An Advanced GIS to Legacy Database Linking Application

Roxanne Gregorio, Rami Raad*, Chris Peluso, Andrew Hu
CH2M HILL
1700 Market Street, Suite 1600
Philadelphia, PA 19103

ABSTRACT

CH2M HILL, Inc., has developed an advanced ArcGIS application for the Philadelphia Water Department which links newly acquired GIS data to legacy valve, inlet, and hydrant data. The application includes a Legacy Data Conversion Toolset, a Legacy GIS Linking Toolset, and a Legacy GIS Matching Geodatabase. Legacy Data Conversion Toolset contains functions to automatically map each legacy record based on an advanced geocoding and migration algorithm. Legacy–GIS Linking Toolset automatically determines possible matches between legacy data and nearby GIS data based on complex matching rules. It also contains manual checking tools for correcting multiple matches and non-matched data records. By using the ArcGIS application, we were able to map 88 to 92 percent of legacy data records and match 74 to 86 percent of them to GIS in a very short time, thus significantly reducing the manual effort required to link the databases and increase the confidence in the matching results.

INTRODUCTION

The mission of the Philadelphia Water Department (PWD) and Water Revenue Bureau is to serve the greater Philadelphia region by providing integrated water, wastewater, and stormwater services. The utility's primary mission is to plan for, operate, and maintain both the infrastructure and the organization necessary to supply high-quality drinking water; to provide an adequate and reliable water supply for all household, commercial, and community needs; and to sustain and enhance the region's watersheds and quality of life by managing wastewater and stormwater effectively.

One important way in which PWD is fulfilling its mission is through this Geographic Information System (GIS) Data Conversion project for water and sewer infrastructure. GIS technology is an increasingly valuable tool for improving overall facility maintenance and strategic utility decision-making. The main objective of the project is supporting the Philadelphia Water Department's needs to fully establish an electronic information management system. The Data Conversion Team is compiling a broad spectrum of engineering data for over 6,620 miles of water, wastewater, stormwater, and high pressure fire infrastructure into citywide GIS coverages. The effort includes extracting data from over 250,000 engineering documents, which currently exist as scanned images or in hard copy. The project has been executed in three phases. The Initiation Phase included a series of workshops designed to ensure that the conversion process properly utilized the 85 different types of source documents maintained by the department. It also included customization of data conversion tools to meet the project's data specifications, the development of a detailed conversion workplan, and conversion of the data for a 2-block area within the city. The Pilot Phase included further definition of the project's data dictionary and conversion tools and applied both to data from 2 of the City's 121 map tiles. The project is currently in the third or Production Phase, which includes conversion of the remaining tiles and the establishment of links between the GIS data, and legacy databases related to valves, hydrants and storm sewer inlets.

The project has been supported through the use of customized conversion tools for data collection, data scrubbing, data entry, graphical placement, and quality control. Conflicts and anomalies in the data are tracked using a web-based tool and database.

Upon completion, PWD expects to utilize the GIS coverages as the foundation for many of their operations including maintenance management, capital improvements, and hydraulic modeling. The single GIS database will help to streamline operations and enhance troubleshooting for its water and sewer systems.

CONVERSION PROCESS

Figure 1 illustrates the data conversion process for all water, sewer, and HPFS systems. Special features of the process include a concurrent but separate system conversion, a scrubbing process sequenced by complexity, direct database entry of facility attributes, automated routines for placement of linear features, and a collective reconciliation of utility systems after the quality control (QC) process.

