

# A Unified Geodatabase Design for Sinkhole Inventories in the United States

by

Ebrahim “Tony” Khan

A Thesis Presented to the  
Faculty of the USC Graduate School  
University of Southern California  
In Partial Fulfillment of the  
Requirements for the Degree  
Master of Science  
(Geographic Information Science and Technology)

May 2016

Copyright© 2016 by Ebrahim “Tony” Khan

# Table of Contents

List of Figures .....	v
List of Tables .....	vi
Acknowledgements.....	vii
List of Abbreviations .....	viii
Abstract.....	ix
Chapter 1 Introduction .....	1
1.1 What are Sinkholes? .....	1
1.2 Sinkholes and Society.....	2
1.3 Sinkholes in the State of Florida.....	4
1.4 Research Objectives.....	7
1.5 Structure of this document.....	8
Chapter 2 Existing Sinkhole Inventories and Related Research.....	9
2.1 National and International Sinkhole Studies.....	9
2.1.1. National Studies .....	9
2.1.2. International Studies .....	12
2.2 SGS Sinkhole Inventories.....	12
2.3 FGS and KGS Databases as Models for This Study.....	15
2.4 The Database Design Process .....	16
2.4.1. Database Design Evaluation .....	17
2.5 The National Address Database – a Unified Database Design Example.....	19
2.6 Summary.....	20
Chapter 3 A Unified Sinkhole Geodatabase Design.....	22
3.1 User Communities .....	22
3.1.1. Expert GIS Users: Geologists .....	22
3.1.2. Intermediate GIS Users: Insurance Fraud Investigators .....	23
3.1.3. Novice GIS Users: General Public .....	23
3.2 Design Principles .....	24
3.2.1. Insight from Sinkhole Studies.....	24
3.2.2. Key Sinkhole Database Models .....	24
3.2.3. Distillation of State Sinkhole Database Models .....	27
3.2.4. Identification of Feature Layers.....	29
3.2.5. Why Use an ArcGIS Geodatabase? .....	29
3.3 Unified Geodatabase Design.....	30
3.3.1. Database Structure .....	30
3.3.2. Attributes.....	32

3.3.3. Geodatabase structure .....	35
3.4 Geodatabase Design Data Integrity .....	37
3.4.1. Integration of Both Sinkholes and Sinkhole Areas Features .....	37
3.4.2. Cross-Referencing VGI Data.....	38
3.4.3. Primary and Foreign Keys .....	39
3.5 Summary of Database Design Steps Completed .....	40
 Chapter 4 Evaluation of the Unified Geodatabase Design .....	 41
4.1 Study Area for Implementation Tests .....	41
4.2 Construction of the Prototype Unified Geodatabase .....	42
4.2.1. Sinkhole Areas Polygon Feature Class .....	43
4.2.2. Sinkhole Point Feature Class .....	44
4.2.3. VGI Tables.....	46
4.2.4. Primary and Foreign Keys .....	46
4.3 Testing the Prototype with User Community Sample Queries .....	47
4.3.1. Geologist Use Case Queries.....	48
4.3.2. Insurance Fraud Investigators Use Case Queries.....	50
4.3.3. General Public Use Case Queries .....	52
4.4 Additional Evaluation Perspectives .....	54
4.4.1. Level of Completeness.....	54
4.4.2. Database Design Quality.....	55
 Chapter 5 Conclusions and Future Work.....	 56
5.1 Suggestions to Improve VGI Collection and Implementation.....	56
5.1.1. A GeoJSON Interface for VGI Data Collection .....	57
5.1.2. OpenStreetMap VGI Data Collection.....	58
5.2 A Web Map Interface.....	59
5.3 National Implementation .....	59
5.4 Conclusion .....	60
 REFERENCES .....	 61
 Appendix A: Attributes Emulated from the FGS Database.....	 64

## List of Figures

Figure 1 Karst topography .....	2
Figure 2 Carbonate outcrops globally, excluding Evaporites .....	2
Figure 3 Existing karst topography in the Contiguous United States.....	4
Figure 4 Aerial view of the Winter Park Sinkhole .....	5
Figure 5 Three types of sinkholes, mechanisms and resulting features.....	6
Figure 6 Database quality dimensions .....	19
Figure 7 Entity-Relationship Diagram .....	31
Figure 8 Entity-Attribute-Relationships Diagram .....	36
Figure 9 The State of Florida and study area.....	42
Figure 10 Sinkhole area locations in study area .....	43
Figure 11 Sinkholes in study area.....	45
Figure 12 Python script for calculating key values.....	47
Figure 13 A sample GeoJSON interface.....	58

## List of Tables

Table 1 SGS sinkhole database sources.....	14
Table 2 Comparison between SGS databases and international studies.....	15
Table 3 Database design steps .....	17
Table 4 SGS databases contributing to the UG design.....	26
Table 5 Sources and common attribute themes in SGS inventories .....	27
Table 6 UG attributes emulated from Florida database .....	28
Table 7 SHA polygon feature class attributes.....	32
Table 8 Sinkhole point feature class attributes .....	33
Table 9 VGI Source table attributes.....	34
Table 10 VGI Duplicate Data table attributes.....	35
Table 11 Attribute-domain associations .....	37
Table 12 Database design steps summary .....	40
Table 13 Geologist use case sample query 1 .....	48
Table 14 Geologist use case sample query 2 .....	49
Table 15 Geologist use case sample query 3 .....	50
Table 16 Insurance fraud investigator use case sample query 1 .....	51
Table 17: Insurance fraud investigator use case sample query 2.....	51
Table 18 Insurance fraud investigator use case sample query 3.....	52
Table 19 General public use case sample query 1 .....	53
Table 20 General public use case sample query 2 .....	53
Table 21 General public use case sample query 3 .....	53

## **Acknowledgements**

I want to thank my thesis advisor, Dr. Kemp, for her guidance, insight, patience and keeping me on track and Dr. Swift and Dr. Vos for their input on possible thesis directions. I would like to thank my supervisor, Doug McKay, who was extremely accommodating when I needed to modify my work schedule throughout the GIST Program at University of Southern California. Dr. Hastings suggestion on a potential approach for the thesis was invaluable. Finally, I am grateful for the support my Mom gave me, without whom it would have been impossible to finish the thesis.

## **List of Abbreviations**

EARD	Entity-Attribute-Relationship Diagram
ERD	Entity-Relationship Diagram
FGS	Florida Geological Survey
KGS	Kentucky Geological Survey
GIS	Geographic Information System
SGS	State Geological Survey
SHA	Sinkhole Area
UG	Unified Geodatabase
USGS	United States Geological Survey
VGI	Volunteered Geographic Information

## **Abstract**

Sinkholes are naturally occurring geologic phenomena which form when karst erosion causes the surface to collapse. Karst formations can be found globally as a result of water eroding soluble bedrock which creates features such as fissures, caves, and sinkholes. In the United States, every state except Rhode Island has the presence of karst terrain and, therefore, the potential of developing sinkholes. Sinkhole formation can negatively impact society, manifesting mostly as property damage, and in some tragic cases, causing a loss of life. There is a lack of protocols for tracking and recording sinkhole events data nationally. The sinkhole inventories that are available do not include all sinkhole activity and are primarily found among different State Geological Surveys (SGS) databases.

The objective of this thesis was to create a single unified geodatabase (UG) schema based on existing SGS sinkhole databases. The majority of SGS sinkhole data is in the public domain and is of an authoritative source while only two states are utilizing Volunteered Geographic Information (VGI). Two states, in particular, Florida and Kentucky, influenced the geodatabase design because of their developed structure and relative completeness respectively. The proposed UG combines authoritative and VGI elements from multiple databases. It is composed of two feature classes and three tables that are joined by primary and foreign keys. Additional design elements stem from database design theory and sinkhole research studies. The geodatabase design was tested by implementing a prototype database for a portion of Florida. The design was evaluated against the needs of three potential user communities: geologists, insurance fraud investigators, and the general public. Based on these fundamentals, a single UG template was created that can be implemented at the SGS level, and lay the foundations for a national geodatabase in the future.

## **Chapter 1 Introduction**

Sinkholes are naturally occurring features in karst landscapes. The unpredictable nature of the ground collapsing suddenly or subsiding over time can result in property destruction, injuries, and in some cases fatalities if the collapse occurs where people reside. Since karst is present throughout the U.S., there is a need for recording sinkhole locations in a comprehensive unified geodatabase as population growth spreads into susceptible karst landscapes. Database models for the recording of sinkholes that have been implemented by multiple state geological surveys in the US all vary significantly. Thus, this thesis presents a unified geodatabase template that can be used to unify existing sinkhole databases within a single standard.

### **1.1 What are Sinkholes?**

Sinkholes are the most recognizable features of karst topography. Karst topography is the result of natural geologic processes that cause soluble bedrock such as Carbonates and Evaporites to dissolve (Fleury, Carson, and Brinkmann 2008). As Figure 1 illustrates, this geologic solution results in caves, fissures, and tubes underground. Often subsidence occurs on the surface, and sometimes the surface collapses resulting in sinkholes.

While karst formations are a global phenomenon indicated by the red areas representing carbonate rock outcrops in Figure 2, the scope of this study is sinkhole activity in the U.S. where every state has karst terrain except Rhode Island (Tobin and Weary 2004). Although sinkholes are a type of geologic hazard, there is no regional or national database tracking such events in the U.S., nor is there a set of standards to record such information. This lack of a central database or common database structure can result in inconsistent data recording of sinkhole formation within different states, developing inefficiencies that affect society's safety and functionality.

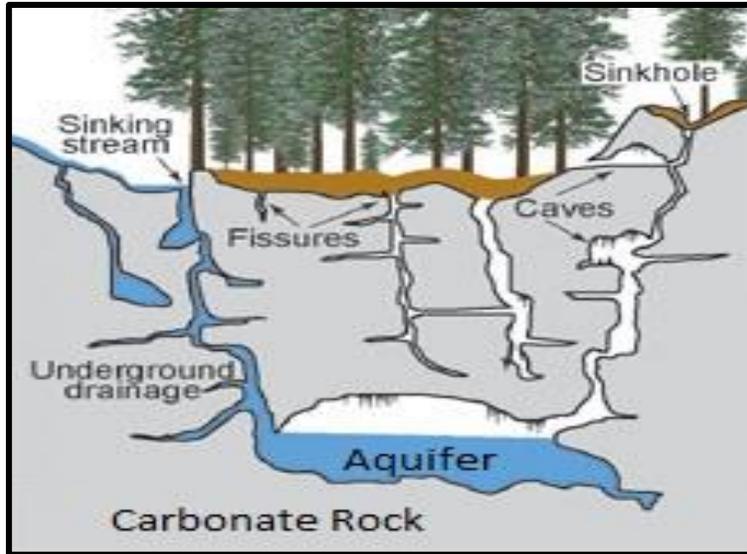


Figure 1 Karst topography (Modified from University of Texas at Austin 2015)

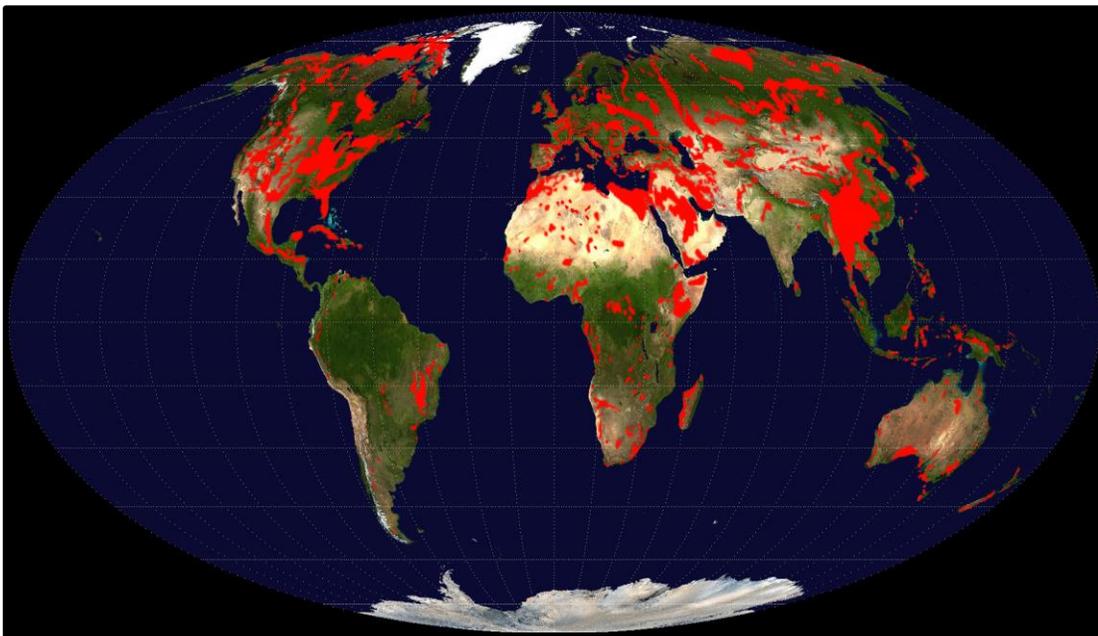


Figure 2 Carbonate outcrops globally, excluding Evaporites (Tichy 2010)

## 1.2 Sinkholes and Society

Sinkhole events occur frequently throughout the U.S., but not all are catastrophic in nature. However, the following cases within the last two years from around the contiguous U.S. gained national media attention, emphasizing the significant threats for society. On June 5, 2015

in Colorado, a police officer was driving on a road when his vehicle was swallowed by a sinkhole. Fortunately, in this case, no deaths occurred (Whitehead and Shiff 2015). On May 23, 2015, a large sinkhole appeared on a golf course in Missouri after heavy rains (Associated Press 2015). Three significant events from 2013 that occurred in Florida illustrate the true danger of sinkholes. The first event in February ended with the death of 37-year-old Jeff Bush whose home partially collapsed into a 20-foot wide sinkhole (Malone 2013). Sadly, his body was never recovered (NOVA 2015). In August, a second incident involved a three-story building of a resort close to Walt Disney World becoming engulfed by a 100-foot wide sinkhole. Luckily, there were no fatalities (Liston 2013). The last event in November 2013 destroyed two homes when a 90-foot wide and 50-foot deep sinkhole collapsed (Malone 2013).

Such dramatic instances are actually rare among the many sinkhole incidents that impact people in their daily lives. The Citizens Property Insurance Company in Florida reported that prior to 2010 less than one percent of the sinkhole claims submitted to Citizens Property were for “catastrophic” ground collapse of the surface below or near the building resulting in significant damage to the building (Zisman 2013). Other sinkhole collapses do not result in astonishing circumstances and, therefore, do not garner media attention, but the danger is ever present.

Considering such unsettling sinkhole incidents, it seems unusual that there is no consolidated database that attempts to track this information at a regional or national level. Having one database would assure many kinds of end users that they have the most complete data for meeting their business or personal needs. In the absence of a single database, a large number of variable state-level collections makes it is very difficult to know the extent of sinkhole activity and its impacts on society in the U.S. Some questions that could be answered with a unified geodatabase include: how much property damage is attributed to sinkholes? What is the

number of people that have been injured or killed by sinkholes? What are the most active areas of sinkhole activity? Is there a specific activity that is correlating with the most active sinkhole areas? The goal of this study is to design a unified sinkhole geodatabase so that such inquiries and countless others can be answered easily and uniformly across the U.S.

