
Figure 54: Training site at Pasir Panjang, Singapore ......................................... 159
Figure 55: Training site at Port Kelang, Malaysia ............................................. 160
Figure 56: Training site at Puerto Barrios, Guatemala ........................................ 160
Figure 57: Training site at Cabello, Venezuela .................................................. 161
Figure 58: Training site at Rotterdam, Netherlands ......................................... 161
Figure 59: Training site at Santos, Brasil .......................................................... 162
Figure 60: Training site at Shanghai, China ...................................................... 162
Figure 61: Training site at Taranto, Italy .......................................................... 163
Figure 62: Training site at Vishakapatnam, India .............................................. 163
Figure 63: Training site at Vladivostok, Russia ................................................. 164
Figure 64: Test site at Hamburg, Germany ....................................................... 165
Figure 65: Test site at Wilmington, USA .......................................................... 166
Figure 66: Test site at Long Beach, USA .......................................................... 166
Figure 67: Test site at Rotterdam, Netherlands ............................................... 167
Figure 68: Typical layout of a container terminal ............................................. 172
ABSTRACT

SPATIAL ARRANGEMENT AS A PART OF GEOSPATIAL FEATURE ONTOLOGIES

Jonathan R. Clark, Ph.D

George Mason University, 2012

Dissertation Director: Dr. Anthony Stefanidis

The geospatial sciences have always employed ontologies to describe spatial data, conduct analysis, and to design the systems and methods for conducting geospatial operations. While research on geospatial ontologies has addressed non-spatial aspects of features on the terrain, space and spatial relationships are usually treated as separate from the nature of geospatial objects. This work presents an approach for making spatial context an integral part of geospatial feature ontologies. Specifically, the spatial arrangement of the components of a complex feature class (features composed of several simple features) was examined as a potential “spatial signature” of that feature class. This work suggests an approach for including spatial relationships and metrics in a prototypical feature ontology, and then using these metrics to judge the similarity of metrics from other feature classes to that prototype. Container terminals, a type of commercial maritime terminal, were used as a test case with a machine-readable
ontology developed using spatial metrics from sample terminals world-wide. Like metrics were also collected for test cases representing several types of facilities, and compared to the prototype ontology. Using a simple similarity model, comparison of the test cases to the prototype resulted in partial success in judging similarity of the test cases to that prototype. Variations in facilities layouts and the simple logic of the similarity model were the likely causes of reduced success in similarity judgments.
CHAPTER 1: Introduction

Human perception of the world around us can be thought of in three ways: “what” we perceive, “where” the perceived objects are, and “when” those objects (or events) are occurring. Our understanding of the meanings of things and events demands that we include all three of these dimensions in our thinking. Any organism which hopes to compete in the physical world learns to think this way, or they risk becoming extinct. A predator, for example, must correctly identify its prey (the “what”), it must correctly identify it’s position (the “where”), and in many cases the predator must be able to predict where the prey will be in the future (the “when”) so it can plot a path to intersect and capture it’s next meal. Without all three of these dimensions of perception, it is difficult to successfully understand and operate in the physical world.

The geospatial sciences have always tried to model the world along these three lines. Photogrammetrists and image interpreters took great care to correctly identify objects (“what”) on imagery during compilation/collection efforts. Cartographers spent considerable time designing clear symbols to identify these items on maps. They also went to exhaustive lengths to develop compilation bases to position these collected items (“where”) on maps and charts. Since the 1970’s, we have extended this to include
topological relationships, increasing our ability to understand the “where” dimensions of
the world around us and extend our analysis capabilities in the spatial sciences.

Yet despite several hundred years of mapping, and almost half a century of
quantitative methods in the spatial sciences, we have only begun to scratch the
technological surface of how we capture and analyze geospatial data. We are still
primarily looking at small, isolated views of the world with little understating of how that
view relates to the larger spatial and temporal context of the world.

We begin by defining the notion of a “feature” as any physical or abstract entity
which can be positioned in space and time, and which has descriptive attributes.
Features can be permanent or transient, they can be static or moving, visible or invisible,
natural or man-made. Examples of physical features in the geospatial sciences include
roads, buildings, mountains, rivers, cities, lakes, etc. Examples of abstract features
include: political/country boundaries, maritime shipping corridors, restricted airspace
boundaries, navigation waypoints, and land parcels. Features can also be “simple” or
“complex”. Simple features are considered those which are the lowest level of
organization within a Geographic Information System (GIS), below which no further
subdivision is used. Composite features are those made up of two or more simple
features or some combination of simple and other complex features. An example of a
simple feature might be a building or a stream. A complex feature might be an airport,
made up of several buildings, a runway, and several roads. The primary difference
between simple and composite features is that of how the user wishes to conceptualize
their data. A “building” might be the lowest divisible unit of data for their application,
and the details of the doors, walls and windows may not be important to the analyst’s task. Attributes of the building might include the number of windows, but there would be no separate (logical) record in the GIS for “window”, only for “building”. If, on the other hand, the building is modeled as a composite feature, there might be several simple features modeled as the roof, another for the door, others for each window, etc. Each simple feature has its own geometry and attributes and the concept of “building” is modeled by linking specific records for the walls, windows, and roof records which comprise that building. The feature called “building” is, in this sense, a new abstraction (a composite feature) made up of several simple features.

In cases where composite feature classes are hierarchical (parent-child relationships where composed-of relations are present) the scale or resolution of the data is important. This is not simply a matter of cartographic rendering where collections of components are represented as discrete geometries at large scales, but only as a point at small scales. Scale of the ontology is very much affected by who the user is, and how they conceptualize the features in question. A shipping company might consider the container terminal as a single point feature with attributes for where it is, the name of the terminal, and its general cargo capacity. A port authority may, on the other hand, need very detailed information about the size and spatial layout of the terminal. The using-domain of the ontology may call for a very different hierarchical view of the feature.

These standards serve as a starting framework for codifying geospatial features, but have not yet matured in the area of complex feature types. Their emphasis reflects the current user focus on simple features though they make provisions for “feature collections” which may be able to address complex feature types.

Whether physical or abstract, simple or complex, all geospatial features need to be defined from the three perspectives:

- Their “what-ness”
- Their “where-ness”
- Their “when-ness”

Historically, mappers have focused on the identifying characteristics of features (their what-ness), including geometry and descriptive attributes. This is due to the cartographic lineage of most GIS technologies where portrayal of basic feature information and locations was sufficient for most uses. One can see this in present geospatial data holdings where most of the data model is dedicated to feature coordinate information, and descriptive attributes. Understanding what a feature “is” and its basic characteristics is important to include in a feature model, but is not enough. Features do not exist as little universes unto themselves, but have spatial (and temporal) characteristics which constrain them or even define them. We next need to model the “where-ness” of features, and include spatial context and spatial arrangement in their ontologies.

Studied extensively in linguistics, context is an essential part of semantics, refining the meaning of a term or phrase through relationships to other concepts. The term “Paris” could refer to a city in France, a city in Texas, or a character in Greek legend (the son of King Priam of Troy). The correct understanding of “Paris” requires that the
speaker/writer include references to these other concepts (i.e., that they place the term Paris in context) for the listener to reach the intended meaning.

In the geospatial sciences, context is often treated as a subjective measure of relationships between features. Image interpreters refer to this as “association”, one of the eight characteristics they use to extract information from image data (color, size, shadow, shape, tone, texture, and pattern being the other 7). Photo-interpretation “keys” may also include association (i.e., context), but is of a subjective, human-readable nature. The OGC and ISO standards for feature modeling (OGC, 2011, ISO/TC211, 2009) focus on the “what-ness” of geospatial data, but are limited in their treatment of “where” and “when” to geographic coordinates and time attributes; neither standard addresses context or spatial arrangement.

One of the more compelling reasons to include spatial context in a formal ontology is the ability to determine similarity between feature instances, or whether a set of observations matches a pattern we might be looking for. Once we add spatial measures to the ontology of a feature class, we can use this ontology as a pattern to search-for and analyze other occurrences of that feature class in geospatial data sets. This leads to the second task: that of quantifying the similarity of observed data to the “type” ontology for a class of features.

Determining whether two objects, or sets of objects, are similar enough that they represent the same set of “things”, is an essential task of our everyday lives. We must make decisions on whether the person we see in front of us is, in fact, the same person we know as a friend, or family member, or colleague at work. We have to determine
whether the route we take to work is the same (or at least similar to) the one we took yesterday. We look for similarities in the storylines of books and newspapers, for familiar and pleasing patterns in art and music, and even for the repeating patterns of behavior in those around us. But how do we determine that two objects are similar enough to be the same, or at least similar enough to be a new variant of a pattern we have seen before? In most studies of similarity, geometry and space play a crucial role in our ability to determine “sameness” of objects. Tversky’s work in psychology (1977) recognized that in addition to a features’ attributes (color, name, etc.), geometric similarity was a part of recognizing facial patterns and letters and many other objects we encounter every day. Gardenfors (2004) extended the notion of geographic space into “conceptual spaces”, asserting that the spatial patterns we see in graphs and diagrams are understood in the same way as for the 3-dimensional space of the physical world.

Spatial context, or the relationship of an object to its surroundings is important in most of the mapping and image processing sciences. Frank (1997) considered spatial relationships in describing “spatial ontologies” as a part of the broader “Ontologies of Entities” which are used to understanding the meaning of geospatial data. Egenhofer (2002) argued that these spatial ontologies should be formalized in a computational model, suggesting how the Structured Query Language (SQL) syntax might be used for this. Image analysts use spatial context in deciding the nature and attributes of both natural and man-made features, GIS analysts employ spatial relationships for site selection, sources of pollution, transportation planning and studies, the Topology in a GIS is a good model for storing spatial relationships of one object to another, but typically
falls short of describing how a feature relates to its neighboring features. Machines can display buildings with their surrounding structures (roads, stores, towers, bridges, utilities, etc.), but an algorithm cannot yet determine that a certain spatial configuration of these is, in fact, a “city”; what the spatial pattern “means”. Quickly measuring distance, proximity and connectivity is done well by machines, but the ability to find spatial patterns in data remains largely a human skill. Once we move this understanding of spatial “semantics” (i.e. what the spatial patterns “mean”) into algorithms, we can begin to unburden the human analyst from manual searches in the growing volume of spatial data on the internet.

Once we understand how to express spatial relationships and context for a feature, temporality (“when-ness”) must be added. This will facilitate change detection, not just at the data-set level as is done today, but at the individual feature level. We can then examine an individual structure (or set of features), or an event through time to see how it changed or grew. Since spatial context is also expressed, we can examine how changes in one feature over time might impact features which are spatially related. Predictive GIS, considering both space and time, will become a reality.

To move towards machines which have spatiotemporal semantic understanding of features, two separate tasks are evident: (1) determining a machine-readable model of spatial context, and (2) developing a measure of spatial similarity. This dissertation is organized around these two task areas.

As the feature ontology is developed, selection of the particular spatial metrics which should become part of the model will be a crucial decision. Contextual metrics
such as proximity of the “subject” feature to instances other feature classes, adjacency/tangency to other objects, angular orientation, relative and absolute sizes, shape, and other spatial measures are all considered for inclusion in the ontology, but final selection must be based on statistical examination.

The hypothesis of this work is that there exists a measurable way to describe how a feature class will typically be arranged on the landscape, and that this spatial arrangement constitutes a “spatial signature” for that feature class. What is “typical” may be determined empirically by collecting spatial metrics from known instances of a feature class, through studies of the function and form of a feature class, or through a combination of methods using both known functions of a feature class and empirical measurements of instances of that class. Empirical metrics from image measurements and from sample GIS datasets can form a “starter” ontology. This can be refined and expanded using inputs from subject matter experts, engineering specifications, and other sources to ensure that the ontology reflects the variability of spatial arrangement for a feature class, thereby making the ontology a “prototypical” pattern for the class.

The prototypical ontology of a feature class will have the following general characteristics:

1. It is driven by the key functional characteristics of a feature class and displays consistency in the observed spatial arrangements of its components,

2. It allows for variability in spatial arrangement to accommodate local spatial conditions (e.g. cultural and terrain variations)
3. The resultant ontology for that feature class has diagnostic potential, and can be used to discover new instances of that class in new datasets.

This is consistent with the approach used to give grounding to image interpretation where terrain features must be identified and characterized (Belcher, 1951, Lillesand, et al., 2007, Olson, 1960).

Therefore, we make the argument that the spatial characteristics of a feature class should not be considered any more or less important than its non-spatial characteristics. Accordingly, this dissertation effort argues that spatial arrangement and spatial context are too important to ignore in the ontologies of feature classes.

Validation of an ontology which incorporates spatial context requires that it can be used as a pattern against which similarity can be measured. The matches between the prototype ontology and new observations need not be exact, but should quantify the level of similarity of candidate observations to the (prototype) feature ontology. The inexact nature of the similarity comparisons means that strict scene-matching or shape-matching approaches must be combined with more flexible techniques of conceptual graph comparisons used in information theory and other semantic research areas.
The hypotheses, both null and alternative, for this research can be stated as follows:

- **Null Hypothesis:** Components within complex feature classes are randomly arranged spatially, therefore this spatial arrangement does not have interpretation value for instances of such classes.

- **Alternative Hypothesis:** The spatial arrangement of the components of a composite feature class as it is dictated by its function is limited, and comprises measurable spatial metrics which can be diagnostic for that class.

Once a feature ontology has incorporated spatial arrangement in a machine-readable way, similarity determinations can be performed using algorithms and will no longer be limited to the slower, human-perception based similarity measurement methods. Algorithms can be developed to search through large volumes of multi-source data to look for matches using the ontologies as patterns. Information discovery can begin to look for complex features, and not simply the piece-part components which make up features. An ontology-based discovery approach will allow for partial matches, meaning that feature discovery algorithms can search large data holdings for “possible” matches which can be further inspected by human analysts for final determination; the resultant time-savings for the analyst can be significant. Partial pattern matches also means that ontology-based algorithms can identify portions of a feature which might be missing or have changed. This will allow for queuing of other collection efforts to find the missing components. It also means that a feature ontology can employed as a pattern for determining changes to features over time.
One of the most compelling reasons for this research is what can be called the “data tsunami” which has already arrived on the desk of many GIS analysts. The online magazine “GOVPRO” cites two studies, one from Daratech Corporation and one from Global Analysts, Inc., of increasing volumes and diversity of data are being created with estimates of from 8 to as much as 10 percent growth per year (Keating, 2011). This is leading to the need for tools to sift and sort these data, identify duplications, deciding which “things” are actually the same objects despite different labels, and look for patterns in the data they did not know were there. Given enough time, a human analyst can disambiguate and integrate these data … but we rarely have the time or understanding of each data source to do this. This is where the use of ontologies can help (Hassell, et. al., 2006, Kuhn, 2005).

By automating portions of the human cognitive process, software will be able to examine geospatial data and “understand” what the data represent in the real world, and how the data objects relate to other data objects. Data in different schemas can be integrated more easily, redundancies found, and patterns within and across data sets examined. Again, humans do this almost instinctively, and we are hopeful that semantic models and algorithms will be able to aid the analyst/researcher in doing this on higher volumes of increasingly diverse data.

The Literature Review (Chapter 2) will examine the past and ongoing research in the area of modeling context, including work done in the mapping sciences, computer vision, and other disciplines. It will also review the various measures of semantic and spatial similarity useful for geospatial feature data. The Methods section (Chapter 3)
describes the steps used to develop feature ontologies, descriptions of the various spatial metrics considered, as well as development of similarity measures. The analysis conducted on feature ontologies, as well as the success of using the ontologies for similarity determinations, are discussed in the Results section (Chapter 4). The Conclusions and Future Research section (Chapter 5) will present the most significant findings of the work, and suggestions for additional research.
CHAPTER 2: Literature Review

2.1. Overview

One way to understand the past and present trends in research in this area of spatial context or spatial “semantics” is to examine how some of the major technical themes relate to each other. Figure 1 graphically depicts the major themes which are closely related to this research effort, and highlights the position of this dissertation in the context of other topics. It also serves as a graphical outline for the Literature Review chapter.

2.2. Semantics

Simply put, “Semantics” is the study of meanings. Usually associated with linguistics and philosophy, semantics is an easily misunderstood field and one often labeled as “too theoretical”, spending too much in trivial academic debates on terminology. It is far from this, and semantics represents one of the most important “mega-trends” today in information systems, including those used for geospatial analysis (Egenhofer, 2002, Janowicz, 2009).
Figure 1: An ontology of the topics in the Literature Review chapter. Numbers indicate paragraph numbers in the document. Ovals highlighted in red indicate the focus areas of this dissertation.
The human mind is a wonder of cognitive power, with the capacity to form relationships between things and concepts, and to use these to understand not just the obvious...but messages which are often subtle, even hidden in what is spoken or written. We have the ability to understand not just the words but the meanings behind those words. In the case of geospatial features, we use definitions in the form of data models and data dictionaries to communicate the characteristics, functions, etc. of the “thing” being described in the model or ontology. The mapping of a given feature model, to the concepts and definitions of a real-world object class is precisely the business of semantics (Frank, 1997, Kuhn, 2005, Ogden and Richards, 1930). The semantics (the “meaning”) of a physical object or group of objects is so ingrained in our daily thinking that we rarely think about how we “do” semantics.

The cognitive process of understanding meaning can be simplified into what has become known as the “Semantic Triangle”. This graphic depiction in Figure 2 is an adaptation of the “Semiotic” triangle first proposed by Ogden and Richards, (1930). In their work in language, the matter of relating ideas to symbols (e.g., words) was essential to prevent ambiguity. “Semiotics”, the study of symbols and how they relate to concepts, remains an important part of semantics.

Recent usage has been to refer to this as the “Semantic triangle” which relates symbols (e.g., words, diagrams, data models), to thoughts (concepts). Both the symbol and the concept then relate to a “referant” which is an actual instance of an object in the world. It is important to note that a “concept” exists only in the mind of the user; it’s
manifestation in the world (physical) is the “instance”, and the explicit definition of that concept is the ontology.

Figure 2: The “Semiotic Triangle”. The graphic on the left is the original triangle published by Ogden and Richards (1930). The graphic on the right is the author’s interpretation of how the semiotic triangle relates to geospatial feature data.

This triangle model was originally applied to philosophy and linguistics, but can also be used to understand the role of feature ontologies in understanding geospatial data. Figure 2 shows that a feature on the terrain is described in a feature ontology, and that the ontology describes the concept in either human or machine readable form. When all three components of the triangle (ontology, concept, and instance) are synchronized, then communication amongst users, domains and algorithms is possible (Ogden and Richards, 1930, Kuhn, 2007).

This is also central to the ideas of the “Semantic Web” where data searches can be based on the what the investigator “meant” in their question, and not simply the terms they used in a query statement.
As noted earlier, semantics as a subject originated in philosophy as a way to describe how major concepts related to each other, as well as how human beings perceived the world around them. Tversky (1977) described how we look for and describe similarities in objects and concepts, and Gardenfors’ seminal “Conceptual Spaces” (2004) described how our perception and description of the world was actually the application of graph-theory in mathematics, and that these perceptions (of both spatial and a-spatial environments) could, in fact, be quantified. Frank (1997) went into more detail about how semantics applies to geospatial information, and pointed out how geospatial data was unique as regards the use of semantic descriptions.

Ontologies, which will be discussed in more detail later (section 2.4) are used to encapsulate basic knowledge about geospatial features. Ontologies become important in several areas of the geospatial sciences, including data sharing, efficient indexing of data (for discovery and retrieval), and for linking data objects to one another, including links between spatial and non-spatial information. Fonseca and Rodriguez (2007) suggest that ontologies of geographic data are important in supporting automated reasoning in GIS. To be practical, the ontology must provide workable bounds on how a feature “could” appear, while also accounting for the variability in how features “actually” appear in the world. This research will use instance data from imagery, GIS, and from subject experts to construct and then refine feature ontologies. This “grounding” of ontologies is something recognized as essential to make the ontologies useable (Kuhn, 2005, Scheider, et. al., 2009).
2.3. **Knowledge Representation**

Developing a feature ontology which incorporates space is an effort to represent knowledge about a feature class in a structure which can be acted upon by software. Unstructured text, music, art and other human-readable structures have historically been used to represent knowledge, but this is often difficult to render in forms which can be acted on in computation. These mechanisms are often used to capture knowledge of a philosophical, ethical, or aesthetic type, but are not suitable for use in algorithms. Sowa (2000) concluded that “Natural Language” was the ultimate in Knowledge Representation (KR) languages, had the highest level of flexibility and expressiveness, but is largely limited to human-readable forms. What is required in the geospatial information sciences are machine-readable forms of feature knowledge.

A good approach to thinking about KR also comes from Sowa’s work (2000) where he suggests the application of three disciplines: Logic, Ontologies, and Computation. Taken together, these three capture object knowledge in a form which can be acted upon by algorithms. Minsky (1975) described several ways in which knowledge could be represented, including: Rules, Frames (stereotypical information of a situation, often in network form), Schematics, Unstructured Text, and Logic. Each was employed by particular domains (e.g., rules are used by sociologists, logic by mathematicians), and have varying suitability for being used in algorithms. Whether intended for human consumption, for processing in algorithms, or for both, the declarative structure (i.e., which provides core statements of truth about an concept), KR structures can be generally referred to as “ontologies”.

18
2.4. Ontologies

The term “ontology” originated in the field of philosophy, and was recently adopted in the geospatial sciences using a basis of the word’s philosophical roots (Egenhofer, 2002. Kuhn, 2005. Smith, 2003, Guarino and Smith, 2003). The simplest way to think of an ontology is that it is the way in which people describe a portion of “reality”.

