Spicing Up Address Points with the ArcGIS Data Interoperability Extension

By

Steven T. Gay, Director of GIS Services
Boone County Planning Commission
Boone County, Kentucky

2007 ESRI International User Conference
June 2007

Paper #1823

Abstract

Data Interoperability in GIS

Track: Database Design, Automation, and Management

Author: Steven T. Gay, Boone County Planning Commission

This paper demonstrates perhaps a new GIS model for managing and publishing address points. For Boone County, the primary focus of this initiative was to streamline the management and updating of address information across all GIS data sets. Another focus of this project was automating the creation of a robust published address point data set that is more efficient and ultimately more effective for end users. A great deal of attention was also given to geodatabase topology rules that can be used to enforce spatial integrity and ensure an accurate address point data set. This paper will also demonstrate how the Data Interoperability extension can be used to easily facilitate the manipulation of data.

I. Boone County, Kentucky

Boone County is located in Northern Kentucky and spans 256 square miles. Occupying the southwestern portion of the Cincinnati Metropolitan Statistical Area, Boone County is the only location where Kentucky, Ohio, and Indiana meet. This multi-state border stretches for 42 miles along the Ohio River.

Having grown in population by eighty-four percent since 1990, Boone County has consistently been one of the fastest growing counties in Kentucky during the last fifteen years. According to a U.S. Census Bureau report released in 2006, Boone County ranks 49th among counties in the United States for fastest growth rate between 2000 and 2005.

II. Statement of the Problem

Boone County's migration to the geodatabase format in 2002 resulted in many dramatic changes regarding the way in which the GIS data layers are maintained. The data schema that had been in existence since the early development of the GIS program was not structured in such a way as to take advantage of some of the benefits of the geodatabase format. Upon close examination, it was discovered that some of the GIS attributes were no longer needed and were therefore not being kept current. In other cases, data entry rules had changed over time with no concerted effort to ensure that historical data had been sufficiently scrubbed to account for these changes. Some business rules were not being applied consistently by all data custodians. Database normalization principles had never been factored into the data design, which resulted in redundancies that ultimately led to errors.

Boone County felt that the data could be organized much better to leverage benefits of the geodatabase format, resolve the afore-mentioned data entry issues, and ultimately provide more useful information to end users. These decisions led to a very successful project which resulted in a complete overhaul and re-design of the schema for all GIS data sets maintained by Boone County.

III. Data Production vs. Data Publishing

There is a key concept underlying the strategies outlined in this paper that must be understood in order to grasp the methodologies currently in place at Boone County. The address point management practices explained in this paper is predicated on the fact that the GIS layers that are maintained by the data custodian (i.e. the production data) and the GIS layers that end users consume (i.e. the published data) *do not* have to have identical schemas. Indeed, the ideas promoted in this paper shows that production and publishing schemas in many cases *should not* be the same.

To help better understand this concept, consider the following definitions and rules for production and published data:

Production Data is considered to be any data set that is designated as the working copy.

RULES FOR PRODUCTION DATA SETS

- Only the person(s) responsible for editing a data set should have access to it in its production environment.
- Regardless of whether or not someone is the designated custodian, a production data set should only be accessed by users if the data set is in a state of being edited.
- Production data sets are to be exported periodically. The resulting *published* data sets are to be made available for consumption by end users.

Published Data is considered to be any data set that has been processed and prepared for consumption by end users.

RULES FOR PUBLISHED DATA SETS

- All non-sensitive published data sets should be available to all end users.
- Published data sets should be read only. (Any edits that need to occur for a layer must be performed by the designated custodian on the production version of the data set).

The *production environment* is a term that collectively refers to all of the hardware, software, data editors, procedures and data layers that factor into the on-going maintenance of data sets. Similarly, the *publishing environment* refers to any hardware, software, staff, procedures and data layers that are necessary to support the consumption of GIS data. A natural parent-child relationship exists between the two in the sense that a published (or child) data set is produced from production (or parent) data set(s).

In some instances, one person's published data may be another person's production data. For example, in order for the data custodian who is responsible for maintaining the zoning layer to do their job, they must utilize parcel boundaries to enforce coincident geometry between the two layers. Since the zoning editor is not making changes to the parcel layer, the rules outlined above dictate that the editor must use the published parcels for this function. The *published* parcels therefore become part of the zoning layer's *production environment*.