### 1.3 Sinkholes in the State of Florida

Karst terrain can be found throughout the U.S., with the exception of Rhode Island. In 16 states, at least 25 percent of their terrain is composed of karst topography. This national extent is shown by the orange areas in Figure 3. Florida is one of the most affected sinkhole regions in the world (Galve et al. 2011) and with approximately 81 percent of the terrain containing karst topography has the highest proportion in the nation (Tobin and Weary 2004). It is also the third most populated state (Census Bureau 2014) thus giving it a high probability to have negative societal consequences from karst terrain, mostly in terms of property damage. As of 2005, it was estimated that sinkholes cost Floridians between 22 to 65 million U.S. dollars annually (Galve et al. 2011).

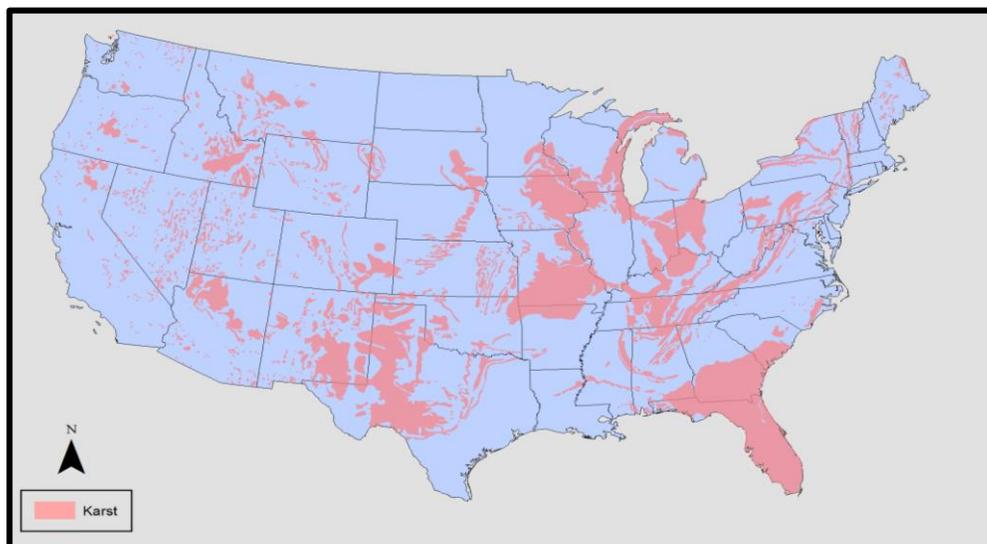


Figure 3 Existing karst topography in the Contiguous United States (Tobin and Weary 2004)

A significant Florida sinkhole incident is the Winter Park Sinkhole. In 1981, a massive sinkhole formed in Winter Park, Florida. It was 320 feet wide and 90 feet deep (Figure 4) and was estimated to cost up to four million U.S. dollars in damages. The Winter Park Sinkhole prompted the State to create the Florida Sinkhole Research Institute (FSRI) to gather information on sinkholes, although the Florida Geological Survey (FGS) had been accumulating data since 1907 (Zisman 2013). The FGS database is still in use today and was the impetus for this study.



Figure 4 Aerial view of the Winter Park Sinkhole (Orlando Sentinel 1981)

While sinkhole data collection began in the early 20<sup>th</sup> century, it was not until 1959 that Florida's Legislature introduced guidelines for insurance claims relating to sinkholes which were arising from a lack of clear descriptions (Zisman 2013). The Florida Sinkhole Statute attempted to clarify the definition of a sinkhole and the types of damage that can be attributed to them as illustrated in Figure 5. The Statute has had several revisions spanning into the mid-2000s to include topics such as sinkhole insurance claim definitions, testing standards for sinkholes and alternative procedures for resolution of disputed sinkhole claims (Zisman 2013).

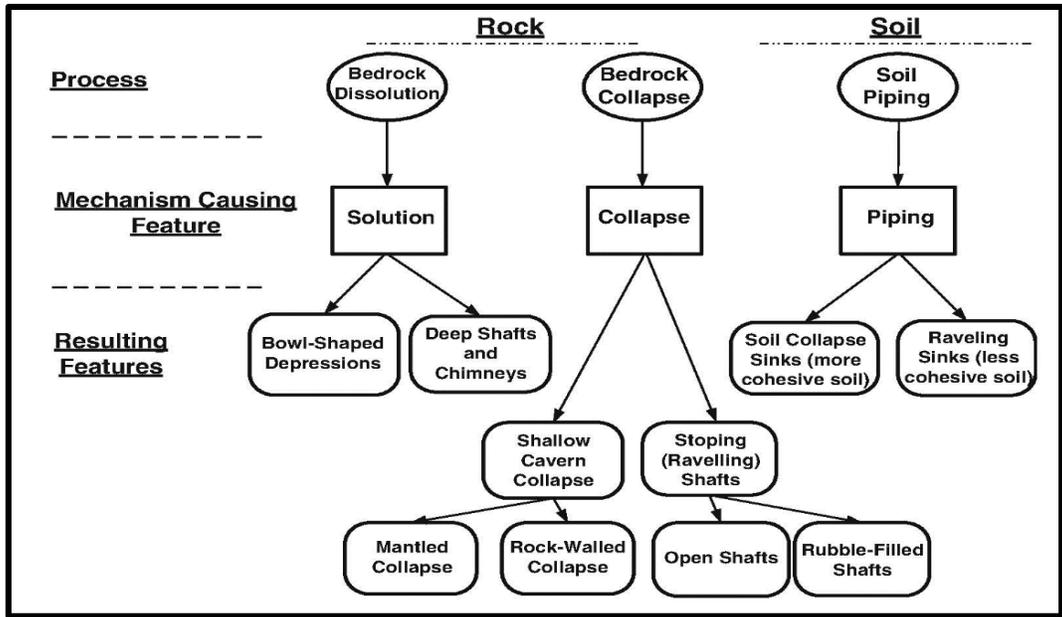


Figure 5 Three types of sinkholes, mechanisms and resulting features (Zisman 2013)

However, because there are no strict procedures to evaluate properties for sinkhole risk, claimants can attribute property damage to sinkholes without proper evidence. Even though damage may have nothing to do with sinkhole activity – but rather poor building construction, other natural causes or neglect – it is usually cheaper for insurance companies to pay the claim instead of disputing it in court (Zisman 2013). It is important to address that after the extensive damage of Hurricane Andrew in 1992, there was a lack of insurance availability. The State of Florida responded by creating the state-run Citizens Property Insurance Corporation. According to Zisman (2013), this entity in 2009 released data showing that it paid out \$97 million U.S. dollars in sinkhole losses while collecting only \$19.6 million U.S. dollars in sinkhole coverage premiums.

Zisman (2013) also states that claimants of sinkhole damage may not even use the funds towards any kind of repair of the associated property, and if they sell the property, the next owner's safety may be in jeopardy if the current owner does not disclose the sinkhole damage in

fear of losing property value. Florida law requires all property owners to purchase sinkhole insurance, which contributes to the higher cost of living even though there may not be significant sinkhole activity in a given area (Fleury, Carson, and Brinkmann 2008). Given all of these situations, it is clear that having access to consistently structured and reliable sinkhole information would be a great asset in Florida and nationwide.

## 1.4 Research Objectives

In his study examining the legislation for sinkholes, Zisman (2013) concisely captures the need for a central database to alleviate insurance costs for the State of Florida. His statement forms the basis of this thesis's objectives:

Data on the occurrence and nature of sinkholes in sinkhole-prone areas will be beneficial in the planning for insurance and development in these areas. Much of this information currently exists but is not shared or is difficult to obtain.... [I]t is recommended that the Florida Geologic Survey be tasked with the responsibility of developing a database base [*sic*] of sinkhole information [obtained] from insurance companies and other sources. This information should be made available to the public.... Such a database would improve knowledge of sinkhole occurrence and lead to reduced insurance costs through a better understanding of the occurrence of karst features. (Zisman 2013, 101)

This project builds on these recommendations, realizing that such a model could be applied to other states with similar sinkhole related challenges, benefitting society at much larger scale. Present sinkhole inventory methodologies lack common data collection and synthesis methods which would be necessary to create a regional or national database. A single data source would eliminate redundant data efforts aimed at reaching similar ends. *The goal of this research is to design a unified geodatabase that can incorporate existing state-level sinkhole inventories into a single standard structure making access to and use of this information more efficient.*

This goal was achieved by locating and evaluating sinkhole inventories developed in seventeen U.S. states. The needs of three end user groups – geologists, insurance fraud

investigators and the public – were integrated into the study. The first two groups would find such a resource to be useful for their business needs while the public would benefit from a heightened awareness of their surroundings. Based on existing inventories and the needs of these user groups, a unified sinkhole geodatabase was designed. This geodatabase design was implemented and tested in a section of Hillsborough County, Florida, an area with a very high number of sinkholes. Given the extensive amount of sinkhole-related research previously undertaken in Florida and the related repercussions society has to contend with there, Florida provides an excellent case study area in which to test the unified geodatabase design.

## **1.5 Structure of this document**

This document is divided into four additional chapters. Chapter Two describes existing state-level sinkhole databases and sinkhole related studies to gain the domain knowledge necessary to inform the design of the unified geodatabase for inventorying sinkholes. In Chapter Three, potential end user needs are summarized, the structure of the FGS database used to guide the attribute design of the geodatabase is outlined, and the Kentucky Geological Survey's methodology for the authoritative collection of sinkhole data, which was used to populate the implemented test geodatabase, is explained. Chapter Four assesses the success and limitations of the geodatabase design by implementing a prototype database and testing it with a series of sample queries from the three potential user communities. Finally, Chapter Five discusses the future directions this project could go beyond the original objectives of this study.

## **Chapter 2 Existing Sinkhole Inventories and Related Research**

The question this thesis seeks to answer is: What combination of design factors for the unified geodatabase (UG) could be integrated that would allow existing state-level sinkhole inventories to be more efficient at the state level? This chapter is divided into three sections. It begins with a brief overview of research about sinkholes carried out within the U.S. and internationally. That is followed with an overview of existing State Geological Survey (SGS) sinkhole inventories and databases, focusing especially on those by the Florida Geological Survey (FGS) and Kentucky Geological Survey (KGS) which are most relevant to this project because of their well-developed status. Next an overview of the database design process and an evaluation framework is provided. Finally, an example of another unified database that is in the process of being implemented concludes this chapter.

### **2.1 National and International Sinkhole Studies**

This section discusses how national and international research on sinkholes aligns with this research. The national studies are very similar; therefore, discussing one in detail is sufficient, allowing focus to shift to the international studies, whose methodologies are more relevant to the UG design. Sinkhole research has primarily focused on delineating sinkhole features or creating models to predict future sinkhole activity. There is no research evidence of a successful long-term development of a database to record sinkhole events in a consistent manner in these studies, something this thesis attempts to resolve.

#### *2.1.1. National Studies*

Sinkhole studies in the U.S. have been conducted in several states including Florida, Illinois, Kentucky, Minnesota, and Texas. The Texas study verified only 20 sinkholes (Harbert

2014) and is therefore excluded from this study due to its small size. Gao et al. (2006) created a Karst Features Database for Minnesota using data from the state geological survey (SGS) and topographic maps focused on organizing and making the sinkhole data accessible publicly in Esri's ArcView along with Microsoft Access tabular formats. According to the authors, the main advantage of using ArcView was the ability to display features in a spatial context rather than only using tables. One important point mentioned was the need for different database permissions where the database administrator has full access while some individuals can only edit attributes and the public can only view the data. This tiered access structure is an important consideration in the design of the UG to ensure data quality by reducing human error.

The other studies were interested in identifying and mapping sinkholes during a limited research period. This approach was understandable since the Areas of Interest (AOI) were large, for example, Monroe County in Illinois is almost 400 square miles and had almost 3,000 sinkholes inventoried (Angel et al. 2004). A survey of the State of Kentucky by the Kentucky Geologic Survey (KGS) from 1999 to 2003 resulted in an inventory of over 100,000 sinkhole areas, depressions where sinkholes or other karst features could occur (Lee 2005). Fleury, Carson, and Brinkmann (2008) recorded about 1,500 sinkholes in four counties on the west coast of Florida.

A large database compiled by an academic researcher in Tennessee reportedly has over 54,000 sinkholes, but only 18,081 of those with a depth of at least three meters or more are available as a table (Dunigan 2015). This restricts the full evaluation of the dataset because the remaining sinkholes were provided only as web maps without a download option. From the subset of data available, the attributes concentrate on the sinkhole size and location – coordinates, perimeter, area, depth, volume, and elevation – attributes that are similar to those

included in the authoritative databases described in the next section. Based on the date of the table file's creation, it appears that this data was compiled in 2013 and has not been updated since that time. Since Dunigan's work cannot be evaluated fully and what could be found did not yield any new information, it was decided that further consideration of this database as input to the design was unnecessary.

The national research studies generally collected the locations of sinkholes using USGS topographic maps and on-screen digitizing in ArcGIS. Angel et al. (2004) mapped sinkholes in Monroe County in Southern Illinois, an area that is defined by 42 Public Land Survey System (PLSS) sections. Using 1:24,000 USGS topographic maps to delineate sinkholes, the goal was to identify sinkholes with hardcopy map manual counting and compare the accuracy of the concurrent GIS count. Researchers classified sinkholes into four categories: simple, complex, compound, and ponded. The latter two were derived in GIS with algorithms. The initial GIS setup took longer to create but was as reliable as hand counting: 2,823 sinkholes to a hand count of 2,830. This shows that topographic map interpretation is an effective method to identify sinkholes, the method which most of the SGS use to create inventories.

The main strength of Angel et al.'s study was the confirmation that using digital topographic maps with a five-foot contour interval allowed the identification of small sinkholes that are the most common. The weakness in this study, as with the others, was it did not consider a long-term database design. If they had done so, the collected data could have been stored and built upon by others in the future. As the focus of that study and the others, with the exception of the KGS inventory, was to test methods for recording sinkholes, consideration of the future use of this data was not recognized as a goal. It seems that an opportunity was missed by not making the preliminary sinkhole data collected available to build upon by creating a standard database

template that others could utilize. While they do not focus on databases structures, these studies do demonstrate the type of data sources that can be used to populate a UG. Thus incorporating Esri's ArcGIS tools and geodatabase into the design of a UG can allow a smooth transition as considerable sinkhole data already exists in shapefiles and scanned maps can be easily digitized.

### *2.1.2. International Studies*

In contrast, international studies that were conducted in Spain, Belgium, Italy, the Island of Crete, and the Dead Sea area between Israel and Jordan had smaller AOIs and added a couple of other methods to identify and generate sinkhole susceptibility maps using Digital Elevation Models (DEMs) and Ground Penetrating Radar (GPR). Three of these studies, those conducted in Belgium (Eeckhaut et al. 2007), Spain (Galve et al. 2009) and Malaysia (Al-Kouri et al. 2013), provide key foundations for the design of the UG. These have been selected because they clearly state what variables contribute to sinkhole formation and thus ones that should be considered for inclusion in the UG attribute design. Since some of these variables coincide well with existing SGS sinkhole inventory attributes which are discussed in the next section, these studies are further examined there.

## **2.2 SGS Sinkhole Inventories**

This section weighs the SGS sinkhole database advantages and disadvantages and how they contribute towards the thesis goals. The FGS and KGS databases are discussed in detail as they form the foundations for the UG template. The research conducted for this thesis was through online scholarly databases and internet web searches to locate enough examples of sinkhole databases to form a consensus for a unified template. Therefore, it is not necessarily an exhaustive list as that was not the objective of this thesis.