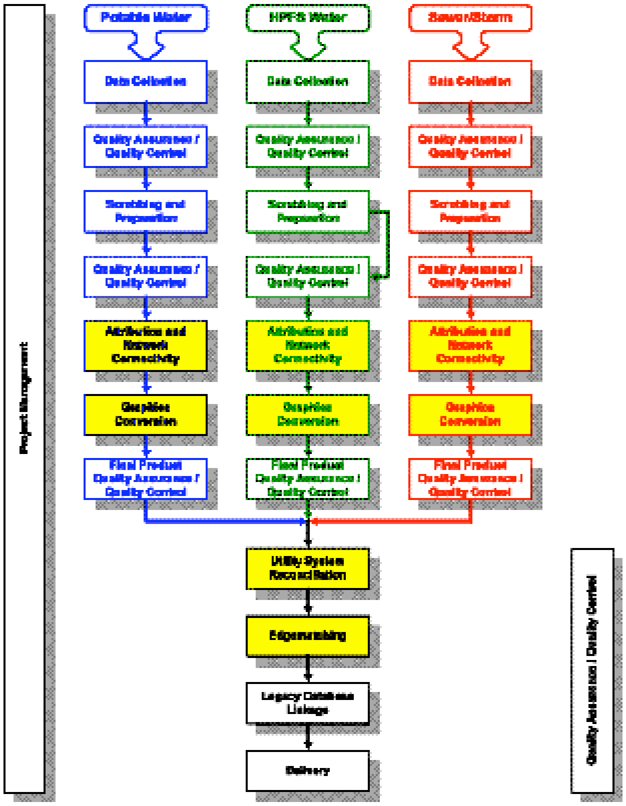

Project Management
Potable Water
HPFS Water
Sewer/Storm
Data Collection
Data Collection
Data Collection
Quality Assurance / Quality Control
Quality Assurance / Quality Control
Quality Assurance / Quality Control
Scrubbing and Preparation
Scrubbing and Preparation
Scrubbing and Preparation
Quality Assurance / Quality Control
Quality Assurance / Quality Control
Quality Assurance / Quality Control
Attribution and Network Connectivity
Attribution and Network Connectivity
Attribution and Network Connectivity
Graphics Conversion
Graphics Conversion
Graphics Conversion
Final Product Quality Assurance / Quality Control
Final Product Quality Assurance / Quality Control
Final Product Quality Assurance / Quality Control
Utility System Reconciliation
Edgematching
Legacy Database Linkage
Delivery
Quality Assurance / Quality Control

Figure 1 - Conversion Process

Data will be converted for all utility systems in each delivery area simultaneously, reconciled to ensure all utilities are placed correctly relative to other utility systems, edge-matched, and then delivered as seamless databases.

DATA COLLECTION

PWD maintains all scanned engineering drawings and supporting documents in a large Image Server. These Images are presented to PWD users via an intranet based viewer called the Engineering Record Viewer (ERV). ERV utilizes a GIS front end-ActiveX Control which is controlled via asp.net and JavaScript code. ERV accesses several City-Wide GIS Datasets and is expandable to hold data conversion data from this project as well as provide reporting capabilities to the users. The Philadelphia Streets Centerline for example is used to tie a majority of the scanned PWD engineering records to a spatial location. The scans can be queried both spatially and non-spatially. Functionality was added to the ERV system to facilitate collection of scanned PWD engineering records by Data Collection Tile (Figure 2).

Data Collection was performed in stages based on delivery areas. PWD divided the city into 121 tile areas. Tiles were grouped into a delivery area based on density of water and sewer infrastructures in that area. Figure 2 illustrates the schedule and delivery areas grouped by color.

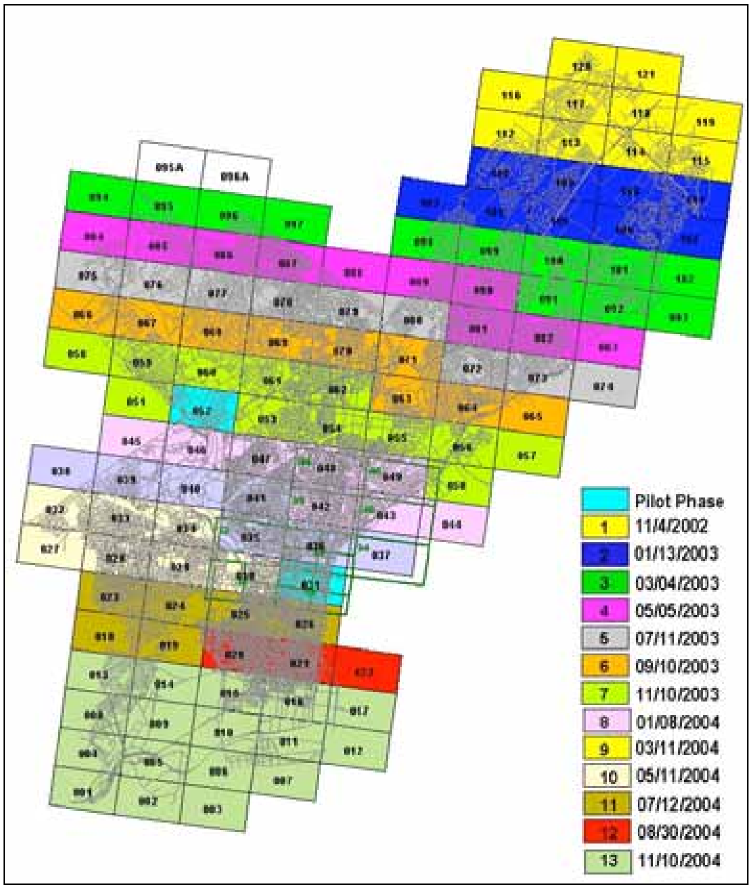


Figure 2 – Data Collection Schedule

Software applications were developed by the Data Conversion Team to duplicate the PWD Source Document Library maintained on the PWD Image Server and support scrubbing activities on the Data Conversion Team server. Three Visual Basic applications were developed:

Data Collection Polisher: This program was developed to format the SQL server Query output by Data Collection Tile of related scan data from the PWD ERV Application into a more usable format. The application replaces PWD file location strings with PWD GIS server strings, replaces blank values with default values, normalizes strings where needed to standardize the data, as well as other steps necessary to load the data into an Access Database.

Data Collection Builder: This program was developed to construct a tile-specific Access Database and the actual document libraries on the PWD team server. The application creates working scrub folders (the term "scrub" is used to refer to the process of digitally marking the scanned images for use by the Digitizing Team—See "Data Scrubbing" Section below) on the team server based on the street names identified in the ERV output. The user also has the option at this point to copy the source documents from the PWD Image Server to the PWD team server, and to copy primary source documents to the scrub folder locations. The last task accomplished is the creation of secondary/tertiary documents lists in HTML format. These are used later in the scrubbing process.

Data Collection Tracking: This program was developed to utilize the Access Databases and HTML documents to track the progress of primary sources through the scrubbing phase of the project. The application is mentioned here because the program uses the data prepared by the two previous programs and was developed under the data collection task.

DATA SCRUBBING

The objective of the data scrubbing process is to:

- Identify what data should be converted with consideration of spatial limits; disregard outdated or confusing information, etc.
- Consolidate all required information into a single (primary) source document to facilitate the attribution and graphical conversion process. All primary documents have secondary and tertiary documents that contain information pertinent to the attribution work. Each feature, whether it is an arc, node, or point, has a list of attributes that need to be culled from the raster image of the marked-up primary source document and placed into a database format. The scrubbing team needs to send the Attribution and Conversion Team enough information on a marked-up primary source document so that they can completely attribute and convert each feature's information according to the project's Valid Value List and established Rules.

Scrubbing Process

The general scrubbing processes for the water, sewer, and high-pressure fire system utilities are based on a seven-step process:

1. Setup Scrubbing Packages – The data collection process previously completed and quality assurance/quality control (QA/QC) checks place the primary source documents (PSDs) in corresponding street folders according to each tile number.

2. Open the Primary Source Document — Open the data collection package for the primary source document using Microsoft Visio Professional.

3. Label Each Feature with a Unique I.D. — Place labels with numeric IDs for each feature in the PSD. This step should be completed moving from left to right (or similar) in the order of abandoned features.

4. Review Attribute Information on Primary Source Document —Review the primary source document and highlight or text edit the drawing to make certain that all attributes for each feature are clearly displayed. Identify attribute data to be obtained from secondary source document(s).

5. Review Secondary Source Documents —Review secondary source document(s) to collect data or features not fully described or annotated on the primary source document.

6. Save the Data—Once the scrub is complete, save the document in Visio and as a .tif image and create an excel report of the scrubbed features.

7. Quality Assurance/Quality Control of the Scrubbed Document—Review the scrubbed document with the original primary and secondary source documents to verify that the source document was properly scrubbed.

DATA ATTRIBUTION

The data attribution process begins once the scrubbed work packages are received and inventoried to ensure completeness. Data scrubbed on the images is entered into Microsoft Access database forms for each facility on the primary source image. ID numbers assigned during the scrub process will link the data conversion graphics to the collected attributed data.

Source data for the PWD project will be delivered in tiles that contain many work packages. A work package contains all sources related to specific scrubbed water block plans (WBP), high pressure fire system (HPFS) plans, or sewer return plans (SRP). The work packages will be inventoried to ensure documents listed on the delivery transmittal are received. Data will be transferred to the production directory once work packages are completed for a tile.