Merriam Webster’s online dictionary (Merriam-Webster, 2012) offers two definitions of “Ontology”

1: a branch of metaphysics concerned with the nature and relations of being

2: a particular theory about the nature of being or the kinds of things that have existence

An ontology is not a portion of reality…it is one group’s (one domain’s) description of that portion of reality. Gruber (1992) suggests that an ontology can be thought of as “.. a specification of a conceptualization. ..a description …. of the concepts and relationships that can exist for an agent or a community of agents”. Gruber (1995) also emphasized the formalization and “portability” of ontologies to support practical implementations, vice theoretical discussions of semantics. His 1995 paper suggests many ideas for text-based ontology structures which appear to have led, in part, to recent RDF (Resource Description Framework), OWL (Ontology Language Web) and other XML (Extensible Markup Language) -based semantic work.

In a practical sense, an ontology is the encapsulation of the theory about a “thing” or “things”, which is used by people and/or software to communicate what a thing “is” or “means”…what is it’s nature. We all use ontologies every day, though we rarely refer to
them using the “o” word. A dictionary is a collection of simple ontologies for each term, a thesaurus is also an ontology portraying the relationship of a word to other words.

There are many forms of ontologies. These may be considered formal or informal, and are often viewed as having varied levels of “robust-ness”. Figure 3 shows how many of the more common forms of ontologies compare in terms of “strength”.

Sorokine (2010) describes stronger ontologies as being those which attempt to describe the relationships between concepts, as opposed to strictly defining each concept.

Which leads us to another important characteristic of more robust ontologies; that they describe the nature of a “thing” (or concept), and the relationship of a thing to other things.

For example, the term “stream” could mean a type of hydrologic structure, or it could mean the manner in which videos are sent to your home computer. To tell which is meant (i.e., to “disambiguate” the term) you need a definition showing relationships to other concepts. You need an ontology which shows how “stream” relates to other concepts and therefore give the user (or an algorithm) contextual information about how the term is being used and what it means. The ontology could indicate a relationship to hydrologic structures, or it could show that “stream” is a related to how data are transmitted to a software application. With this information, the user can disambiguate the term.

Such ambiguity is important to designers and engineers when implementing geospatial data systems. Fonseca, et. al. (2003) suggest that ontologies for data are an important precursor to implementing data models for GIS. An ontology becomes the
starting document for a data engineer to build a conceptual database schema, and then a logical and finally physical data model. Without an ontology to guide their designs, the data modeler may overlook key aspects of, and relationships between, the geospatial data resources.

![Ontologies by Strength](image.png)

Figure 3: Ontologies by strength, with weaker ontologies on the left, stronger on the right. Stronger ontologies are those which describe relationships between concepts.

### 2.4.1. Conceptual Graphs

One of the more useful ways to depict ontologies is that of a graph. Widdows (2004) and Gardenfors (2002) both argue that graphic renderings of ontologies closely match the way that many human-cognitive processes work and, therefore, are ideal as a mechanism to capture, portray and analyze the nature and relationships amongst
concepts. A “graph” depicts connects as nodes, and the relationships between the nodes as edges. The resultant network graphs can be “complete” (relationships shown between all nodes) or, more often, are partial with irrelevant relationships left out.

Transportation and utility networks in GIS are stored as graphs of physical objects, but may be extended to include a-spatial information, thereby turning the physical graph into a conceptual graph.

The application of Dijkstra’s (1959) algorithm to shortest-path analysis is a well-known example of the application of conceptual graphs. A physical network is captured from GIS datasets, and rendered as nodes and edges. Each node and edge can be assigned weights which denote almost any type of relationship (distances between nodes, elevation changes, etc.). The cost to traverse the network may be a function of both spatial and a-spatial costs on the edges. Whether the physical distances or the non-spatial weights are employed in the shortest path calculations, the network being used has become a type of conceptual graph. Another way to think about this is that all physical networks in GIS are special cases of conceptual graphs, with edge values depicting physical distances.

The advantages of conceptual graphs over graphic depictions of physical space is at least two-fold:

1. Conceptual graphs can capture both spatial relationships and non-spatial relationships, thereby serving as an integrating framework for all knowledge about a feature.
2. The edges on the conceptual graph may be assigned values or weights which describe many types of relationships between connected nodes on the graph.

Sowa (2000) also sees benefits of using conceptual graphs as a medium for communication between knowledge engineers (those who build the ontologies) and subject-matter experts (SMEs) knowledgeable about the features being described by an ontology.

When attribution is added to the nodes and edges of conceptual graphs, they are sometimes called “attributed graphs” (Gardenfors, 2004). Use of attributed graphs has been common in computer vision to codify both scene information, as well as information about objects within the scene. Ligozat and Condotta (2005) argues that the use of conceptual spaces are important in qualitative spatial reasoning, allowing more robust understanding of geometry, and also suggesting that conceptual space can aid in temporal reasoning. Hofman and Jarvis (2000) demonstrated that attributed graphs where useful in computer vision where a pattern graph, developed a-priori, could aid in identification of 3-D objects. Attributed graphs were central to Stefanidis, et. al.’s (2009) work in scene-matching for comparing geometries of known road/building groupings, to feature occurrences in aerial imagery.

2.4.2. Geospatial Semantics

The geospatial sciences have always employed ontologies to describe spatial data, conduct analysis, and to design the systems and methods for conducting geospatial operations (Egenhofer, 2002, Mark, et. al., 1999, 2004). Historically these ontologies
have appeared in forms which were easily read by humans (e.g., dictionaries, glossaries of terms, photo-interpretation keys, facilities diagrams) but machine-readable forms (e.g., database and XML schemas, UML (Unified Modeling Language) data models, RDF, etc.) have begun to appear to allow machine-processing of large quantities of data. This allows machines to understand not just the values of attributes in geospatial data, but the relationships of data to other data and the meaning of a data object in the context of a broader data landscape on the internet.

Considerable research has been done on the non-spatial aspects of context (Kessler, 2007, Janowicz, 2008, Lee, et. a., 2008, Schwering, 2008) but the geospatial perspectives have remained less formal in GIS research (though this is changing rapidly). This recent research into spatial similarity is focused on single objects or features, and how similar they are to other singular objects. Shape, size and orientation are compared, but research into spatial context (the relationships between and amongst numerous objects on the terrain) have not often been done.

In mid/late 1990s, several key research centers formed to address matters of geospatial semantics or “geo-semantics”, including the MUSIL (Muenster Semantic Interoperability Lab) in Germany and the NCGIA (National Center for Geographic Information and Analysis) Research Initiative # 10 (Spatio-Temporal Reasoning in GIS) lead out of the University of Maine (http://www.ncgia.ucsb.edu/research/initiatives.html). Kuhn (2005) suggested that describing the semantics of geospatial data was vital in enabling data interoperability. He made a strong case for inclusion of spatial relationships, and suggested a geodetic processing analogy where object descriptions
needed to have semantic coordinate systems and datums which would allow such things as the transformation of one domain’s conceptualizations of data to that of another domain. Janowicz (2010) also emphasized space, but also time, as an important part of organizing and tagging information for use on the web. He suggests not only improvements in data discovery, but also in data understanding through the inclusion of semantic information in data models designed for use on the future semantic web.

Topological relationships were the focus of work between industry and academia (Egenhofer and Herring, 1990) outlining an initial framework mathematical descriptions for spatial relationships in geospatial data. Similar work was then formalized in what became known as “4-Intersection” and “9-Intersection” models describing spatial relationships between several objects on the terrain (Egenhofer and Herrring, 1990, Egenhofer and Franzosa, 1993). Machine-readable models for spatial-relationships were addressed by Shariff, et. al. (1998) who developed a taxonomy of the possible relationships between area and linear features and set the stage for a standard set of descriptors for spatial context. Egenhofer (2002) suggested that semantics would play an increasingly important role in several areas of geospatial analysis and web applications such as data discovery and feature understanding, and outlined some basic approaches to structuring ontologies about spatial information.

The use of “conceptual space”, or the application of graph theory to portraying and analyzing just about any form of data measure, has been a basic approach in cognitive science and was summarized in Gardenfor’s “Conceptual Spaces” (2000). In his work, Gardenfors submits that rendering ideas and concepts in graphic form, aside from being
one of the basic ways we think about the world, allows us to apply mathematics to such questions as whether two concepts are similar, how different they might be, what are the major trends in our data, etc. He also argues that we consider spatial ideas (i.e., geographical) in the same manner that we consider conceptual spaces and that the analytical methods of one, should be directly applicable to the other.

2.4.3. System Ontologies

These are ontologies which guide system architectures and software development. The ontology of a system is often referred to as it’s “architecture”, documenting both the components of a GIS and the relationships and interactions between/amongst the components. Systems engineers employ many standard tools and methods to develop these architectures at many levels of detail from conceptual architectures, to specifications for individual software and hardware components. Ontologies for system processes and the data handled by a system are often considered part of the system architecture. DoDAF (Department of Defense Architectural Framework) (DoD, 2011) FEA (Federal Enterprise Architecture (OMB, 2011), and the Zachman Framework (http://zachman.com/) are well known examples of formal system architecture frameworks available to systems engineers. Fonseca, et. al., (2002) described the importance of ontologies in designing GIS technologies and processes by reverse engineering the GIS from the analytical questions and the nature of the spatial data, to select/design the best technologies to address the users’ questions. Most National Mapping Agencies (e.g., US NGA and USGS, UK Ordnance Survey, Germany AMilGeo, Netherlands Topografische Dienst, etc.) employ ontologies in the form of
system architectures in their acquisition and development of systems used in geospatial analysis and production.

2.4.4. Process Ontologies

Processes and workflows are often documented as part of a “Business Process Re-engineering” (BPR) effort and captured in several standard forms which may be motivated by efficiencies (and therefore profitability) of manufacturing, business, and engineering efforts. This need to optimize processes has led to the formalization of process ontologies in such standard forms as the Process Specification Language, Cyc, SUPER/DDPO, and XPDL. Process ontologies may appear in GIS in the form of workflow definitions where processes and process chaining need some level of automation. Fonseca, et. al. (2002) developed ontologies for an analysis process by “reverse engineering” from the question posed by a user. This approach was also suggested as beneficial for developing ontologies for structuring a GIS (a system ontology) and for an ontology of the data required to answer the posed question.

2.4.5. Data Ontologies

Current interest in ontologies in the geospatial sciences is highest for defining data and datasets. This is important to support development of algorithms to act on those data, and for the exchange of data between users. Kuhn (2005) and Mark, et. al. (2004) argue that data interoperability is a primary reason for ontologies, as is more effective data queries and discovery. Interoperability may be between users, groups of users (aka domains), and also between software components.
Data models can be thought of as ontologies, though models tend to be system specific. The primary difference between a data model, and an ontology for a feature class is that models are lead to implementations, and ontologies serve as patterns against which data models are designed. This is not a crucial issue and many people consider ontologies as the prototypical model for a data class.

Current data engineering practices (DoD, 2011) generally group data models into three “levels”

1. Conceptual Data models: Definition of data objects at a general level
2. Logical Data Models: Definition of data objects at a level that engineers can build databases against these models
3. Physical data models: Detailed data structure specifications, usually technology-specific

Ontologies are appropriate and necessary for all types of geospatial data, particularly those data entities exchanged between users and applications.

Early versions of geospatial data ontologies were at the general, multi-theme level, and typically appeared in the form of data taxonomies, data dictionaries, and other information necessary to allow application developers to design software to act on those data, and data management specialists to design databases to store geospatial data.

Some of the early documented ontologies were published by national agencies in the United States (US), the United Kingdom (UK) and elsewhere, where spatial datasets are available for public use.
The US Census Bureau began publishing TIGER spatial data in 1980 (US Census, 2012) issuing a data dictionary along with the vector files to explain the nature and attribution of each data type (i.e. data “class”) in the datasets. The US Geological Survey publishes standards for Digital Line Graph (DLG) data, including definitions of feature classes and attributes (USGS, 1998). The UK Ordnance Survey (OrdSurvey) publishes ontologies for several feature classes, as well as “Ontology Modules” describing spatial, network, and mereological relationships which could be used with feature ontologies (UK Ordnance Survey, 2011). An important distinction of the OrdSurvey’s effort is that the ontologies are described in XML-based RDF (Resource Description Framework) documents which are readable by software algorithms.

While most people would agree that many these data artifacts are forms of ontologies, they still have several common characteristics which limited their use as full ontologies:

1. They did not describe relationships between and amongst the data classes described within the ontology

2. They usually did not try to address the different ways that feature classes were defined by different communities of users.

One could argue that point #2 is not the role of an ontology and that intra-ontology mappings should remain as separate documents. This raises the question of whether there exists a common, over-arching, “uber-ontology” which all user-level ontologies should reference themselves to. An “Uber-Ontology” defining all ideas and relationships is a laudable goal, but may never be possible. Ontologists can usually agree about the very “top” level of a taxonomy where the concept “thing” appears. This “thing” includes
everything in the physical universe, and well as abstractions such as “love”, processes, organizations, political parties, administrative regions, etc. Once we cascade below the concept of “thing”, we move to definitions and concepts which tend to be unique for a group of people or “domain”. This has been well accepted in most fields and described in such landmark books as Ogden and Richards’ “The Meaning of Meaning” (1930). So, while it is important to operate within a given domain, the advent of the internet has accelerated the need and ability to work across domains; but how do these domains communicate without a common ontology or lexicon?

Two distinct philosophies are apparent, plus a hybrid of the two (Calavanese, et. al., 2001, Wache, et. al. 2001):

**GAV (global as a view of the local).** In this case, concepts for all entities and relationships are captured in a single ontology, to which all data holdings and query statements for all domains must adhere. The advantage of this approach is that all domains reflected in that ontology are easily able to communicate and exchange information. The disadvantage is that if a new domain or concept appears, then the entire ontology, and any query statement written against that ontology, will need to be re-evaluated and possibly re-engineered.

**LAV (local as a view of the global).** Each domain is free to define their own domain-specific ontologies and evolve and change them as required. For two domains which never have occasion to interact, this is the preferred approach. A key advantage is that it tends to be quicker to implement and maximizes flexibility for a given domain. The disadvantage of this is that it puts a larger
burden (i.e., expense) on a domain to coordinate their changes with domains they interact with and possibly re-engineer things like data exchange services and interfaces to account for changed concept definitions.

The hybrid approach (sometimes referred to as “GLAV”) is where local domains define many portions of their ontology, and then work with other domains to evolve a central “common” ontology where all agree on definitions. This may be supplemented by cross-domain concept mappings to aid in those portions of their data which need to be exchanged with other domains. The advantage here is that it gives each domain an appropriate level of control over changes to their ontology, and supports cross-domain data exchange where needed. It also recognizes the reality that a domain usually has data which is rarely if ever of interest to other domains. The disadvantage of GLAV is similar to the issue with LAV where each domain must invest some level of management and engineering effort to keep portions of their ontology coordinated with other domains.

This author adheres to a variant of the GLAV approach. Ontological control at the local domain level is a practical approach, but there also needs to be a common forum or working body of interested domains to monitor cross-domain exchanges to find common definitions of concepts. This working forum could, for example, observe that many domains are exchanging transportation data, and that they are all using the same definition of “traffic light”. In this case, the forum could agree to turn the authority for the traffic light ontology over to a common (i.e., cross-domain) engineering team to
maintain that ontology and service definitions for exchanging that feature class. The
disadvantage of this is an ongoing investment of staff time for the common forum. The
advantage is that all domains can look to stability in concept definitions, against which they can collect data and build analysis applications.

Cruz and Xiao (2005) expands and refines this GLAV/LAV approach with example ontology comparisons, and by showing how the approach can actually be implemented using standardized structures (e.g., RDF, OWL, etc.) for enhancing data interoperability. Other researchers have also focused on the importance of ontologies for interoperability (Buccella, et. al. 2001, Calnan and Cruz, 2001, Janowicz, 2009, Cruz, et. al., 2005) and Laskey et. al. (2008, 2010) introduced probabilistic dimensions to ontologies to address uncertainties in the data during interoperability and data integration.

2.4.6. Feature Ontologies

The data type central to this research is that of feature data. Geospatial “features” are considered those entities which can be ordinated in space and time, and represent physical or abstract’ objects. Buildings, roads, rivers, and factories are examples of physical features; political boundaries, economic zones, placenames, and restricted airspace areas are examples of abstract feature. Abstract features are distinguished by affecting human activities, have geometry and attributes, but can only be visualized through the use of a GIS display (or printed map product).

The term “feature” is also used in information science to refer to the characteristics of a data object. Schwering (2008) uses the term in this fashion when discussing feature-
based similarity methods for geospatial ontologies. In this document, such “features” are considered the attributes of geospatial features.

2.4.7. Feature Conceptualizations and Models

Organization of a feature ontology is strongly dependent on how the using domain views or “conceptualizes” a given feature class. Smith and Mark’s (2001) work in how non-expert persons (non-expert in the geoinformation sciences) actually perceive geographic phenomena shows how difficult it can be to reach a simple definition in an ontology. This will necessarily cause feature ontologies to be highly diverse, but there are some commonalities which should be considered. Geospatial features need to be defined from three distinct perspectives:

i. Their “what-ness”

ii. Their “where-ness”

iii. Their “when-ness”

This approach is quite common and can be seen (with the exception of including spatial relationships) in standardized approaches such as the feature model of the OGC (2011) and the geospatial ontologies of the W3C (2012). These have focused on the characteristics of features (their what-ness), including attributes and geometric measures, and included markings for temporality, but did not address context or spatial arrangement. While these standards are an important step in feature ontologies, they are only the first step not. Features do not exist as little universes unto themselves, but have spatial characteristics beyond their geometries which constrain them or even define them.
This is where this research work in adding spatial dimensions to ontologies is so important.

Published feature ontologies do not often include spatial relationships. Those at the UK Ordnance Survey (2011) are available in RDF/OWL as templates or “starter” ontologies, but spatial relationships are held in completely separate definitions, and not considered a part of a feature. Likewise, the feature definitions for the National Map (at the US/USGS) include geometry and attribution but not spatial relationships. This effort is an important area of research in adding spatial concepts to feature ontologies for an operational environment (Varanka, 2011).

The importance of this research is best understood when viewed in the context of how the geospatial sciences have thought of “features” since the 1970’s when we began the switch from analog cartography to GIS. Figure 4 is a summary of this evolution with each “phase” (the author’s construct) progressing from left to right in the diagram. Each phase is distinguished by different available technologies, different user/market drivers, and different ways that features are conceptualized.

The author’s research is part of the larger effort to move to phase-IV where features are not simply thought of as geometry plus attributes, but have spatial characteristics which actually define them. Features do not simply occur at a coordinate on the landscape but have spatial relationships which serve to define them, and to constrain them. A road, for example, must be spatially coincident with a bridge when it passes over water; it cannot float and this bridge/road relationship must be true for all instances
of roads. Similarly, airports never have buildings positioned in the middle of the paved runways; this topology condition should be reflected in the airport’s ontology.

The “where-ness” of a feature is not simply its map coordinate, but includes how it relates in space to other features (i.e., its spatial context), and how its components are organized in space (local spatial arrangements). Since the model will include both spatial and a-spatial information, this will move us towards what Mulligan, et al. (2011) have termed a “semantic signature” for a feature.
2.5. **Space and Context**

Context is as important in understanding geospatial data, just as it is in understanding the meaning of written and spoken language. Stalnaker (1999) states that to understand any linguistic assertion, the listener (or reader) must have access to information surrounding the assertion and not just the asserted fact itself. Ideally an assertion will be accompanied with information about the person making the assertion, and about facts related to the core assertion, i.e., the context of the statement. For example, the assertion of “I am travelling to Paris” could be misunderstood, without the context of also knowing whether travel will be to France or Texas.

Context is likewise important in understanding spatial information. A building takes on very different meanings depending upon whether it is related to residential functions, or manufacturing functions. Context can be spatial, a-spatial, conceptual, and/or temporal. In most GIS data, context only becomes apparent when a data object is examined along with other objects in a common coordinate framework; this is a common and valuable function of geographical displays. However, the contextual information exists only in that collective display and not explicitly in the data object’s structure. This might not seem to be an issue, but could become problematic if a group of data objects (i.e., geospatial features) are extracted from a larger dataset and exchanged with other users. The contextual information of the original dataset could be lost and the meaning of the data objects misinterpreted. We must, therefore, determine methods for including contextual information in the ontology of the features which users exchange. Rodriguez and Egenhofer’s (2004) used semantic context to demonstrate an improvement in
similarity assessments, though they observed that it was people’s conceptualization of feature classes which most affected their similarity measures. Even without considered spatial context, Janowicz (2008) argued that there are several types of context which need to be considered when judging similarity (which he generally equates with a query. He suggests six “types” of context:

1. User context: Cognitive capabilities and social background of the user posing the question
2. Noise context: How much extraneous information is present in the queried data?
3. Application context: The number and types of arguments which the user may pass to the application during the query process
4. Discourse context: The contexts encapsulated in the database being searched. This is determined by the conceptualizations of the community of interest which formed that database
5. Representation context: Limitations to semantic context borne of the way the searched database is structured
6. Interpretation context: The manner in which, and the degree to which the resulting similarity metric is and can be applied.