The 16 U.S. states with 25 percent or more of their terrain in karst topography are included in this study. Although Colorado only has 14 percent of its area in karst terrain, the state tracks and offers sinkhole data publicly, and is therefore included here, making a total of 17 states examined for their handling of sinkhole information.

Each of these states was placed into one of four categories based on their method of distributing their inventories of sinkhole occurrences. First, some SGS such as Alabama, Indiana, Iowa and Kentucky have publicly available GIS layers. Second, others like Missouri and Ohio have kept their GIS data internal but published static electronic or web maps. Third, Colorado, Florida and Pennsylvania distribute a combination of GIS layers and static maps. Lastly, Georgia, Illinois, New Mexico, Tennessee, Texas, Vermont and Wisconsin, after extensive searching turned up no SGS sinkhole inventories, do not appear to have any publicly available sinkhole inventory data.

Table 1 summarizes techniques used by the various SGS to generate databases or maps along with other relevant information about the states. These databases and studies include digitized karst features from sources such as United States Geological Survey (USGS) topographic maps, LiDAR data, field site photographs and Volunteered Geographic Information (VGI). The number of features is the number of records found in the databases. The percent land surface as karst is derived from Tobin and Weary's (2004) research, which is depicted in Figure 3 in Chapter 1.

Table 1 SGS sinkhole database sources.

Category	State	% of Land Surface as Karst	# of Features in the database	Sources	Format
<b>GIS Layers</b>	Alabama	29.4	6,460	USGS Topo Map	Point Shapefile
	Iowa	47.4	37,548	SSURGO spot data, LiDAR, Photos	Point and Polygon Shapefile
	Indiana	43.0	1,201	Field Sketch Maps	Polygon Shapefile
	Kentucky	42.3	101,176	USGS Topo Map	Polygon Shapefile
<b>Maps</b>	Florida	81.2	3,580; 43 verified	Mostly VGI	Web map (Points) and Point Shapefile
	Missouri	58.1	15,981	USGS Topo Map	State map JPEG
	Ohio	31.7	975	LiDAR, Photos	County maps PDF
	Pennsylvania	27.0	141,274	USGS Topo Map, Aerial Imagery	Web map (Points)
<b>GIS Layers and Maps</b>	Colorado	14.1	1,557	Hardcopy records, National Agriculture Imagery Program	Point Shapefile and PDF Map
	South Carolina	27.3	28	SGS Investigation, Maps	Point and Polygon Shapefile, State map – PDF
<b>No Public Digital Data</b>	Georgia	56.4	-	-	-
	Illinois	26.0	-	Hardcopy Maps	Static County Maps, Polygon Feature Class
	New Mexico	32.0	-	-	-
	Tennessee	50.8	-	-	-
	Texas	28.1	-	-	-
	Vermont	29.3	-	-	-
	Wisconsin	26.1	-	-	Mostly VGI

Returning to the discussion in the previous section about attributes included in national research studies on sinkholes, Table 2 shows attribute categories (“Variable”) listed in some of the SGS databases that match those in three key national research studies (Eeckhaut et al. 2007, Galve et al. 2009, and Al-Kouri et al. 2013). It can be seen here that Florida’s database has the most compatibility with the susceptibility studies variables.

Table 2 Comparison between SGS databases and international studies

Variable	Geological Survey	Eeckhaut	Galve	Al-Kouri
Sinkhole Dimensions	Florida, Iowa, Kentucky	✓	✓	✓
Time	Florida	✓	✓	✗
Human Impact/ Land Use	Florida	✓	✓	✓
Topography	Florida	✓	✓	✓
Hydrology – Water Bodies, Precipitation	Florida	✓	✓	✓
Geology – Soil Type, Bedrock	Colorado, Florida	✓	✓	✓
Hydro Chemistry	✗	✗	✓	✗

For the UG to be successful, the existing databases used as design models had to be as complete as possible in both the number of sinkholes included and the range of attributes used. Building upon sinkhole databases designed by various SGS should allow the UG to be replicated within different states easily and create a smoother transition to a national universal geodatabase in the future.

### 2.3 FGS and KGS Databases as Models for This Study

For various reasons, the sinkhole databases created by Florida and Kentucky became key examples for the design and assessment of the UG developed in this project. Florida has the highest amount of karst topography in the nation at 81% of its terrain covered and Kentucky has 42% (Tobin and Weary 2004).

The FGS data is composed largely of volunteered geographic information (VGI). VGI's advantage is access to a large, voluntary labor force capable of collecting geographic data that would otherwise take official agencies using traditional methods long periods of time to perform (Goodchild 2007). While this means that a large amount of data can be quickly collected, VGI can have considerable data quality problems. The major issues with VGI arise from the collection of data by private citizens with no formal qualifications or rigorous quality assurance mechanisms in place to ensure result accuracy. The FGS database suffers from these problems

but also exists because of the unique advantages of VGI that allow rapid data collection through many volunteers. It provides an important design key for the UG and a data source for a test implementation of the UG.

The most significant concern of the FGS's database is how the VGI data is collected. There is a two page PDF form used to report a sinkhole event found on the FGS Website (<http://www.dep.state.fl.us/geology/default.htm>). As will be seen in the next chapter, the form is far too complex for a private citizen to fill out completely and accurately without specialized geological knowledge and access to GIS layers. That is likely why many of the attribute fields in the database are empty. Also, on many occasions, individuals want to remain anonymous, and so the contact information attributes are often also blank. Improving their reporting form is beyond the scope of this study but is discussed in Chapter 5.

The KGS database covers the entire State of Kentucky and has over 100,000 sinkhole areas recorded making it the most complete sinkhole areas inventory in the U.S. In contrast to the Florida VGI inventory, the value of this inventory for the purposes of this project is the data collection method. Here the complete survey of the state was conducted by digitizing as polygons from USGS topographic maps all depressions indicating areas of potential sinkhole activity. As discussed in Chapter 4, this methodology provides a means to validate the FGS VGI data used in the case study.

## **2.4 The Database Design Process**

While the type of database used for the UG is a spatial database, there is a common set of steps used to create any database. These are described in detail by multiple authors. Microsoft (2015), Sekstrin (2015) and sources as far back as the University of California, Berkeley (1997) provide similar useful frameworks. Based on these sources, Table 3 summarizes the ten steps in

database design that were applied to the UG design process in this thesis. The completion of each of these steps is described in the next two chapters, with steps one to nine outlined in Chapter 3 and Chapter 4 dedicated to testing the geodatabase as noted in step ten.

Table 3 Database design steps

<b>Step</b>	<b>Description</b>
1	Determine the purpose of your database
2	Find and organize the required data
3	Create a simplified E – R Diagram
4	Divide the data into entities and attributes
5	Decide which entities and attributes you want to store in each layer
6	Specify primary and foreign keys
7	List cardinal relationships between layers
8	Create detailed ER Diagram with keys and cardinal relationships
9	Convert detailed ER Diagram into geodatabase format
10	Test geodatabase and refine design as needed

*Source:* Adapted from Microsoft (2015) and Sekstrin (2015)

#### 2.4.1. Database Design Evaluation

The methodology for the evaluation of a database design lacks a set of agreed upon standards (Lukyanenko and Parsons 2012). Hoxmeier (1998) outlined one possible framework that continues to be widely used (see for example Cherfi, Akoka and Comyn-Wattiau 2011 and Singh et al. 2011). He suggests there are four dimensions — process, data, model and behavior — that can be used to assess database quality for a specific problem domain. Each of these encompasses various aspects of database design (Figure 6):

- The process dimension is where the domain knowledge used to build the database design occurs both conceptually and physically. This is also the step where database implementation and performance evaluation occurs, thus ensuring the database is not being hindered by poor design features and maintenance operations.

- The database data quality dimension addresses the data itself that makes up the database. These are factors ranging from accuracy, source data, accounting for temporal changes and security safeguards that any standalone data or database must account for to ensure integrity.
- The database model dimension is concerned with how well users and the database model interact. Factors such as representation, scope, and consistency help answer questions such as is it easy to use, understandable, and consistent and how much of the problem domain does it respond to?
- Finally, the database behavior dimension assesses how effectively the database responds to the problem domain. The effectiveness of the solution can vary based on multiple factors such as: the experience of the database designer, scope of the problem domain and the three previous dimensions of database design quality.

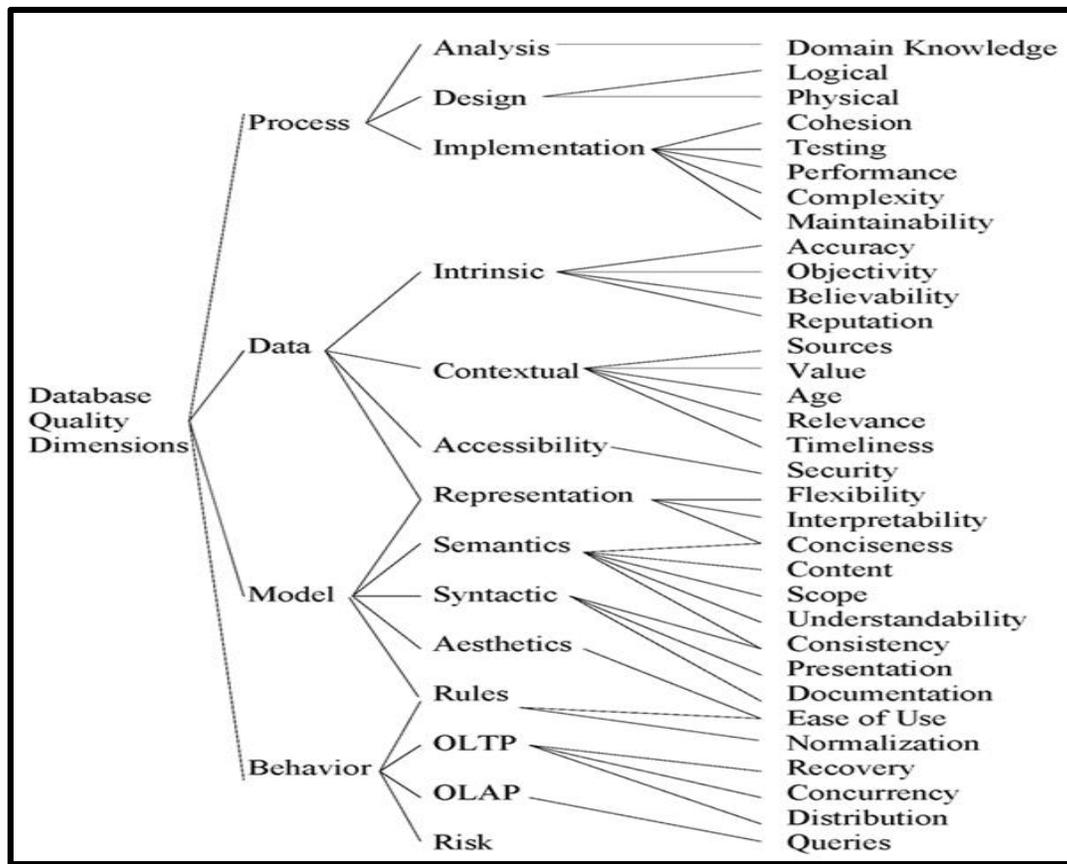


Figure 6 Database quality dimensions (Hoxmeier 1998)

As discussed later in Chapter 4, these database design quality evaluation aspects provide a useful template by which to assess the final design of the UG.

## 2.5 The National Address Database – a Unified Database Design Example

The UG design undertaken in this research is not unique in its attempt to unify information compiled by multiple entities into a single database. A current example of an opportunity for a unified database is being addressed by the U.S. Department of Transportation (Otto 2015). There is a widely-recognized need to have a single national database with every address in it. Groups or activities that would benefit include emergency responders, Census tracking, income tax collection, and natural disaster planning. These are just a few of the applications that a centralized address database would support (McKinney 2015). Similar to the

case of the various SGS sinkhole databases having their own designs and varying levels of data collection, address data collection relies on local and state entities to gather this information. As of 2013, only 22 states had some type of address data collection plan implemented and of those 22, only eight had data for the entire state (Otto 2015).

A National Address Database Summit was held at the Maritime Institute in Linthicum, MD on April 8-9, 2015 where 58 participants attended from all levels of government as well as private and non-profit organizations. Four main points were agreed upon at the summit, which included: (1) local authorities would be responsible for data collection; (2) state agencies would aggregate the local level datasets; (3) Tribal Nations, US Territories and the District of Columbia will have a role in the NAD; and (4) federal leadership and support is required for a national implementation of NAD. Although this summit did not provide specific information about database design, it did recognize the need for standardization of data, taking into account logistics when so many agencies are involved and deciding on using data collection methods that would be implementable by agencies with limited resources (Applied Geographics 2015). These final three objectives were taken into consideration for the UG because they are applicable to the work performed in this thesis.

## **2.6 Summary**

After reviewing the database design process, international and national sinkhole research studies, as well as the SGS databases, a strategy for developing the design of the UG emerged. While the national studies and three SGS examples provided context as to why sinkhole areas should also be included as a separate feature layer, the international studies argued the need to include some of the sinkhole development factors as attributes in the UG layers.

The scale of the AOIs present in the studies led to evaluation of the geodatabase design by doing a prototype implementation in only a portion of Hillsborough County, Florida, a county that had the most reported sinkholes in the state's database. Testing and refining the UG design at this scale should reveal problems relevant at the state scale of use that is the initial goal for the UG. As this thesis focuses on a standard geodatabase design incorporating data from various sources, it is essential to include attributes common to various database models and consider the differences that exist to ensure that most people benefit when using the UG. As explained next in Chapter 3, using the most comprehensive existing SGS databases as templates provided the foundation on which to proceed through the database design steps.

## **Chapter 3 A Unified Sinkhole Geodatabase Design**

The primary objective of this thesis is to design a unified geodatabase template to consistently manage sinkhole data collections. Adapted from existing SGS sinkhole database models, the unified template is designed to incorporate the key attributes and functions exhibited by them. The design also considers the needs of anticipated user communities, specifically geologists, insurance fraud investigators and the general public who have a range of uses that the UG should be able to resolve. This chapter is divided into three sections: a description of the key user groups, an overview and summary of the characteristics of the various SGS database models, and a description of the steps and decisions leading to the final design of the UG template.

### **3.1 User Communities**

To ensure the design of the UG is useful, three user communities were identified as potential consumers of the UG. Each user community has a different level of GIS expertise; geologists are considered to be experts, insurance fraud investigators are intermediate, and the general public are novices. In this section, an overview of what each user community's needs are and how the UG should fulfill them is provided. Later in Chapter 4, possible queries for each user community are tested as one step in the evaluation of the design.

#### *3.1.1. Expert GIS Users: Geologists*

Geologists would use the UG as an additional resource in combination with their own GIS data for research similar to the sinkhole studies occurring around the world and for hazard mapping. This group would be doing advanced spatial analysis, aerial interpretation, and independent sinkhole mapping studies, thereby members of this group would not be limited by the contents of the UG nor lack of GIS abilities to meet their objectives. Realizing that such

advanced users would be one of the communities using the UG, the decision not to include attributes such as “slope,” “land use,” and “trigger mechanism” is valid because they can easily locate and add layers with this information to their analysis. Specifically, attributes such as “trigger mechanism” cannot be easily acquired without expert knowledge and would require a field check. In addition, if the final database were to be populated from VGI reporting (i.e. non-authoritative sources), such attributes may be incorrectly reported.