Attribution Process

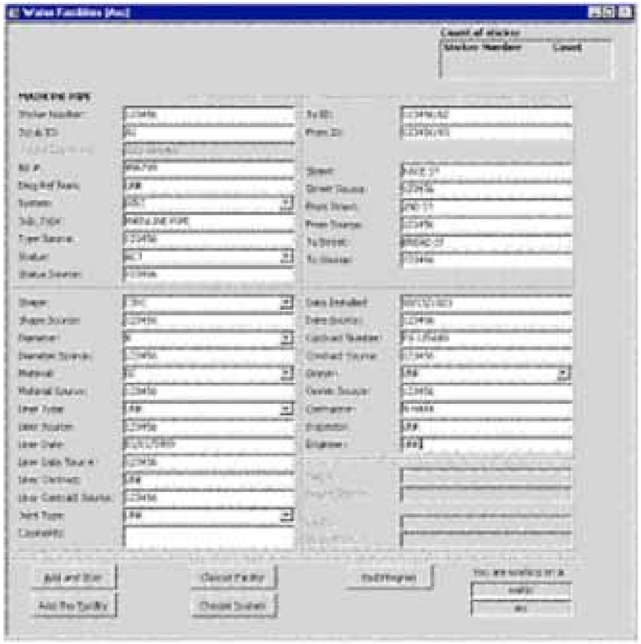
The general data attribution process for water, sewer, and HPFS utilities is based on the three-step process that will be further detailed in this document. The three steps for data attribution are:

1. Open Primary Source Documents

Open primary source documents annotated during the scrubbing process.

2. Access Data Attribution Application

Open the data entry application, illustrated in Figure 3, for the system found on the source document, choose facility type, and enter required data in sequential order by unique ID number.



Water Facilities (Arc)

Count of sticker

Sticker Number	Count

MAINLINE PIPE

Sticker Number:	123456	To ID:	123456/02
Sub & ID:	01	From ID:	123456/03
[illegible]	[illegible]		
WI #:	[illegible]	Street:	MACE ST
Dwg Ref Num:	UNK	Street Source:	123456
System:	[illegible]	From Street:	2ND ST
Sub_Type:	MAINLINE PIPE	From Source:	123456
Type Source:	123456	To Street:	BROAD ST
Status:	ACT	To Source:	123456
Status Source:	123456		
Shape:	CIRC	Date Installed:	[illegible]
Shape Source:	123456	Date Source:	123456
Diameter:	8	Contract Number:	[illegible]
Diameter Source:	123456	Contract Source:	123456
Material:	CI	Design:	UNK
Material Source:	123456	Design Source:	123456
Liner Type:	UNK	Contractor:	[illegible]
Liner Source:	123456	Inspector:	UNK
Liner Date:	[illegible]	Engineer:	UNK
Liner Date Source:	123456		
Liner Contract:	UNK		
Liner Contract Source:	123456		
Joint Type:	UNK		
Comments:			

You are working in a:
water
arc

Figure 3 - Attribution Form

3. Quality Assurance/Quality Control of the Data Attribution

When all attribute data has been collected for the work package, run validation programs and correct all errors. Generate reports listing number of facilities (arc, node and point).

Attributes for each facility will be collected in Microsoft Access from the scrubbed source image. The primary sources will consist of a Water Block Plan (WBP), Sewer return Plan (SRP), and High Pressure Fire System Plan (HPFS). Scrubbed images will contain the information required to attribute the data tables for each facility.

The facilities on the primary source have been scrubbed using Microsoft Visio software. Each facility that will require capture and data attribution collection has been assigned an ID number. Numbering for facilities on the source will be sequential for each type (unless edits are made after initial delivery). The sticker number and scrub ID number assigned during scrub will be concatenated to create the unique ID number for each facility captured.

DATA GRAPHIC

The general data graphic conversion process for water, waste, and HPFS utilities are based on the seven-step process that will be further detailed in this document. The seven steps for data graphic conversions are:

1. **Projection and Land Base -** The operator will define the projection metadata constraints and create the land base area with the necessary layers for the work area.

2. **Library, Tile, and Theme** - Operator will define the library, tile and theme to create designated working layers in Arc.

3. **Database Linkage Information -** Operator will define Feature-to-Database linkages for all Features that are defined to have them. Operator will also import unique IDs for the point facilities being captured in Arc. This ID extraction will actually be taken from both the appropriate point and node PWD data entry tables. The converted facilities will be coded with this field for linkage to the correct database records.

4. **Rectification of Image** - Operator will rectify source image for the plan in which he or she is working.

5. **Capture of Facilities/Arc Segment Generation** - Capture all point and node facilities using customized placement tools in Arc Info. Operator will code these facilities with the unique identifier from the Access database. The pipe (arc) facilities will be generated automatically using the "from-source" and "to-source" identifier that was placed in the Access data table during data attribution.

