

## **BIOGRAPHICAL INFORMATION**

**Emmy Andrews**  
**GIS Analyst**  
**Farallon Geographics, Inc.**

### **Specific Responsibilities**

Ms. Andrews joined Farallon Geographics, a small GIS consulting firm with headquarters in San Francisco, in April 2003. Ms. Andrews's technical experience spans geospatial data modeling and development, technical GIS analyses, and web design. Recently, she has focused on GIS technical procedure design including the development of quality control procedures, data maintenance procedures and technology transfer workflows. Her current projects include implementing an enterprise GIS at Fairfield-Suisun Sewer District, building a modern linear referencing system at Caltrans, and developing CAD (computer aided dispatch) data for the California Department of Forestry.

### **Past Experience**

Prior to joining Farallon, Ms. Andrews was employed as a telecommunications Data Analyst, where she gained extensive experience in data management, forecasting and analysis. She has additional career experience as a professional graphic designer.

### **Educational Information**

BA, *cum laude*, Duke University, 1998.

GIS certification, San Francisco State University, 2003.

Graduate studies in Environmental Management, University of San Francisco, 2003 – present.

### **Professional Memberships**

BAAMA, Bay Area Automated Mapping Association, 2003 – present.

# **Strategies for Success in Large-Scale Data Development Projects**

**Emmy Andrews  
Farallon Geographics, Inc.  
609 Mission Street, 2<sup>nd</sup> floor  
San Francisco CA 94105**

## **Abstract**

When an organization gets serious about using and stewarding GIS data, a concerted effort at data development is often necessary to create authoritative datasets. This paper discusses step-by-step approaches to data development using parcel fabric and utility network case studies. Topics include estimating the data development effort, combining disparate datasets, automating tasks via scripting, avoiding wasted time, creating repeatable workflows for non-expert staff, and protecting your investment through continuing data maintenance.

---

## **INTRODUCTION**

Planning and needs analysis

## **FIRST STEPS**

Obtain any existing GIS data

Define your coordinate system, projection, measurement unit and data precision

Create attributes for audit tracking and flagging anomalous data

## **GEOMETRY DEVELOPMENT**

Workflow-based geometry development

Script-based geometry development

Order of operations

Developing parcel layer geometry

Developing utility network geometry

## **ATTRIBUTE DEVELOPMENT**

Implement a naming convention

Join existing attributes

Plan future attribution

## **FINAL STEPS**

Quality Control (QC) and data statistics

## **PROTECT YOUR INVESTMENT – USE YOUR DATA!**

## **CONCLUSION**

---

## **INTRODUCTION**

Successful data development projects depend on both technical knowledge and a keen understanding of project planning and management. If the steps of the effort are not well thought out and logically implemented, time and effort will be wasted and the quality of data may even decline.

Investing in properly planned data development can yield high returns. Accurate data is the cornerstone of an organization's ability to use its data. Developing data can seem slow and frustrating, and the results are not always flashy, but the ROI is high when the data is developed to meet defined user needs.

This paper discusses technical and management considerations at each stage of a data development project. Large-scale projects are emphasized and illustrated with examples from transportation, parcel fabric and utilities applications.

### **Planning and needs analysis**

Planning is absolutely essential to successful data development. If the steps are not crystal clear or the estimated effort is more than a day's worth of work, **STOP! STEP BACK and PLAN!** It would be disastrous to introduce untraceable errors to geometries, attributes or both through haphazard editing. Before editing data, (a) make a backup so you can return to a previous state if the data become inaccurate or corrupted and (b) consider adding audit tracking fields to track who, what, when, and why of edits.

Needs analysis is also essential. Determine the audience for the data. How will users be using the data? How accurate must the data be? When do the users need the data? Are some aspects of data development higher priority than others? The biggest risk of a data development project is that the resulting data does not meet the needs of the users. Think of the effort in terms of cost/benefit. Interview users and managers and resolve conflicting organizational priorities before charging forward with a project that will improve the data, but not the organization's ability to use it to support decision-making.

The larger the project, the more crucial an accurate budget, timeline, and an estimation of staff and technical (hardware/software) resources will be to the project's success. These items may even be required for project approval or to justify additional staffing. Benchmark time intensive processes to accurately estimate the scope of the project. Remember, management will judge the project as much by adherence to the budget and timeline as by improvements to data quality.