### *3.1.2. Intermediate GIS Users: Insurance Fraud Investigators*

Insurance fraud investigators would find the UG useful for querying specific information and combining it with some of their own data to meet their business needs. In comparison to the geologists, insurance fraud investigators would be relying on this geodatabase to complement their own data and make it more complete. There would be less abstract and analytical use of this data and more practical applications such as tracking how many properties are damaged by sinkhole events. From an insurance fraud investigator’s perspective, attributes concerning if a reported sinkhole is a true sinkhole, whether property damage occurred and if repairs are planned are among the most pressing information needed for them to perform their business transactions. The UG should allow them to cross-reference with their own information to reduce fraudulent and duplicate claims.

### *3.1.3. Novice GIS Users: General Public*

The general public would have the simplest GIS queries and visualization needs for the UG. Curiosity and safety concerns are the main reasons the public would be viewing this data and thereby raising awareness of their surroundings. People would be curious to know how many sinkholes have been reported in the city where they live, which could then lead to awareness of a safety concern as to where they are. However, their queries and explorations of the data may not

have a specific goal like research in a geologist's case or a part of their occupation as is the case for insurance fraud investigators. Therefore, the information included in the UG for the public is for awareness purposes only and does not have a precise use when compared to the other user communities.

## **3.2 Design Principles**

The UG was designed by integrating three components: potential needs of three user communities, insight from previous research on sinkhole phenomena, and existing state sinkhole database models. The potential user communities were discussed in the previous section. This section summarizes the choices made about the fundamental principles that guided the database design.

### *3.2.1. Insight from Sinkhole Studies*

Previous sinkhole research studies, such as those discussed in Chapter 2 which explore risks and causes of sinkholes, suggest that given the multitude of factors that may be incorporated into such studies, there is no need to include in a sinkhole database a great deal of additional information that geologists and others might have or use from other sources. Adding this information into the main sinkhole database tables would greatly and unnecessarily expand their size. Therefore, layers such as land use, precipitation, slope, karst formations, and many more are not included in the UG template. Such data is easily found and can be related spatially as needed.

### *3.2.2. Key Sinkhole Database Models*

From the initial 17 states included in this study, 12 had some type of published sinkhole data. These 12 karst-rich states with published data were categorized in Table 1 by those who

provide: (1) public GIS data about sinkholes, (2) published maps showing sinkholes, either web based or static, and (3) those using a combination of the first two categories. As Table 1 shows, while they did have published data, Missouri and Ohio only had maps of the point locations of sinkholes and besides the spatial display did not provide additional information that would be useful here to inform UG attribute choices. This reduced the set of states with sinkhole databases of interest in the design phase to ten.

Table 4 shows the ten databases were organized mostly using ArcGIS shapefiles and, in a couple of cases, ArcGIS feature classes. This lends support for the use of an ArcGIS geodatabase as the technology for the database implementation. Most of the GIS data used a point format to represent a single sinkhole event. In a few instances, in Iowa, Kentucky, Pennsylvania and South Carolina, polygons were utilized to capture *sinkhole areas* and not individual sinkholes. As explained earlier, sinkhole areas are land depressions where sinkholes or other karst features could occur. After examining these two approaches that the state databases were using for recording features, it was determined that the UG would need to have both a sinkhole point feature class and a sinkhole areas polygon feature class, a multi-feature model used in Iowa.

Table 4 SGS databases contributing to the UG design. \* Indicates no data found.

<b>State</b>	<b>Data Format</b>	<b>Spatial Display</b>	<b>Data Source</b>
Alabama	Point Shapefile	Point	USGS Topo Map
Colorado	Point Shapefile and PDF Map	Point	Hardcopy Records and USDA Aerial Imagery
Florida	Point Shapefile and Web Map	Point	Mostly VGI based on available data
Illinois*	Polygon Feature Class	Polygon	Metadata - Hardcopy Maps
Indiana	Polygon Shapefile	Polygon	Hardcopy Maps
Iowa	Point and Polygon Shapefile	Polygon and Point	SSURGO spot data, LiDAR, Field Site Photographs
Kentucky	Polygon Shapefile	Polygon	USGS Topo Map
Pennsylvania	Web Map with shapefile/geodatabase and raster download options	Polygon and Point	USGS Topo Map and USDA Aerial Imagery
South Carolina	Point and Polygon Shapefile, Coastal area map - PDF	Polygon and Point	SGS Field Studies
Wisconsin*	Unknown	Unknown	Mostly VGI based on online reporting form

Table 4 also shows that all state geological survey sinkhole databases except Florida and Wisconsin were compiled authoritatively, relying primarily on expert map interpretation. Florida and Wisconsin were the only two states that relied on Volunteered Geographic Information as a primary source of information.

Table 5 summarizes the attribute categories that the ten states with detailed databases had developed, showing which ones they had in common. It is worth pointing out that the attribute themes of Florida and Wisconsin are very similar because they both use VGI, while among the authoritative sources, Alabama and Pennsylvania have several in common with the VGI source states. To accommodate both VGI and authoritative sources, it was decided that the UG should emulate the attributes used in Florida and include a combination of attributes from the authoritative databases as well. Florida's database was chosen as the foundation of the UG because existing databases kept VGI and authoritative methodologies separate which would complement each other if merged. It is important to note that Florida's set of attributes were

chosen over Wisconsin's since, although their sinkhole online reporting form was located and potential attributes could be surmised, no actual data could be accessed.

Table 5 Sources and common attribute themes in SGS inventories

	Alabama	Colorado	Florida	Illinois	Indiana	Iowa	Kentucky	Pennsylvania	South Carolina	Wisconsin
<b>Data Source</b>										
Topo Map	✓	✓		✓	✓		✓	✓	✓	
VGI			✓							✓
Field Site Photograph						✓				
Imagery		✓				✓		✓		
<b>Attributes</b>										
Lat/Long coordinates	✓							✓		✓
Sinkhole/SHA Size			✓	✓			✓			✓
County Name	✓	✓	✓					✓		
Geologic Comments		✓	✓	✓	✓			✓		✓
General Comments			✓						✓	✓
Occurrence Date	✓		✓						✓	✓

### 3.2.3. Distillation of State Sinkhole Database Models

The Florida Geological Survey database table served as the primary attribute model for the unified geodatabase. Modifications made for the UG included removing redundant attribute fields for location, such as county and zip code which can easily be referenced with other GIS layers and including features such as a cave, fissure or general subsidence even though they are not strictly sinkholes. To illustrate how state database models were incorporated into the design, Table 6 shows the attributes emulated from the Florida database into the UG database. The complete list of attributes from the Florida database is shown in Appendix A.

Table 6 UG attributes emulated from Florida database

<b>Attribute Name from Florida Database</b>	<b>Florida Description</b>	<b>Attribute description in UG</b>
TRUE_SINK	Verified sinkhole	Separates sinkholes from other karst features
DATE_REV	Date Revised	Record Update
COUNTY	County	Location reference attribute
RPT_SOURCE	Source of report	Who reported it
RPT_PHONE	Report phone number	Contact information
RPT_NAME	Report name	Contact information
EVENT_DATE	Date of reported event	Date of event
SINSHAPE	Shape	Sinkhole Shape
SINLENGTH	Length	Dimension
SINWIDTH	Width	Dimension
SINDEPTH	Depth	Dimension
PROPDAM	Property damage	Was there damage?
REPAIRED	Feature repaired	Sinkhole repaired – Y/N
PLANNED	Repairs planned	Repairs planned – Y/N
COMMENTS	Other comments	Comments about event
COMMENTS2	Additional comments	Comments about event
WITNAM	Witness Name	Name of person who saw it
WITPHONE	Witness Phone	Contact witness about event

The 12 SGS with some type of data collections had relied on either an authoritative data collection method or VGI for compiling sinkhole locations. These distinct approaches were merged in the unified geodatabase template in order to allow the public to be involved in reporting potential sinkholes that an SGS would not be able to collect with its own staff in the same amount of time. This would allow SGS staff to prioritize verifying VGI sinkhole reports in urban settings where the risk is greater and free them to document sinkholes in more rural and remote areas where VGI reporting may be less frequent.

In the UG, a separate VGI Source table contains information about the person reporting the sinkhole. This was put into a different, linked table to separate contact information from the primary point feature class. This not only allows that information to be kept confidential for some users, but it also keeps the main sinkhole table more compact when compared with the bulky FGS database table.

#### *3.2.4. Identification of Feature Layers*

The UG has two separate layers for the geographic features. Replicating many of the SGS examples, the most important feature layer in the database is the sinkhole point feature layer. However, two states using authoritative methods, Kentucky, and Pennsylvania, did not compile individual sinkhole locations but rather they recorded areas of sinkhole activity. These areas were derived from USGS topographic maps by identifying the upper-most closed contours around depressions. These sinkhole areas (SHAs) represent places where sinkholes are or could be occurring. Therefore, it was decided to include such SHAs as a layer in the unified geodatabase to show possible sinkhole prone locations in addition to individual sinkholes. As described later in Chapter 4, this method of identifying SHAs was used as a means of populating the prototype database.

#### *3.2.5. Why Use an ArcGIS Geodatabase?*

An ArcGIS geodatabase was chosen as the technology in which to implement the design. A file geodatabase was chosen primarily because it can store spatial and tabular information and for three additional reasons: customization options, file organization, and storage capacity. The geodatabase offers a range of options to customize attributes and to organize different data types. The use of domains and subtypes for attributes helps eliminate human error when entering data and provides a fixed number of responses to choose from, thus encouraging record completion.

Importantly, all data is stored in a single file that is a huge advantage because multiple data types are all in one location. With standalone shapefiles, which are composed of a collection of files with the same name and different file extensions, component files can get lost or deleted by accident as some users may not use ArcCatalog to gather them all in one folder, relying on Windows Explorer. The projection file, attribute table or index file could be misplaced resulting

in data not loading and, depending on the user's GIS ability, could make this task difficult to resolve. Finally, a file geodatabase has no data storage limit while a personal geodatabase is limited to two Gigabytes (GB). This leaves one fewer issue to worry about as a national geodatabase in the future combined with other layers could easily exceed two GB.

### **3.3 Unified Geodatabase Design**

This section reviews the overall structure of the geodatabase design. It begins with a summary of the basic database structure, followed by a description of the attributes included in each database component. It concludes with a description of the full geodatabase with relationships and domains outlined.

#### *3.3.1. Database Structure*

The database is composed of two tables and two feature classes. The database structure is represented by the basic Entity-Relationship Diagram (ERD) shown in Figure 7. The polygon feature class is the digitized depressions that show SHAs while the point feature class contains the sinkholes. The two tables are for any sinkhole events that came from a VGI source. The ERD also shows the primary and foreign keys and the types of relationships between the different datasets.

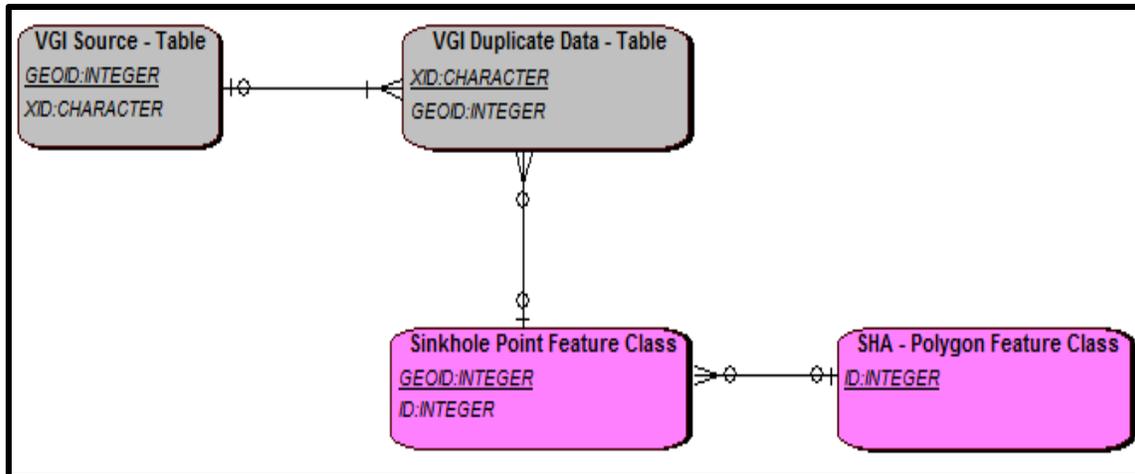


Figure 7 Entity-Relationship Diagram of the UG structure

Data to populate the sinkhole point feature class may come from both authoritative sources and VGI sources. Authoritative data will populate only the point feature class. If the data comes from a VGI source, the sinkhole point in the feature class will have additional information related to it in the two VGI tables.

The VGI Duplicate Data table is important because when a sinkhole event occurs, there is a chance that multiple reports could be describing the same feature, and this data needs to be validated by the SGS as one record in the point feature class. The multiple representations are deleted from the point feature class after validation occurs. At the same time, it is important to save the multiple reports of the same feature to the VGI Duplicate Data table. They can be useful for end users to validate entries themselves because the SGS may have not yet verified if it is a true sinkhole or a single event. This type of a situation could be very helpful for insurance fraud investigators who would want to accurately track sinkhole claims.

The relationships between the feature classes and tables are as follows: a sinkhole area can have one or more sinkholes intersecting it, a sinkhole may not intersect a sinkhole area at all; a sinkhole point record can have multiple VGI duplicate data records linked to it, a VGI

duplicate record may not have a link to any sinkhole point; a VGI source record can have links to multiple VGI duplicate data records and finally, a VGI duplicate record must have a link to a VGI source record. Sinkhole points do not always coincide with sinkhole areas because they are derived from different data sources and sinkhole areas in most cases represent significantly larger features than the majority of sinkholes.

### 3.3.2. Attributes

The following tables (Tables 7 to 10) show the attributes that are included in each component of the UG along with a description of what they represent. For each of the components, the primary and foreign keys are combined with the U.S. Census Bureau state and county FIPS codes. These are unique numbers assigned to every state and county in the U.S. making them ideal for identifying each record uniquely at the state level and providing a way to organize records at the national level in the future. The source column indicates which SGS contributed to the particular attribute design, if the designation is “Multiple” then three or more SGS databases had the same or very similar attribute in their designs. Finally, if the designation is “New”, then it was created for this design.