Boone County utilizes a centralized data maintenance environment, with all production layers stored via ArcSDE in a SQL Server relational database. Any editing of ArcSDE data is performed by staff members whom are equipped with ArcInfo licenses of ArcGIS Desktop. Editing of ArcSDE layers is accomplished with out-of-the-box editing tools included within ArcMap. Boone County has established versioning protocols to assist with conflict resolution that may result from multi-user editing.

The majority of Boone County's end users are located at remote sites (defined as buildings that are not interconnected with the computer network used by the Boone County Planning Commission). As such, there is no easy way to configure their custom GIS desktop application to acquire its data directly from ArcSDE. Therefore, in order to accommodate these remote sites while maintaining consistency with regard to the publishing process, Boone County utilizes ESRI shapefiles as its standard publishing format. The custom GIS applications that are built and deployed by Boone County are therefore designed to read shapefiles. These shapefiles are exported periodically out of ArcSDE and delivered to each remote site on a monthly basis. To promote consistency with which all published data is organized and managed, a standard directory structure is established by Boone County. Each data consumer that utilizes custom GIS applications deployed by Boone County must have a publishing environment that conforms to this directory structure.

IV. Data Interoperability Extension

The ability to quickly manipulate production data into published data that is ready for consumption is central to the ideas promoted in this paper. Boone County anticipated this need and sought an easy-to-use tool that would enable such manipulation. ESRI's Data Interoperability extension proved to be exactly the solution for what Boone County needed. Boone County is committed to the ArcGIS family of software and the Data Interoperability extension fits very nicely into that framework.

Using the Data Interoperability extension, Boone County creates models in ArcToolbox that retrieve data from the ArcSDE production geodatabase. The Data Interoperability models re-organize the ArcSDE data to match a client's requested file schema and format (personal geodatabase, shapefile, CAD, etc.). Once the model is saved in ArcToolbox, generating an update for the client is simply a matter of re-running the model. Boone County refers to these ArcToolbox models as *publishing models*.

V. Data Design Strategies

Domain Tables

An important part of geodatabase design is the use of domain tables. These tables contain a list of values that constrain what is acceptable to enter into a field. Wherever possible, domain tables were utilized to ensure the consistency of the information in the GIS. Coded values in the domain table can be of any data type, but integer values were used most of the time.

Upon publishing a GIS layer, most geodatabase attribute values that originate from a domain table's list of coded values are converted into its corresponding description value in the published data. In certain situations where an attribute's coded value could be used as a label (e.g. SR-1 : Suburban Residential 1), both values – the code *and* description, are exported to different attributes in the published data.

Atomic Attribution

Concatenation of multiple values is very easy to do, whereas parsing an attribute's value into its component parts is much more complicated. A data design that will support the need to concatenate attributes with the maximum amount of flexibility calls for storage of attribute values at their most atomic level. In the case of addresses, Boone County chose to store each item that comprises a fully qualified address in its own attribute in the ArcSDE production environment. The attribute values stored in ArcSDE can then be combined to build complex attribute values in the published data according to the client's requested schema. The table below describes the various address components modeled by Boone County, using the following address example:

123 N MAIN ST NW, SUITE #14; BURLINGTON, KY 41005

Field Name	Field Alias	Value
SITENUM	Address Number	123
SITEDIRPRE	Prefix Direction	N
SITESTNAME	Street Name	MAIN
SITESTTYPE	Street Type	ST
SITEDIRSUF	Suffix Direction	NW
SITEADD2	Sub-address	SUITE #14
SITEPOST	Postal Designation	BURLINGTON
SITESTATE	State	KY
SITEZIP	Zip Code	41005

A simple example of the usefulness of such a design can be seen with address labels. It is extremely common to label buildings on a small scale map with *either* a fully qualified address *or* merely the address number (i.e. for maps that already have street centerlines labeled). Since it is not practical to extract the address number from an attribute value containing a fully qualified address, most data designs encourage storing the address number in its own attribute. Additionally storing the address number in another attribute with the rest of the address components contradicts database normalization principles since the data custodian cannot guarantee that the address number stored in both of these attributes will always be synchronized. Although it is possible to label a building with the fully qualified address by concatenating the address number stored in one attribute with additional address information stored in other attributes, inexperienced GIS users may not be able to create a complex expression such as this on the fly.