## **FIRST STEPS**

### **Obtain any existing GIS data**

Most likely, the project is not starting with a blank slate. Gather together any existing datasets that can act as a starting point. A bit of detective work, such as contacting other departments or sister organizations, may turn up more existing data than you might think. For example, if you are a city, contact your county office. If you are a sewer district, contact the cities in your district. If you are an environmental health department, contact the departments responsible for emergency response and planning (police, fire). Always contact your IT and/or IS department, as

these departments often act as a repository for data. For example, the following projects would need the layers described:

#### Parcel layer development

<del>/</del> Control network	Points of known (surveyed) location that anchor the layer
<del>/</del> Existing parcels	An outdated layer is better than starting from scratch
<del>/</del> Aerial photography	Reference features that is especially useful in QC
<del>/</del> Assessor maps	Reference for digitizing
<del>/</del> Recorded documents	Reference for digitizing
<del>/</del> Street centerlines	Reference feature, can be used to QC address ranges

#### Utility network development

<del>/</del> Lines	Sewer, water or storm lines
<del>/</del> Manholes	Manholes are the nodes connecting your line segments
<del>/</del> Ancillary features	Valves, fire hydrants, rod inlets, cleanouts, catch basins, etc.
<del>/</del> Hydrography	Rivers, lakes, streams, canals, ditches, etc.
<del>/</del> Street centerlines	Reference feature
<del>/</del> Parcels	Reference feature

#### **Define the coordinate system, projection, measurement unit and data precision**

Prepare the existing data you found by choosing a coordinate system, projection and a unit of measurement (miles, meters, survey-feet, etc.) and converting all data to this format. Make sure stakeholders agree that new data will be developed with this standard. This step may be more time consuming than you think and creating a standard now will undoubtedly save time throughout the project.

Setting and maintaining data precision is also critical. Precision is the exactness of your measurements or the number of significant digits in your measurements. All major GIS software allows you to set and manage data precision. Each time you convert data to a different format, check the precision. If you allow precision to be lost from your data at any step in the data development process, the former precision is cannot be regained.

#### **Create attributes for audit tracking and flagging anomalous data**

Audit tracking fields create row-level metadata for features. They are useful both in data construction and data maintenance. Typical audit tracking includes the following:

OPERATOR	Initials or name of the person who performed the edit
DATE	Date the edit was performed
TYPE	Edit type – such as geometry, attribute or both
SOURCE	Source data that prompted the edit –orthophoto, engineering drawing, etc.

Consider your project and determine if additional or custom audit tracking fields are required. Audit tracking fields enable easy location of edits for QC and other post-edit data processing. They also make it easier for multiple users to edit data without losing track of who edited which records. By tracking the date of edits, you can also gauge data currency.

If you require (a) tracking of multiple edits to a single spatial feature or (b) tracking the lineage of geometry across multiple edits, a more complex geodatabase is required. To track multiple edits to a single feature, you could create an audit tracking table which links to spatial features by a unique ID. To track the geometry of a single feature through multiple edits, you could develop an audit tracking method that *retired* features each time they changed shape but recorded the retire date in order to reconstruct a feature lineage.

Developing methods to flag anomalous data is especially important. Flagging data anomalies allows you to differentiate accuracy levels within a feature class. Less accurate data is flagged as such and therefore does not undermine users' confidence in the data. Additionally, when a follow-up data development effort is undertaken, less accurate features can be easily identified so they may be researched and their accuracy improved.

Flag fields should be customized to meet the needs of your organization. In a recent parcel development project, the parcel boundary lines feature class included two flag fields: ERROR\_TYPE and COMMENTS. ERROR\_TYPE was coded with one of the following values:

- 1 Does not match orthophoto
- 2 Assessor map not sufficient to define parcel boundary
- 3 Line length does not match assessed length
- 4 Other, fill in COMMENTS field

### **GEOMETRY DEVELOPMENT**

The majority of the project time and effort will probably be spent on this step (geometry development) and the next (attribute development). The two steps are commonly done in parallel rather than in series; I separate them here only to provide a more logical framework for discussion.

Development of geometry and attributes is likely to be iterative. For example, in a utility network, you may only have the time and resources to develop a few feature classes in your first effort. If you develop lines and manholes first, the data will meet many user needs immediately. The users' increased productivity will drive support for the development of additional data. Then, a subsequent effort can be undertaken to develop additional features (valves, fire hydrants, etc.) and populate maintenance data.

Two primary methods of creating and editing data will be discussed here: (a) using workflows for largely repetitive digitizing or editing tasks and (b) batch processing of edits with scripts.