Table 7 SHA polygon feature class attributes

Attribute Name	Definition	Source	Type	Character Length
ID*	Primary Key using State FIPS Code and number of up to 999,999. To link Point and Polygon Layers	Iowa	Long Integer	8
Area	Area of SHA in acres	Kentucky	Double	4
D_Type	SHA Depression Type (Dry, Lake, Marsh or Water)	New	Text	20
Quad_Name	USGS Topographic Map name	Alabama	Text	20
Year_Pub	Year USGS Map was published	Alabama	Short	4
Year_Rev	Year USGS Map was revised	Alabama	Short	4
Comments	Additional comments about the SHA	New	Text	50

Table 8 Sinkhole point feature class attributes

<b>Attribute Name</b>	<b>Definition</b>	<b>Source</b>	<b>Type</b>	<b>Character Length</b>
ID*	Foreign Key using State FIPS Code and number of up to 999,999. To link Point and Polygon Layers	Iowa	Long Integer	8
GEOID*	Primary Key using State FIPS Code, County FIPS Code and a number up to 999,999. To link Point layer to VGI Source Table <b>and</b> VGI Duplicate Data Table	Iowa	Long Integer	11
Feature_Type	Type of feature - sinkhole, fissure, cave, etc.	Illinois	Text	20
Sink_Shp	Sinkhole Shape –circular, elongated, unknown	Florida	Text	20
Sink_Verified	SGS verification of event– Y, N or P (pending)	Florida	Text	1
Source	Data Source – VGI, Topo, NAIP	Alabama, Florida	Text	15
Date	Date of occurrence	Florida	Date	n/a
Date_Revised	Date record updated	Florida	Date	n/a
Address	Address associated with sinkhole	Florida	Text	50
City	City name of event	Multiple	Text	20
State	State in which event occurred	Multiple	Text	20
Length	Feature length (ft)	Florida	Short Integer	3
Width	Feature width (ft)	Florida	Short Integer	3
Depth	Feature Depth (ft)	Florida	Short Integer	3
Prop_Dmg	Did property damage occur – Y,N, Unk	Florida	Text	3
Repairs	Were features and property repairs planned – both Y, Both N, Feature Y & Prop N, vice-versa, Unk	Florida	Text	10
Drainage	Any drainage in or around feature – Y, N or Unk	Illinois	Text	3
Comments	Additional comments about feature	Multiple	Text	250

Table 9 VGI Source table attributes

<b>Attribute Name</b>	<b>Definition</b>	<b>Source</b>	<b>Type</b>	<b>Character Length</b>
GEOID*	Primary Key using State FIPS Code, County FIPS Code and a number up to 999,999. To link Point layer to VGI table	Iowa	Long Integer	11
XID*	Foreign Key using an "X" then State FIPS Code, County FIPS Code and a number up to 999,999. To link VGI Duplicate Data Table to VGI Source Table	Iowa	Text	12
Contact	Y/N answer if the person(s)/ organization want their information to be public knowledge and are open to being contacted regarding event(s)	Florida	Text	1
Name	First and Last	Florida	Text	30
Organization	If individual not available perhaps an organization is provided	Florida	Text	40
Address	Mailing address for contact	Florida	Text	50
Email	Email for contact	Florida	Text	30
Phone	Phone number for contact	Florida	Text	20
Property_Owner	Is report source, property owner –Y/N	Florida	Text	1

Table 10 VGI Duplicate Data table attributes

Attribute Name	Definition	Source	Type	Character Length
XID*	Primary Key using an “X” then State FIPS Code, County FIPS Code and a number up to 999,999. To link VGI Duplicate Data Table to VGI Source Table	Iowa	Text	12
GEOID*	Foreign Key using State FIPS Code, County FIPS Code and a number up to 999,999. To link Point layer to VGI Duplicate Data Table	Iowa	Long Integer	11
Feature_Type	Type of Feature – sinkhole, fissure, cave, etc.	Florida	Text	20
Sink_Shp	Sinkhole Shape –circular, elongated, unknown	Florida	Text	20
Sink_Verified	SGS verification of event Dropdown menu – Y, N or P (pending)	Florida	Text	1
Date	Date of occurrence	Florida	Date	n/a
Date_Revised	Date record updated	Florida	Date	n/a
Address	Address associated with sinkhole	Florida	Text	50
City	City name of event	Multiple	Text	20
State	State in which event occurred	Multiple	Text	20
Length	Feature length (ft)	Florida	Short Integer	3
Width	Feature width (ft)	Florida	Short Integer	3
Depth	Feature Depth (ft)	Florida	Short Integer	3
Prop_Dmg	Did property damage occur – Y,N, Unk	Florida	Text	3
Repairs	Were features and property repairs planned – both Y, Both N, Feature Y & Prop N, vice-versa, Unk	Florida	Text	10
Drainage	Any drainage in or around feature – Y, N or Unk	Illinois	Text	3
Comments	Additional comments about feature	Multiple	Text	250

### 3.3.3. Geodatabase structure

Having defined the included attributes, Figure 8 shows an expanded Entity-Attribute-Relationship Diagram (EARD) of the two feature classes and two tables that make up the UG along with their corresponding attributes and associated domain values. The attributes to which the domains, shown in the lower portion of Figure 8, are applied are listed in Table 11. Domains denoted with a “\*” or “\*\*” are shared by more than one attribute while all other domains apply to an attribute of the same name. Finally, all the keys and relationships between the components are displayed in Figure 8 encompassing the complete structure of the UG.

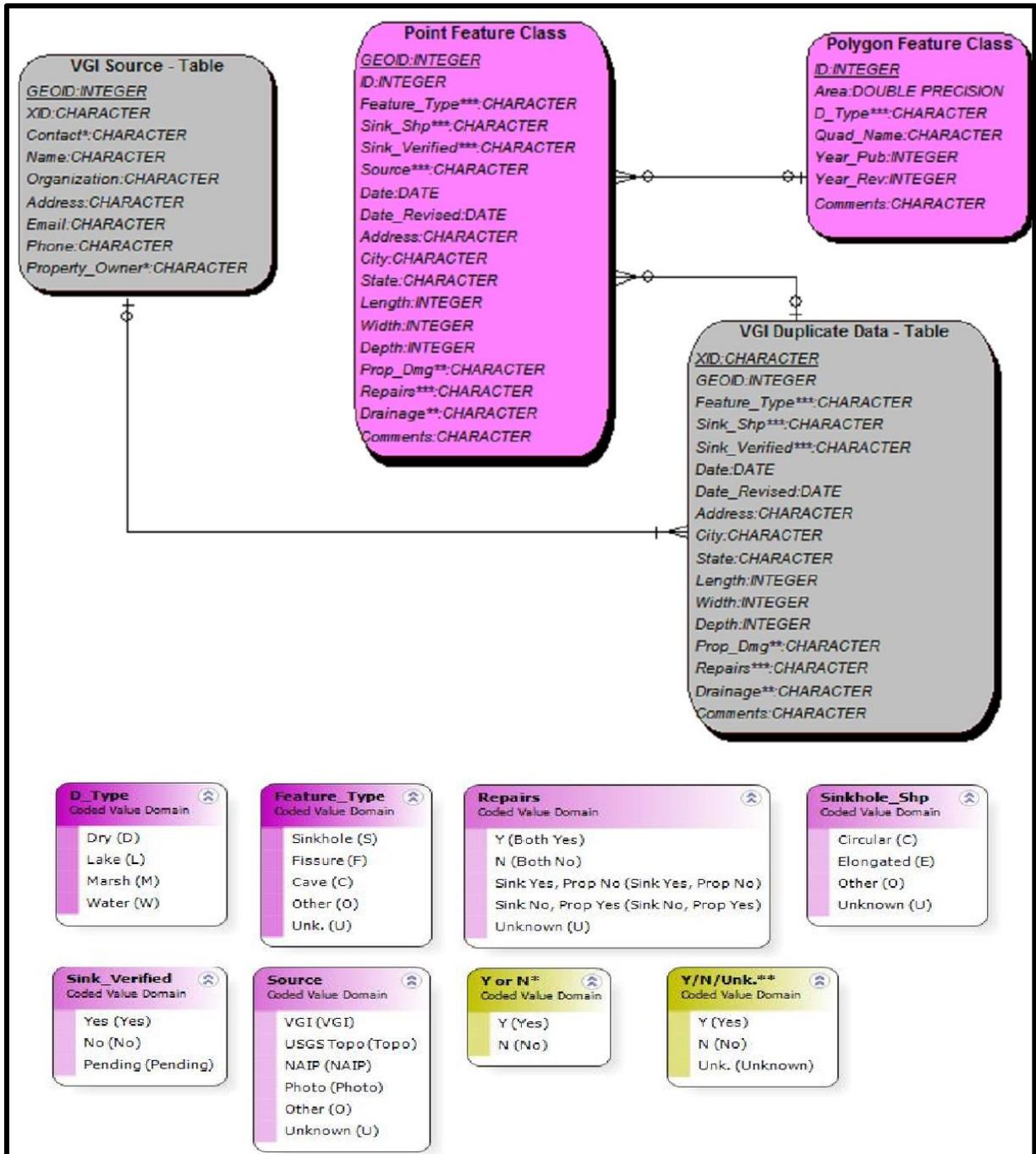


Figure 8 Entity-Attribute-Relationships Diagram of the UG structure

Table11 Attribute-domain associations

<b>Domain Name</b>	<b>Component</b>	<b>Associated Attributes</b>
Y/N	VGI Source	Contact, Property_Owner
Y/N/Unk.	Point and VGI Duplicate Data	Drainage, Property_Dmg
All others	Point, Polygon, and VGI Duplicate Data	Domain and Attribute have the same name

### **3.4 Geodatabase Design Data Integrity**

There are at least three ways that the design of the unified geodatabase manages to increase data integrity: (1) the use of the SGS methodologies that generated the sinkhole areas layer; (2) cross-referencing several VGI records to accurately as possible identify possible sinkhole feature(s); and (3) having primary and foreign keys between components is an essential feature to reduce data entry errors and duplicate records.

#### *3.4.1. Integration of Both Sinkholes and Sinkhole Areas Features*

The majority of the SGS databases were populated using some type of terrain analysis. In some cases, they used USGS topographic maps to identify closed contours, LiDAR to generate DEMs and identify depressions, or other remote sensing techniques such examining NAIP Imagery to identify individual sinkhole features or SHAs. The use of such techniques is most helpful in mapping SHAs as these are features with areas in acres that are larger than most individual sinkholes. In turn, these locations provide an indication of where sinkholes could be occurring. As demonstrated later in the implementation tests (Chapter 4), of the 211 VGI sinkholes from the Florida Database located in the portion of Hillsborough County, 195 were within half a mile of an SHA and 210 were within a mile. In fact, flagging sinkhole features located in areas lacking SHAs might be used as an indication of errors, or possibly areas that should be early targets for a field check by SGS staff.

### *3.4.2. Cross-Referencing VGI Data*

The separation of data collected as VGI from that collected by authoritative techniques in the SGS databases sometimes weakened data integrity of existing designs since the components could not be easily integrated. That is why it was decided to combine both methods of data collection in the UG. VGI accounts can cover large areas in a fraction of the time that SGS staff can survey them. VGI also identifies features that the SGS may want to investigate further.

Perhaps VGI's most essential quality is having multiple witness accounts about the same feature. Having more than one witness account provides a wealth of intelligence about a single reported feature from different people that the SGS staff can filter through before finalizing the record in the sinkhole feature class in the UG.

For example, consider a scenario where three people have reported through a VGI reporting interface details about the same sinkhole. It is likely that all reported locations will differ slightly, so their duplication will not be immediately apparent. When an SGS staff member examines the reports and deduces that they represent the same sinkhole feature, only one spatial representation will be determined to be the most accurate and kept in the point feature class. However, just as the three spatial locations can vary, so can the attributes and the spatial point chosen as the most accurate location wise may not have the most accurate and complete attributes. In this situation, the attributes from another record are copied to the one with the best spatial accuracy. Finally, all the records are kept in the VGI Duplicate Data table with a foreign key relating them to the spatial record in the point feature class.

When appropriate, multiple accounts will become associated with a single sinkhole point feature while the multiple accounts continue to be preserved in the VGI Duplicate Table, whether they have or have not been verified by the SGS. This would allow interested parties to

pursue their own investigations of these reports, possibly uncovering additional information or verifying if a feature is a sinkhole or not, and informing SGS staff for their attention.

However, multiple reports of the same feature create a potential problem resulting in duplicate records. This is addressed by the use primary and foreign keys which provide a way to identify related records. The design of these keys is discussed in the next section.

### *3.4.3. Primary and Foreign Keys*

Multiple VGI witness accounts can easily create duplicate records and other errors in data entry if there is no way to identify them separately. That is why the UG uses a simple structure to organize VGI accounts in separate tables to show how VGI source information was derived for the sinkhole point feature class. Using a set of primary and foreign keys to identify records uniquely and allow them to be related between components is an important means to ensure data quality. This is valuable in the situations when multiple VGI records exist for the same feature or when the SGS have their own data and choose to combine it with some VGI to update an existing record.

### 3.5 Summary of Database Design Steps Completed

To summarize the design process accomplished in this project and described in this chapter, Table 12 matches each of the database design steps outlined in Chapter 2 with the task completed in this project that best fits the description.

Table 12 Database design steps summary

<b>Step</b>	<b>Description</b>	<b>Task Completed</b>	<b>Relevant Section or Table</b>
1	Determine the purpose of your database	Acquired domain knowledge	Section 1.1, 1.2
2	Find and organize the required data	Investigated SGS databases	Section 2.2, 2.3, Table 1
3	Create a simplified ER Diagram	Determined database model	Figure 7
4	Divide the data into entities and attributes	Decided which entities and attributes are essential to design	Table 5
5	Decide which entities and attributes you want to store in each layer	Decided which entities and attributes are not redundant	Section 3.3.2
6	Specify primary and foreign keys	Created keys based on U.S. Census Bureau FIPS codes	Section 3.3.2
7	List cardinal relationships between layers	Determined how the four layers relate to each other	Section 3.3.1
8	Create detailed ER Diagram with keys and cardinal relationships	Insured consistency of process steps 4 – 7	Section 3.3, 3.4, Figure 8
9	Convert detailed ER Diagram into geodatabase format	Created geodatabase in ArcGIS 10.3.1	Section 4.2
10	Test geodatabase and refine design as needed	Discussed in next chapter	Section 4.3, 4.4

Having described the design of the UG, the next chapter describes how a prototype of the geodatabase was implemented and populated as a means of evaluating the success of the design.

## **Chapter 4 Evaluation of the Unified Geodatabase Design**

In order to evaluate the unified geodatabase design, it was necessary create, populate and test a prototype implementation. This chapter discusses four related topics: (1) the study area used for the test implementation of the geodatabase design, (2) the construction and population of the prototype geodatabase, (3) the results of a sample of queries for each of the user communities that can be answered using the prototype sinkhole geodatabase, and (4) additional database design qualities.

### **4.1 Study Area for Implementation Tests**

The study area chosen is a portion of Hillsborough County in Florida. This County had 556 reported sinkholes as of the August 31, 2015, release of the Florida database, the most of any Florida County. It was, therefore, considered a good testing location for the geodatabase design. In Figure 9, the map on the left shows the State of Florida and Hillsborough County. The map on the right shows the study area highlighted by a black rectangle that was defined by four 1: 24,000 scale USGS topographic quad maps: Citrus Park, Gandy Bridge, Sulphur Springs, and Tampa. The points represent reported sinkholes from the Florida Geological Survey sinkhole database. There are two main clusters of sinkholes in Hillsborough County: the one that was chosen as the study area is within the major metropolitan area of Tampa and had 211 sinkholes recorded in the existing database.