Using the design illustrated above, the data custodians can design custom publishing models that export the address number into its own attribute in the published data *and* concatenate the many disparate attribute storing the address components (including the address number) into another attribute in the published data. In doing so, the address number that is entered into two separate attributes in the published data will never be out of sync since it originates from the same single attribute in the production data.

Leveraging Spatial Relationships to Promote Geodatabase Normalization

Custom built legacy applications used by Boone County users were initially designed to expect mailing addresses to be stored as parcel polygon attributes and site addresses to be stored as building polygon attributes. At the time that the data re-design effort was undertaken, Boone County had no plans to re-program these legacy applications. When considering that database normalization dictates that address information should not be stored in two different locations (i.e. addresses and parcels; and/or addresses and buildings), the question arose regarding how to transfer address attributes from a point layer to polygon layers in an automated and easily repeatable fashion. ESRI's Data Interoperability extension proved once again to have the solution to this dilemma.

The Point-On-Polygon transformer provided by the Data Interoperability extension allows attributes from a layer of one geometry type (point or polygon) to be transferred to coincident features in another layer of the other geometry type. This transformer essentially allows the parcel and building polygon features to "absorb" the address attributes from address points that exist on top of the parcel and building polygons. This satisfies Boone County's desire to utilize the ArcSDE address point layer as the master source for all address information, while continuing support for Boone County's legacy applications by publishing the parcels with mailing address attributes and the buildings with site address attributes.

Published Address Points

The discovery of the Point-on-Polygon transformer opened up a whole new way of looking at published address points. A large portion of queries that are solved with GIS have their results linked to addresses. Believing that addresses would be a focal point for many end user's queries, Boone County realized that the Point-on-Polygon transformer could be used to add value to the published address points.

Boone County's address point publishing models utilize eighteen different Point-On-Polygon transformers. For each point in the address layer, each transformer performs a spatial query to determine coincident polygons from another ArcSDE layer. These spatially related polygons provide information that is attached to the coincident address points as additional attributes. The attributes that are transferred to the published address points are detailed in the table below.

Address Point Attribute	Address Point Attribute Alias	Source Feature Class in ArcSDE	Source Attribute in ArcSDE
BASEGRID	Base Mapping Grid	sde.BaseGrid	GRID
ZONINGDIST	Zoning District	sde.ZoningDistrict	ZONINGCODE
WATSERVICE	Water Service Provider	sde.WaterServiceArea	PROVIDER
SEWSERVICE	Sewer Service Provider	sde.SewageServiceArea	PROVIDER
STMSERVICE	Storm Service Provider	sde.StormwaterServiceArea	PROVIDER
FIREDIST	Fire District	sde.FireDistrict	FRDISTNAME
FIREESZ	Fire Emergency Service Zone	sde.FireESZ	FIREESZNAME
LAWENFRCMT	Law Enforcement Provider	sde.LawEnforcementArea	LAWENFRCMT
POLICEBEAT	Police Beat	sde.PoliceBeat	POLICEBEAT
VOTEPRCNCT	Voting Precinct	sde.VotingPrecinct	PRCNCTNAME
VOTELOC	Polling Location	sde.VotingPrecinct	POLLINGLOC
SCHOOLDIST	School District	sde.SchoolDistrict	SCDISTNAME
ELEMSCHOOL	Elementary School	sde.SchoolDistrictE	ELDISTNAME
MIDLSCHOOL	Middle School	sde.SchoolDistrictM	MDDISTNAME
HIGHSCHOOL	High School	sde.SchoolDistrictH	HGDISTNAME
TAXDISTRCT	Tax District	sde.TaxDistrict	DISTNUM
LOTNUMBER	Lot Number	sde.Parcel	LOTNUMBER
PLATID	Plat ID	sde..Parcel	PLATID

The polygon layers that supply attribute values to the published address points reside in their respective production environment (i.e. ArcSDE). The beauty of this methodology is that any edits that are made to any polygon layer in the production environment will automatically be reflected in the published address points after the next address point update is performed (usually every week).