#### **Workflow-based geometry development**

A workflow is a document that describes a process in detail, step-by-step. Good GIS workflows describe processes in enough detail that anyone with basic computer skills (no GIS skills required) could follow the workflow and complete the work described by that process. Workflows are also great for technical processes that are not performed often and thus are forgotten between repetitions – such as restarting a server – or for tasks that are easy enough to describe that almost anyone in the organization could perform them when necessary.

Workflows have been a key component of every data development project I have participated in. The general process of developing a workflow is as follows:

1. Identify the editing process to be performed.
2. Verify that it can be broken down into repeatable, explainable steps (benchmark; use test data).
3. Technical lead or other experienced project team member writes workflow.
4. Train staff to execute workflow.
5. Divide work into sections for each staff member to complete (don't duplicate work).
6. Technical lead or other experienced project team member continually reviews.
7. Make periodic backups of data.
8. Technical lead or other experienced project team member performs formal QC and finalizes data deliverable.

Below are some examples of generic workflow titles/subject areas:

#### Parcel layer development

- ~~///~~ Connect to parcel data on GIS server
- ~~///~~ Adjust lot line
- ~~///~~ Split parcels
- ~~///~~ Merge parcels
- ~~///~~ Add new parcel

#### Utility network development

- ~~///~~ Flip line vectors to standardize flow direction
- ~~///~~ Add fire hydrant
- ~~///~~ Convert utility network data from CAD

#### **Script-based development**

Scripting data development tasks has a romantic appeal, but the attraction to scripts must be tempered by an analysis of the costs versus the benefits of developing scripts. If you are working with a fairly small dataset, perhaps the effort and expertise needed to develop a script is out of proportion with the task at hand. If you are likely to have to repeat the process on multiple datasets or at regular time intervals (say, every year), investing in script development may be worth your while.

On the plus side, scripting can automate almost any task, such as connecting lines to points, analyzing flow direction and flipping vectors as needed, assigning attributes, and even developing map layouts. Scripts may be linked together such that the output of one process becomes the input of another, automating a string of data development tasks. On the minus side, scripts require the (often expensive) time of computer programmers and may not be useful outside the scope of a single project.

Many scriptwriters magnanimously post completed scripts on the web, which you may download for free. Definitely search for an existing script before developing one of your own. You may

find just what you need or a script you can modify slightly for your purposes. A simple web search for “GIS scripts” returns a wealth of information.

### **Order of operations**

When creating a workflow, and when planning your data development in general, it is important to strategize on the order of operations. You need to identify each step that will be necessary in the data development process and perform them in logical order so that no step has to be repeated and the highest data quality is insured after each step.

Also, remember to make a backup copy of your data each night and between each data development process so that if you lose data, discover a flawed process, or mix up the order of operations, you can return to a previous state of your data cleaning process without losing too much invested time.

Consider the example of developing a basic utility network with four features: water lines, manholes, fire hydrants, and fire hydrant runs (lines connecting fire hydrants to water mains). Assume that fire hydrant runs have no spatial features and are currently stored only in an Excel spreadsheet. Assume that all other features have existing geometry but it has not been updated in several years. The order of operations would be something like this:

1. Clean point layers (fire hydrants and manholes).
  - ~~///~~ Remove duplicate features
  - ~~///~~ Remove incorrect features
  - ~~///~~ Add additional features by referring to engineering drawings, field knowledge, etc.
2. Clean line layer (water lines).
  - ~~///~~ Remove duplicate features
  - ~~///~~ Remove incorrect features
  - ~~///~~ Add additional features by referring to engineering drawings, field knowledge, etc.
3. Snap water lines to manholes.
  - ~~///~~ Note that this will change the length and position of your water lines. Review the distance the lines are likely to move to be certain that you can live with this amount.
4. Create fire hydrant runs between fire hydrants and water lines.
5. Update layer attributes in parallel or after geometry development.

Think about the logic behind the order of operations. If you had created the fire hydrant runs first, what would have happened? First, you would have invalid hydrant runs if you had not already cleaned your fire hydrant layer. Second, if you had not already snapped your water mains to your manholes, the fire hydrant runs might become ‘unattached’ from their associated water mains.

At worst, incorrect order of operations creates extra work or introduces errors into your data. At best, a well-considered order of operations saves time and prevents costly mistakes. Consider the development of a parcel layer. By using a control network of points and then digitizing the rights-of-way, you can save time when you fill in lot lines because you have set up a careful framework for lot line development.

### **Developing parcel layer geometry**

It is worth taking a moment to point out some of the specific aspects of developing the geometry for a parcel layer. First, if you are developing parcels, it is of paramount importance that you discuss how this layer will be used within your organization and thus how the level of accuracy that is required.