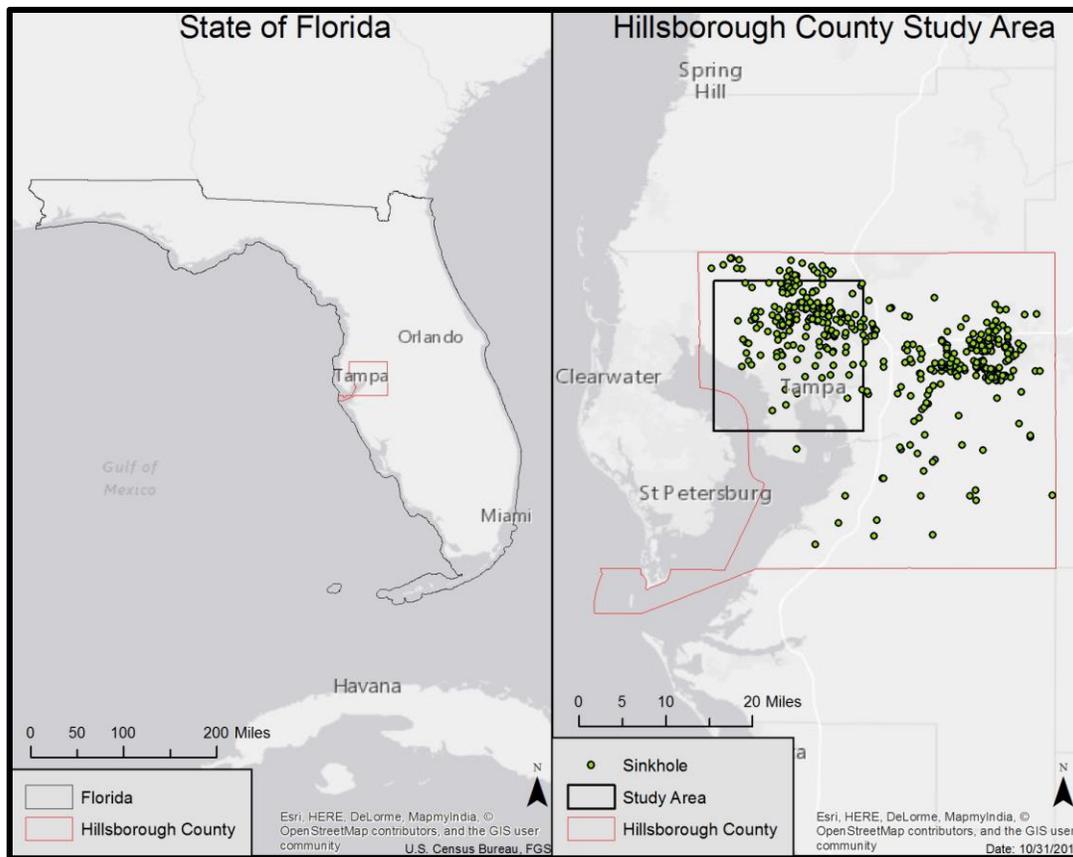


Figure 9 The State of Florida and study area

## 4.2 Construction of the Prototype Unified Geodatabase

An empty prototype geodatabase was created using the schema shown in Figure 8. It consists of four components: a polygon feature class, a point feature class, and two tables to house the VGI information. The projected coordinate system was set to the appropriate Florida State Plane system. The data from the Florida VGI database was then loaded into the point feature class and tables while the polygon features were created using on-screen digitizing from raster topographic map images. Finally, primary and foreign keys were created using a python script. The population of each component with data and the creation of the keys are discussed in separate subsections below.

#### 4.2.1. Sinkhole Areas Polygon Feature Class

Emulating the Kentucky Geological Survey sinkhole area data collection method, first a polygon feature class was created and then the uppermost closed contour depressions visible in raster images of the four USGS topographic maps encompassing the study area were captured using on-screen digitizing. Figure 10 shows the resulting 1,083 sinkhole areas that were digitized in this study area.

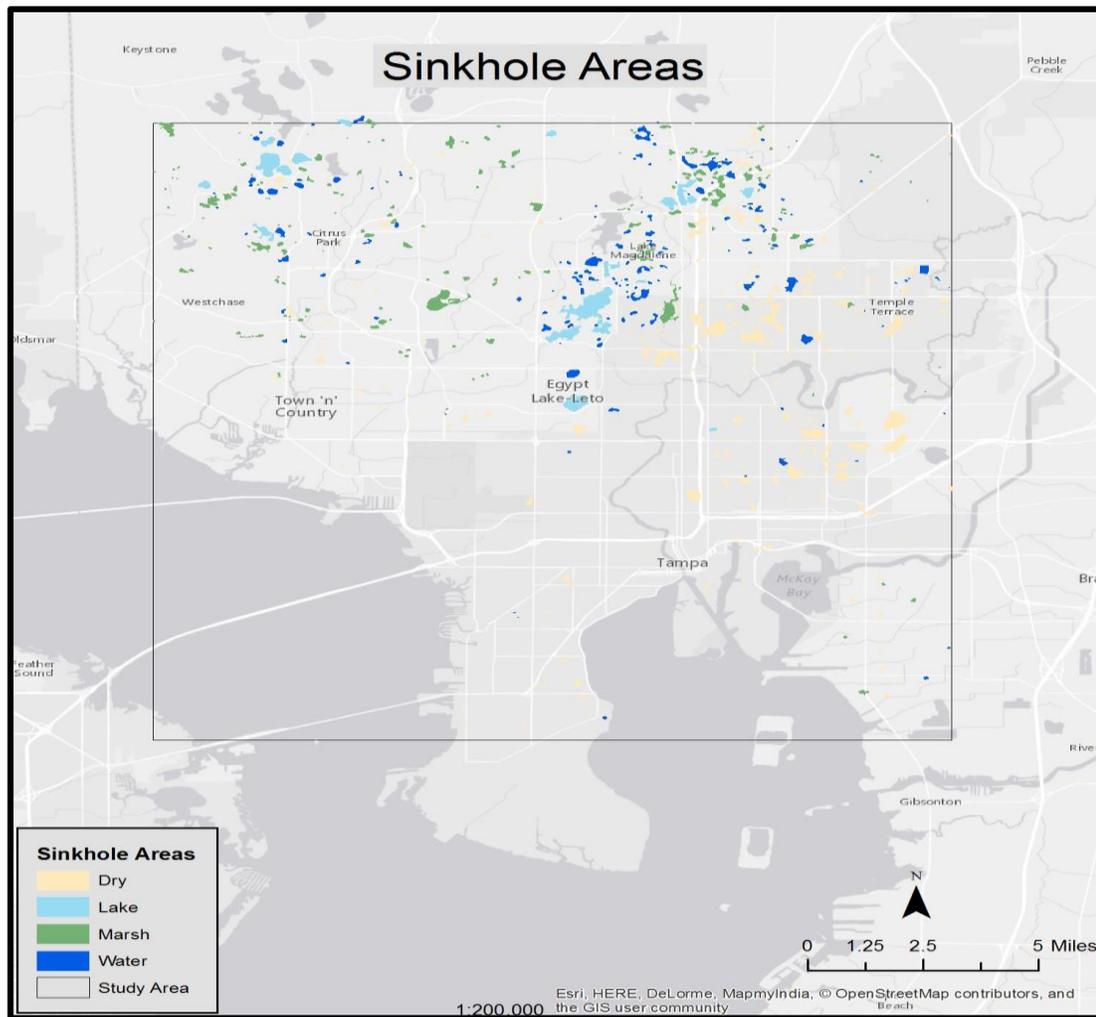


Figure 10 Sinkhole area locations in study area

These areas were categorized into four types of depressions based on what was found on the topographic maps: dry, lake, marsh, and water. The “D\_Type” or “depression type” attribute

was created to take account of these different designations. The difference between lake and water designations is that lakes are named, and water depressions have no names on the topographic maps used to digitize them. While Kentucky's sinkhole areas data structure does not include such a type attribute, this enhancement was inspired by the Indiana Geological Survey. In the attribute table of their sinkhole layer, Indiana differentiated depressions that had hydrological features such as sinking streams from sinkhole areas that were dry. Since such information is available on the topographic maps, it was decided to capture this information during the on-screen digitizing process. Thus, sinkhole area depression features were digitized and this information was added to the "D\_Type" attribute. In this step, the iterative nature of the geodatabase design process is illustrated since the process of populating the polygon feature class informed this aspect of the database design.

#### *4.2.2. Sinkhole Point Feature Class*

The spatial locations and the majority of the attribute data for the sinkhole feature class in the prototype geodatabase were acquired from the downloaded Florida sinkhole shapefile (<http://www.dep.state.fl.us/gis/datadir.htm>). An empty point feature class was first created in ArcCatalog with attributes of similar names but the same data type as the original shapefile because there were a few cases where the names in the UG point layer were modified to make them less cryptic. The next step, while still in ArcCatalog, was to load the spatial data from the original shapefile and match the attributes correctly before transferring. This process was accomplished by right-clicking on the point feature class and selecting the "load data" option. This is how the attributes with modified names were matched correctly to one another.

Once all the data from the Florida source were loaded, a final step was to select only the sinkholes within the study area. First, a polygon boundary was created from the four topo maps

that represent the study area. Then, the “Select by Location” function was used to select sinkholes within the polygon study area boundary. This yielded 211 sinkholes making up the prototype UG sinkhole point feature class (Figure 11). Any sinkholes located outside of the study area were deleted.

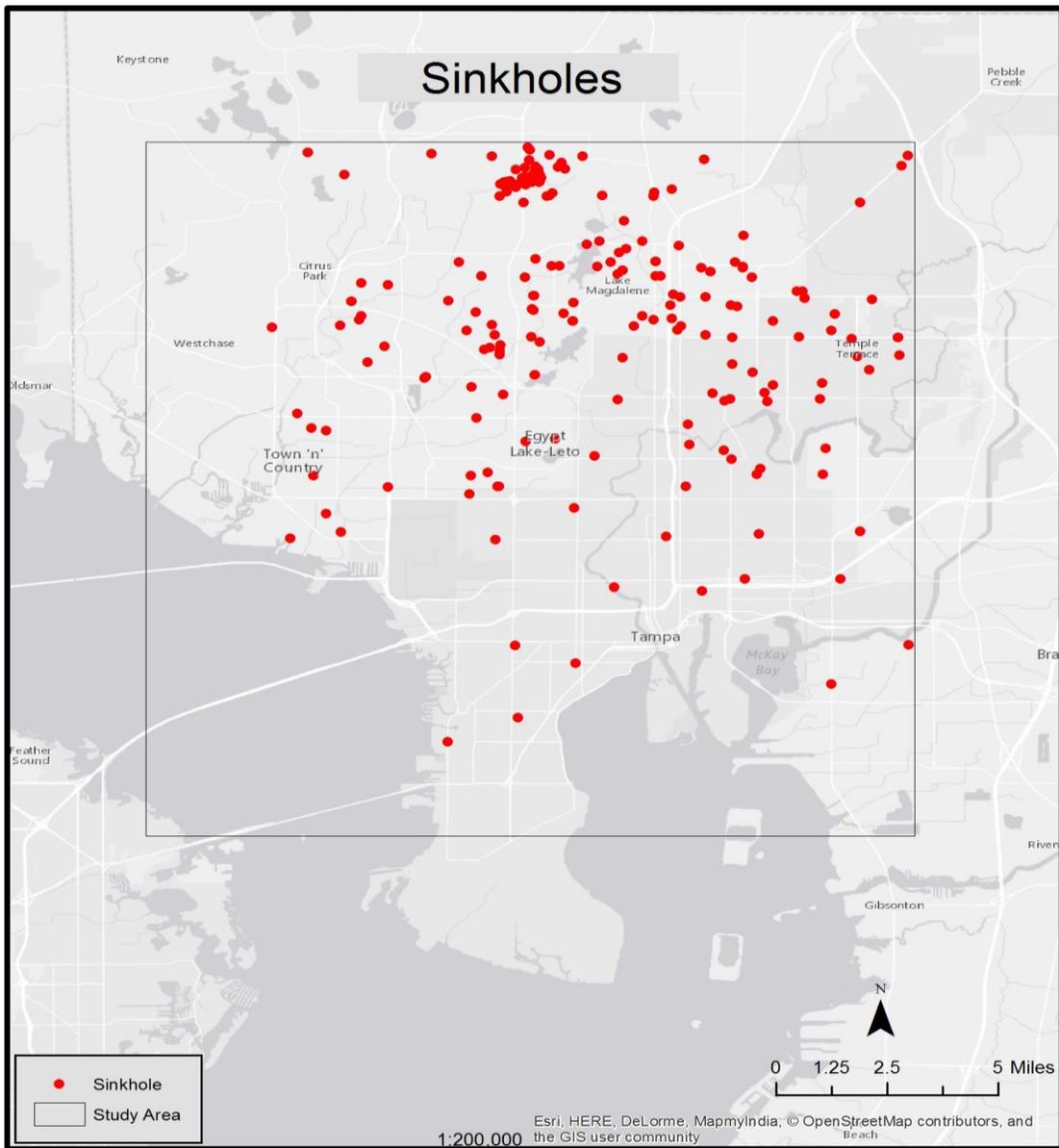


Figure 11 Sinkholes in study area

#### *4.2.3. VGI Tables*

The original Florida shapefile contained VGI related attribute fields and sinkhole characteristic attribute fields in a single table containing 51 attributes. Each record in the Florida shapefile was considered VGI data unless otherwise noted by the “True\_Sink” attribute. If the value was “Y” or “N” that indicated that the Florida SGS staff had verified this record and determined if it was a true sinkhole or not, making this authoritative. In the final database, these records are included as “Sink\_Verified” = Y. If the record had a “U” value in the “True\_Sink” attribute designating the feature as unknown or if the field was blank indicating that it has not been authoritatively verified by the SGS and is of VGI origin, the data is deemed unreliable as it has not been checked yet (thus, “Sink\_Verified” = N). The table was exported from the shapefile in ArcMap and copied: one copy to be used as the source for the VGI Source table and a second copy for the VGI Duplicate Data table. The attributes that were not incorporated in the prototype UG were deleted, including from the point feature class. New ones that were going to be the primary and foreign keys were added. Finally, both tables were imported into the prototype geodatabase.

#### *4.2.4. Primary and Foreign Keys*

Multiple primary and foreign keys were needed to link the two feature classes and VGI tables. Before any keys were calculated, the original primary key in the Florida database “REF\_NUM” attribute column was loaded initially into the sinkhole feature class and retained in the two tables. This ensured all records could be matched before migrating to the new primary and foreign keys.

The keys were then calculated by using a python script modified from Duggan (2013) that utilizes the FIPS codes in the calculation as demonstrated in Figure 12 for the SHA polygon

feature class in ArcMap. The majority of the script was inserted in the Pre-Logic Script Code box to calculate each record and for each of the two feature classes and two tables. The final script code: “autoIncrement()” was inserted in the smaller box below the Pre-Logic Script Code one. In the case of the SHA polygon feature class, the “ID” attribute was designated as the primary key and was calculated. The “pStart” value is made up of the State FIPS code for Florida which is twelve and a six digit number. The “pStart” value is set to begin at 12,000,000 because it is a combination of the State FIPS code and a six digit number. The “pInterval” is set to 1, to incrementally increase each record by 1. A similar approach was used to calculate the primary and foreign key values in the remaining point feature class and two tables individually using this same script but with different values. The final step was to remove the join and delete the “REF\_NUM” attribute column.

```
Paste the following into the Pre-Logic Script Code:

rec=0
def autoIncrement():
    global rec
    pStart = 12000000
    pInterval = 1
    if (rec == 0):
        rec = pStart
    else:
        rec += pInterval
    return rec

Paste the following code in the smaller box below the Pre-Logic Script Code:

autoIncrement()|
```

Figure 12 Python script demonstration for calculating key values for SHA feature class (after Duggan 2013)

### 4.3 Testing the Prototype with User Community Sample Queries

The usability of the UG design was tested by running several sample queries from the three user communities previously identified. For each of the user communities, three sample

queries were explored. Each set of queries consists of at least one spatial and one attribute based scenario. Below, results of the queries are reported based on the prototype geodatabase constructed.

To keep this evaluation of the design uncomplicated, for these queries, it is assumed each user has downloaded the geodatabase and has access to and is capable of using ArcMap. In the future, it is anticipated that the UG would be implemented in a web application through which the database could be queried directly, and the selected data downloaded as shapefiles, tables or displayed on a web map. This future work is discussed further in Chapter 5.

#### 4.3.1. Geologist Use Case Queries

Tables 13 to 15 provide sample queries that geologists may perform on the UG to extract data to be used in ArcMap. Additional data needed for processing these queries is assumed to have been obtained from locally stored collections or public domain data portals.