Creating a single layer that contains such a wealth of information benefits end users in two ways: 1) it enables inexperienced GIS users to focus their query efforts on one layer, and 2) allows the end user to construct a single attribute query rather than many consecutive spatial queries. The Point-on-Polygon transformers essentially perform the "heavy lifting" of the spatial queries before-hand, thus minimizing the possibility that inexperienced end users will make mistakes during their spatial analyses.

To illustrate the benefits of this, consider what a user would need to do to determine which addresses meet the following criteria: 1) located in the Burlington Fire District, 2) not in a subdivision, and 3) serviced by the Boone County Water District. With Boone County's current methodology, users only need to know how to construct a single attribute query and can stay focused on a single layer – the address points. Before the current methodology was in place, users could have still found the answer to their question, but it would have involved a series of *spatial* queries using four separate data sets (address point, fire district, subdivision and water service area). The former methodology introduces quite a bit of room for error and leaves much of the workflow up to the end user to determine.

Geodatabase Topology
A significant advantage of ESRI's geodatabase storage model is the increased ability to assure topological integrity between disparate feature classes. ESRI's geodatabase topology implementation allows for different rules to be established for a layer's defined subtypes. Boone County has created topology rules that relate the address points to building and parcel polygons. These rules help guarantee the quality of all three layers. The following subtypes and corresponding topology rules have been established for the buildings, parcels, and address point layers:

Layer	**Subtype**	**Topology Rule**
Address Points	Improved	"Must be properly inside" of a *Primary* Building polygon
	Unimproved	
Buildings	Primary	"Must contain" an *Improved* Address Point
	Secondary	
Parcels		"Must contain" an Address Point

Periodically, the topology is validated and corresponding updates are made to ensure these layers conform to the topology implementations specified.

The following is broken down into three sections. The order of the sections – production, transformation, and publishing follow the natural progression of the production-publication paradigm promoted in this paper. The first section – *Production Schema*, details the address point schema that is stored in ArcSDE. The second section – *Transformation Model*, diagrams the data interoperability model that facilitates the creation of the published address point layer. The third section – *Publishing Schema*, details the address point schema of the published address points (i.e. the layer intended for end user consumption).