Janowicz’ (2008) key point is that context is not simply a characteristic of or in the data, it is also a characteristic of the user asking the question, and of the systems and structures used to address that question.
2.5.1. Spatial Relationships and Spatial Reasoning

Formalizing a machine-readable feature ontology which includes space as a consideration requires that we first agree on ways and methods to express spatial relationships, and then reason on those methods. Important work was done in the early 1990's to formalize spatial relationships (Egenhofer and Herring, 1990, Egenhofer and Franzosa, 1991, 1995) resulting in a point-set approach generally known as the “9-intersect” model. This was an important step, focusing on the basic relationships between two geometric objects. This led to the development of approaches to defining topological neighborhoods for GIS applications (Egenhofer and Mark, 1995). This work addressed combining binary topological relationships (i.e., between two geographic objects) to express groupings of objects in conceptual space though not in geographic space.

Reasoning about space and spatial relationships is an important topic in human cognition and is a very large area of research in many fields including geospatial science and elsewhere. A comprehensive overview of human perception of space, discussed from the perspective of the mapping sciences, is presented by Mark, et. al. (1999).

Of the many research efforts to formalize spatial relationships in forms amenable to machine-based algorithms, some of the key efforts are shown in Figure 5. This graphic depicts a very small number of the more important papers, with older work at the bottom and more recent research at the top of the diagram.

Frank (1996) used direction as an example of “Qualitative Spatial Reasoning” which attempts to approximate human cognitive processes about space, and Renz (2002) who
proposed approaches for how RCC-8 (Region Connection Calculus-8) would apply to geospatial applications. Interestingly, both Renz’s RCC-8 work and Egenhofer’s “9-intersect” approaches for the geospatial sciences were both based on work in the 1980’s in artificial intelligence and computer science by Clarke (1981, 1985), simplified by Randell and Cohn (1989) and Randell et., al. (1992) into what would be the basis for RCC-8. This early work was in the area of topologic “regions” and how they related in 2-dimensional space with 2 and 4-intersect models (Egenhofer and Franzosa, 1991) eventually leading to the 9-intersect model which addressed more relationships. The relationships between these regions were generally equivalent to polygon features in GIS, and the extensions to linear and point primitives would not take place until the early 1990’s when line/region relationships were added to the theory (Randell, et., al., 1992, Egenhofer and Mark, 1995, Cohn, et. al., 1997). The 9-intersect model and RCC-8 are generally the same approach to expressing spatial relationships. A comparison of the two systems was discussed by Knauff, et.(1997).

Using the 9-intersect model as a starting point, Papadias and Egenhofer (1997) worked to extend formalized spatial reasoning to include directional relations (east, north, etc.) amongst topological point and area primitives via hierarchical algorithms which worked outward from a starting location to determine the direction of a relationship (e.g., “A” is east of “B”) through a broader framework of groups of relationships. Papadias continued some of this work towards “fuzzy” relationships amongst topological objects (Papadias, et. al., 1999), which worked towards more flexibility in relationships amongst groups of objects.
All of these efforts have been crucial building blocks in describing spatial relationships in machine-readable forms, but they are limited in two areas as regards to “features” on the terrain:

1. Both the 9-intersect model and RCC-8 are focused on area/area, area/line, and line/line relationships. Application of these to feature ontologies will require that they be expanded to include relationships involving point features (area/point and line point relationships).
2. These approaches address topological primitives, but have not yet been applied to the level of abstraction of “features”, particularly features made up of groupings of topological primitives.

The spatial relationships in these research efforts have been considered a framework within which geospatial data reside. While this is appropriate in many analysis efforts, we also need to consider space as being an integral characteristic of “things” on the terrain.

One potentially valuable approach taken by Baglioni, et. al. (2008) has been to “enrich” a basic feature ontology with spatial relationships generated by a natural language processing (NLP) algorithm. The NLP algorithm acts on a query statement, and generates SQL query statements which can then be used to query spatial data. The algorithm includes semantic statement in the SQL query, some of them spatial in nature. The idea that space is actually a characteristic of a “thing” (i.e., a feature on the terrain), and not simply a coordinate framework, will be discussed more in the following sections.

2.5.2. Spatial Context

Kessler (2007, 2008) describes the importance and general importance of similarity measures using “context”, including spatial dimensions. His work is centered in the idea of “conceptual space” (versus geographic space) and how context is described using a-spatial information. Measures of Spatial Context have gained recent interest in adding additional depth to the process of conflation and similarity, when combined with “semantic” similarity. Adams, et al. (2010) showed that the “Hausdorff Distance” can be used to quantify topological relationships between two networks of points in geospatial
data, and are using this as a way to compare spatial configurations and context in data. This is part of their broader conflation research and is the start of efforts to model spatial ontologies of vector feature data. Min, et. al (2007) extended the Hausdorff distance metric to better understand the subtleties and differences in similar (though not identical) shapes in a GIS. They demonstrate that this can help with judging similarity of both simple geographic objects, and two collections of objects on the terrain.

Measures of spatial context have been used to aid in image classification (Qi, et. al., 2010), and shape and context template approaches to feature recognition have proven feasible in locating faces and even entire persons in imagery (Huttenlocher and Rucklidge, 1992, Huttenlocher, et. al., 1993). Such applications within imagery processing and computer vision tend to be limited to spatial patterns and context for the target object alone, and do not consider non-spatial relationships to other features on the terrain.

Context has long been recognized by image interpreters as important for understanding terrain and man-made features on imagery. Of the eight photo-interpretation “characteristics” introduced by Olson (1960) for imagery analysts, three addressed context. (These three contextual clues are underlined.)

1. Shape
2. Size
3. Pattern (layout and arrangement of objects)
4. Tone/Hue
5. Texture
6. Shadows
7. Site (nature of the locale where an object is located)
8. Association (occurrence in relation to other features in the area)
This emphasis parallels the importance of context in documented photo-interpretation “keys”, which are a type of human-readable ontologies for features of interest. An example of such keys are those for landforms are those of Belcher (1951) developed for interpreting geomorphological features on black & white aerial imagery. These keys provided the imagery interpreter with characteristics and spatial context knowledge with which to identify objects and their functional characteristics. Oliva and Torralba (2007) used spatial context surrounding candidate objects in an image to not simply find an object, but identify the type object in an image. One of their conclusions was that simple representations of the idea of context was one of the key problems in image understanding.

Recent work by Janowicz et. al. (2010) and Janowicz (2006) have focused on the semantics of spatial objects, with their emphasis being on what those objects “are” and how similar one object is to another from a conceptual and lexical sense. While both papers recognize the importance of space and time in determining similarity, working with detailed spatial relationships has not been prominent in their research.

Fonseca, et. al. (2006) recognized the importance of space and spatial relationships to geospatial data interoperability, but stops short of demonstrating how these can actually be incorporated into ontologies when comparing ontologies. Comparisons of ontologies (for interoperability) took place in conceptual space and focused on used of the more “classic” semantic comparisons of concepts and the geometry of the ontology graphs.
2.5.3. Spatial Arrangement of Feature Components

As important as “pattern” was in Olson’s (1960) description of image-interpretation characteristics, the concept of including spatial arrangement/context in feature descriptions appears to be limited to human-readable models for much of the mapping sciences. This is due to two reasons:

1. Most features in mapping are conceptualized as simple objects, and not as multi-part objects where individual components are discretely modeled. While there are exceptions to this, many commercial GIS have difficulty modeling complex features (features composed of other multiple features). An airport, for example, can be modeled as being composed of runways, buildings, fences, roads, and other structures. Where such a model is implemented, the data tend to be useable by a limited number of users within a specific domain.

2. Spatial arrangement and context can be highly variable, and methods to model this variability are not yet mature in the mapping sciences. Computer-vision research does address the arrangement of feature components, though these applications tend towards edge-detection and “stick-figure” object recognition needed for robotic vehicle navigation. For example, Felzenszwalb, and Huttenlocher (2005) employed edge-detection methods to find basic shapes in imagery, including face-recognition uses where the subject’s eyes, nose and lips made up nodes in simple graphs as shown in Figure 6.
They also applied these techniques to finding entire human subjects in images using what they called “cardboard-people” models. (Figure 7).

What distinguishes the use of spatial arrangement of components is that the domain conceptualizes a “feature” as being made up of multiple components, and so long as the
network graph (i.e., the stick-figure) exhibits topological isomorphism with a model pattern, then the detection is considered successful.

2.5.4. Spatial Relationships to Surrounding Features

Spatial relationships between objects has been addressed by Schwering (2008) and Schwering and Raubel (2005) in work which follows their major research themes begun under Kuhn at the MUSIL lab in Germany. Schwering (2008) presents a survey of approaches to semantic similarity for two sets of data, including network and graph theory, but her inclusion of spatial relationships was minor. Schwering and Raubel’s work (2005) included formal spatial relationships in context and similarity descriptions, examining actual instance data in the UK Ordinance Survey’s “Master Map” to develop “as-was” spatial ontologies for select linear and area feature classes. They showed that such ontologies were able to help improve similarity matches of spatial patterns in two sets of data.

2.6. Similarity

Judgments about whether objects are similar is essentially a classification task (Tversky, 1977). We compare commonality of object characteristics, evaluate symmetry of the two objects or datasets, and examine the common and distinctive characteristics of two sets of information. Our goal may be to find exact matchings but, often our goal is to judge whether two objects have enough common characteristics that we can decide that the two come from the same parent class of objects; we look for similarity. Rosche
(1978) also views similarity judgments as primarily a classification task. Attributes are used to place things into categories in a taxonomy which then allows us to examine whether these taxa are “close” or “far” from each other in that taxonomy. This semantic proximity is in concept space (versus geographic space) and leads us to qualitative decisions about degrees of similarity.

Humans are very skilled at similarity judgments, though we are only now beginning to understand the cognitive process involved enough to allow development of algorithms to make similarity decisions. To move similarity decisions into an algorithm requires that we develop numerical measures of how similar two objects are. This is important in the spatial sciences as we develop data discovery tools where large and often distributed data holdings might overwhelm the human analyst.

2.6.1. General Measures of Similarity

The simplest form of similarity determination is that of set comparisons; to what degree are the members of one dataset also found in another set? Several methods have been developed for the comparison of data samples using principles of set theory to determine the degree of commonality of members of one set to members of another set of data. The Jaccard Distance and Dice’s Coefficient, Sorenson’s quotient of similarity, and Mondford’s Index are similarity measures used in scientific sampling efforts to determine the degree of commonality in two data samples. Geographic space is not included in these measures and the techniques are not employed in the mapping sciences.
Cosine Similarity is a very common mathematic model for similarity, and has been applied extensively in comparing the contents of documents. In this application, the approach looks at the cosine of the angle between two vectors, where the vector describes the frequency of occurrence of terms between documents. Searching web-based documents and large document databases is a common application of the cosine similarity measure (Ankerst, et al, 1999). Inclusion of geographic space in the vector could make this approach viable in the spatial sciences.

2.6.2. Semantic Similarity

Determination of similarity of word meanings (i.e., semantics) is an extensive research area in linguistics and has been greatly helped by the development of the “WordNet” lexical database at Princeton University (2011). This database may be thought of as a large thesaurus-like environment in a relational database, which links words (i.e., “concepts”) to like words. The system is built around the idea of synsets (cognitive synonyms) but also addresses what it calls “senses” of words which are the different ways that communities of interest use the same word. It also includes such relations as hyponymy (super-subordinate relationships). Approaches to (linguistic) semantic similarity in WordNet are often based on edge-based schemes (how many edges are traversed when going from one word to another in Wordnet’s taxonomy of concepts), or node-based comparisons of word attributes (words are considered nodes in the WordNet taxonomy). The relative position of a word within the WordNet taxonomy was used by Banerjee and Pedersen (2002) to determine the degree of similarity in concepts. Jiang and Conrath (1997) combined the two approaches and compared them
with human-based similarity judgments and found improvements over use of separate similarity metrics (node versus edge approaches). Leacock and Chodorow (1998) were able to improve on similarity judgments by inclusion of surrounding words as a context metric, using nearby words in the Wordnet taxonomy to refine similarity judgments.

In the geospatial sciences, Schwering (2008) proposed a very useful framework for geo-semantic similarity, suggesting 5 different approaches or “models” for comparing two datasets:

1. Geometric Model: comparison of spatial data classes rendered in conceptual graphs
2. Feature Model: comparison of spatial data based upon attributes (“features” are considered to be attributes by Schwering (2008))
3. Network Model: comparison of two networks of words (e.g., WordNet), suggesting shortest path measures for traversing the two networks beginning with a common term.
4. Alignment Model: Comparing the relative position of concepts in two taxonomies
5. Transformation Model: How much effort is involved in “deforming” one idea into another. The two sets of ideas are typically rendered in graphs/networks and a measure describing the divergence of one graph from the other (at the node-to-node comparison level) is evaluated.

These approaches were combined with use of a semantic network (Schwering and Kuhn, 2009) to show improvements in data retrieval for hydrographic data.

Kavouras and Kokla (2008) present a framework of similarity types, including the Geometric and Feature models which Schwering (2008) suggests, but then proposing an
Edge-counting method for a shortest path comparison between like concepts in the two ontologies, and then an Information Content measure. This later measure examines the degree of overlap in the information contained in the two ontologies.

Li (2010) suggested a three-type approach to similarity which was organized as:

1. Edge-counting: Compares the edges of two ontologies rendered as conceptual graphs
2. Information Theory/Content: Largely an examination of the attributes for each node and edge of the graph to assess the richness of information in the two graphs.
3. Feature Based: Comparisons of “features” (a.k.a.: attributes) of the objects portrayed in the two ontologies.

While Schwering (2008), Li (2010), and Kavouras and Kokla (2008) all focused on semantic similarity as applied to the geospatial sciences, none included geographic space as an element in their similarity approaches. The “Geometric” model of Schwering’s work was the geometry of a conceptual graph of concepts, and not the geometry of the physical objects in the world. Li allowed for geometric and spatial measures in his algorithm, but these were secondary to his focus on the logic in the semantic search engine.

Research in semantic similarity often employs taxonomies (i.e., hierarchical ontologies) but many domains use concepts which are related in non-hierarchical graphs. Maguitman, et. al. (2005) suggest a more generalized algorithmic approach to semantic similarity which employs non-hierarchical networks (i.e., graphs) to evaluate similarity
amongst different ontologies. Also, Li (2010) uses neural networks which do not require hierarchical graphs to measure similarity. While both approaches hold more flexibility for similarity measures, neither employs spatial relationships to any degree as part of the similarity model.

Many other researchers have also examined ontology-based similarity, including efforts in the medical profession for genomics and for medical records searching (Lee, et al., 2008) as well as in the information and geospatial sciences (Rodriguez and Egenhofer, 2003, Thiagarajan, et al., 2008). Adams, et. al. (2010) demonstrated ontology comparisons as a key step in conflation to avoid category mis-matching, and some use of spatial relations for general alignment of two data sets. Ahlqvist (2005) examined the use of ontologies in detecting land cover changes through changes in an ontology over time. This work focused on changes in the thematic categories, but did not incorporate changes in spatial extents of the vegetative map units. Janowicz (2005) showed how the addition of thematic roles to each element within an ontology can extend the ability to perform similarity measurements.

As important as all of these research efforts have been (and continue to be), the inclusion of geographic space and spatial relationships has not been prominent. Most of the ontology-based similarity research has focused on the comparisons of the concepts and concept definitions using descriptive definitions and lexical clues to determine how similar two groups of data are. Space and spatial relationships are seldom considered in such semantic comparisons.
2.6.3. Spatial Similarity

When space (2 or 3 dimensional) is added to the question of similarity, the inclusion of geometric and/or topological metrics are required. This is clearly seen in the geospatial sciences in the form of registering one set of data to another. With image-to-map registration, image-to-image registration, and map-to-map (or data-to-data) registration, affine transformations are often used to achieve a “best” fit between the two data items. Most often this involves matching accordant data points with scale, translation, and rotation of the two coordinate systems being addressed. This method is appropriate for registering two sets of spatial data, but is not normally considered as a “similarity” metric. Three of the better approaches to spatial similarity are matching two graphs, matching two scenes of data, and matching (or recognizing) shapes.

Holt (1999) suggested that spatial similarity is governed by the following factors:

1. Context: Frame-of-mind, spatial relationships
2. Scale: spatial
3. Repository: application area, local domains
4. Techniques: available technology
5. Measure and Ranking Systems: available taxonomies (a-priori ontologies)

He presents a GIS experiment using “case-based-reasoning” (CBR) which is an approach from psychology which allows for evolution of both the ontology and the similarity reasoning algorithm, using new situation data fed into the algorithm. Spatial relationships were clearly demonstrated in Holt’s ontologies and allowed some improvements in spatial reasoning on vector data. CBR-based systems actually began in
the legal profession (Web-Lee, et. al., 1997) and have been applied in the medical field (Begum, et. al., 2011) for improvements in diagnosis efforts. The application of CBR techniques to the spatial sciences has been limited, though Holt and Benwell’s (1999) work showed improvements in soils classification using spatial relationships.

2.6.4. Graph Matching

When data are rendered as “graphs” (nodes connected by edges in a network structure), techniques of graph matching can be employed to numerically describe the level of similarity between graphs. Graph matching is often viewed as either “exact” or “inexact” (Bengoetxea, 2002), with several types of measures employed for each. Figure 8 depicts the general types of graph matching techniques. It is important to understand that “exact” refers to isomorphism of the compared graphs, typically in a topological sense and not the exact geometric shape of the graph. Inexact matching means that one or more nodes (vertices) of the compared graphs could not be matched.

![Graph Matching Diagram](image)

Figure 8: Types of graph matching (Bengoetxea, 2002).
Graph matching has been shown to be useful in comparing groups of objects in an image, with objects depicted in line maps for the purpose of determining whether terrain objects could be used to align imagery to a map (Wang, 2009). This involved the use of attributed graphs and showed that sub-graph matching could be used to register image-based data with map information.

2.6.5. Scene Matching

Finding matches in spatial patterns of sets of data in two separate images is an important research area since it looks to formal, machine-readable descriptions (i.e., “spatial ontologies) to determine whether two “scenes” can and should be spatially aligned. The 9-intersect model was shown help to determine similarity of two features (Bruns and Egenhofer, 1996) suggesting that topological, directional and geometric similarities between objects could be used to quantify a similarity measure, thus aid in deciding whether two sets of data ought to be integrated or represent the same set of objects from two perspectives. Nedas and Egenhofer (2008) used actual maps over the University of Maine and tried to (spatially) match imagery of the area to the same coordinate system. While much of their work was subjective, they introduced the idea of “relaxation” of matching constraints to scene matching, an idea important to handling variation of data sources. Scene-matching work by Stefanidis et. al. (2009) continues to refine and evolve the geometric and topological metrics in matching scenes and, while using imagery as a key test input, uses an approach where vector spatial data from any
source may be used. This work is also designed to address multiple feature classes in the data being matched; a significant improvement over previous research.

While these efforts in scene-matching add valuable methods for formalizing spatial measurements, they all take a data-instance-specific approach; they compute spatial matches between each source beginning with new measurements. None of them use or store a semantic “pattern” against which data are compared. This is reasonable since scene matching undertakes to look for spatial alignment between data sets, and is not intended to be used in data discovery where pattern recognition could be important. The formality (topologically and geometrically) of these scene-matching efforts, together with the multi-source capability introduced by Stefanidis, et. al. (2009) does make scene-matching a good framework to begin with.

2.6.6. Shape Matching/Recognition

Heitz, et. al., (2008) used a “LOOPS” approach to shape recognition. This was an effort to detect object instances in imagery based upon shapes described through significant points. Shapes could be deformed. Spatial arrangement..i.e., the shape of the points was considered, but conditions proximal to the shape (as seen in imagery) were not considered. In ”Learning Spatial Context: Using Stuff to Find things” (Heitz and Koller, 2008) “stuff” is defined as areas in an image which have repetitive fine-scale patterns (like road surfaces). “Things” are objects with specific size and shape (like geospatial features). This research attempted to refine the detection of “things”, using the context of candidate things within image fields of “stuff”. They do call out thing-to-
thing context, and that spatial relationships there are important, but this was not part of the research.

A modified version of the Hausdorff distance (discussed earlier in section 2.5.2) has been applied to shape recognition as part of a machine-readable model which would more quickly detect shape similarity (Yu and Leung, 2006). This work in character and logo recognition was working towards a most-efficient algorithmic approach, but did not attempt to include shape metrics in an object ontology. The Hausdorff metric has recently been used to aid in identifying candidate matches at the feature level between two feature instances for the purpose of dataset conflation (Li, 2010). This was done in a dataset-to-dataset comparison and showed improvements in the conflation effort, but the metric was not included in any sort of ontology.

Stefanidis, et. al. (2002) extended the work by Agouris, et al. (1999) in shape matching for scene understanding. The later work moved from single objects, to include the configuration of multiple objects within a scene. Direction and object orientation are combined with position and size to create a numerical description of how the important multiple objects in an image are configured. This was applied to the problem of image retrieval where an image might contain a group of objects which are arranged in the manner of interest to the user. In a sense, this work was trying to retrieve an image from a data store which “looked” like another image or a map of an area of interest. While not immediately useable on feature ontologies, the mathematical models proposed in the work hold promise for encapsulating spatial metrics in the ontologies of feature data.
2.6.7. Feature Recognition

Feature recognition differs from scene matching and shape recognition primarily in the nature of the variability of features. Scene and shape matching techniques look mainly for exact matches in geometry, though sometimes obscuration and different visual perspectives (translation, rotation, shearing) are accounted for. The assumption is that the two scenes or the two shapes are basically of the same geometry and that the matching process need only to account for different user perspectives and/or obscured portions of the two datasets, and a match will be successful. It asks the question: “…does this set of geometry match that set of geometry…?”