Table 13 Geologist use case sample query 1

<i>Query</i>	List the sinkholes that are pending confirmation, that are within a quarter mile of an SHA and that caused property damage.
<i>Objective</i>	Prioritize investigating sinkhole reports that are pending based on other factors such as if they are within a quarter mile of an SHA and caused property damage.
<i>Additional data needed</i>	No additional data needed.
<i>Procedure</i>	In ArcMap, 1. Buffer features in the SHA feature class by ½ mile. 2. Select by Location to select features from the sinkhole feature class that intersect the SHA buffers.
<i>Result</i>	47 sinkholes selected

Table 14 Geologist use case sample query 2

<i>Query</i>	Is there any relationship between the type of soil and aquifers present for sinkholes of a certain shape and a depth of ten feet or more?	
<i>Objective</i>	Investigate some of the geologic properties around sinkholes.	
<i>Additional data needed</i>	Soil and Aquifer layers.	
<i>Procedure</i>	In ArcMap, 1. Intersect Sinkhole layer with soil and aquifer layers. 2. Select by Attributes to select features from the sinkhole layer that meet the following SQL statement: "Depth >= 10." 3. Summarize the number of sinkholes within each type of soil and aquifer.	
<i>Sample Result</i>	<b>Layer</b>	<b>Classification</b>
	<b>Soil</b>	1 sinkholes in Limestone
		27 sinkholes in Medium Fine Sand and Silt
	<b>Aquifers</b>	8 sinkholes in Carbonate Rock
		20 sinkholes Other Rocks

It is interesting to see that more sinkholes did not occur in the carbonate rock type aquifer or limestone type soil. This could indicate VGI reporting errors in the original FGS database that was imported or errors in the boundaries on the other layers. It could certainly lead the geologist to further investigations.

Table 15 Geologist use case sample query 3

<i>Query</i>	Create a risk assessment map of sinkhole occurrence.
<i>Objective</i>	Implement a ranking system in which sinkholes are given a higher value the closer they are to sinkhole areas. Ranks are defined as: <ol style="list-style-type: none"> <li>1. Highest risk, sinkholes within SHAs.</li> <li>2. Sinkholes more than 0 and less than ¼ mile from SHAs.</li> <li>3. Sinkholes more than ¼ mile and less than ½ miles from SHAs.</li> <li>4. Sinkholes more than ½ miles from SHAs.</li> </ol>
<i>Additional data needed</i>	No additional data needed.
<i>Procedure</i>	In ArcMap, <ol style="list-style-type: none"> <li>1. Open the “Generate Near Table” tool in ArcToolbox.</li> <li>2. The ‘Input Feature’ is the ‘Sinkhole’ feature class and the ‘Near Feature’ is the ‘SHA’ feature class.</li> <li>3. Open the resulting table and create an attribute column ‘Miles.’</li> <li>4. Open the Field Calculator for the Miles column and enter the following: ‘NEAR_DIST/5,280’ to convert feet to miles.</li> <li>5. Create an attribute column ‘Rank.’</li> <li>6. While still in the table, click on the ‘select by attributes’ option</li> <li>7. To calculate Rank, select for each class sequentially and enter the appropriate value in the Rank column as follows:  Rank 1: ‘Mile = 0’  Rank 2: ‘Mile &gt;0 AND Mile &lt;= .25’  Rank 3: ‘Mile &gt; .25 AND Mile &lt;= .5’  Rank 4: ‘Mile &gt; .5’</li> </ol>
<i>Result</i>	16 sinkholes as rank 1 or the highest risk of sinkhole occurrence, 194 as rank 2 (zero-quarter mile from SHA), 1 as rank 3 (quarter-half mile) and 0 as rank 4 (half-one mile) or minimal risk of sinkhole occurrence.

#### 4.3.2. Insurance Fraud Investigators Use Case Queries

Tables 16 to 18 provide sample queries that insurance fraud investigators may generate on the UG. Again, it is assumed that the geodatabase has been downloaded and accessed through ArcMap. Any additional data is assumed to have been acquired by the user from their internal collection or public domain sources.

Table 16 Insurance fraud investigators use case sample query 1

<i>Query</i>	Compare the sinkholes in the state database that have not been reported to have caused property damage with those for which the insurance firm has paid compensation.
<i>Objective</i>	Reduce fraudulent claims.
<i>Additional data needed</i>	Insurance firm’s data table that lists all sinkholes on which they have paid claims. Geographic reference is by UG geodatabase “GEOID” attribute.
<i>Procedure</i>	In ArcMap, <ol style="list-style-type: none"> <li>1. Import the company table and join to the sinkhole point feature class by “GEOID” attribute.</li> <li>2. Select sinkholes for which the company data shows compensation paid and the state database shows “Property_Dmg = N.”</li> <li>3. Export the resulting attribute table for further review.</li> </ol>
<i>Sample Result</i>	62 sinkholes had caused property damage according to the state database while 51 had <b>not</b> . The remaining results are dependent on the insurance firm’s data to analyze which were paid though are not considered to have caused damage in the state database.

Table 17: Insurance fraud investigators use case sample query 2

<i>Query</i>	List zip code areas that contain ten or more sinkhole occurrences.
<i>Objective</i>	Calculate monthly premiums on perceived sinkhole risk.
<i>Additional data needed</i>	Zip Code layer found on U.S. Census Bureau Data Portal.
<i>Procedure</i>	<ol style="list-style-type: none"> <li>1. Intersect Zip Code and Sinkhole layers.</li> <li>2. Use Summarize to list by zip code the number of sinkhole occurrences.</li> </ol>
<i>Result</i>	There were 25 zip codes in the study area, and six had greater than ten sinkholes: 33618 – 58, 33613 – 28, 33612 – 22, 33624 – 15, 33617 – 13, and 33614 – 11.

Table 18 Insurance fraud investigators use case sample query 3

<i>Query</i>	Locate sinkhole VGI records with a phone contact related to the sinkhole claim.
<i>Objective</i>	An insurance fraud investigator wants to conduct interviews regarding a sinkhole claim and needs to gather additional details about this particular feature.
<i>Additional data needed</i>	Insurance firm’s data table of pending sinkhole claims. Locations are indicated with lat/long coordinates.
<i>Procedure</i>	<p>In ArcMap,</p> <ol style="list-style-type: none"> <li>1. Right-click the “VGI Duplicate Data” table and choose the ‘Relate’ option.</li> <li>2. Choose the “GEOID” attribute from the first drop down menu and then select the “sinkhole” feature class to relate in the second drop down menu and finally select the “GEOID” attribute in the third dropdown menu.</li> <li>3. Load insurance firm’s data table and make a point feature class from X, Y coordinates.</li> <li>4. Using the “Buffer” tool in ArcToolbox, make a ½ mile buffer for the sinkhole feature of interest.</li> <li>5. Intersect the buffer result with the UG sinkhole feature class points.</li> <li>6. Click the ‘Relate’ function in the point feature class to see which VGI records in the VGI Duplicate Data table have adequate contact information.</li> </ol>
<i>Sample Result</i>	2 sinkhole VGI reports are selected, but only one has contact information for conducting an interview.

#### 4.3.3. General Public Use Case Queries

Tables 19 to 21 provide sample queries that the general public may generate using the sinkhole geodatabase in ArcMap. This group has the simplest queries and so rather than actually doing three individual queries, the first query is intended to represent many of the most common questions that could be asked. This might include questions such as how many sinkholes have occurred in a county or how many sinkholes are within a certain distance from a location. These can be spatial or attribute queries. Any additional data is assumed to be acquired and processed by the user. The third query (Table 20) represents one that might be made by an individual who

has some GIS experience and can perform some additional steps beyond those needed for the first two query examples.

Table 19 General public use case sample query 1

<i>Query</i>	List number of sinkholes in a specified area.
<i>Objective</i>	Discover how many sinkholes are in the city that an individual resides in.
<i>Additional data needed</i>	No additional data needed.
<i>Procedure</i>	In ArcMap, 1. Select by Attributes where “City = Tampa.”
<i>Result</i>	110 sinkholes were selected.

Table 20 General public use case sample query 2

<i>Query</i>	Where are any sinkholes that caused property damage, were at least ten feet deep, and occurred over the last decade?
<i>Objective</i>	Curiosity to see how many sinkholes meet this criteria and where they are.
<i>Additional data needed</i>	No additional data needed.
<i>Procedure</i>	In ArcMap, 1. Select by Attributes where “Property_Dmg = Y.” 2. Select from Current Selection , “Depth >= 10” 3. Select from Current Selection , “Date > date 2005-01-01 00:00:00”
<i>Result</i>	Two sinkholes were selected.

Table 21 General public use case sample query 3

<i>Query</i>	List number of sinkholes occurring in forest or wetland areas.
<i>Objective</i>	What is the magnitude of sinkhole incidence in non-urban areas?
<i>Additional data needed</i>	National Land Cover Data (NLCD)
<i>Procedure</i>	In ArcMap, 1. Open the “Extract Values to Points” in ArcToolbox. 2. Select the ‘Input Point Features’ as the ‘sinkhole’ feature class and select the ‘Input Raster’ to NLCD. 3. Open the generated point layer from step 2 and scroll to the “RASTERVALU” attribute and right-click it to ‘Summarize.’ 4. Open the Summary table where the “Count_ RASTERVALU” indicates the number of sinkholes per value, refer to the NLCD key that shows what each value represents.
<i>Result</i>	For this query, Forest is represented by values of 41-43 and Wetlands as 90 and 95. 35 sinkholes were selected; 9 have a value of 42 or Evergreen Forest,

	25 have a value of 90 or Woody Wetlands, and 1 had a value of 95 or Emergent Herbaceous Wetlands.
--	---

After conducting possible test queries for the three user communities, it can be concluded that the UG is able to deliver on the demands of real world needs. The geodatabase can be used in standalone queries or in conjunction with additional datasets that are available in the public domain, such as the USGS and U.S. Census Bureau data portals. A geodatabase created from the UG template was capable of being used to perform tabular, spatial or a combination of both kinds of queries to meet an objective for all three user communities.

#### **4.4 Additional Evaluation Perspectives**

In addition to the usability evaluation discussed in the previous section, there are other perspectives by which the UG design can be evaluated. These include assessments of the database level of completeness and the database design quality.

##### *4.4.1. Level of Completeness*

When it comes to any sinkhole inventory database, it will never be complete because sinkholes continue to occur. It is rather an ongoing exercise to be as complete as possible at any point in time. The decision to merge data collected by authoritative sources and data from VGI collection techniques for the UG is one way of making sinkhole inventories more complete. Authoritative data, for example, may be published on an annual basis when satellite imagery is used for collection or there may be several years between publications if collection depends on new topographic maps becoming available. In the time period between data publications, many sinkholes may go recorded.

In addition, collecting possible sinkhole features from volunteered public reports enhances the traditional data sources. VGI allows for monthly if not weekly updates to the

database. Of course the SGS may not be able to verify immediately if each VGI report is a true sinkhole. However, the UG structure allows such data to be included and information about the VGI sources are also listed allowing users to contact the sources and perform their own investigations. This dual source approach provides an improvement over the authoritative-only approach which would have much longer periods between updates.

#### *4.4.2. Database Design Quality*

Finally, it is useful to revisit the database design quality evaluation aspects outlined in Figure 6. These provide a template by which to assess the overall UG design. The design process utilized the four dimensions of database design, focusing especially on the process, data, and model aspects. From the process quality, gaining the domain knowledge about sinkholes through existing research and finding databases with the SGS provided a solid platform of organizing the data. In regards to the data aspect, designing the UG in a way to represent authoritative and VGI collection methods with attribute accountability was important. Finally, the model aspect of choosing to house the data in a geodatabase in an easily understandable format as most of the SGS data are using shapefiles allows a smoother transition of information to the new UG platform.

Having demonstrated through various perspectives that the UG design is sound and appropriate, this report now concludes in Chapter 5 with a consideration of the possibilities for future development and use of the UG including support for a web map viewing and download platform.

## **Chapter 5 Conclusions and Future Work**

By beginning with an effort to understand the existing state of SGS sinkhole databases and then finding a way to bring VGI and authoritative methods together, this thesis builds the case that the unified sinkhole geodatabase designed here provides an effective database template for implementation at the state level. This chapter discusses ways in which the UG may be further enhanced in the future. Building upon the possibility of implementing the UG at the state level, three additional directions that could be the next steps in the evolution of the UG are:

(1) improvements in methods for collection of VGI data about sinkholes, (2) development of a web map interface to display and download the sinkhole inventory data, and (3) implementation of the UG at the national level.

### **5.1 Suggestions to Improve VGI Collection and Implementation**

Only two SGS are using VGI as part of their data collection methods. Florida has a two-page PDF form (<http://www.dep.state.fl.us/geology/geologictopics/sinkhole.htm>) while Wisconsin has a two-page web based form (<http://www.tfaforms.com/209523>). A recommendation to improve VGI data collection in general is to simplify the VGI reporting form, a revision that would be particularly helpful in the case of Florida's sinkhole reporting. Also, both states rely on tabular reporting that could be greatly enhanced by adding functionality to record spatial features. Rather than collecting the location information and terrain characteristics by data entry from the keyboard, location details could be easily captured through a web interface. Two methods for this are proposed below.

### *5.1.1. A GeoJSON Interface for VGI Data Collection*

One simple mechanism for collecting the spatial data in conjunction with the simplified reporting forms discussed above could be through GeoJSON. The SGS website where the reporting form is could also have a link and quick tutorial of how GeoJSON functions. The simple GeoJSON interface could allow a contributor to easily digitize features using available imagery as a reference. Figure 13 shows an example of this interface. The digitized features could then be exported as a shapefile and sent to the SGS along with the reporting form. This is one way to increase the accuracy of the sinkhole database should it rely on a VGI collection methodology and also as a way to have a stream of updates that the SGS could verify more efficiently. Although this approach may appear to create a bottleneck in updates given the ease with which they can be submitted, it would ensure more data integrity when compared to a method where many users can edit the same features simultaneously or sequentially, such as the OpenStreetMap project.

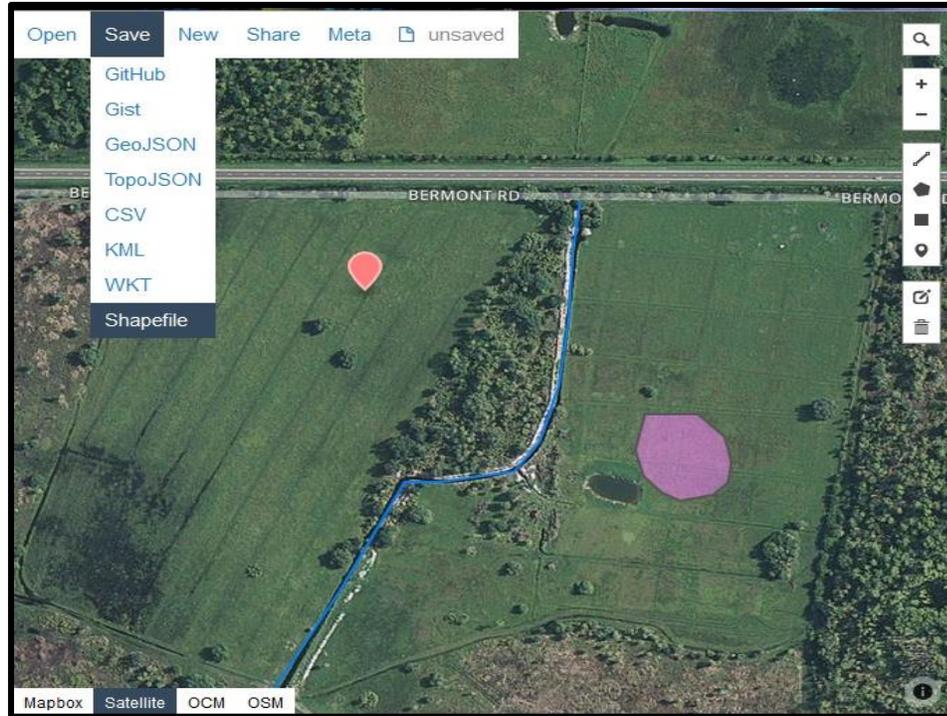


Figure 13 A sample GeoJSON interface (<http://geojson.io/#map=2/20.0/0.0>)

### 5.1.2. OpenStreetMap VGI Data Collection

The OpenStreetMap (OSM) project is a successful example of VGI contributors around the world digitizing and updating features on a web map interactively. The users contributing to OSM have personal knowledge about locations allowing them to correct features more effectively than the traditional mapping agency that would not be as aware of localities in other countries or remote locations.