6. **Appending Coverage's** - Operator will append all coverage's of the same utility first and create the pipes (arcs) that fall in between the plan areas. The individual appended coverage's will be overlaid together and checked to ensure that utilities do not overlap unless sources indicate otherwise.

7. **QA/QC of Captured Facilities and Linkage** - QA editors will review the captured data and linkages to data table records to ensure that every facility has been captured. Quality process including Dog Creek will be run to check correctness, connectivity and data integrity prior to delivery.

FINAL QC PROCESS

Quality Assurance (QA) depends on the development and implementation of Quality Control (QC) and Quality Assurance procedures capable of ensuring accurate and consistent infrastructure mapping.

Quality assurance and quality control of the GIS data is vital to assure the data PWD receives is consistent with PWD expectations and will support their intended applications. The QC process provides for data rework if needed, and ensures that data rework and associated schedule delays are minimized.

During the Prototype and Pilot phases of the project, specific QC procedures were developed, tested, and implemented. As the project has moved through the Production phase, the QC checks have been expanded and continue to change to include new procedures and modifications as they are implemented, providing the QC staff with a handbook for performing rigorous QC review on each of the project deliverables prior to delivery. Most importantly the QA/QC process is implemented at each step in the data conversion process to ensure quality data.

Standard QA/QC Workflow for PWD GIS Database

QC workflow involves a variety of information inputs, evaluation tasks, report generation, information transfer, and quality review meetings. Together, these tasks make up the QC workflow as illustrated in the Figure 4.

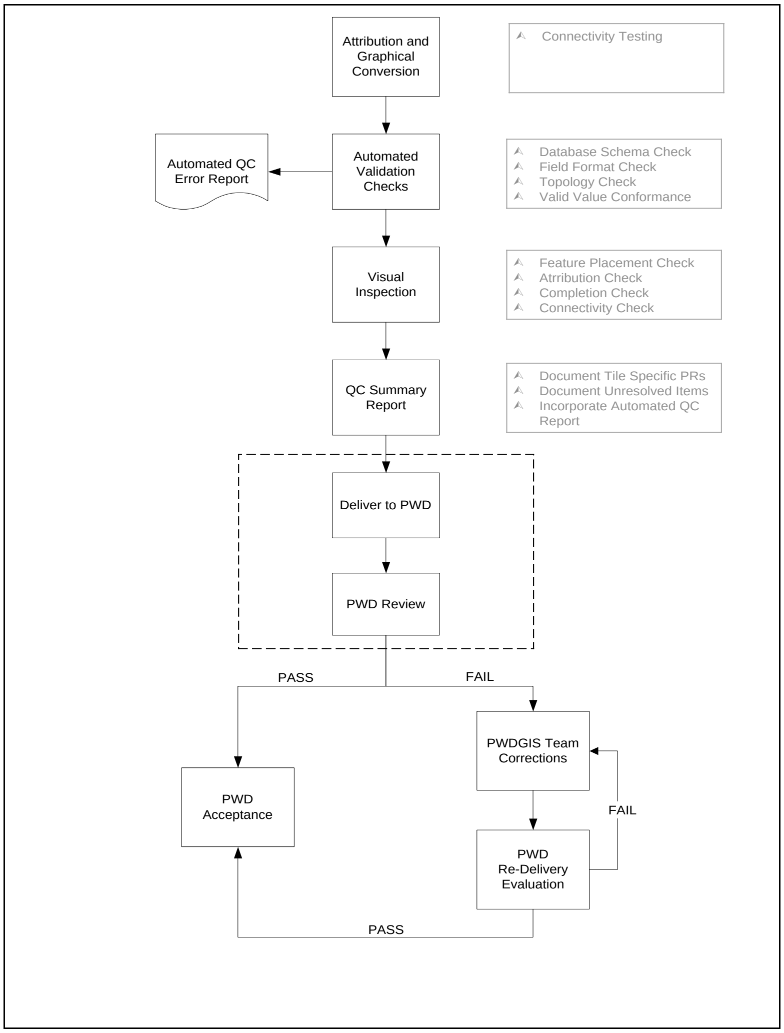


Figure 4 - QA/QC Workflow

Three general categories of checks are performed by the Conversion Team prior to PWD delivery.

1. Automated Validation Checks are performed to compare deliverables with project wide GIS specifications and database design standards.

2. Visual Inspection Checks are performed to validate issues such as feature placement, data entry, connectivity, and overall completeness. These checks can not be automated and require manual review.