This is not just a concern in the geographical sciences. Automation in manufacturing often depends on recognition of shapes for drilling, cutting, milling and assembling processes. Brunetti and Grimm (2005) discuss several methods for explicite representation of shapes using different levels of detail and abstraction in object descriptions. They also refer to physical objects as “features” in the same manner as for GIS’, though their application (manufacturing) benefits from features being much more geometrically regular than in the mapping sciences.

Feature recognition, on the other hand, attempts to match an observed set of properties, to a model of a prototypical object in that class of features. The matching process involves using some sort of “pattern” or template model, against which new observations are compared. Variability must be accounted for in the way that a feature class appears on the earth, and an observed “thing” must be compared with a prototypical
definition and similarity computed. Feature recognition methods conclude a match when that similarity metric reaches a certain threshold.

2.7. **What Gap Areas Will This Research Address and Fill?**

The research reported in this paper builds on all these areas, with a particular focus on the spatial configuration of a “thing” (how it exists on the landscape), and how it relates spatially to objects around it. This is a new area for research which will extend the framework of research on a prototype spatial context model using real-world feature as calibrating inputs. A recent article by Shi (2011) complained that too many recent “geospatial ontologies” failed to include the notion of space in their treatment of geospatial objects.

“If spatial ontology has nothing about spatial relationships but only deals with the conceptual matchmaking through logic, then the disciplinary identity of spatial science is missing...” (Shi, 2011)

The spatial context approach of this research, while limited in scope, will help future investigators to determine the spatial characteristics of a context model which are machine-readable and can then be engineered into systems where algorithms, and not just human analysts, can look for geospatial patterns in large volumes of data.
CHAPTER 3: Methods

3.1. The Approach

The approach used first determines a spatially-based prototypical ontology (a model) for a geospatial feature class. This prototypical ontology is built using observed metrics of known instances of that feature class. The metrics include both spatial and a-spatial measures, each described as mean and standard deviation. This prototypical ontology is then compared to the ontology of newly observed features to the prototype. The comparison of the two ontologies is numerically-based, and generates a score which quantifies the degree of similarity between the new observations and the prototype ontology for a feature class. This provides a quantified description of how similar the new observations are to the prototypical feature class.

This study uses the complex feature class “Container Terminal”, which consists of several simple “component” features functioning together as a single facility for cargo processing at maritime ports. A more detailed description of this feature class is found at Appendix A.
3.2. A Workflow

The study can be viewed as a generalized workflow as shown in Figure 9. In this flow, we begin with selection of sample features and work towards a “type” ontology which includes spatial context. A basic feature definition which includes attributes and basic descriptions of each component of the complex class, is appended with spatial metrics. Basic spatial metrics of each component are captured and then processed into the contextual spatial/topologic measures needed to describe the complex feature class. This involves descriptive statistics of each measure for the sample group of feature instances. These measures are then added to the complex class as descriptors of spatial context. These descriptors and metrics in the expanded class model are considered a “type” ontology for that class, and now serve as a pattern for similarity comparisons with new observations (i.e., data believed to contain instances of the feature class). This type ontology is considered an attributed conceptual graph as a pattern for matching with candidate observations.

The same spatial metrics are collected for candidate features in source data and converted to the same structure developed using the training samples. The two graphs (pattern ontology, and the ontology graph from the candidate observations) are compared using a similarity algorithm and generates a similarity score.
Figure 9: The study effort shown as a general workflow.

3.3. Geospatial Information

3.3.1. Feature Class

The research uses container terminals, a type of commercial maritime port facility, as its feature class. Container terminals, an example of which is shown in Figure 10, are one of several key types of cargo facilities at maritime ports, and are designed to process and move cargo in and out of that port facility using standardized, metal containers for
various types of cargo. These terminals are characterized by specialized, purpose-built equipment and structures (e.g., cranes, loaders, storage areas) to optimize the movement of cargo containers between land transportation networks (road and rail), and maritime cargo ships. Container terminals are present in almost every major port city around the world, and their use is expanding with the globalization of many nations’ economies. Container terminals were selected for this study because of their large size (and therefore ease of detection on commercial imagery), relative ease of obtaining supplementary information about the terminals (from commercial Ports Authorities), and high degree of standardization in facilities due to the standardization of the cargo containers.

A more detailed description of these terminals may be found in Appendix A.

This study examines the spatial arrangement of the components of a complex feature class as a potential “signature” for that type feature. A complex feature is a geospatial feature composed of other, simple features. As shown in Figure 11, an airport can be viewed (or “conceptualized” as ontologists like to say) as being composed of runways, taxiways, buildings, roads, and fences. Each of these structures is considered a “component” which, when combined with other components in various types of relationships, function together as an airport.

The relationships between components may include parent/child relationships, peer-to-peer relations, and spatial arrangements such as adjacency, overlap, proximity, etc.
The proposed approach accommodates these types of relationships, with a focus on spatial relationships, while still allowing for other types of relationships amongst the components of a complex class. This has several advantages for modeling a (spatially-enhanced) ontology:

1. Classes may be defined as composed of many different types of component classes.
2. Metrics related to spatial context may be specifically designed for the abstract-level of class, and the user is not obliged to use these measures at the component level.
3. Component objects may be “re-used” by any number of abstraction classes. A communications tower, for example, could be a component of a cell phone system, or a component of an airfield. Either complex feature class could re-use the ontology of the simple feature class “tower” in their ontology.

Complex objects are an important conceptualization and are how we view much, if not most, of the world. We talk about “airports”, not runways, for travel. When we drive to work, we use a “car” and not a list of component parts. It is important, then, that we have ways to model these complex objects and communicate their existence to each other; we need ontologies for them.
Complex features do not exist as unlimited spatial arrangements of their components; wheels are always positioned near the corners of a car, airport runways never run through the center of a building. These and other spatial arrangements need to be included in ontologies as local context, in functional topology terms. This provides bounding on how components might occur, and will help define variability in the spatial metrics within the feature ontology.

3.3.2. Study Areas

The World Port Index (WPI) was used as the master list of ports to bound the selection of sites for this research. The WPI publication (National Geospatial-Intelligence Agency, 2011) lists over 3,700 maritime ports world-wide, and is updated on a monthly basis. This listing is intended as a navigation aid for mariners in general, and is not limited to ports with particular types of commercial terminals.

Of the 3700+ ports listed in the WPI (Oct 2011 edition), 22 ports were selected for study. These included a mix of large and small ports across the world, representing most continents and a variety of cultural situation. The assumption was that different cultures probably had variations in the way that container terminals were constructed, and that this variability should be included in the feature ontologies built from these sites. The study sites were also picked to include some of the largest container terminals based on cargo throughput each year. Table 1 provides a listing of the 22 selected ports.
Table 1: List of training and test sites for the study.

<table>
<thead>
<tr>
<th>Training Site Location</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busan</td>
<td>Korea</td>
</tr>
<tr>
<td>Cape Town</td>
<td>South Africa</td>
</tr>
<tr>
<td>Durban</td>
<td>South Africa</td>
</tr>
<tr>
<td>Fremantle</td>
<td>Australia</td>
</tr>
<tr>
<td>Hamburg</td>
<td>Germany</td>
</tr>
<tr>
<td>Jawaharil Nehru</td>
<td>India</td>
</tr>
<tr>
<td>Long Beach</td>
<td>United States</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>United States</td>
</tr>
<tr>
<td>Luanda</td>
<td>Angola</td>
</tr>
<tr>
<td>Napoli</td>
<td>Italy</td>
</tr>
<tr>
<td>New York</td>
<td>United States</td>
</tr>
<tr>
<td>Newark</td>
<td>United States</td>
</tr>
<tr>
<td>Pasir Panjang (Singapore)</td>
<td>Singapore</td>
</tr>
<tr>
<td>Port Kelang</td>
<td>Malaysia</td>
</tr>
<tr>
<td>Puerto Barrios</td>
<td>Venezuela</td>
</tr>
<tr>
<td>Puerto Cabello</td>
<td>Guatemala</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Santos</td>
<td>Brasil</td>
</tr>
<tr>
<td>Shanghai</td>
<td>China</td>
</tr>
<tr>
<td>Taranto</td>
<td>Italy</td>
</tr>
<tr>
<td>Vishakhapatnam</td>
<td>India</td>
</tr>
<tr>
<td>Vladivostok</td>
<td>Russia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Site Location</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamburg</td>
<td>Germany</td>
</tr>
<tr>
<td>Wilmington, California</td>
<td>United States</td>
</tr>
<tr>
<td>Long Beach, Califonia</td>
<td>United States</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>

Twenty two sites were selected as “training” sites, and four sites were selected as “test” sites. Training sites were used for construction of the prototype ontology and the
test sites were used to test the similarity model. These were selected from 22 separate port areas around the world (Figure 12).

![Study Sites Map](image)

Figure 12: Study site locations.

The largest ports were considered, as well as smaller container terminals from around the world. These largest ports are expressed in terms of container throughput, with “TEUs” (Twenty-foot Equivalent Units) describing a 20-foot-long cargo container box, which is the international convention for measuring the number of containers passing through a commercial port each year. The most common container boxes (often seen loaded on a truck or railcar), is 40 ft. in length, or 2 TEUs. Detailed views of the 22 training sample sites may be found in Appendix B.
Only one container terminal was selected at each port city; in many cases a major port area has several container terminals, but only one was selected at each city. For example, Hamburg, Germany hosts 4 separate container terminals, Newark, New Jersey has 3, and the port city of Singapore is the home to 6 separate terminals. Using a single terminal from each city allowed for a broader sample of cultural differences in port configuration, while still making for a manageable data collection effort.

Input from persons with expertise in ports and harbors was helpful in determining the spatial metrics of the layout of port facilities. This helped define the component features which need to be included in the feature ontologies (e.g., number and type of structures, their characteristics, etc.), and also the spatial arrangement of these components (e.g., how far apart are the structures?, are certain structures always adjacent to others?, etc.). The subject matter experts also provided some of the business rules and limits for the spatial metrics for the ontologies.

Commercial satellite imagery was used in collection of geometries of the feature components. Attributes of the features were derived in part from the imagery, and supplemented with attribution from open sources on the internet. SME input was obtained primarily through interviews and workshops, supplemented by SME reviews of documents and of the ontologies themselves.

3.3.3. Imagery and Image Processing

Panchromatic and multi-spectral imagery acquired by the three DigitalGlobe commercial imaging satellites (Quickbird, Worldview-1, and Worldview-2) was used as
the primary source of information for collecting the geometries of the features of the study.

For most of the 22 study sites, panchromatic imagery was sufficient to collect the feature data but for some areas, pan-sharpened multi-spectral imagery was used. In these cases, MSS bands 2,3,&4 were combined to synthesize a true-color image and then sharpened with ~.5 m GSD pan imagery. These images were displayed in a GIS (ArcGIS) in their correct geographic locations, where feature geometries were collected.

The only pixel processing done on each image (aside from pan sharpening) was contrast and brightness adjustments, and cubic-convolution filtering during display. These steps were taken only to improve image display during digitizing, and did not alter the original spectral values of the pixels in the imagery.

In addition to imagery from DigitalGlobe, oblique and ground photography of selected port areas was used to verify feature structures at some terminals. For some port cities (e.g., Rotterdam, NE, New York, USA) Google-Maps makes such imagery available for viewing in a web-browser. While these type images were helpful at some port areas, such imagery is not widely available and could not be used as a primary source; it was, however, helpful in confirming the structures at some ports. Ground-based photography was occasionally available from the Ports Authorities websites and also was used to verify structures at selected ports. Port area maps were also available from some Ports Authority sites and helped to determine the extent of some terminals.
These data sources (oblique and ground images) were not consistent in extent or quality, and could only be used to supplement the commercial satellite imagery from DigitalGlobe.

3.3.4. Digitizing Feature Data

Three different groups of feature data about container terminals were collected or synthesized:

**Training Data**: Twenty two different container terminals from various locations around the world were collected from satellite imagery. These data were used to develop the prototype feature ontology, and are referred to as “Training Data” or “Training Sites” throughout this document.

**Test Data**: Four additional facilities were collected from satellite imagery, and used to test whether the similarity algorithm could discriminate between container terminals and areas which were other types of marine terminals.

**Synthesized Data**: Due to the small number of “test” sites, 40 additional sets of metrics representing 40 synthesized test sites were created. Actual geometries were not collected for these data, but instead, the metrics were created synthetically using the statistics (mean, standard deviation, probability distributions) of the sites actually measured.
The geometries of the structures at the training and test facilities were collected via manual digitizing in ArcGIS using DigitalGlobe imagery (panchromatic) and pansharpened multi-spectral imagery (also DigitalGlobe). The imagery files are delivered to customers in UTM coordinates by DigitalGlobe and the imagery was not re-projected in this study.

No re-projection of the imagery or the resultant feature geometries were performed during any step of the analysis, i.e., all angles and distances used for expression of spatial relationships were based on the UTM coordinates in the original source imagery. Collected feature geometries and spatial metrics were left in their original UTM-based coordinates for the following reasons:

1. The spatial relationships used were dependent upon angular measures, which are preserved in the (conformal) UTM coordinate system

2. Distances used for the feature geometries were small (less than 4 km as mentioned above), and remained entirely within a single UTM zone. This meant that the scale variation would only change by 0.9996 at worst (scale distortion at the central meridian of the UTM zone)
3. The error introduced by the imagery was larger than the largest distance distortion over the largest anticipated facility (~ 4 km square) from the scale distortion in a given UTM zone (4,000m X .9996 = 3998.4 or 1.6 m distortion of distance metrics for a given port area).

The largest source of error was the circular error of the original imagery, and not the distortions introduced by the UTM coordinate system, data collection, or other data processing. No angles or distances between sites from different harbor areas (i.e., across UTM zone boundaries) were used. In other words, all spatial relationships between feature components were based on local geometries.

During digitization in ArcGIS, the (GIS) map file coordinate system was set to the UTM zone of the image being used; all data collection (of the feature geometries) was performed in the coordinates of that UTM zone. Once the "raw" geometries for the features were obtained, these data were moved into MATLAB for subsequent generation of spatial metrics which ArcGIS did not compute. Calculations in MATLAB were based on plane, Euclidean geometry. Geometric calculations in MATLAB used meters as the unit of measure; these were the units-of storage in the ArcGIS geodatabases for the port areas.

Digitizing was organized around each of the 22 training sample sites, with all geometries and attribution for a site collected into a single ArcGIS personal geodatabase (GDB). Keeping each site in a separate GDB was done to make subsequent processing in MATLAB easier.
Geographic boundaries of each terminal were pre-defined for each site in order to limit the collection effort for certain components (buildings, apron, storage yards, container stuffing facilities (CSFs) and cargo transfer areas (CTAs)). The collection of shorelines, roads, and railroad tracks was allowed to go beyond the defined site boundaries.

All digitizing was performed manually, using the definitions outlined in the ontology at Appendix A. To help ensure consistency in data collection, the feature data for each site was inspected by a second person. If corrections were made, they were done so in consultation with the person who did the original collection.

Feature data in ArcGIS were then imported into the MATLAB environment where additional metrics were generated. This was done because computations of some spatial relationships were more efficient in the generalized MATLAB application, versus a GIS.

3.4. Feature Ontology

3.4.1. Spatial Context in the Feature Ontology

The idea of “context” in verbal expression is crucial to understanding (Rosche, 1978). In language, context is considered the words and phrasings which surround a word, which contribute to its meaning. The same principle can apply to terrain features: the “meaning” (or use or function) of an object on the terrain, can be determined in part by the objects around it. In remote sensing, this is referred to as “association” (Lillesand, et. al., 2007) and is a key discriminator in visual interpretation of aerial photography. We have a prototypical pattern for a feature class which we learn and then use to decide when
we’ve seen that class on an image. What this approach does is instantiate such a spatial pattern as part of a machine-readable ontology.

There are a large number of spatial metrics which could be included in such ontology. These generally fall into two categories: geometric measures, and topologic measures. Geometric measures include those which describe the location; size, orientation, and shape (rectangularity, elongation) are readily described in numerical terms (meters, angular degrees, etc.) Topological measures attempt to describe relationships such as adjacency and connectivity, and are usually described as “true” or “false” for a given relationship.

This research employs geometric measures, but also suggests that “Contextual Topology” in the form of discrete metrics (versus true/false conditions) are important in understanding features. Contextual Topology, as used here, can be thought of as a hybrid between geometric and topologic measures. It includes descriptions of such things as relative sizes and distances, the arrangements, proximity, adjacency, and juxtaposition of things on the terrain.

Expressions of spatial context will differ depending upon geographic scale. For example, at the global level, the Mississippi River runs north-to-south through the central portion of the North American continent. At the national level we can describe its course as running along the edge of certain states. At the local level we can describe the river’s spatial context as being between certain counties. As there are many levels of scale, so too there are many levels of detail in describing spatial context. For this approach, we are focusing on the large-scale spatial context where the geometry and
topology of components of a single feature class need to be described. This might also be called “local” context, and is intended to describe the arrangement of “nearby” features on the terrain. The definition of “nearby” depends as much on the functional influence of the feature, as it does on geometric distances.

A distance of 100 meters between an elementary school and a residential area means something very different than that same 100 meter distance to a munitions factory. Spatial relations in an ontology must be selected with function in mind and then adjusted for significance to that feature class.

Inclusion of spatial context in an ontology is important not only to understand how a feature can occur on the terrain, but also gives us additional metrics and attributes which can be used in feature searches and analysis.

3.4.2. Ontologies as Conceptual Graphs

Ontologies can be rendered in many forms; this research will use attributed conceptual graphs to portray the ontology of a feature class. While the focus of the research will be on spatial metrics, the use of graphs will allow the inclusion of both spatial and a-spatial information about a class. This will set the stage for future research where other aspects of a feature (temporality, other non-spatial relationships to other feature classes, etc.) may be incorporated. The use of conceptual graphs also means that the similarity determinations can be based on algorithms which may consider conceptual, versus geographic distances in comparing the graphs. Conceptual graphs are simply a more general environment for comparing ontologies, and an environment amenable to computational methods.
3.4.3. Developing the Basic Ontology

In order to facilitate data collection, and to provide a beginning framework for development of a machine-readable ontology, a “starter” description for container terminals was developed. This task can be considered as the “knowledge engineering” portion of the study.

Port Engineering Documents, interviews with experts in port operations, examination of aerial views of known terminals, and information obtained from Trade and Industrial Journals were used to identify the key functions of a container terminal, and the key structures (components) involved in terminal operations. This resulted in a human-readable ontology (Appendix A) which described the major functions and how these functions appeared in the physical components of a typical container terminal.

The port engineering documents (Agerschou, et. al., 2004, Tsinker, 2004) were helpful in identifying the range of structures and key equipment used at these type facilities, but focused primarily on detailed engineering standards and specifications at the individual structure level (e.g., materials standards, construction methods, etc.). These documents did address site layout to a limited degree and identified key functions and structures, but described facilities layouts as being heavily site-specific.

Several known container terminals were selected from cargo volumes data indicating some of the larger terminals from around the world (AAPA, 2010). Port areas were examined using Google Earth/Google Maps imagery to get a sense of how many of the larger terminals tended to be built, types of structures, facilities layouts, and
whether there might be regional differences in port construction. This non-rigorous examination of select facilities did provide insight into the variety of spatial arrangements used in actual instances of the terminals. It also allowed an initial sampling of some of spatial metrics (distances between structures, sizes of facilities, etc.) which would be considered in the final ontology.

Discussions with subject matter experts in port engineering also provided important insights into terminal layout, and served as a key “grounding” mechanism for the ontology. Discussions centered on the review of a first draft of a container terminal ontology, prepared using inputs from engineering documents, examination of aerial images over known terminals, and trade journals. They provided input in the form of suggestions for additional metrics, ranges for the values of some metrics, what structures were optional for terminals, and reasons for some of the regional variability in facilities layouts.

Combining all of these inputs resulted in an initial human-readable ontology (Appendix A) which was then used to construct the initial machine-readable ontology (i.e., a GIS feature definition) in the form of a personal geospatial database (GDB) into which geometries and other metrics for instances of container terminals could be collected.

Three important points should be noted from this initial ontology-construction effort:

- The effort involved both quantitative and qualitative examination of the features
• Considerable effort was made to “ground” the resultant ontology (Kuhn, 2005) in the important functional characteristics of a container terminal

• The resulting ontology was not final, and served only as a starting structure into which a sampling of metrics would then be loaded from a more rigorous data collection effort over known terminals

3.4.4. From Function to Form

For most physical objects, their form or structure is driven by function. Plants grow upright for a reason, wheels are round for a reason, steps on a stairway are a certain height and width for a reason. When architects and engineers develop a plan or blueprint (i.e. an ontology), they blend the key functions for an intended structure, together with materiel and ergonomic constraints to develop a physical design.

A building architect works from function to form, but it is also valuable to work from form to function. Given a feature ontology (the form), we use the ontology as a “pattern” to find instances of the feature in new observations.

The suggested approach takes advantage of this principle to determine which measures of spatial context are important to include in an ontology. Many geometric and topologic conditions can be measured in a GIS for a feature; but which of these are important for describing that feature class? The particular metrics of contextual topology which should be included in the ontology must be related to key functions of the feature class. While the height of a building may be easy to measure, we should only include that metric if the height tells us something about the purpose of the structure. Ideally, if a contextual metric reflects a function which is diagnostic for the feature class, that metric
can be part of a litmus test in the ontology for whether or not an instance of that class is present in a set of observations. A distance of 0 meters between a building and a parking lot could indicate that the parking lot is used by the building’s occupiers, but a distance of 1000 meters might indicate that the two structures are unrelated.

The key point is that for most man-made structures, the old adage of “form-follows-function” applies. Said another way, spatial arrangement (contextual topology)–indicates–meaning.