Your GIS parcel layer will not be as accurate as the assessor maps and recorded documents from which it is digitized. Generally, a parcel layer considered *very accurate* has +/- 2 feet of accuracy. Any parcel boundaries that do not meet this standard have explanatory attributes and can thus be symbolized in a map.

Use a parcel-editing tool to digitize your parcels accurately. All GIS software packages that I have worked with have tools that allow you to enter metes and bounds and lines will be drawn accordingly. Digitize rights-of-way first and then draw parcel lines. The rights-of-way serve as a method to double check the accuracy of lot lines and they keep lot line errors from spreading from one right-of-way block to another.

Most parcel geodatabases I have worked with contain two features classes – Parcel boundaries (lines) and Parcels (polygons). This allows you to associate attributes such as APN or address with a polygon geometry while maintaining lot line lengths where it makes sense to – in a line class. The two feature classes should have a topological relationship to insure that when one is edited, the edits are mirrored in the other.

### **Developing utility network geometry**

Most utility network development projects I have been involved with began with translating data from a CAD program (AutoCAD, Microstation or MGE) to the organization's choice of GIS software. CAD is a common engineering medium, and there are several straightforward methods for translating this data; one method is a drawing interchange file (.dxf). When translating data, remember to set your projection and data precision, and be careful not to lose features that aren't properly referenced in the CAD drawing (such as features on the wrong layer). A visual overlay of the old data and the new data can help you identify shifts or gaps in data.

A potential use of pressurized utility network geometry is flow analysis and hydraulic modeling. To prepare your network for this type of analysis, two important conditions must be met.

First, all lines must be snapped together at endpoints, or snapped to nodes (valves, manholes, etc.) which serve as junctions. To avoid digitizing gaps into your network, make sure snapping is *turned on* before you start digitizing. In most GIS software, this is a very simple process; consult your software help files for specifics. The importance of snapping in all digitizing tasks cannot be understated.

Second, flow direction must be correct and standardized. Each line segment has a start point and end point; the flow direction, or polarity, of the line is from the start point toward the end point. Symbolize your lines with arrows pointing in the direction of polarity in order to analyze the consistency of your network flow direction. Lines that flow in the wrong direction can be "flipped" (consult your software help files). Use a *flip* command rather than trying to physically

swap the endpoints of the line by moving them for speed and accuracy. There are other ways to improve your network flow direction. I have used scripts successfully to improve network flow. In addition, some linear referencing tools allow you to create *routes* from line segments, bypassing polarity issues altogether.

A variety of engineering drawings will still need to be referenced regularly even after building your GIS. Many GIS programs allow you to store hyperlinks to these documents in the attribution so that you can essentially select a feature and jump right to the necessary drawing. Web-based GIS applications are also a great way to provide access to engineering drawings. Users can find and select the feature they are interested in from the map and then get a list of associated drawings that they can view online, saving the time that would be necessary to retrieve a paper map from a map library.

### **ATTRIBUTE DEVELOPMENT**

Sometimes attribute development is just downright painful. It's not as much fun as developing spatial data, and progress is often slow and tedious. But just remember that useful attributes are what animate your spatial features. Attribution turns meaningless points, lines and polygons into dynamic features that can be used in ever more complex analyses. If you want a map showing traffic volume versus road capacity, you can't make it without the underlying attribute data. So, bite the bullet and get to work populating attributes in your geodatabase!

#### **Implement a naming convention**

The primary function of a naming convention is to serve as a unique identifier for your data. The NAME field will be used to make relationships between attribute data stored in tables, as opposed to inline with your geometry.

The naming convention may also tell the user something about your features. APN is the classic example of a naming convention for parcels; it tells the user the book, page and parcel number on the associated assessor map. In utility networks, pipe names are often a concatenation of the names of their upstream and downstream nodes. Thus, the pipe name tells a user what nodes the pipe should be connected to and the flow direction of the line.

If your organization already has a naming convention in place, interview data users to decide if a new naming convention is warranted. Users get to know features by their unique names, and a change to the naming convention may be seen as unnecessary by users familiar with an existing system, even if that system seems arbitrary. If a new naming convention is implemented, retain the old name for as long as necessary (often indefinitely) so data that refer to the old name can still be referenced to the proper feature.

#### **Join existing attributes**

Remember that non-spatial tables can be part of a geodatabase. If you have attributes stored in dBASE, Excel, Access, Text, or any other reasonably common table file type, you can add these attributes to your geodatabase. All you need is the unique name (see 6.1) in each table record and you can associate the attributes from the table with the geometries in the geodatabase.