Address Points

Address Point Production Schema

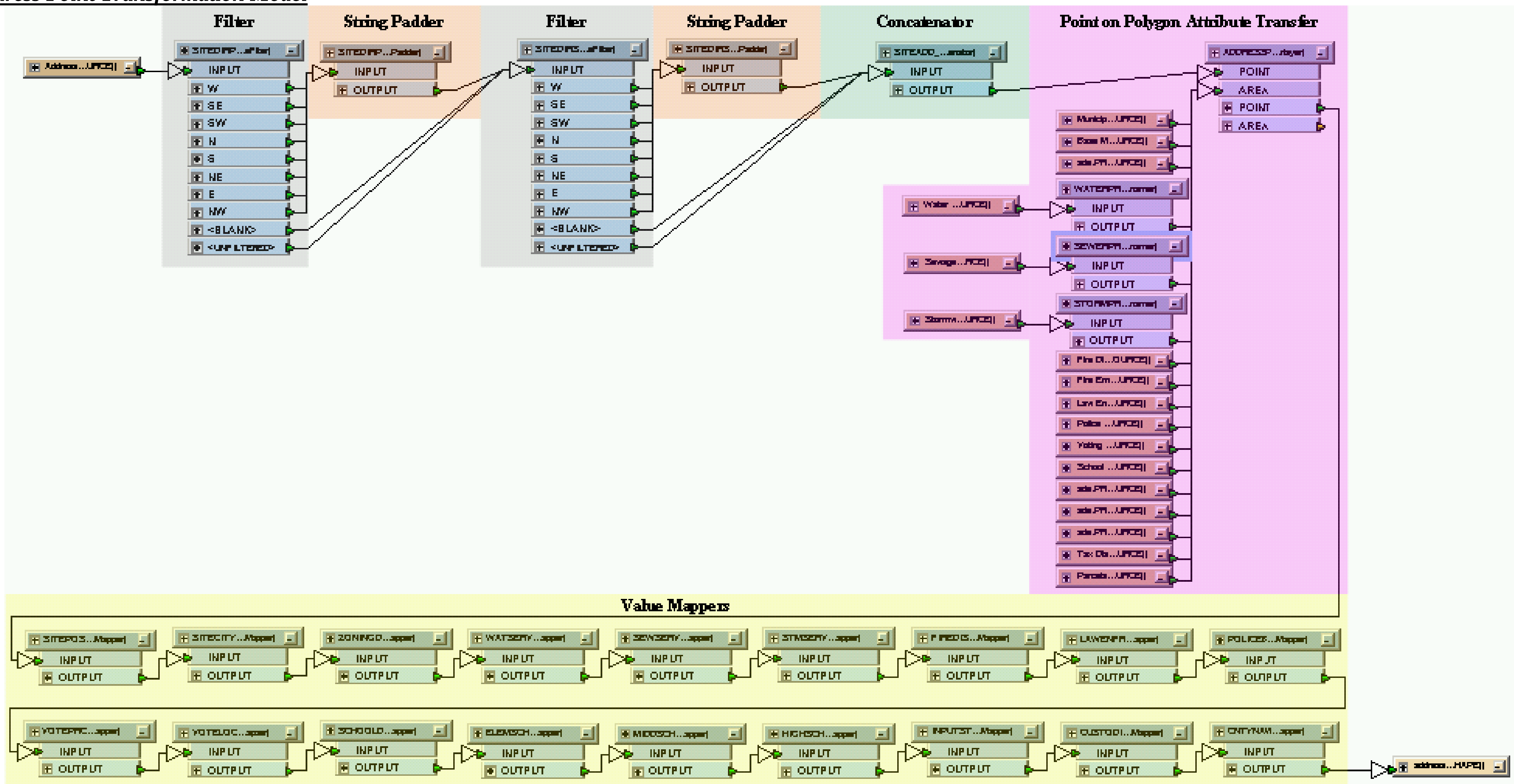
Feature Class Name: Address
Geometry Type: Point
Spatial Index Grid 1: 545
Production Platform: ArcSDE
Feature Dataset: Structure

Description:
The Address Point layer's primary purpose is to store all address-related data. There are two addresses that have the most value – site address and mailing address. During the transformation from production to publishing, each address point is overlaid virtually onto several polygon layers. During these overlays, certain polygon attributes are transferred onto each address shape and published as attributes of the point features.

Field Name	*Data Type*	*Precision*	*Scale*	*Width*	*Field Alias*	*Nullable*	*Required*	*Domain Table*
PIDN	text			17	Parcel ID	Y	N	
OWNER	text			40	Owner	Y	N	
MAILADD1	text			40	Mailing Address	Y	N	
MAILADD2	text			25	Mailing Sub-Address	Y	N	
MAILPOST	text			30	Mailing Postal Zone	Y	N	
MAILSTATE	text			2	Mailing State	Y	N	StateCodesUSPS
MAILCNTRY	text			2	Mailing Country	Y	N	CountryCodesUSPS
MAILZIP	text			7	Mailing Zip Code	Y	N	
MAILZIPEXT	text			7	Mailing Zip Extension	Y	N	
SITEOCC	text			50	Occupant	Y	N	
SITENUM	long	10			Site Number	Y	N	
SITEDIRPRE	text			3	Directional Prefix	Y	N	Octants
SITESTNAME	text			25	Site Street Name	Y	N	
SITESTTYPE	text			4	Site Street Type	Y	N	StreetTypesUSPS

SITEDIRSUF	text			3	Directional Suffix	Y	N	Octants
SITEADD2	text			25	Site Sub-Address	Y	N	
SITEPOST	long	10			Site Postal Zone	Y	N	PostOfficeZones
SITESTATE	text			2	Site State	Y	N	StateCodesUSPS
SITEZIP	long	10			Site Zip Code	Y	N	ZipCodes
INPUTSTAT	short	5			Status	Y	N	ConstructionStatus
CUSTODIAN	short	5			Custodian	Y	N	CustodianCodes
CNTYNAME	text			3	County	Y	N	CountyCodes
CREATIONDT	date				Created in GIS	Y	N	
REVISIONDT	date				Revised in GIS	Y	N	
BCGISUSER	text			15	Data Technician	Y	N	
BLDGTOPLGY	short	5			Building Topology	Y	N	AddressSubtypes