The proposed approach will examine the key functions of the selected feature classes and select spatial measures which are reflections of those key functions. These measures will be added to the ontology for that feature class as the spatial aspect(s) of that class’ ontology. This ontology will then serve as a prototypical pattern against which new instance data (new observations) are compared. If the ontologies are similar enough to be called a “match”, then the assumption is that the functions of the features, and therefore the feature classes themselves, also match. Function is capture in the form of spatial metrics in the ontology, which is then used to determine the function (and therefore the feature class) in new observations.

3.4.5. The Nature and Structure of the “Prototypical” Ontology

The set of metrics selected for collection and analysis are shown in Table 2. These are the metrics initially selected after consultation with the literature and with subject matter experts. This list was reduced for the final ontology as a result of statistical analysis which detected colinearity amongst several of the metrics. That analysis is
discussed further in section 3.2.5. A more detailed table describing the relationship of each metric to the important functions at a container terminal may be found at Appendix A.

Table 2: List of initial metrics. Blue shading indicates metrics about components; green shading is for metrics about relationships between the feature components.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apron Confidence</td>
<td>Confidence of proper identification</td>
</tr>
<tr>
<td>Apron Width</td>
<td>Width of loading apron (meters)</td>
</tr>
<tr>
<td>Apron Length</td>
<td>Length of loading apron (meters)</td>
</tr>
<tr>
<td>Apron l/w ratio</td>
<td>Ratio of length to width</td>
</tr>
<tr>
<td>Apron # Cranes</td>
<td>Number of permanent loading cranes on the apron</td>
</tr>
<tr>
<td>Yard Confidence</td>
<td>Confidence of proper identification</td>
</tr>
<tr>
<td>Yard Size</td>
<td>Size of storage area (square meters)</td>
</tr>
<tr>
<td>Yard Longest Axis</td>
<td>Length of longest axis of storage yard (meters)</td>
</tr>
<tr>
<td>Building Confidence</td>
<td>Confidence of proper identification</td>
</tr>
<tr>
<td>Number of Buildings</td>
<td>Count of number of non-CSF buildings at the terminal</td>
</tr>
<tr>
<td>Largest Bldg Size</td>
<td>Size (square meters) of the largest non-CSF building</td>
</tr>
<tr>
<td>Avg Bldg Size</td>
<td>Average size of non-CSF buildings</td>
</tr>
<tr>
<td>Cargo Transfer Area (CTA) Confidence</td>
<td>Confidence of proper identification</td>
</tr>
<tr>
<td>CTA Longest Axis</td>
<td>Length (meters) of longest axis of the Cargo Transfer Area</td>
</tr>
<tr>
<td>Number of CTAs</td>
<td>Count of the number of Cargo Transfer Areas</td>
</tr>
<tr>
<td>CSF Confidence</td>
<td>Confidence of proper identification</td>
</tr>
<tr>
<td>Number of Container Stuffing Facilities (CSFs)</td>
<td>Count of the number of Container Stuffing Facilities</td>
</tr>
<tr>
<td>CSF Width</td>
<td>Width (meters) of each Container Stuffing Facility building</td>
</tr>
<tr>
<td>CSF Length</td>
<td>Length (meters) of each Container Stuffing Facility building</td>
</tr>
<tr>
<td>Apron-Yard-CTA Angle</td>
<td>Angle (degrees) described by the Apron, Storage Yard, CTA centroids</td>
</tr>
<tr>
<td>Yard/Building Distances</td>
<td>Average distance (meters) from the center of the storage yard, to the</td>
</tr>
<tr>
<td></td>
<td>buildings at the terminal</td>
</tr>
<tr>
<td>Yard/CSF Distance</td>
<td>Distance (meters) from the center of the storage yard, to the center of the</td>
</tr>
<tr>
<td></td>
<td>CSF. Distance is an average where more than one yard or CSF occurs.</td>
</tr>
<tr>
<td>Yard/CTA Distance</td>
<td>Distance (meters) from the center of the storage yard, to the center of the</td>
</tr>
<tr>
<td></td>
<td>CTA. Distance is an average where more than one yard or CTA occurs.</td>
</tr>
<tr>
<td>Yard/CTA Angle between Major Axes</td>
<td>Angle described by the major axes of the yard and the CTA</td>
</tr>
<tr>
<td>Road Yard Intersects</td>
<td>Number of roads intersecting the storage yard(s)</td>
</tr>
<tr>
<td>CTA-Apron Distance</td>
<td>Distance (meters) from center of CTA, to center of apron</td>
</tr>
</tbody>
</table>
Figure 13 shows the container terminal ontology as a graph. This is a conceptual graph in that the positions of each node and edge are ordinated in “conceptual space” and not in geographic space.

As discussed in section 3.1.3, this feature is a “complex” feature, with each node being a component of the feature. The full container terminal feature is shown in the graph as a shaded polygon in the background (the polygon is a visual indicator in the figure only, and is not used as a part of the feature ontology itself). Nodes and edged which fall entirely within this polygon are components of the container terminal feature class. Those nodes and edges which fall at least partially outside the shaded polygon are other parts of the graph which were used in the similarity model as contextual relationships for the container terminal as whole. Note also that the graph allows for more than one node for several components (e.g., Aprons and Yards).

This is not a complete graph and not all possible edges are represented; only those edges which reflect spatial node-to-node relationships are shown. Most of the edges are “simple” edges reflecting a relationship between only two nodes. One of the edges (the one showing that a storage yard should lie between the apron and a transfer area) is considered a “hyper-edge”, which is an edge which touches two or more nodes in the graph. This type of edge posed challenges to the study since most GIS applications can only express relationships between two spatial objects. As such, a mechanism needed to be designed and implemented outside of a GIS to address this. In this case, a method in MATLAB was developed which examined the angle described by the two vectors:
Apron<>Yard, and Yard<> Transfer area, using the centroids of each component as the endpoints of the vectors. If the angle was less than 90 degrees, then

![Diagram of Container Terminal Ontology](image)

Components within the blue shaded area are considered part of the complex feature "Container Terminal"

Figure 13: Container terminal ontology shown as a graph.

the storage yard was considered to be “between” the CTA and Apron, and satisfied the functional requirement of that edge (for cargo movement). A typical layout for a container terminal, and how each metric appears in a notional terminal is shown in Appendix A
3.5. Numerical Analysis

Metrics data were then exported into Excel spreadsheets for inspection and final quality control before statistical analysis. These analyses were intended to address two areas or questions:

- What is the relative importance of each metric in the similarity model?
- Is there any redundancy (i.e., colinearity) amongst the metrics?

These analyses used only the 22 training sites and not any test cases which would be examined in the later similarity scoring.

3.5.1. Sensitivity Analysis and Adjusting/Weighting Metrics

An initial sensitivity analysis was run on the as-collected metrics from the training sites. This analysis, run using the “Sensit” add-on to Excel, was a parameter sweep where each metric was taken in steps through a series of values, and the effect on the overall site score was evaluated. This score was an accumulation of the z-values for each of the metrics, using the mean and standard deviation of the training samples as the basis for each z-value.

Sensitivity analysis holds all but one metric constant, and steps that metric through 10 incremental values, computing a site score for that step. This first metric is then stepped to the second value (holding all others constant), and another site score is computed. The process continues for all combinations of all 10 steps for all metrics, and all site scores are retained. These site scores are then ordered and summarized to show the relative effect of the changes of each metric on the overall site score.
The analysis generates two products: (1) a tabular summary of the effect of each metric and (2) a “Tornado” chart which displays horizontal bands depicting the effect of each metric on the overall score. Figure 15 shows an example of these two products.

The tabular summary shows each of the metrics in rank-order with those at the top of the table having the largest effect on the score and decreasing towards the bottom of the table. The effect is expressed as the “% swing” of the score as generated by the range (10 different values in our example) of each metric (column on right side of figure).

Figure 14: Results of sensitivity analysis.
The Tornado chart in the center of Figure 14 also shows the metrics in rank-order of effect from top to bottom, but depicts the effect of each as a horizontal bar whose width is proportional to the effect of that metric on the overall score.

Sensitivity analysis calls for three inputs:

- **The allowable range for the values for each of the metrics.** For this effort, the metrics were allowed to range from 2 standard deviations above and 2 standard deviations below the mean for that metric.

- **The number of increments or steps for each metric.** Ten metrics values (between the +/- 2 standard deviations range) were used for each metric. An early test of 5 and 20 steps were also tried, but had no effect on the results of the sensitivity analysis.

- **The “model” or algorithm on which the effect of the metrics changes is being evaluated.** For this study, the scoring model (aka, the similarity model) is a simple, additive model of the value for each metric.

All sensitivity analysis was performed on the weighted metrics. Weights for each metric were applied to the mean for each metric and not to the individual site values for the metrics.

As the reader will see, in the “Results” chapter, the original “raw” values for the metrics produced a fairly extreme sensitivity result, with those metrics with very large magnitude values (e.g., storage yard size in m²) numerically dominating the analysis result. As such, the values for all of the metrics were normalized (within each metric range) from 0 to 1 for that metric. The sensitivity analysis was then run on the normalized and weighted metrics.
3.5.2. Correlation and Regression Analysis

To understand whether there was any redundancy in the metrics, correlation analysis and then multiple regression tests were run against the metrics. These tests were run in the Minitab statistical package, using data exported from Excel into the table structure of Minitab.

Correlation between each of the 22 metrics from the training sites were used to detect whether a positive correlation existed between the metrics. If found, this correlation could cause a metric to appear significant in the similarity model when, it was not.

Multiple regression was used in addition to correlation to detect whether colinearity was present, an indication of where two or more metrics might be actually measuring the same aspect of similarity.

3.5.3. Other Numerical Tests

Two other numerical tests in Minitab were used in the study:

- Tests for whether the original metrics were distributed according to a normal probability distribution or some other distribution. The Anderson-Darling test for normalcy was used for this.

- Also, a 2-t test was used to determine whether two groups of metrics for synthesized test sites generated different similarity scores. The similarity scores from these two synthesized sites were subjected to this test to help determine if the null hypothesis could be rejected.
3.6. Similarity

3.6.1. Matching and Similarity

Matching is often applied where the two sets of data represent the same instance of a “thing”, and that the differences between the data represent differences in perspective in the observed data, differences in data quality or scale, or perhaps data completeness. It is also suitable where the subject of the data represent two different instances of a type object, and where these object types always occur with identical (or near identical) geometry. Matching usually results in a “true/false” conclusion about whether the two different data sets represent the same real-world object(s).

In cases where two objects have common functional characteristics, but do not always occur with identical geometries; evaluating “similarity” is more appropriate. Similarity measures usually have a higher level of uncertainty than matching approaches, but have the ability to compare two objects which have some number of “like” characteristics, but may differ in these characteristics. In this research, this appears as different spatial arrangement of feature components, while the overall complex feature is still (functionally) the same.

In much of our daily lives, we are constantly looking for similarities. Is that a person I know or have seen before? Is that song one which I have heard before? How is a new problem or challenge, like one which I have solved before? Occasionally we will come across exact matches, but most of our time is spent looking for similarities between things. A vital skill in dealing with similarities is having have both qualitative and quantitative measures so that we may decide whether to take action as information is
gained. “It looks like rain” might cause us to take shelter in a building, “that looks like my friend approaching” might cause us to smile. In the case of approaching rain, we might be comparing some combination of observed temperature, wind and cloud conditions to conditions we’ve seen in the past. We see the right facial patterns and the gait of a person’s walk to conclude that a good friend is arriving. Instinctively, we are comparing new observations to a learned pattern (i.e., an ontology) and, when we see enough similarities, we take action.

Taking a geospatial example, and the one which is the subject of this research effort, we examine two commercial maritime port facilities (“Container Terminals”). In the example from the Port of Long Beach, CA shown in Figure 15, both facilities are functionally the same yet the spatial arrangement of structures within each terminal are quite different. On the left we see two cargo storage areas (outlined in red) but only one on the right. The terminal on the left has an equipment storage area on the west side of the facility whereas the terminal on the right positions the equipment storage. There are also differing geometries for the railroad tracks (blue) and for the cargo storage areas as well. Yet these are two functionally identical types of facilities; they are both container terminals.
Figure 15: Two container terminals with different physical layouts, but the same functions.

If we were to take a strict shape-matching approach, the differing geometries would drive us to the conclusion that the facilities do not match (which is technically correct)

However, the question for this research is not whether or not the two geometries match. Rather, we wish to determine whether the two facilities are members of the feature class we define as “container terminals”, while still allowing for differences in spatial arrangement.

The outcome of this comparison could be to answer the question:

“…are the two facilities actually the same container terminal…?”

Or

“….are both of these facilities examples of a ‘container terminal’ ?”

The second question must allow for variations in attributes and in spatial arrangement of the facility, and yet, we hope to find characteristics which allow consistent similarity measures

To do this, we look to the principle of “form-follows-function” we hope to compare the instances of these terminals in such a way that we examine their “functional arrangement” and not their strict geometric arrangement. We must first determine the functional characteristics which are unique for the type object class we are examining.
This might be a single characteristic, or a unique combination of characteristics combined into a machine-readable description or “ontology” of a “typical” container terminal. We may then use that typical ontology as a pattern against which new instances may be compared.

The distinguishing aspect of this research is that we incorporate functionally-grounded spatial metrics into the ontology, but do not attempt to codify the entire geometry of an object. Instead, we use those spatial dimensions and relationships in defining a feature class, which are related to the unique functions of that feature class. These distinguishing spatial metrics are added to a model of the feature.

By taking a sample of known feature class instances and measuring their key spatial metrics (as well as non-spatial characteristics), a “type” or prototypical ontology for that feature class may be built up. This prototypical ontology may then serve as a pattern for searching and analyzing other data sets to locate and understand other instances of that feature class.

Perhaps the most important difference of this work is that spatial metrics are used as a distinguishing characteristic of “what” it is, and not simply “where” a feature is.

This approach also differs from historical feature data modeling approaches (e.g., OGC, 2012, USGS 1998), which codify static descriptive attributes describing the feature, in that we incorporate metrics which relate to important functional aspects of a feature class. Working from function to form, we ask:

- What are the key functional characteristics of a feature class which distinguish it from other classes?
• What are the best metrics to measure those functional characteristics?

The third difference is that this approach attempts to address “complex” objects, i.e., objects composed of two or more other objects.

3.6.2. Graph-matching in Determining Spatial Similarity

Similarity determinations are not a binary, true/false process; they result in degrees of descriptions of “same-ness” between two objects or sets of objects. Linguistic approaches to concept similarity are common (Resnik, 1999) but rarely consider spatial dimensions of the objects. Shape-to-shape approaches such as those used for computer vision (Chen, et. al., 2003) tend to focus on geometry and topology which looks for exact point and edge matches for such exact shape or scene matching.

In this approach, we use a similarity “score” to compare two ontologies:

• The type ontology derived from known feature samples
• An instance ontology derived from observations of suspected cases of a feature class

The similarity score used in the experiment was intentionally simple, consisting of Z values for the several spatial metrics used. Many similarity models are possible and could be used instead of the one described, but the suggested approach now includes measures of spatial context in the comparison
3.6.3. Similarity Scoring

The similarity model used in this study examined used an aggregate divergence from the mean for the metrics, compared with the mean and standard deviation for a set of metrics from 22 selected “training” sites. Each metric for a test case was compared with the accordant metrics for the training set as shown in Figure 16, and a z-value for that test metric was computed. The Z-values for all metrics for a test case were added to produce a single “site” similarity score for that test site.

![Diagram](image.png)

Figure 16: Approach used for comparing the ontologies of candidate test sites, with a prototypical ontology built using training sites.

The Z values for each metric was always considered as a positive value since the similarity model was trying to detect differences from the training set (i.e., the
prototypical ontology), and not the direction of the difference. Whether each metric the test case was higher or lower than the prototype was helpful in understanding why the differences were occurring, but not used in the similarity score. If the sign of the Z value were considered then two different metrics which were divergent from their mean, but in two different directions as shown in Figure 17, and would have the effect of cancelling each other out in the final similarity model. This was not desired so, after computing the Z value for each metric, only the absolute value of that Z score was used.

![Diagram](image)

**Figure 17:** Similarity is expressed as accumulated Z values for the candidate site metrics, as compared with the distribution of each metric from the prototype ontology.

The resultant similarity score for a given test case can be compared to other test cases, but cannot be compared alone to a definitive “standard” score from the prototype.
ontology. The score is a relative indicator of “how similar” one test case is to the prototype ontology, in comparison to other test cases.

A single score by itself is not meaningful; a history of scores is needed. This build-up of a “scores experience” is much like the human learning process involved in developing an experience base which can be used for subsequent judgments on similarity. The methods used in the study began with a “seed” or “starter” ontology based in large part on reviews of engineering literature. While this represented an experience base of engineering experts for port facilities, such literature necessarily focuses on what “should” be built, but cannot always include the variety of conditions at a port which determine what “was” built; that is the role of the empirical measures (i.e., the spatial metrics) collected in this study. Once collected, those metrics were subjected to numerical analysis to understand how they related to the starter ontology and, eventually, were used to refine or “ground” the ontology in instance data as Kuhn (2005) reminds us as being so important in semantics of spatial data as we build our experience base.

In the matter of similarity, this study used a relative scoring method to determine not whether new observations exactly “matched” the prototypical ontology, but rather we allowed for variability in the observations. Instead of a “matching” approach, we used a similarity approach which tried to quantify the level of agreement of the test-site metrics to the prototype metrics. The intention was to approximate the qualitative approach to similarity which is used in human cognitive processes for judging similarity much like those mentioned by Rosche (1978) and Tversky (1977).
CHAPTER 4: Results

4.1. The Metrics and Ontology

4.1.1. Training Sites: Basic Characteristics

Metrics from the 22 training sites presented in Table 3 included distances, surface areas, angular measurements, and counts. Units of measure for distances are in meters, areas are in square meters, and angles are in degrees. Across metrics, the numerical values varied substantially, with yard sizes in the thousands, whereas the numbers of buildings were typically in the single digits. There were also 5 “confidence” metrics which indicate the level of certainty that a given component had been identified properly in the source data. These were numerically graded from 0 (uncertain), to 1 (absolutely certain). These confidence metrics were placed into the ontology with the thought that identification on imagery might be questionable, and that this uncertainty should be considered in the similarity model. In an operational setting, this would be important but in this study all feature components were identified with absolute certainty. All confidence values used were “1”, which meant that these 5 metrics had no effect on the model. As such, the confidence metrics were removed from consideration in the study.
Thirty six percent of the training 22 sites have missing metrics due to the absence of either (or both) CFSs or CTAs. These missing metrics are shown as darkened cells in Table 3. The presence of these two components was, at the advice of subject matter experts (SMEs), considered optional; some port areas had them, others did not. We did consider eliminating CSFs and CTAs from the model but after consultation with SMEs decided to retain them. Their presence or absence did contribute to why a terminal was similar to the ontology. The basic descriptive statistics for the original metrics are shown in Table 4.

The initial statistics shown in Table 4 were all based on the assumption of normal distributions for each metric. The large variance (and standard deviations) for some metrics (e.g., Apron Length) was unexpected but turned out to be the result of this assumption of normality. This assumption was wrong for some of the metrics and required re-computation of the standard deviation; this will be discussed in section 4.3.1.

4.1.2. Initial Examination of Which Metrics Were Most Important

The 22 training locations were combined into a single site-level “score” in order to express what a “typical” container terminal ontology would be. Since test sites were not being compared to this prototypical ontology, a composite Z score could not be used. Instead, the impact of each metric was assessed in terms of its mean, and variance, and how these impacted what that single composite score might be.
Table 3: Metrics collected from the 22 sampled training sites. Blue shading indicates metrics for feature components, green shading for relationships between components. Darkened cells are where metrics were missing. Lengths, widths and distances are in meters, area in square meters, and angles are in degrees.
Table 4: Descriptive statistics for the metrics from the 22 sampled training sites

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Avg</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apron Width</td>
<td>46.14</td>
<td>28.37</td>
</tr>
<tr>
<td>Apron Length</td>
<td>875.72</td>
<td>628.83</td>
</tr>
<tr>
<td>Apron L/W ratio</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Apron # Cranes</td>
<td>13.64</td>
<td>16.19</td>
</tr>
<tr>
<td>Yard Size</td>
<td>637717.18</td>
<td>595093</td>
</tr>
<tr>
<td>Yard Longest Axis</td>
<td>1996.45</td>
<td>1045</td>
</tr>
<tr>
<td>Number of Buildings</td>
<td>12.95</td>
<td>6.30</td>
</tr>
<tr>
<td>Largest Building Size</td>
<td>6842.50</td>
<td>7682.93</td>
</tr>
<tr>
<td>Average Building Size</td>
<td>1227.14</td>
<td>1358.92</td>
</tr>
<tr>
<td>CTA Longest Axis</td>
<td>766.75</td>
<td>264.85</td>
</tr>
<tr>
<td>Number of CTA's</td>
<td>1.00</td>
<td>0.62</td>
</tr>
<tr>
<td>CSF Confidence</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Number of CSFs</td>
<td>2.82</td>
<td>3.11</td>
</tr>
<tr>
<td>CSF Width</td>
<td>68.59</td>
<td>32.24</td>
</tr>
<tr>
<td>CSF Length</td>
<td>151.08</td>
<td>57.46</td>
</tr>
<tr>
<td>Apron-Yard-CTA angle</td>
<td>121.33</td>
<td>30.75</td>
</tr>
<tr>
<td>Yard/Building Distance</td>
<td>688.91</td>
<td>376.86</td>
</tr>
<tr>
<td>Yard/CSF Distance</td>
<td>384.86</td>
<td>305.63</td>
</tr>
<tr>
<td>Yard/CTA Distance</td>
<td>14.89</td>
<td>14.26</td>
</tr>
<tr>
<td>Yard/CTA Angle between Major Axes</td>
<td>12.44</td>
<td>16.74</td>
</tr>
<tr>
<td>Road/Yard Intersects</td>
<td>2.41</td>
<td>1.75</td>
</tr>
<tr>
<td>CTA-Apron Distance</td>
<td>676.33</td>
<td>389.70</td>
</tr>
<tr>
<td>CTA-Yard Separation</td>
<td>80.44</td>
<td>97.46</td>
</tr>
</tbody>
</table>

As expected, an initial sensitivity analysis showed that if the original values and units of measure are used as-is, that metrics with large magnitude numbers will appear more dominant simply due to their magnitude. This was expected and shows why some form of normalization is required. In the final similarity model this is achieved through the use
of the Z value but in this early assessment, the data needed to be normalized in this initial analysis using a simpler approach.