Spend time searching for non-spatial data sources within your organization. Remember that users often store a wealth of data that could be useful to the whole organization on the unshared C drives of their computers. If you find a table with useful attributes that does not contain the NAME of your features, spend time populating the NAME field and then join the two sources. It may take a while to populate the NAME field, but it is better than losing the opportunity to leverage the pre-existing information source.

That said, older data from non-central sources is likely to contain more errors than newly created data. Use audit tracking fields (4.1) to track verification of the data as your data development efforts progress. For example, say you are developing a utility network and found an old file containing the diameter and pipe material for 80% of your lines. You should (a) record the origin of this 80% (b) field verify the remaining 20% and record the fact that this is newly collected data and (c) field verify the remaining 80% when crews have reason to interact with these pipe sections.

### **Plan future attribution**

Plan now for the robust set of attribution you would like to have in the future. Add tables with these fields to your geodatabase and encourage users to populate this information as they are able. Eventually, you will have a phase of data development during which you address these gaps in your data in a systematic manner. But if you make space for these attributes now, some of the data will already be populated when this time comes.

For example, in developing a utility network, in the first phase you will gather basic information such as pipe diameter and pipe material. Over time, you can build records about pipe history and maintenance, which knowledgeable users can help populate.

### **FINAL STEPS: QUALITY CONTROL AND DATA STATISTICS**

During the course of data development, many operators may touch the data. At the end of the project, there should be some level of QC of the data in order to (a) estimate the level of error in your data and (b) look for any systematic errors that should be addressed.

For example, the QC of a parcel layer might proceed as follows:

1. Assign an amount of budget or a number of hours for QC
2. Create a list of checks to perform to the overall data, such as
  - ~~☒~~ Check for agreement with orthophotography
  - ~~☒~~ Check for basic topology – no gaps, slivers, dangling linework, missing parcels
3. Pick a percent or number of parcels for more detailed review that can be QC'd within the specified timeframe
4. Check this subset thoroughly against assessor parcel maps.

Another useful QC trick is to use thematic visualization to check relationships in your data. For example, you could symbolize parcels with a different color representing each unique book/page combination. In this way, you could easily identify parcels within a block that have a possible incorrect APN because they would be a different color. You could symbolize pipe diameter by thickness of line and easily identify areas where pipe diameter was not consistent.

To finalize a phase of data development, you should document the basic statistics concerning your data. Basic statistics about your data include:

- ~~✍~~ Number of features
- ~~✍~~ Percent of data that is Null (in the overall dataset or in each field)
- ~~✍~~ Estimate overall accuracy of data
- ~~✍~~ Estimate effort for next phase of data development

### **PROTECT YOUR INVESTMENT – USE YOUR DATA!**

As discussed in the previous section, data quality degrades quickly when it is not continually maintained and updated. The best way to protect your investment in data development is to make sure that people in your organization are using the data daily. If accessing the data is a daily task for people for multiple departments, GIS personnel will have broad and vocal support for spending the money necessary to maintain and improve the data over time.

As a GIS manager, you should also be continually brainstorming new ways to use the data within your organization to serve the organization's needs. GIS is generally developed for asset tracking and management, but by developing a few simple applications, it can support many other functions. For example, data analysis can support community outreach programs, web applications can allow the public access to portions of your data, and web services can allow data sharing between organizations such as between a city and a county. In a utility network, the same data that is managed for asset tracking can be used as a basis for organizing environmental sampling data.

Geographical location is a feature that all your organizational data has in common. Exploit this fact to its full potential to gain increasing benefit from your GIS.

### **CONCLUSION**

The take-home message of this paper is that large-scale data development requires a good project manager who can break the effort into logical, sequential processes as much or more than it requires technical skills. At the outset of the project, it is important to define user needs and design the effort to fit these needs. It is important to scope the effort properly. As the project progresses, it is important to track edits and to double-check (QC) the results.

There are many details of data development that could not be covered fully or were not covered at all in this short paper. For example, I did not discuss methods for scoping projects such as benchmarking. I did not discuss data models, a hot topic in GIS today that you can research by simply searching for "GIS data models" on the web. Remember, there is no reason to reinvent the wheel. If you are looking for project support and did not find what you were looking for in this paper, there is a plethora of information available via the web and via networking with other GIS professionals. Regardless of your experience level, always be open to good advice and new methods, and good luck!