The states could use the OSM model and have contributors digitize sinkhole features online which that could then be verified by geological survey staff. Such a system seems like a more efficient way to handle sinkhole inventories compared to the GeoJSON approach given the existing OSM infrastructure, but there are some drawbacks. Having all the data available instantly online before all the features are verified is an important issue. In addition, digital vandalism where people make deliberately erroneous features may be difficult to detect by

contributors or geological staff compromising data integrity an issue recognized by Gao et al. (2006) who argued that a tiered structure of database access permissions would aid in combatting such occurrences. The OSM approach would work well if the geological surveys could dedicate staff to monitor the online dataset otherwise it does not make sinkhole inventories more efficient.

## **5.2 A Web Map Interface**

It has been mentioned that a web map may be the most effective way to display and possibly distribute the sinkhole data. Florida and Pennsylvania are examples of states that have sinkhole web maps, and Pennsylvania includes an option allowing user communities to download the data. Being able to view, query and download the data would make it accessible for all the different user communities, including those without access to ArcGIS. Using a web map in conjunction with an updated VGI collection method would allow updates to become available faster and seamlessly compared to publishing a new dataset every few months.

## **5.3 National Implementation**

As demonstrated by the National Address Database project, when a unified database implementation is needed, it requires a complex set of tasks to be undertaken involving many interest groups. In order for the effort to succeed, these groups must be coordinated, agreeing on the database structure, data collection procedures and clear deliverable milestones. This would apply to a national sinkhole database because of the multiple SGS involved who will have to agree upon similar standards for their state databases while simultaneously laying the foundation for a national database to be implemented in the future. As the geological surveys are the primary data collectors and compilers for sinkholes, it would make sense to have the coordination for a national database to be managed by the USGS at the federal level since they

are in a position to provide a feedback loop with the states, enabling a common database design template to be reached.

There are several aspects that would need resolution in a national version. Decisions would need to be made regarding the inclusion of database fields that are applicable in only some states. Definitions would need to be debated. For example, the definition of a sinkhole may vary among geologists. What other karst features should be included or defined differently for inclusion in the database would need discussion. Does the VGI contact information get omitted at the national level because it is only relevant at the state level? Clearly, there are many important choices to make before a template that can be used at the national level is determined.

Once a consensus is reached for the database design, existing collections can be transferred to the new platform from the state level. It would then be the responsibility of the multiple SGS to engage local government agencies, private institutions, universities and the general public to upload sinkhole related information using a common collection method. This approach is similar to the U.S. Census Bureau's Community TIGER portal discussed by Otto (2015) for the National Address Database.

## **5.4 Conclusion**

The UG proposed by the author of this thesis is only one possible design for a standard template for sinkhole inventory purposes. It was devised in the context of a wide range of different approaches to a common problem. Researching existing databases yielded design improvements that are included in the proposed universal sinkhole geodatabase template outlined here. It is a possible first step to creating a national sinkhole database, but realizing that objective remains a task for the institutions responsible for managing sinkhole data who will be able to build upon the framework presented herein.

## REFERENCES

- Angel, Julie C., Daniel O. Nelson, and Samuel V. Panno. 2004. "Comparison of a new GIS-based technique and a manual method for determining sinkhole density: An example from Illinois' sinkhole plain." *Journal of Cave and Karst Studies* 66, no. 1: 9-17. Accessed: November 14, 2013.
- Al-Kouri, Omar, A. Al-Fugara, Samih Al-Rawashdeh, Balqies Sadoun, and Biswajeet Pradhan. 2013. "Geospatial Modeling for Sinkholes Hazard Map Based on GIS & RS Data." *Journal of Geographic Information System* 5 (6): 584-592. <http://search.proquest.com/docview/1490579576?accountid=14749>.
- Applied Geographics. 2015. "National Address Database Summit Report." Accessed: January, 2016. [https://www.nsgic.org/public\\_resources/NAD\\_Summit\\_Final\\_Report\\_070815.pdf](https://www.nsgic.org/public_resources/NAD_Summit_Final_Report_070815.pdf).
- Associated Press 2015. "Huge sinkhole damages golf course in Missouri." *ABC13*, May 23. Accessed June 16, 2015. <http://abc13.com/news/huge-sinkhole-damages-golf-course-in-missouri/737917/>.
- Cherfi, Si-saïd Samira, Jacky Akoka and Isabelle Comyn-Wattiau. "Federating Information System Quality Frameworks using a Common Ontology." Paper presented at the Proceedings of the 16th International Conference on Information Quality, pp. 160-173. Adelaide, Australia, November 18 - 20, 2011. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84871555845&partnerID=40&md5=c564fd5479e97b9a40fb02be12286e82>.
- Chicago Tribune. N.D. "Pictures: Winter Park sinkhole." *Orlando Sentinel*, 1981. Accessed: December 6, 2013. <http://www.chicagotribune.com/news/nationworld/os-fla360-pictures-winter-park-sinkhole-20121113,0,5366877.photogallery>.
- Duggan, Nicholas. 2013. The Spatial Blog. "GIS Tips – ArcGIS, Create Sequential Numbers in the Attribute table (without using = FID!!)." Aug 15. Accessed November 17, 2015. <https://dragons8mycat.wordpress.com/2013/08/15/gis-tips-arcgis-create-sequential-numbers-in-the-attribute-table-without-using-fid/comment-page-1/>.
- Dunigan, Tom. 2013. "Tennessee Sinkholes." Accessed: June 10, 2015. <http://tnlandforms.us/landforms/sinks.php>.
- Eeckhaut, M. Van Den, J. Poesen, M. Dugar, V. Martens and Ph. Duchateau. 2007. "Sinkhole formation above underground limestone quarries: A case study in South Limburg (Belgium)." *Geomorphology* 91, no. 1–2:19-37. Accessed: November 13, 2013. <http://dx.doi.org/10.1016/j.geomorph.2007.01.016>.
- Fleury, E. Spencer, Steve Carson and Robert Brinkmann. 2008. "Testing Reporting Bias in the Florida Sinkhole Database: An Analysis of Sinkhole Occurrences in the Tampa Metropolitan Statistical Area." *Southeastern Geographer* 48, no. 1: 38-52. Accessed: November 15, 2013.

- Galve, Jorge P., Juan Remondo and Francisco Gutiérrez, 2011. "Improving sinkhole hazard models incorporating magnitude–frequency relationships and nearest neighbor analysis." *Geomorphology* 134, no.1–2: 157-170. Accessed: November 14, 2013. <http://dx.doi.org/10.1016/j.geomorph.2011.05.020>.
- Galve, Jorge P., F. Gutiérrez, J. Remondo, J. Bonachea, P. Lucha and A. Cendrero. 2009. "Evaluating and comparing methods of sinkhole susceptibility mapping in the Ebro Valley evaporite karst (NE Spain)." *Geomorphology* 111, no. 3–4: 160-172. Accessed: November 14, 2013. <http://dx.doi.org/10.1016/j.geomorph.2009.04.017>.
- Gao, Yongli, Robert G. Tipping and E. Calvin Alexander Jr. 2006. "Applications of GIS and database technologies to manage a karst feature database." *Journal of Cave and Karst Studies* 68, no. 3: 144-152. Accessed: November 24, 2013.
- Harbert, Artemis. 2013. "Texas Sinkholes." Accessed: June 10, 2015. <http://artemisharbert.com/gis-projects/texas-sinkholes/>.
- Hoxmeier, John A. 1998. "Typology of database quality factors." *Software Quality Journal* 7, no. 3: 179-193. Accessed December 11, 2015. <http://dx.doi.org/10.1023/A:1008923120973>.
- Liston, Barbra. 2013. "UPDATE 3-Guests saved as Florida resort building falls into sinkhole." *Reuters*, August 12. Accessed: November 13, 2013. <http://www.reuters.com/article/2013/08/12/usa-florida-sinkhole-idUSL2N0GD0HO20130812>.
- Lukyanenko, Roman, and Jeffrey Parsons. 2012. "Impact of Conceptual Modeling Approaches on Information Quality: Theory and Empirical Evidence." *17th International Conference on Information Quality*, p. 5. Accessed December 10, 2015. <http://dx.doi.org/10.13140/2.1.3942.9442>.
- Malone, Shyann. 2013. WTSP-TV, Tampa-St. Petersburg, Fla. "2 houses likely lost because of Florida sinkhole." *USA Today*, November 14. Accessed: November 14, 2013. <http://www.usatoday.com/story/news/nation/2013/11/14/florida-sinkhole-dunedin-homes/3524877/>.
- McKinney, Richard. 2015. "DOT presses forward on National Address Database." July 17. Accessed November 27, 2015. <https://www.transportation.gov/fastlane/pressing-forward-on-national-address-database>.
- Microsoft. 2015. "Database design basics." Accessed December 10, 2015. <https://support.office.com/en-US/article/Database-design-basics-EB2159CF-1E30-401A-8084-BD4F9C9CA1F5>.
- NOVA. 2015. "Sinkholes—Buried Alive." by Larry Klein. Aired January 28, 2015 on PBS. Accessed June 16, 2015. <http://www.pbs.org/wgbh/nova/earth/sinkholes.html>.

- Otto, Greg. 2015. Fed Scoop. "Coming Soon: An open database of every U.S. address." September 28. Accessed November 27, 2015. <http://fedscoop.com/national-address-database>.
- Sekstrin Limited. 2015. "The Sekstrin Project." Accessed December 10, 2015. <http://sekstrin.org/theproject.html>.
- Singh, A. J., Balvir Singh Thakur, Rajesh Chauhan, and Vikas Sharma. 2011. "Comprehensive Framework for quality Database Applications." Accessed December 10, 2015. <http://ijoes.vidyapublications.com/paper/SI2/52.pdf>.
- Tichy, Ulrich. 2010. Wikimedia Commons. "Global distribution of major outcrops of carbonate rocks." February 6. Accessed June 19, 2015. [https://commons.wikimedia.org/wiki/File:Carbonate-outcrops\\_world.jpg](https://commons.wikimedia.org/wiki/File:Carbonate-outcrops_world.jpg).
- Tobin, Bret D. and David J. Weary. 2004. Digital Engineering Aspects of Karst Map: A GIS version of Davies, W.E., Simpson, J.H., Ohlmacher, G.C., Kirk, W.S., and Newton, E.G. 1984. "Engineering aspects of karst: U.S. Geological Survey, National Atlas of the United States of America." Accessed: April 12, 2015. <http://pubs.usgs.gov/of/2004/1352/>.
- United States Census Bureau. 2014. "Florida Passes New York to Become the Nation's Third Most Populous State, Census Bureau Reports." U.S. Department of Commerce. Last Modified: December 23, 2014. Accessed: June 16, 2015. <http://www.census.gov/newsroom/press-releases/2014/cb14-232.html>.
- University of California, Berkeley. School of Information Management and Systems. 1997. "Files and Databases." August 28. Accessed December 10, 2015. <http://courses.ischool.berkeley.edu/i202/f97/Lecture6/sld001.htm>.
- University of Texas at Austin. N.D. "What is Karst?" Accessed May 4, 2015. <http://www.esi.utexas.edu/outreach/caves/karst.php>.
- Whitehead, Darren, and Blair Shiff. 2015. "Police SUV swallowed by sinkhole in Colorado." *USA Today*, June 5. Accessed June 16, 2015. <http://www.usatoday.com/story/news/nation/2015/06/05/colorado-police-suv-swallowed-sinkhole/28531325/>.
- Zisman, E.D. 2013. "The Florida sinkhole statute: its evolution, impacts and needed improvements." *Carbonates and Evaporites* 28, no. 1-2: 95-102. Accessed: November 14, 2013. <http://dx.doi.org/10.1007/s13146-013-0144-7>.

## Appendix A: Attributes Emulated from the FGS Database

Attribute Name	Description	Emulated	Why or Why Not Emulated
REF_NUM	FGS Assigned Reference Number	✓*	*Temporary unique ID
ADDED_BY	Person who added the record		Redundant
TRUE_SINK	Verified sinkhole	✓	Separates sinkholes from other karst features
DATE_ADD	Date Added		Redundant
DATE_REV	Date Revised	✓	Record Update
LONG_DD	Longitude Degrees		Redundant
LAT_DD	Latitude Degrees		Redundant
COUNTY	County	✓	Require a reference attribute
TWNSHP	PLSS Township		Redundant
RANGE	PLSS Range		Redundant
SECTION	PLSS Section		Redundant
QTRSECT1	Quarter section		Redundant
QTRSECT2	Quarter section		Redundant
RPT_SOURCE	Source of report	✓	Who reported it
RPT_PHONE	Report phone number	✓	Contact information
RPT_NAME	Report name	✓	Contact information
EVENT_DATE	Date of reported event	✓	Date of event
OSTREET	Owner's address		Redundant
OCITY	Owner's city		Redundant
OZIP	Owner's zip code		Redundant
EVT_ADDR	Event address		Redundant
SIZDIM	Dimensions		Redundant
SINSHAPE	Shape	✓	Sinkhole Shape
SINLNGTH	Length	✓	Dimension
SINWIDTH	Width	✓	Dimension
SINDEPTH	Depth	✓	Dimension
SLOPE	Slope of Sides		Complex for VGI Report Source
WATSIN	Water visible		Complex for VGI Report Source
WATBLS	Water below land surface (FT)		Complex for VGI Report Source
LIMVIS	Limestone visible		Complex for VGI Report Source
CAVVIS	Cave visible		Complex for VGI Report Source
SUBRATE	Subsidence rate		Complex for VGI Report Source
TRIGGERS	Triggering mechanisms		Complex for VGI Report Source
COL_CODE	Pre-collapse indicators		Complex for VGI Report Source
PROPDAM	Property damage	✓	Was there damage?

<b>Attribute Name</b>	<b>Description</b>	<b>Emulated</b>	<b>Why or Why Not Emulated</b>
REPAIRED	Feature repaired	✓	Sinkhole repaired – Y/N
PLANNED	Repairs planned	✓	Repairs planned – Y/N
DRAINSTR	Drainage structures present		Complex for VGI Report Source
LUCODE	Land use code		Complex for VGI Report Source
SOILTYPE	Soil type		Complex for VGI Report Source
COMMENTS	Other comments	✓	Comments about event
COMMENTS2	Additional comments	✓	Comments about event
ACCESS	Access to sink		Complex for VGI Report Source
WITNAM	Witness Name	✓	Name of person who saw it
WITADDRE	Witness Address		Redundant
WITCZIP	Witness Zip Code		Redundant
WITPHONE	Witness Phone	✓	Contact witness about event
EM_Hard_Co	Unknown		n/a
ACCURACY	Unknown		n/a

Source: Florida Geological Survey (2015)