Normalization of the original metrics was performed on the data from the 22 training sites by scaling each measurement from 0 to 1, based on the range of values of that metric across the training sites. For example, if the range of values for the number of buildings across the training sites were 1 to 20, then the building number metric for a site with one building would be assigned a “0”, sites with 20 buildings would be assigned a “1”, and sites with 10 buildings would be assigned a metric value of 0.5.

This normalization had the effect of removing the effect of differing units of measure in the data, and allowing a more “balanced” sensitivity analysis. Figure 18 shows the result of a sensitivity analysis test of the original metrics before and after normalization. In the left graphic of the Figure the metric for surface area of the storage yard dominated the other metrics, accounting for almost all of the variability in the ontology similarity score. The graphic on the right shows the relative importance of each metric after normalization.
Sensitivity analysis expresses the level of contribution each metric makes to the similarity model in terms of how much of a predicted outcome is due to the allowed range of values for each metric. This describes how much of an effect each parameter (metric) has in terms of what percentage of the score differences that metric’s range has on the model. This is shown graphically in Figure 18 as “Tornado Charts” for each set of data. The Y axis shows each of the metrics, and the X axis shows the effect on the overall site score, with the high and low score values showing the range of the site score as impacted by each metric. In this example, the Storage Yard Area metric (outlined in
red) dominated the score outcome in the original metrics, but fell to 6th place when the
metrics are normalized in the range 0 to 1 (right side of Figure 18).

4.1.3. Weighting

Weighting each metric had a large impact on its importance to the similarity scoring.
This is shown by the sensitivity analysis results in Figure 19 where the un-weighted and
then weighted metrics for the training sites were used to generate composite scores. The
test showed that applying weights caused some metrics to drop in importance and others
to increase in the similarity model.

Weighting also caused the relative importance of each metric to change. This is
evident in the overall shape of the Tornado charts in Figure 19. For example, in the
unweighted metrics, the metric “Yard/CSF Distance” (outlined in red) was most
important but in the weighted data, this metric became least important. Likewise, the
metric “Number of CTA’s” was medium in the unweighted data but became higher in
importance when weightings were applied to the model. While the similarity model is
easily adjusted using weights on the metrics, and these weights can be effective in
inserting the knowledge of subject matter experts, caution must be exercised. These
weights are highly subjective and cannot increase the (quantitative) precision of a
similarity model. At best, they can indicate that a metric appears to be “more” important
or “less” important, but weights cannot tell us that metric “A” is, for example, 2.5 times
more important than metric “B”.

Figure 19: Effect of weighting each metric on the relative importance of each metric in site scoring.
4.2. Importance of Each Metric

The contribution of each metric to the ontology score variability ranged from 9.6% for the number of buildings, down to 1.3% for the average building size as shown in Figure 20. The contributions appeared to occur in 3 subjective groupings with two metrics contributing nearly 8 or more percent each, 13 metrics contributing 3-7% each, and 6 metrics contributing 1-2 percent each.

![Diagram](image_url)

Figure 20: Results of initial sensitivity analysis showing the contribution of each metric to the variability in site scoring.
Metrics for the edges of the ontology’s graph (shaded in green in Figure 20 tended to occur in the middle range of the importance listing with only one spatial relationship metric occurring in the top five, most important metrics. This seems to indicate that while spatial arrangement is important in describing a prototypical feature, spatial metrics alone may not be sufficient to describe a feature’s “type” ontology. The three most important non-spatial metrics were those describing the numbers of components (buildings, CTAs and CSFs). This was only an initial examination of importance of each metric, further analysis of the role of each metric is discussed in the following sections.

4.3. Statistical Examination of the Metrics

Two important questions about the metrics needed to be addressed before using them in the similarity scoring:

- How are the metrics distributed in terms of probability?
- Is there any redundancy amongst the metrics in terms of how they contribute to the similarity score?

4.3.1. Distributions for the Metrics

Since the similarity algorithm uses standard scores (Z values) to generate a similarity score, it was important to understand how each metric was distributed. This was because the Z value is computed using the standard deviation of a metric, and the standard deviations for different distributions may be determined quite differently. For example, the standard deviation for the normal distribution is given by the formula:
Whereas the standard deviation for a metric with a Poisson distribution is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n} - \left(\frac{\sum_{i=1}^{n} a_i}{n}\right)^2}.$$ 

If the wrong distribution is assumed, then the Standard Deviation values (and thus the Z values and resultant similarity scoring) will be incorrectly computed.

The metrics were subjected to the “Anderson-Darling” (AD) test for normalcy. Of these, 7 were distributed normally (P value of .05 or greater, AD statistic less than 1.0) and the other 15 were non-normal. These later metrics were concluded to be of a Poisson nature for the following reasons:

- They were rejected as normal in the AD test
- Their probability distributions (Figure 21) do not appear (graphically) as normal distributions.
- They satisfied the major assumptions of the Poisson distribution
- Their metrics were all non-zero

Based on these normalcy tests, the standard deviations for 15 of the metrics were re-computed for a Poisson distribution. Table 5 shows how the standard deviations changed as a result of this re-computation. The highlighted rows of the table are these variables which were judged to be of a Poisson distribution based on the AD test. For
these metrics, the standard deviations under both distributions (normal and Poisson) are shown in two separate columns. Notice that the two values are quite different, which would result in very different Z values (and therefore different similarity scores). It was this mixed group of standard deviation values, i.e., those based on the “Measured” distributions, which were used in the final similarity scoring.
Table 5: Descriptive statistics for the original 22 metrics, showing how standard deviation will differ depending upon assumptions of the metric's probability distribution (two columns on the right side of the table). The “Measured…” distribution was empirically derived from the collected data, and followed the probability curve listed in the center column (i.e.” Distribution”).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Avg</th>
<th>Distribution</th>
<th>Std Dev Assuming Normal</th>
<th>Std Dev using Measured Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apron Width</td>
<td>46.14</td>
<td>Poisson</td>
<td>28.37</td>
<td>6.79</td>
</tr>
<tr>
<td>Apron Length</td>
<td>875.72</td>
<td>Poisson</td>
<td>628.83</td>
<td>29.59</td>
</tr>
<tr>
<td>Apron L/W ratio</td>
<td>0.06</td>
<td>Poisson</td>
<td>0.03</td>
<td>0.25</td>
</tr>
<tr>
<td>Apron # Cranes</td>
<td>13.64</td>
<td>Poisson</td>
<td>16.19</td>
<td>3.69</td>
</tr>
<tr>
<td>Yard Size</td>
<td>637717.18</td>
<td>Poisson</td>
<td>595093</td>
<td>798.57</td>
</tr>
<tr>
<td>Yard Longest Axis</td>
<td>1996.45</td>
<td>Normal</td>
<td>1045</td>
<td>1045</td>
</tr>
<tr>
<td>Number of Buildings</td>
<td>12.95</td>
<td>Normal</td>
<td>6.30</td>
<td>6.30</td>
</tr>
<tr>
<td>Largest Bldg Size</td>
<td>6842.50</td>
<td>Poisson</td>
<td>7682.93</td>
<td>82.72</td>
</tr>
<tr>
<td>Avg Bldg Size</td>
<td>1227.14</td>
<td>Poisson</td>
<td>1358.92</td>
<td>35.03</td>
</tr>
<tr>
<td>CTA Longest Axis</td>
<td>766.75</td>
<td>Normal</td>
<td>264.85</td>
<td>264.85</td>
</tr>
<tr>
<td>Number of CTAs</td>
<td>1.00</td>
<td>Poisson</td>
<td>0.62</td>
<td>1.00</td>
</tr>
<tr>
<td>Number of CSFs</td>
<td>2.82</td>
<td>Poisson</td>
<td>3.11</td>
<td>1.68</td>
</tr>
<tr>
<td>CSF Width</td>
<td>68.59</td>
<td>Normal</td>
<td>32.24</td>
<td>32.24</td>
</tr>
<tr>
<td>CSF Length</td>
<td>151.08</td>
<td>Normal</td>
<td>57.46</td>
<td>57.46</td>
</tr>
<tr>
<td>Apron-Yard-CTA angle</td>
<td>121.33</td>
<td>Normal</td>
<td>30.75</td>
<td>30.75</td>
</tr>
<tr>
<td>Yard/Building Average Distance</td>
<td>628.91</td>
<td>Poisson</td>
<td>376.86</td>
<td>26.25</td>
</tr>
<tr>
<td>Yard/CSF Distance</td>
<td>384.86</td>
<td>Poisson</td>
<td>305.63</td>
<td>19.62</td>
</tr>
<tr>
<td>Yard/CTA Distance</td>
<td>14.89</td>
<td>Poisson</td>
<td>14.26</td>
<td>3.86</td>
</tr>
<tr>
<td>Road/Yard Intersects</td>
<td>2.41</td>
<td>Poisson</td>
<td>1.75</td>
<td>1.55</td>
</tr>
<tr>
<td>CTA-Apron Distance</td>
<td>676.33</td>
<td>Normal</td>
<td>389.70</td>
<td>389.70</td>
</tr>
<tr>
<td>CTA-Yard Separation</td>
<td>80.44</td>
<td>Poisson</td>
<td>97.46</td>
<td>8.97</td>
</tr>
</tbody>
</table>

4.3.2. Are all the Metrics Necessary?

As with any multi-variate model, the question of whether any of the metrics contribute to the similarity score in a duplicative way is important. A specific metric should contribute to the similarity model in a unique way, and not duplicate the effect of other variables in the model; “double-counting” should be avoided. To test for “colinearity”, a correlation test followed by multiple regression was used.
4.3.3. Statistical Assessment of Colinearity

Correlations were run iteratively on several combinations of the metrics, beginning with all 22 of the metrics. Each correlation result was inspected, looking for values of the Pearson correlation statistic which were at or below 0.010. This was an indicator of high colinearity between metrics pairs. When a low Pearson value was found, the involved metric became a candidate for elimination from the model. A multiple regression test was immediately run on the full complement of 22 metrics to look for the “Variance Inflation Factor” (VIF) which is a strong indicator under regression for colinearity. The correlation results, together with the VIF on each metric then served as a basis for removing a metric from the similarity model. The final decision to remove a metric was also based on whether any metrics would remain in the model which measured some aspect of a key component or relationship. The goal was to remove colinearity, but retain at least some metrics about key parts of the ontology. For example, the original 22 metrics included 3 different measures of the loading apron: apron length, apron width, and the apron’s length/width ratio. Testing indicated strong colinearity within these three, with width and the ratio accounting for most of the redundancy. In this case, the apron length was retained and the other two removed from the model. This still left one metric, the apron length, as a measure of this important component of the container terminal. The results of this first iteration of tests (for the 22 metrics) are shown at Figure 22 and 23.
Figure 22: A portion of the initial correlation results for all 22 metrics. The top number for each pair is the Pearson Correlation statistic, and the bottom number is the p-value.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apron Width</td>
<td>-0.1317</td>
<td>0.7970</td>
<td>-0.17</td>
<td>0.877</td>
<td>42.330</td>
</tr>
<tr>
<td>Apron Length</td>
<td>1.8117</td>
<td>0.6071</td>
<td>2.98</td>
<td>0.041</td>
<td>20.755</td>
</tr>
<tr>
<td>Yard Size</td>
<td>1.8076</td>
<td>0.7369</td>
<td>2.45</td>
<td>0.070</td>
<td>35.879</td>
</tr>
<tr>
<td>Yard Longest Axis</td>
<td>1.6123</td>
<td>0.5944</td>
<td>2.71</td>
<td>0.043</td>
<td>23.484</td>
</tr>
<tr>
<td>Number of Buildings</td>
<td>0.6365</td>
<td>0.3662</td>
<td>1.73</td>
<td>0.158</td>
<td>8.911</td>
</tr>
<tr>
<td>Largest Bldg Size</td>
<td>1.5228</td>
<td>0.4913</td>
<td>3.10</td>
<td>0.036</td>
<td>8.770</td>
</tr>
<tr>
<td>Avg Bldg Size</td>
<td>0.5665</td>
<td>0.4555</td>
<td>1.24</td>
<td>0.223</td>
<td>7.528</td>
</tr>
<tr>
<td>CTA Longest Axis</td>
<td>1.5576</td>
<td>0.3231</td>
<td>4.82</td>
<td>0.009</td>
<td>6.892</td>
</tr>
<tr>
<td>Number of CTAs</td>
<td>0.9127</td>
<td>0.3599</td>
<td>2.56</td>
<td>0.060</td>
<td>5.131</td>
</tr>
<tr>
<td>Number of CSFs</td>
<td>1.1564</td>
<td>0.5707</td>
<td>2.03</td>
<td>0.113</td>
<td>22.911</td>
</tr>
<tr>
<td>CTA Width</td>
<td>1.1045</td>
<td>0.5614</td>
<td>1.97</td>
<td>0.121</td>
<td>20.941</td>
</tr>
<tr>
<td>CTA Length</td>
<td>1.6856</td>
<td>0.4700</td>
<td>3.52</td>
<td>0.024</td>
<td>14.688</td>
</tr>
<tr>
<td>Apron-Yard-CTA angle</td>
<td>0.8313</td>
<td>0.5593</td>
<td>1.49</td>
<td>0.183</td>
<td>14.418</td>
</tr>
<tr>
<td>Yard/Building avg .distance</td>
<td>0.6501</td>
<td>0.5542</td>
<td>1.17</td>
<td>0.306</td>
<td>13.421</td>
</tr>
<tr>
<td>Yard/CTA distance</td>
<td>0.2922</td>
<td>0.6657</td>
<td>0.45</td>
<td>0.677</td>
<td>32.151</td>
</tr>
<tr>
<td>Yard/CTA distance (2)</td>
<td>0.3339</td>
<td>0.7350</td>
<td>0.45</td>
<td>0.673</td>
<td>30.998</td>
</tr>
<tr>
<td>CTA-Apron distance (3)</td>
<td>2.7447</td>
<td>0.9356</td>
<td>2.74</td>
<td>0.062</td>
<td>43.955</td>
</tr>
<tr>
<td>CTA-Yard Separation (d2-d1)</td>
<td>1.3166</td>
<td>0.4219</td>
<td>3.12</td>
<td>0.036</td>
<td>15.026</td>
</tr>
</tbody>
</table>

S = 0.575316

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>18</td>
<td>5165.62</td>
<td>286.98</td>
<td>867.04</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>4</td>
<td>1.32</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>5166.94</td>
<td>286.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 23: Multiple Regression results for the second iteration of the original metrics. In this figure, three of the 22 metrics had already been removed due to high colinearity indicated in the VIF statistic.
This process was repeated two more times, resulting in several more metrics being removed and the VIF metric being lowered to 10 or less. Once a correlation and regression test reached a total of 13 metrics, the indicators for colinearity (Pearson metric and VIF) reached acceptable ranges, the tests were stopped and the final 13 metrics were retained for use in computation of similarity.

4.4. The Final Ontology and Metrics

The final metrics in the ontology are shown in Figure 24 as both a table and a conceptual graph. This ontology is a reduced version of that shown in the methods chapter with metrics with high colinearity removed from the model. While 40% of the original metrics were rejected from the model, those that remained did not introduce colinearity into the similarity model. The ontology was also made up primarily of spatial metrics (10 of the 13 metrics were spatial), which kept the emphasis of the research on the spatial aspects of the feature ontology.
4.5. Issues with the Ontology

4.5.1. Feature Conceptualization

Two issues in feature definitions (i.e. “conceptualizations” in the ontology) were recognized during the data-collection effort. While great care was taken in defining the initial ontology, it was only after attempting to collect feature geometries that issues with the conceptualization of container terminals became apparent. These both impacted the way that data were collected, but also in the way that certain spatial metrics were expressed.

Figure 24: Final ontology shown in tabular and graph forms.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apron Length</td>
<td>875.72</td>
<td>29.59</td>
</tr>
<tr>
<td>Yard Longest Axis</td>
<td>5996.45</td>
<td>44.68</td>
</tr>
<tr>
<td>Number of Buildings</td>
<td>12.95</td>
<td>6.30</td>
</tr>
<tr>
<td>Largest Etd Size</td>
<td>6842.50</td>
<td>82.77</td>
</tr>
<tr>
<td>Avg Etd Size</td>
<td>1277.14</td>
<td>35.03</td>
</tr>
<tr>
<td>CTA Longest Axis</td>
<td>766.75</td>
<td>264.85</td>
</tr>
<tr>
<td>Number of CTAs</td>
<td>3.00</td>
<td>0.62</td>
</tr>
<tr>
<td>Number of CSFs</td>
<td>2.82</td>
<td>1.68</td>
</tr>
<tr>
<td>CSF Width</td>
<td>68.59</td>
<td>32.24</td>
</tr>
<tr>
<td>CSF Length</td>
<td>151.08</td>
<td>57.46</td>
</tr>
<tr>
<td>Apron-Yard CTA angle</td>
<td>121.33</td>
<td>32.24</td>
</tr>
<tr>
<td>Yard/Building avg.</td>
<td>688.91</td>
<td>57.46</td>
</tr>
<tr>
<td>CTA-Yard Separation</td>
<td>80.44</td>
<td>8.97</td>
</tr>
</tbody>
</table>
4.5.1.1. The Container Stuffing Facility (CSF) Definition

The geometries of these facilities included footprints of the buildings where container packing/unpacking took place, but did not include the paved areas immediately surrounding these buildings. These paved areas should have been included, since this is where trucks can bring loose and palletized materials to the CSF buildings, transport materials unpacked from the containers to the local economy, and where the containers are positioned awaiting their turns to be packed/unpacked. These paved areas are dedicated to the CSF functions, and are not considered part of the general storage yard; they are functionally part of the CSF. Figure 25 is an example of one CSF facility in Busan, Korea. The geometry of the buildings was collected for this study, but the larger paved area (dashed lines) was not.

Figure 25: Container Stuffing Facility (CSF) at a container terminal at the Port of Busan, Korea.
Since this issue was not realized until after data capture had occurred, the decision was made to proceed with data analysis using only the building footprints. However, we suspect that had the entire CSF outline, including paved areas, been included, that the similarity scoring might have generated different results.

4.5.1.2. The Cargo Transfer Area (CTA) Definition

The movement of containerized cargo into and out of the terminal is a key function, and is executed using wheeled trucks or railcars (trains). Cargo may eventually be moved to airports for further transit, but transporting the cargo from the terminal itself is done primarily through road or rail. In some cases, cargo is moved from the terminal’s storage yards, directly to smaller boats and barges for movement inland, but these cases are relatively rare. The importance to the similarity model is that these intermodal areas are important in understanding the nature of the local economy and transportation modalities, and may be a strong indicator of terminal type. In some regions of the world, for example, the presence of a rail intermodal area at a harbor may be a strong indicator of the presence of a container terminal.

CTA’s were originally defined to be any area within the terminal where containers entered or left the terminal. For terminals using rail, the geometry of the CTA’s were easy to demark, for terminals where cargo moved in and out of the area only on trucks, the entire storage yard was functionally a CTA. Trucks entered the terminal and were directed to the area within the storage yard where they would receive or drop off their containers. In effect, there was no designated CTA; the entire storage yard was one large CTA.
The definition of a CTA was adjusted to mean only rail-based cargo transfer areas. Since intermodal rail transfers still constitute an important function of container operations, this definition change was probably still beneficial in the similarity scoring. Future studies should examine whether this limited definition of cargo transfer functions (i.e., rail-only) is sufficient, or if other metrics about the truck-based cargo movement are needed.

4.5.2. Multiple Components and their Arrangement

Several assumptions were made early in the project about how many components might exist for a container terminal and how these might be arranged. Metrics were then selected based on these assumptions. In most cases this did not cause any issues but in a few, it was difficult to determine what the metric ought to be for the resultant ontology.

For example, the Pasir Panjang site (Singapore) had loading aprons, storage yards, and CTAs arrangement in a horseshoe as shown in Figure 26, including one of the CTAs actually inside one of the storage yards. This generated a large number of metrics for the separation between the storage yard and the CTA. It had been assumed that terminal arrangements would be much more simple, with the terminal forming a rectangular arrangement with a single apron and single CTA on opposite sides of the yard.

This arrangement of components at the Pasir Panjang terminal resulted in not one, but many possible metrics for the CTA/Yard separations, some crossing the water portion of the terminal site. Geometrically they were all valid metrics but from a functional
perspective (i.e., how cargo actually moves between these two components), some of the computed metrics made little sense.