Address Point Transformation Model



Address Point Publishing Schema

Shapefile Name: address

Field Name	*Data Type*	*Precision*	*Scale*	*Width*	*Source Attribute(s)*
PIDN	text			17	sde.PRO.Address.[PIDN]
OWNER	text			40	sde.PRO.Address.[OWNER]
MAILADD1	text			40	sde.PRO.Address.[MAILADD1]
MAILADD2	text			25	sde.PRO.Address.[MAILADD2]
MAILPOST	text			30	sde.PRO.Address.[MAILPOST]
MAILSTATE	text			2	sde.PRO.Address.[MAILSTATE]
MAILCNTRY	text			2	sde.PRO.Address.[MAILCNTRY]
MAILZIP	text			7	sde.PRO.Address.[MAILZIP]
MAILZIPEXT	text			7	sde.PRO.Address.[MAILZIPEXT]
SITEOCC	text			50	sde.PRO.Address.[SITEOCC]
SITENUM	long	7		7	sde.PRO.Address.[SITENUM]
SITEADD	text			45	sde.PRO.Address.[SITENUM]+“ ”+[_sitedirpre]+[SITESTNAME] +“ ”+ [SITESTTYPE]+[_sitedirsuf]
SITEADD2	text			25	sde.PRO.Address.[SITEADD2]
SITEPOST	text			12	sde.PRO.Address.[SITEPOST]
SITESTATE	text			2	sde.PRO.Address.[SITESTATE]
SITEZIP	long	5		5	sde.PRO.Address.[SITEZIP]
SITECITY	text			32	sde.PRO.Address.[SITECITY]
BASEGRID	text			7	sde.PRO.BaseGrid.[GRID]
ZONINGDIST	text			35	sde.PRO.ZoningDistrict.[ZONINGCODE]
WATSERVICE	text			35	sde.PRO.WaterServiceArea.[PROVIDER]
SEWSERVICE	text			24	sde.PRO.SewageServiceArea.[PROVIDER]
STMSERVICE	text			24	sde.PRO.StormwaterServiceArea.[PROVIDER]
FIREDIST	text			31	sde.PRO.FireDistrict.[FRDISTNAME]
FIREESZ	text			5	sde.PRO.FireESZ.[FIREESZNAME]
LAWENFRCMT	text			31	sde.PRO.LawEnforcementArea.[LAWENFRCMT]

POLICEBEAT	text			11	sde.PRO.PoliceBeat.[POLICEBEAT]
VOTEPRCNCT	text			35	sde.PRO.VotingPrecinct.[PRCNCTNAME]
VOTELOC	text			40	sde.PRO.VotingPrecinct.[POLLINGLOC]
SCHOOLDIST	text			25	sde.PRO.SchoolDistrict.[SCDISTNAME]
ELEMSCHOOL	text			35	sde.PRO.SchoolDistrictE.[ELDISTNAME]
MIDLSCHOOL	text			35	sde.PRO.SchoolDistrictM.[MDDISTNAME]
HIGHSCHOOL	text			35	sde.PRO.SchoolDistrictH.[HGDISTNAME]
TAXDISTRCT	text			5	sde.PRO.TaxDistrict.[DISTNUM]
LOTNUMBER	text			5	sde.PRO.Parcel.[LOTNUMBER]
PLATID	text			7	sde.PRO.Parcel.[PLATID]
INPUTSTAT	text			18	sde.PRO.Address.[INPUTSTAT]
CUSTODIAN	text			10	sde.PRO.Address.[CUSTODIAN]
CNTYNAME	text			16	sde.PRO.Address.[CNTYNAME]
CREATIONDT	text			20	sde.PRO.Address.[CREATIONDT]
REVISIONDT	text			20	sde.PRO.Address.[REVISIONDT]
BCGISUSER	text			15	sde.PRO.Address.[BCGISUSER]