For example, in Figure 26 the Pasir Panjang terminal includes two Cargo transfer Areas (CTAs) and two separate Storage Yards. This resulted in multiple CTA-Yard distance measurements, two of which are illustrated in Figure 26. Since the original ontology did not envision multiple such measures, a decision had to be made about which of the several distance values would be retained. In this case, only the smallest of the distances were retained, but this leaves open the possibility that some information (i.e., the other distance measures) might have actually been better descriptors of the terminal.

The possibilities for such varied arrangements of multiple components need to be accommodated for in the ontology as well as in the logic of the similarity algorithm. This may be accomplished through the use of logic to decide how to select the correct metric where multiples exist. This may improve the ability to discriminate sites.
Figure 26: Geometry for the components of the Pasir Panjag terminal complex in Singapore. Graphics at A and B indicate the locations where two different "CTA-Yard" distances were computed.
4.6. Assessing Similarity

4.6.1. Overall Success of the Similarity Approach

Using the spatially-based feature ontology, the similarity algorithm of this research had partial success in scoring test-case container terminals as more similar to a prototypical container terminal than were marine terminals of other types.

This conclusion is based on a small sample size (4 terminals), but supports the hypothesis that machine-readable ontologies that include spatial relationships can be used to determine algorithmically whether machine-readable ontologies can determine whether unknown feature instances are “more” or “less” similar to a prototypical pattern (an ontology) for that feature class.

The Hamburg site is a general cargo or “break-bulk” terminal where cargo exists as loose, often “palletized” units which are stored and processed in warehouses in preparation for loading/off-loading to ships. These type terminals are characterized by large numbers of medium/large buildings, and very little storage yard areas. An apron is present but is usually of smaller width than for container terminals.

The Wilmington site is a “Ro-Ro” (Roll-on, Roll-off) terminal, servicing the shipment of automobiles and light trucks. This type terminal is characterized by large storage yards, small numbers of buildings (one of which may be very large), and a rail intermodal area (Cargo Transfer Area). The quay-side loading apron tends to be fairly long and narrow, and almost never contains cranes, since the vehicles drive onto/off-of the ships under their own power using ramps.
The Ro-Ro type terminal (Wilmington) was included in the test cases since it’s layout tends to be much like that of a container terminal; it was assumed that if the similarity model could be confused, it would be by a Ro-Ro terminal.

The Long Beach and Rotterdam sites are both known container terminals. The Rotterdam site contains no CSFs CTAs. The Long Beach site has a CTA, but no CSF. This site’s shape is interesting and includes two storage yards and a long-narrow CTA arranged in such a way that several spatial metrics could not be computed.

Figure 27 shows the final similarity scoring for the four test sites.

![Figure 27: Final similarity scores for the four test sites. The Y axis is the similarity score.](image)

The similarity scores are relative numbers with lower scores indicating that a site was more similar to the prototype ontology, than were those test sites with higher scores.
The Long Beach test site (a known container terminal), was judged to be least-similar to the prototype of all four of the test cases. This was unexpected, but appears to be due to two conditions in the metrics:

- The largest building was only 2148 m$^2$, compared with an average of 6842 m$^2$ in the prototype
- A storage yard longest axis was 901 m, which was less than half of the average axis length of 1997 m in the prototype.

The largest building size metric was intended to reflect the types of functions at port terminals where enclosed/covered structures are needed: (1) processing/security of inbound and outbound trucks and trains or (2) equipment and vehicle maintenance. As expected, the Los Angeles terminal’s largest building is used for truck in/out processing (Figure 28), but its size, even considering the covered drive-through areas, was small compared with the prototype ontology. This smaller size may be due to three reasons at the port area:

- Vehicle and equipment maintenance functions, which often involve larger buildings, for this terminal, such functions may be contracted-out and be located elsewhere in the Port.
- The mild weather in the Southern California area may allow for maintenance and other related functions to take place outside of a covered structure.
- Real estate values may be higher than other port cities, limiting the available square meters of land available for terminal construction.
The two non-container terminals were both scored as less-similar to the prototype container terminal, but for different reasons:

- The Hamburg test site, a general cargo terminal, had smaller storage yards (243 m$^2$) than the prototype with almost 2000 m$^2$, which increased the similarity score. This was expected since general cargo terminals normally hold cargo items in covered warehouses, versus open storage yards for container terminals. A smaller apron length (563m versus 876m in the prototype), and a smaller largest building size (4851m$^2$ versus 6843m$^2$ for the prototype) also caused a “less-similar” score. The average yard-to-building distance was less (151m) than
the prototype (689m) which is explained by the need at general cargo terminals to position buildings closer to storage yards and loading aprons; this supports more efficient movement of loose and palletized cargo items from warehouses to/from the ships.

- The Wilmington test site is an auto shipping center and from a structural standpoint, looks much like a container terminal. It has large, open storage yards, long and wide quay-side ship loading aprons, a few buildings, and is connected to land transportation networks by roads and rail-intermodal facilities. This site, however, had a very large building (11,222 m²) which increased the similarity score. This building serves as the main in/out processing center for commercial automobiles at the terminal and, as expected, requires enclosed areas for vehicle cleaning, servicing, maintenance, and inspection prior-to and after ocean transits. The average size of the other buildings was smaller (131m²) than the prototype average (1227m²) which also made the terminal less similar to the prototype.

4.6.2. Sample Size and Synthesizing Test Sites

Due to the small number of test sites, two groups of test cases were synthesized and compared with the prototype ontology to determine if a larger sample size would lead to a better understanding of whether the similarity comparison is able to distinguish between container, and non-container terminals.
Synthesizing 40 test cases, 20 container terminals and 20 non-container terminals, was done using the mean and standard deviations from the measured training and test sites (the ones actually collected in ArcGIS). The synthesized metrics were created in the Minitab software according to the probability distributions found in the original 22 training sites. These 40 sites were compared with the prototype ontology, and similarity scores generated. The resultant scores were then subjected to a 2-t test to see whether they represented two populations or one.

Graphic comparison of the two synthesized datasets (Figure 29) appeared to be different, but a 2-t test was not significant at the .95 level. The scores from the two sets of synthesized test cases could not be shown as coming from two populations of similarity scores. In other words, this attempt to increase the sample size could not show enough of a scoring difference to allow rejection of the null hypothesis.

The similarity scores for both groups of synthesized sites followed normal distributions as shown in Figure 30. The AD statistics (from the Anderson-Darling Normality tests) were less than 1.0 which indicates that for these data, their distribution follows a normal distribution. Figure 30 also shows the 95% confidence interval (CI) bands for these distributions.

The CI bands overlaps and the Mean and Standard Deviation metrics, helps to explain why the two groups of scores cannot be confirmed to have come from different populations of test sites.
Figure 29: Graphic plots and statistics for the two sets of synthesized test sites. Score histograms are shown in the upper left, simple point plots of the scores for the synthesized sites are in the upper right, and the results of a 2-t test for the two sets of synthesized sites.

Figure 30: Probability plots for similarity scores for both groups of synthesized sites.
4.6.3. Missing Components and Missing Metrics

Container terminals will not always include the “optional” CSFs, or CTAs. These feature components may be missing because the terminal might not be serviced by a railroad (in the case of missing CTAs), or containers may all arrive at the port area pre-packed and do not require facilities in pack cargo materials into the containers. The presence or absence of these items are differences in local terminal operations methods, but does not mean that the facility ceases to be a container terminal.

Certain of the spatial metrics in this study assumed that the CSFs and the CTAs were present; where they are not, then some of the metrics used in the similarity model are missing. One challenge we must address is how determine whether the two groups are spatially “similar” where members of the group (and therefore metrics) are missing.

Missing feature components should indeed reduce the similarity since the absence of one or more components correctly means that the test case feature is different than the prototype ontology. When considered as a graph, the presence or absence of a “node” (a feature component) correctly changes the similarity score.

However, the disappearance of a node in a graph also causes one or more edges (relationships) to also disappear. In this study, the variability in the makeup of container terminals resulted in test cases where certain metrics were missing. Therefore, a decision had to be made regarding how to compute similarity for cases with these missing metrics.

The challenge was that since the similarity score is based on a composite “Z-score”, large numbers of missing metrics would lead to a computationally lower similarity score.
The lower score (meaning “more similar”) would be due to missing values for the metrics, and not due to metrics which were different than the prototype.

Four approaches to handling missing metrics were considered. These involved using all metrics, using only the metrics which were present in the test cases, and two different ways to adjust the available metrics according to the present metrics for a particular test site.

1. **Use all metrics**, regardless of whether the values are missing or not.

   Missing values are considered “0” values in the model, and generates Z-scores of “0” for the missing metrics. These zero Z values for the missing metrics are added to the composite score, along with the Z scores for the metrics which are present.

   \[
   \text{score}_{\text{all}} = \sum_{i=1}^{n} \text{weighted Z scores}_{\text{all metrics}}
   \]

   This approach makes little sense as it treats missing metrics as having Z-scores of “0”, which means that that the missing metrics are treated as identical to those of the prototype. The effect will be that missing metrics will cause the similarity score to decrease and give the impression that the test case is more similar to the prototype ontology.

2. **Use only present metrics**, ignore those which are missing

   Missing metrics in the test cases are not considered and the composite score is the sum of the weighted Z scores for the present values only. This is the
simplest of the four approaches, allowing those metrics which do exist, to contribute to the similarity score but missing values are not considered.

\[
\text{score}_{\text{present}} = \sum_{i=1}^{n} \text{weighted } Z \text{ scores}_{\text{presentmetricsonly}}
\]

One potential problem with this approach is that if the missing metrics are all important (have high weights), and the present metrics are of lower importance, then the metrics of lower importance could unduly bias the score.

3. **Adjust score according the present versus absent metrics**

The score for the “present-only” (#2 above) results are adjusted for the fact that some metrics are missing. The adjustment uses the number of all possible metrics, divided by the number of present metrics and multiplies the present-only score by that ratio. The method is as follows:

\[
\text{score}_{\text{adjusted for metrics}} = \frac{\text{Max } \# \text{ metrics}}{\# \text{ Present Metrics}} \times \text{score}_{\text{present}}
\]

The problem with this method is that a simple ratio of the number of possible to present metrics means that each missing metric is assumed to be the same level of importance as the metrics which are present. If the missing metrics were of high importance, their higher importance would not be included in the score and the present metrics would take on a higher performance.

4. **Adjust score by ratio of total/present weights**

This is a slightly different form of adjustment of the results from the “present-only” score (#2 above). The score is adjusted according to the sum of the total possible
weights, ratio-ed against the total present weights. This ratio is then used as a multiplier on the present-only score to generate the adjusted score. The method appears as:

\[
\text{Score}_{\text{adjusted for weights}} = \frac{\sum_i^m \text{All Possible Weights}}{\sum_i^n \text{Weights Present Metrics}} \times \text{score}_{\text{present}}
\]

This method also suffers from ignorance of the importance of the missing values, though not as much as in the third method since the aggregate importance of the missing metrics are considered in the numerator of the ratio.

To further help determine which method to use, these four different scoring methods were applied to the four test sites; the results are shown in Figure 31.

Figure 31: Similarity of test sites under different scoring methods.
The two test sites in red (Hamburg and Wilmington) are non-container terminals, and the two in green (Long Beach and Rotterdam) are known container terminals. When the first strategy is used (consider only the metrics which are present) the score differences between sites is fairly small, and one of the container sites (Long Beach) was scored as least similar to the prototype. The second strategy (consider all metrics, even when values are absent) resulted in site differences which were even less but now the Hamburg site was scored as least similar. The last two ratio-based methods to adjust for missing metrics both produced the same order of scoring as for the first strategy (i.e., Long Beach was less similar), but the score differences between sites was greatly increased (Figure 14).

The second method (use only the metrics generated from the collected feature geometries) was selected for use in the model. This was because the other three methods allowed larger numbers of missing values to influence the score, and would ignore site dissimilarity due to missing components. In other words, if a particular component was missing (i.e., CSF or CTA) we want their absence to influence the similarity judgment.

We have shown that by working from function to physical form, spatial arrangement of a feature can be captured in a machine-readable prototypical ontology in the form of spatial metrics. We also demonstrated that a similarity model which compared like spatial metrics for test cases, to that prototype ontology, showed partial success in judging similarity. We found that careful statistical examination of the spatial metrics helped minimize colinearity in the similarity model, thereby lending statistical realism to
the similarity scores. The number of test sites collected for the similarity test was insufficient, and called for the use of synthesized sites. These synthesized sites, however, were not sufficient to improve the performance of the similarity model. We also observed that by examining how individual metrics contributed numerically to the similarity scores, we had a better understanding of why certain test sites were judged as less similar to the prototype than expected.
CHAPTER 5. Conclusions and Future Research

This research demonstrated that machine-readable, ontology-based feature recognition can use spatial metrics to help determine similarity. The hypothesis tested was that a set of spatial metrics could be captured and placed into a machine-readable ontology for a type (a “class”) of geospatial feature, and that this prototypical ontology could then serve as a pattern against which new sets of observations could be compared. This comparison would show the degree of similarity between a prototype ontology for a feature class, using a numerical expression of that similarity. This would allow for automated determination of the similarity of new observations to a prototype ontology and lead to improved object discovery algorithms. The research did not, however, attempt to make exact matches between a feature class ontology and new instances of that feature class.

This research employed a complex feature class, i.e., a super-ordinate feature made up of several distinct physical components, and showed that the spatial arrangement of these components could be codified in a numerical model and used to evaluate the spatial arrangement of candidate features.
In addition to being useful for describing spatial arrangement of components within a single feature entity, this approach can also be applied to questions of the spatial context of a simple, one-component feature and how it relates to other features on the terrain. This would allow separate feature ontologies to be given spatial constraints relative to other ontologies which would be used in feature discovery and identification.

While partial success in similarity determinations was shown, the null hypothesis (that spatial arrangement of components was random and not discriminate for a feature class) could not be rejected. Variability in spatial metrics used to describe arrangement of feature components appeared to be a major reason for this. This variability was due primarily to small number of container terminals used to build the prototype ontology. A computational similarity model used a single numerical score to determine the degree to which metrics from a candidate feature conformed to a prototypical ontology developed from previous “training” observations. While a single score was used to judge similarity of new observations to the prototype ontology, all of the input metrics were retained as separate items in order to understand the reasons for the measured similarity levels and which spatial metrics were contributing to the similarity scores. Partial success may have also been due to the manner in which certain components were conceptualized. The CSF (container stuffing facilities) definition was, for example, limited to the geometry of the CSF building, but did not include the vehicle/container holding areas surrounding the building. Had these additional areas been included in the CSF geometry, some of the spatial relationships in the ontology may have been affected and the similarity scoring changed.
Related to the matter of conceptualization and metrics variability was whether there was a single type of container terminal, or several subtypes. This study assumed that all container terminals were of a single type and that regional differences expressed variability due to terrain and local business practices; all container terminals were thought to be from a single taxa but with some variability. If a single taxa was present in the training sites sample, then the resultant variability in the spatial metrics would correctly reflect the variability of container terminals across the world. If, however, there were actually several different types of terminals in the training sample, then the variability in the spatial metrics may have come from inclusion of these differing taxa in the same statistical values. Examples of this possibility were the terminals in riverine situations where causeways were used to connect the storage yard to the loading aprons positioned several hundred meters out in the flowing rivers (to allow deep-draft ships to be moored next to the aprons). These terminals may have actually represented a subclass of container terminals versus terminals where the aprons were immediately adjacent to the storage yard. The causeway may have caused the yard/apron distance to be much greater and potentially affect the similarity scoring.

The matter of whether there was one, or multiple types of container terminals is somewhat subjective and would require considerable discussion with domain experts to understand whether they consider different subtypes of container terminals or a single taxa. It could also require that sampling the training sites be stratified according these multiple taxa to keep the samples separate and thereby reduce the variability in spatial metrics.
The spatial metrics were all tied to key functions of the container terminal feature class. Instead of attempting to find immutable physical metrics to place into the ontology, this research first identified key, presumably immutable, functions for a feature class and then selected measurable physical characteristics which reflected those functions of that feature class. This approach “grounded” the ontology in key functions for that class while still allowing for spatial variability in the spatial arrangement of a facility due to local terrain conditions. The specific spatial metrics used in the ontology were selected based upon input from subject matter experts in ports and harbor engineering, research in engineering specifications for CT construction. This was an important step in early knowledge engineering of the ontology, that helped to ensure that key functions of container terminals were captured in the ontology.

Examination of colinearity between the initial set of spatial metrics was an important step in the selection of the final metrics for inclusion in the ontology. Significant numbers of metrics (nearly half) exhibited high colinearity and were therefore removed from the similarity scoring. Had these (statistically) redundant metrics been retained, the final similarity assessment would have been biased towards the redundant metrics.

The use of conceptual graphs as a structure for the ontology provided a coherent and flexible way to instantiate a feature ontology. Unlike the more common approaches to shape-recognition, the use of conceptual graphs to encapsulate terrain features allows both spatial and non-spatial characteristics to be incorporated. The nodes of the graph are used to codify the physical components of a feature and the edges are used to capture both spatial and a-spatial relationships between graph nodes. This means that non-
spatial information and relationships about terrain features may be included in their ontologies.

Two areas of future research should be considered; one is in the area of basic research and other can be considered as applied research.

Alternate approaches to the similarity algorithm should be tried, including Decision-tree and Bayesian approaches. Neural networks as a framework also hold promise in this area since they may be able to detect subtle patterns in the ontology comparisons, and also have the ability to evolve or “learn” based on the input of new observations over time (Li, 2010). Incorporation of Bayesian approaches into decision-tree methods also hold promise for improving geospatial reasoning (Blackmon Laskey, et. al., 2008 and 2009, da Costa, 2005) and should be considered in future algorithm designs.

Applied research in this area of spatial aspects of feature ontologies actually includes a number of sub-topics which can be pursued in parallel with the suggested basic research. Ontologies of features, particularly complex entities of the type used in this research, should be componentized such that they may be re-used in other studies and projects. The sub graphs describing the buildings, storage yards, etc., in this study could form separate ontologies and be re-used in a modular fashion in larger ontologies for other feature classes. Transportable structures for the ontologies should also be investigated to include RDF/OWL and other XML-based methods. While the graphic and tabular approaches of this study were effective during the actual (run-time) analysis steps, they would not be suitable for exchanging with other researchers and publishing to the geospatial sciences community.
While many alternative approaches to similarity logic and ontology codification can be attempted, the fact that we have demonstrated that spatial arrangement can be included in feature ontologies shows that this research has moved us towards better machine reasoning of geospatial features.
Appendices
Appendix A: Ontology of Container Terminals

Introduction

To understand the nature of container terminals, one must first appreciate the diversity of different types of maritime terminals. Terminals are but one type of facility at maritime ports, and are where passengers and cargo of various types are processed and moved in and out of the port area. This ontology begins with an overview of the commonly recognized types of terminals, and then goes on to describe container terminals in more detail. The ontology includes a discussion of the key functions of a container terminal and how they tend to manifest themselves in the structure and layout of a typical terminal.

This ontology is considered a fairly general “human-readable” ontology. It was developed as a first step in understanding the functions and structures of container terminals, and formed a basis for building the machine-readable structures which comprised the more formal ontology used in this study.

Types of Ports and Terminals

Maritime ports are collections of facilities, equipment and services dedicated to the movement of personnel, goods, and materials from water-borne transportation means, to land and air based transportation networks, and well as transfer of cargo to vessels bound for other maritime ports. Maritime ports also provide important services related to
shipment, including ship-building and maintenance, ship mooring/berthing, cargo storage, security, food and lodging for passengers and employees, and piloting/traffic control.

Ports are usually organized into several types of “terminals” designed to handle a major cargo type. Port engineers consider the following types of terminals (Agerschou, et. al., 2004, Tsinker, 2004):

- **Oil and/or Gas Terminals**: Petroleum and Gas in liquid form
- **Bulk Terminals**: Various dry and liquid materials such as grains, coal, minerals, and chemicals
- **Break-bulk**: This is the oldest form of terminal functions in maritime shipping. It involves receiving goods or materials in large quantities, which are “broken” into smaller packages and re-organized and re-packaged for continued shipping. Break-bulk operations are relatively labor-intensive and slow, and are increasingly being replaced by container shipping methods. Break-bulk is still quite popular and suited to smaller ports where the consuming market is close to the port.
- **Ro/Ro Terminals** (Roll-on/Roll-off): Vehicular cargo such as cars, trucks, and other wheeled or tracked vehicles. “Ro-Ro” refers to the method of driving the cargo on and off the ships.
- **Passenger Terminals**: Passenger traffic for local or long-distance travel.
- **Container Terminals**: Originally developed in the 1800’s for coal shipments in the UK, this method of shipping evolved rapidly in the 1950s and employs 40,
and 20-foot long closed metal “containers” which are on/off-loaded to/from medium and large ships using cranes of several types. Specialized cranes and lifting vehicles move the containers to and from storage yards, and sort and stack them according the shipping plans and schedules. The containers are trans-shipped to/from both trucks and trains for onward shipping. Standards for the containers have developed which greatly reduces the time and cost of moving cargo between carriers and modes. Containers are standardized as 20 ft long units (“TEUs” or twenty-foot-equivalent-units) 40ft long containers (i.e., 2 TEUs).

Ship-building and repair/maintenance areas are usually called “yards” and, like terminals, include specialized facilities and equipment for building and/or servicing and repairing ships.

**Container Terminals**

These are dedicated terminals, with specialized equipment. They can occupy large areas (many hectares) and have distinctive spatial layouts which tend to be consistent across the world as shown in Figure 32.
A typical terminal includes a gated inprocessing area for land transportation (vehicles, trains) bringing cargo into the terminal, a local network of roads/tracks within the terminal for moving and sorting the containers, storage areas ("yards") for temporary storage of filled inbound/outbound containers, and storage yards for empty containers awaiting future use. Yards are also present for temporary storage of trailers used by trucks for moving the containers within the yard and by commercial shipping companies. Specialized "gantry" type cranes of varied numbers are present and are positioned along trackways positioned along the wharfs/quays of the terminal. These large cranes are designed for loading/offloading the ships, and can move latterally along the quay to accommodate the number and size of ships being serviced. Container terminals have several (1-4) small/medium buildings for offices, security, maintenance shops for the
terminal’s equipment, and other services needed by the terminal’s personnel. An example of the terminal on Pier J of the Port of Long Beach, CA is shown in Figure 33. A second example of a large terminal in South Korea is at Figure 34.

Figure 33: Container Terminal at the Port of Long Beach, California.
Key Functions of a Container Terminal

Container terminals are designed around function, and optimized for container throughput. Medium and large terminals tend to be fully dedicated to containerized cargo operations (versus multiple cargo types), and have very much the same basic layout worldwide. There are several exceptions to this which will be discussed.

The following summary of container operations highlights some of the key functions of this type terminal, and provides a basis for selecting spatial metrics used in this study. These functions are summarized from Agerschou, et. al.(2004) and Tsinker (2004)

Inbound, and Outbound Shipping: Containers arrive at the terminal from land transportation systems (rail and road) and are moved to the terminal’s storage yards and/or cargo preparation facilities prior to loading onto ships. In parts of the world
where navigable rivers converge on the port, cargo may arrive via boats and barges. Cargo arriving at the port from ocean-going cargo ships will be moved into the storage yards, and then depart the terminal via these same land (or riverine) transportation networks.

**Cargo Preparation:** Cargo may arrive at the terminal in “loose” or palletized modes and need to be packed into the shipping containers at the terminal. This “container stuffing” is most often accomplished inside covered warehouse facilities to protect the cargo from the elements during the stuffing operation. Similarly, containers may be unpacked at these facilities and separated and sorted for transloading onto trucks and rail.

**Storage:** Containerized storage is held in open, paved yards while awaiting loading to ships or movement to the road and rail transportation networks. These yards are where containers are sorted into ship-specific groupings for the loading operations.

**Loading and Off-loading:** Containers are loaded or off-loaded from the ocean-going ships using specialized cranes to minimize ship-loading time

**Structures and Components**

**Storage Yard:** These are the open, paved areas as shown in Figure 35 where the cargo containers are held temporarily as they move from ships to and from the land-based transportation systems. In some parts of the world, containers are also moved from the ocean-going ships to and from smaller ships for transportation along river systems which converge with the port. The organization of the containers within the yard is controlled
by the loading plans for each ship, as well as the plans and schedules for cargo movement
to and from the ground transportation networks.

Containers are normally held in the yard for 3 days or less. This is to optimize
throughput time for cargo at the terminal. The storage yard also contains specialized
container-handling equipment (lifts and cranes) for moving, sorting and stacking the
containers.

![Image of a storage yard with containers and handling equipment]

Figure 35: A portion of a storage yard showing containers, and specialized handling
equipment.

**Apron and Cranes:** Aprons are long, paved area, immediately adjacent to the water
and is the place where containers are moved on and off of ships as shown in Figure 36.
Larger terminals typically have specialized gantry cranes permanently located along the
apron for loading/off-loading the ships. These cranes will be track-mounted and can
move along the length of the apron to position themselves according to the ship(s) being loaded. The apron is typically 20m or more in width to allow for multiple lanes of vehicles to drive under the cranes as they deliver containers for loading. Medium sized terminals may have “jib” type cranes since they are less costly than wheeled gantry cranes. Some smaller and multi-use terminals may also have general purpose cranes or even no cranes, relying on ship-board cranes for container handling.

**Buildings:** Container terminals are optimized for cargo through-put and, together with the fact that the cargo is inside the weatherproof containers, need relatively few buildings. The container “stuffing” facilities are the one exception, and will be discussed.
shortly. Any buildings which are present will tend to be 1 ha or less in size, and be dedicated to cargo in/out processing for the terminal, equipment maintenance (cranes, cargo handling equipment, etc.), and some staff and administrative functions.

**Container Transfer Areas (CTAs):** These are the rail-intermodal areas at the terminal where rail cars move the containers into and out of the terminal; an example of such a facility is shown in Figure 37. They are most common at larger terminals where rail car movement of cargo is more cost-effective than truck-only movement. CTAs are characterized by multiple, parallel rail tracks (“spurs”), and specialized gantry cranes for rapid loading/unloading of trains. Containers are usually moved from the storage yard to the CTA by local, dedicated trucks.

![Figure 37: Cargo transfer Area (CTA) in Hamburg, Germany showing specialized gantry cranes for unloading trains.](image)

**Container Stuffing Facilities (CSFs):** These are large, warehouse-type buildings where loose and palletized cargo and materials are packed into containers (Figure 38); they are also where full containers from the ships are unpacked and broken-down into
smaller units for further movement on trucks into the local economy. They tend to occur in areas where the local economy is based in small to medium sized companies which do not have the cargo volume to warrant dedicated container packing facilities at their home locations. Instead, the local manufacturers and farmers will bring their “loose” or palletized items to the CSF where they are inspected and placed into the containers for shipping. CSFs may be wholly dedicated to a single container terminal, or may service multiple terminals.

Figure 38: A Container Stuffing Facility (CSF) in Gdansk, Poland showing a container being packed with palletized cargo
Regional Variations

While the internal arrangements of terminal components may vary somewhat, there are two interesting spatial configurations of these terminals which have been observed. While these are termed “regional”, they are due primarily to the nature of the local terrain, and the nature of the local economy.

Terminals with causeways leading to the (quayside) loading areas. There are several very large terminals which use causeways for movement of containers between ships and the storage yards. These seem to appear at terminals located on rivers where large ships cannot be positioned closer to shoreward quays due to water depth; in effect, the quay is moved outward from the shore to deeper water. This is because maintaining channel depths in some river situations is cost prohibitive and use of causeways to move the loading functions out to the ships is more cost-effective. An example of this situation is found in Figure 39.

Figure 39: Example of a container terminal where loading aprons (on quays) have been built out into a river and are accessed via causeways. This is a large terminal in Shanghai, China.
Terminals which use piers, versus quays, for loading. Smaller ports occasionally use piers (perpendicular to the shoreline) as opposed to wharves/quays (parallel to the shore) for container loading. This may be due to available lands to shoreward for a container terminal, or may be due to use of the pier for multiple types of cargo from the local economy. The port may not sustain enough container traffic to warrant a specialized, fully dedicated (and more expensive) container terminal. Container loading/offloading will be slower with pier-based operations and storage yards may be small. Figure 40 is an example of this situation. This particular terminal is operated by the Dole fruit corporation for export of fresh produce from Guatemala. Produce must be processed quickly to avoid spoilage and as such, relatively few containers are held in the terminal for more than 24-48 hours. Outside storage yards are relatively small and containers are moved through the terminal as quickly as possible. This is also because the containers are refrigerated ("reefers") and require continuous electrical power which can be expensive.

![Pier-based container operations at Puerto Barrios, Guatemala.](image)
Appendix B: Study Sites

The location of all study sites are shown in the Figure 41.

Figure 41: Map of study site locations.
The locations of the training sites are the centroid locations of the ports where each container terminal is located. Those coordinates were taken from the World Port Index (WPI) publication from the US Government, is issued to the general public on a monthly basis. The WPI lists approximate geographic centers of each port city in its database. Locations of test sites on the map are approximate. A full listing of training and test sites is shown in table 6. For exact locations of the terminals studied within each port area, the reader should refer to the site overviews in pages 153 through 167 of this appendix.
Table 6: List of training and test sites

<table>
<thead>
<tr>
<th>Training Site Location</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busan</td>
<td>Korea</td>
</tr>
<tr>
<td>Cape Town</td>
<td>South Africa</td>
</tr>
<tr>
<td>Durban</td>
<td>South Africa</td>
</tr>
<tr>
<td>Fremantle</td>
<td>Australia</td>
</tr>
<tr>
<td>Hamburg</td>
<td>Germany</td>
</tr>
<tr>
<td>Jawaharial Nehru</td>
<td>India</td>
</tr>
<tr>
<td>Long Beach</td>
<td>United States</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>United States</td>
</tr>
<tr>
<td>Luanda</td>
<td>Angola</td>
</tr>
<tr>
<td>Napoli</td>
<td>Italy</td>
</tr>
<tr>
<td>New York</td>
<td>United States</td>
</tr>
<tr>
<td>Newark</td>
<td>United States</td>
</tr>
<tr>
<td>Pasir Panjang</td>
<td>Singapore</td>
</tr>
<tr>
<td>Port Kelang</td>
<td>Malaysia</td>
</tr>
<tr>
<td>Puerto Barrios</td>
<td>Venezuela</td>
</tr>
<tr>
<td>Puerto Cabello</td>
<td>Guatemala</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Santos</td>
<td>Brasil</td>
</tr>
<tr>
<td>Shanghai</td>
<td>China</td>
</tr>
<tr>
<td>Taranto</td>
<td>Italy</td>
</tr>
<tr>
<td>Vishakhapatnam</td>
<td>India</td>
</tr>
<tr>
<td>Vladivostok</td>
<td>Russia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Site Location</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamburg</td>
<td>Germany</td>
</tr>
<tr>
<td>Wilmington, California</td>
<td>United States</td>
</tr>
<tr>
<td>Long Beach, California</td>
<td>United States</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>Netherlands</td>
</tr>
</tbody>
</table>
Training Sites

An annotated graphic for each training site is shown in the following pages. Using imagery from Google Maps as a background, the area around each training site is shown, along with an image inset to the upper left, showing the general region around the site. Also shown is a red outline polygon which was the general limit of the data collection for that site. These overview graphics were used to guide data collection at each site.

Busan, South Korea

Figure 42: Training site at Busan, South Korea (imagery from Google Maps).
Figure 43: Training site at Cape Town, South Africa (imagery from Google Maps).

Figure 44: Training site at Durban, South Africa (imagery from Google Maps).
Figure 45: Training site at Fremantle, Australia (imagery from Google Maps).

Figure 46: Training site at Hamburg, Germany (imagery from Google Maps).
Figure 47: Training site at Jawaharial, India (imagery from Google Maps).

Figure 48: Training site at Long Beach, USA (imagery from Google Maps).
Figure 49: Training site at Los Angeles, USA (imagery from Google Maps).

Figure 50: Training site at Luanda, Angola (imagery from Google Maps).
Figure 51: Training site at Napoli, Italy (imagery from Google Maps).

Figure 52: Training site at New York, USA (imagery from Google Maps).
Figure 53: Training site at Newark, USA (imagery from Google Maps).

Figure 54: Training site at Pasir Panjang, Singapore (imagery from Google Maps).
Figure 55: Training site at Port Kelang, Malaysia (imagery from Google Maps).

Figure 56: Training site at Puerto Barrios, Guatemala (imagery from Google Maps).
Figure 57: Training site at Cabello, Venezuela (imagery from Google Maps).

Puerto Cabello, Venezuela

Figure 58: Training site at Rotterdam, Netherlands (imagery from Google Maps).

Rotterdam, Netherlands
Figure 59: Training site at Santos, Brasil (imagery from Google Maps).

Figure 60: Training site at Shanghai, China (imagery from Google Maps).
Figure 61: Training site at Taranto, Italy (imagery from Google Maps).

Figure 62: Training site at Vishakhapatnam, India (imagery from Google Maps).
Vladivostok, Russia

Figure 63: Training site at Vladivostok, Russia (imagery from Google Maps).
Test Sites

Overviews of the four test sites are shown on the following pages. The limits of each test site (for data collection purposes) are shown as a red outline on Google Maps imagery. A general area map is shown to the upper left of the image, and general descriptive information about the site is included as text below the map. These overviews were used to guide the data collection over each of the test sites. Addition information (addresses, links, company names, etc.) were used to verify whether each test site was or was not a container terminal.

Figure 64: Test site at Hamburg, Germany (imagery from Google Maps).
Figure 65: Test site at Wilmington, USA (imagery from Google Maps).

Figure 66: Test site at Long Beach, USA (imagery from Google Maps).
Rotterdam, Netherlands
Test Site
(Container Terminal)

Figure 67: Test site at Rotterdam, Netherlands (imagery from Google Maps).
Appendix C: Metrics Definitions

This appendix includes descriptions of each of the metrics used in this study as shown in Table 7, along with a diagram of a typical container terminal showing where each metric occurs (spatially) in a terminal layout as shown in Figure 68.

The reader is advised to review Appendix A for a basic description of a container terminal, along with descriptions of the basic structures and components of such a terminal.
Table 7: Metrics used in the study, along with their functional basis.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
<th>Functional Basis of the Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Apron Width</td>
<td>Width of loading apron (meters)</td>
<td>Cargo loading</td>
</tr>
<tr>
<td>2 Apron Length</td>
<td>Length of loading apron (meters)</td>
<td>Cargo loading</td>
</tr>
<tr>
<td>3 Apron l/w ratio</td>
<td>Ratio of length to width</td>
<td>Cargo loading</td>
</tr>
<tr>
<td>4 Number of Cranes</td>
<td>Number of permanent loading cranes on the apron</td>
<td>Loading/unloading of ships.</td>
</tr>
<tr>
<td>5 Yard Size</td>
<td>Size of storage area (square meters)</td>
<td>Cargo storage volume for the terminal</td>
</tr>
<tr>
<td>6 Yard Longest Axis</td>
<td>Length of longest axis of storage yard (meters)</td>
<td>Cargo storage volume for the terminal</td>
</tr>
<tr>
<td>7 Number of Buildings</td>
<td>Count of number of non-CSF buildings at the terminal</td>
<td>Indicator of the amount of administrative processing and equipment maintenance which takes place at the terminal</td>
</tr>
<tr>
<td>8 Largest Building Size</td>
<td>Size (square meters) of the largest non-CSF building</td>
<td>Indicative of the amount of I/O processing (security and customs) which is taking place. Assumes that largest building is dedicated to security and customs.</td>
</tr>
<tr>
<td>9 Average Building Size</td>
<td>Average size of non-CSF buildings</td>
<td>Indicator of the amount of administrative processing and equipment maintenance which takes place at the terminal.</td>
</tr>
<tr>
<td></td>
<td><strong>CTA Longest Axis</strong></td>
<td>Length (meters) of longest axis of the Cargo Transfer Area</td>
</tr>
<tr>
<td>---</td>
<td>----------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>11</td>
<td><strong>Number of CTAs</strong></td>
<td>Count of the number of Cargo Transfer Areas</td>
</tr>
<tr>
<td>12</td>
<td><strong>Number of CSFs</strong></td>
<td>Count of the number of Container Stuffing Facilities</td>
</tr>
<tr>
<td>13</td>
<td><strong>CSF Width</strong></td>
<td>Width (meters) of each Container Stuffing Facility building</td>
</tr>
<tr>
<td>14</td>
<td><strong>CSF Length</strong></td>
<td>Length (meters) of each Container Stuffing Facility building</td>
</tr>
<tr>
<td>15</td>
<td><strong>Apron-Yard-CTA Angle</strong></td>
<td>Angle (degrees) described by the Apron, Storage Yard, CTA centroids</td>
</tr>
<tr>
<td>16</td>
<td><strong>Yard/Building Average Distance</strong></td>
<td>Average distance (meters) from the center of the storage yard, to the buildings at the terminal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17</td>
<td>Yard/CSF Distance</td>
<td>Distance (meters) from the center of the storage yard, to the center of the CSF. Distance is an average where more than one yard or CSF occurs.</td>
</tr>
<tr>
<td>18</td>
<td>Yard/CTA Distance</td>
<td>Distance (meters) from the center of the storage yard, to the center of the CTA. Distance is an average where more than one yard or CTA occurs.</td>
</tr>
<tr>
<td>19</td>
<td>Yard/CTA Angle between Major Axes</td>
<td>Angle described by the major axes of the yard and the CTA</td>
</tr>
<tr>
<td>20</td>
<td>Road/Yard Intersects</td>
<td>Number of roads intersecting the storage yard(s)</td>
</tr>
<tr>
<td>21</td>
<td>CTA-Apron Distance</td>
<td>Distance (meters) from center of CTA, to center of apron</td>
</tr>
<tr>
<td>22</td>
<td>CTA-Yard Separation</td>
<td>Distance (meters) between two closest edges of the CTA and the Storage Yard</td>
</tr>
</tbody>
</table>
Figure 68: Typical layout of a container terminal showing components (left graphic). Graphic on the right shows a selection of the metrics used in the study. The numbered items on the right graphic match the numbers of the metrics shown in Table 7.
### Appendix D: Glossary of Terms

These are terms which are used repeatedly in this document and describe the manner in which they are used in this dissertation.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantics</td>
<td>The study of the “meaning”(s) in data. The term is also used to refer to the meaning itself.</td>
</tr>
<tr>
<td>Ontology</td>
<td>A description of a portion of the real world, as defined by a particular domain.</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>A hierarchical ontology.</td>
</tr>
<tr>
<td>Conceptualization</td>
<td>The manner in which a domain perceives and defines “things” in the world. This is a key determinant on what is codified in an ontology.</td>
</tr>
<tr>
<td>Domain</td>
<td>Any group of persons who have similar interests or practices. May also be referred to as a “community of interest” or “community of practice”. A domain may also be described as a particular body of knowledge, e.g.: mathematics, health sciences, sociology, etc. Domains are the knowledge entities which define and use ontologies.</td>
</tr>
<tr>
<td>Thing</td>
<td>Any physical or abstract entity in the world. Things include “features” but also include processes, events, and other concepts which, unlike features…may not be ordinated in space and time.</td>
</tr>
<tr>
<td>Feature</td>
<td>Any entity, physical or abstract, which may be ordinated in space and time. From a semantics perspective, features are a type of “thing”. Also referred to as geographic features.</td>
</tr>
<tr>
<td>Feature Class</td>
<td>A particular type of feature which are defined as a distinct taxa within an ontology.</td>
</tr>
<tr>
<td>Complex Feature</td>
<td>A super-ordinate feature made up of 2 or more features, or some combination of features or other complex features.</td>
</tr>
<tr>
<td>Object</td>
<td>Synonymous with “Feature”.</td>
</tr>
<tr>
<td>Component</td>
<td>The simple features which make up a complex feature.</td>
</tr>
<tr>
<td>Spatial relationships</td>
<td>The physical position of an object relative to a reference object. Close in meaning to spatial arrangement, but usually used to describe the physical position of one object to a second object. Commonly recognized spatial relationships in the GIS.</td>
</tr>
</tbody>
</table>
community include: distance and proximity, overlap, adjacency, overlap, contains, meets, disjoint and others.

| **Context** | The things surrounding a central concept of interest, which contribute to the meaning of that concept. Context may be in conceptual space (e.g.: family relationships, political organizations, etc.) or in physical space (e.g.: a building is physically next to a shopping center). Context is important in semantics since the meaning (function, role, etc.) of a thing is often understood only when its relationships to other things are examined. |
| **Spatial Arrangement** | The physical position of a specific set of features and components relative to each other. In port facilities and other industrial areas, this may be referred to as the “layout” of the structures. Spatial arrangement in facilities is function-driven. This arrangement usually results in a spatial pattern which is repeated in some form for other facilities with the same functions. |
| **Similarity** | The collection of characteristics which multiple objects have, which leads to a subjective/qualitative decision that those multiple objects are in some respect, the same. |
| **Matching** | The equivalency of 1 to n metrics about 2 or more objects which leads to a decision that the objects are identical. This is a more stringent form of similarity evaluations, and is used where absolute equivalency (geometric or conceptual) is required. |
| **Conceptual Graph** | Any graph made up of nodes and edges, where the lengths of the edges, and the angles described by those edges, are non-geographic in nature. The coordinate system of such a graph is often described as “concept space” where distances between nodes and the arrangement of nodes may be determined by non-spatial metrics. Spatial metrics can be included as attributes of the nodes and edges, but the graph itself is not organized according to physical space. |
REFERENCES


http://www.personal.psu.edu/kuj13/sage_janowicz_semantic_interoperability_draft.pdf.


http://www.ordnancesurvey.co.uk/oswebsite/ontology/.


Cartography and Geographic Information Science v. 38, n. 2, p. 126-136


Curriculum Vitae

Mr. Clark studied at Cornell University earning a Bachelor of Science in Biology in 1973, where he also worked as a cartographer for the New York State wetlands inventory. He then entered military service where he served as a terrain analyst with the US Army. He continued serving in the Army as an Engineering Officer in the Reserves while also studying digital image processing at Mississippi State University (MSU). Mr. Clark’s graduate studies resulted in a Master of Science from the School of Forest Resources at MSU in 1980. His Thesis was an evaluation of various techniques for spectral classification of forested areas, using supervised, unsupervised, and hybrid spectral clustering techniques.

Mr. Clark became Assistant Director of the Mississippi Remote Sensing Center at MSU, directing various studies of forest resources for State and Federal organizations. He was also certified as a Photogrammetrist by the American Society of Photogrammetry and Remote Sensing for his work in applications of remote sensing and digital image processing.

He then continued his military service as a Topographic Engineering officer until 1990. Mr. Clark then joined Intergraph Corporation and worked as a systems engineer and program manager, supervising the development of digital cartographic and image processing systems. He also conducted cartographic and GIS studies and projects in support of various customers around the world.

Mr. Clark presently serves as a data architect, systems engineer, and program director for the U.S Government.