3. ARC/INFO AML program Checks are performed to validate connectivity and network flow testing on the coverages before final delivery.

Automated Validation Checks

The first set of checks on delivered datasets is a series of automated validations processed by the Dog Creek QC Software program. Dog Creek has been customized to check the following QC validations. The validations fall into two Sub-Categories. The result of these checks is a summary report showing Pass/Fail status of each check.

Coverage Integrity Checks

- Coverage Existence – Verify that coverage exists in delivery area folder and coverage name is correct.

- Missing Coverage Files **-** All coverages include a specific set of files that are essential to their use in ARC/INFO. Verify that all of these files are present for each given feature type.

- Projection Errors **-** Check for proper coverage projection definition.

- Coverage Feature Errors - Verify that no invalid feature types occur within a coverage (e.g., lines in a point coverage). This problem sometimes occurs when data is developed in one environment and then converted into ARC/INFO at a later processing phase.

- Topology Errors - Check for valid topology types and verify that coverage attribute tables are properly sorted.

- Table Format Errors **-** Verify that coverage INFO tables exactly match the physical specification in the data dictionary files.

Database Specific Checks

- Duplicate Item Values **-** Verify uniqueness of attribute values within a user-specified item (such as Feature IDS).

- Invalid Item Value **-** Check for invalid codes using discrete values and ranges defined in the appropriate database definition files.

- Attribute Consistency Error. **-** Check for attribute value consistency between various items in a coverage.
- Items not populated. **-** Determine whether attribute values are present or absent within a specific item.
- Frequencies. **-** Generate attribute code frequencies for a specified item or combination of items.
- Relational Errors. - Check the integrity of links between related tables.

Visual Inspection Checks

The second set of checks on delivered datasets is a series of visual inspections performed by the PWD Project Team. Custom GIS tools were created to inspect the data. The validations fall into the following categories, and the results of these checks are also summarized in the summary report showing Pass/Fail status of each check.

- Feature Placement **-** Verify that features are placed in the appropriate location in regard to curblines and other landbase features.
- Feature Attribution **-** Verify features are appropriately coded with the correct attribute value in reference to data sources.
- Feature Completion **-** Verify that all features are captured in the appropriate coverage for the selected tile.
- Connectivity & Node Errors **-** Verify that there are no dangling arcs in polygon coverages, or gaps between nodes that should be coincident.
- Overlay Consistency Errors **-** Verify that graphic features in one coverage are logically placed with respect to a separate polygon coverage. For example, a point coverage of hydrants may be combined with a polygon coverage of carriageways to verify that no hydrants fall within carriageways.
- Random Sampling **–** Feature specific checks, such as Placement and Attribution, will be performed for 100% of the features during the Pilot phase of the project. During the production phase, features will be reviewed by random sampling techniques that comply with ANSI standards.

ARC/INFO AML Programs

A series of ARC/INFO AML programs are used to perform the next series of checks. Project files were created in ArcView 3.2 for review of the errors. These files have relative path references and standard extensions so that they can be used in any delivery subdirectory created in the QC process. A description of each AML is summarized as follows:

NODEVALENCE.AML – Performs a series of queries to identify potential valence errors (dangles) that require further inspection. Specific feature types, such as hydrants, are eliminated from further inspection based on rules incorporated into the program. The following are some of the node valence checks for Water, HPFS and Waste Systems:

WATER AND HPFS SYSTEM CHECKS

- Tee Subtypes should have no more or less than three pipe connections
- Valve Subtypes should have no more or less than two pipe connections
- Sleeves should have no more or less than two pipe connections
- Abandoned lines should not be connected active lines
- Raw water lines should not be connected distribution or transmission lines
- Hydrants should be connected to hydrant lines
- Hydrants should only have one pipe connection
- A Cross should have four pipe connections
- Dead End should only have one pipe connection.
- Wye fitting should have three pipe connections.
- Reducers should be connected to 2 pipes with different sizes.

WASTE SYSTEM CHECKS

- Abandoned lines should not be connected active lines
- Stormwater lines should not be connected to Sanitary lines.
- Bulkhead should have one pipe connections
- Dispersion Chamber should have three pipe connections
- COLLAR should have two pipe connections
- Inlet Connection should have three pipe connections
- No Information Connection node should have one pipe connection

- Pipe End should have only one pipe connection
- Plug should have only one pipe connection
- Sleeve should have two pipe connections
- Tee field constructed should have three pipe connections
- Tee manufactured should have three pipe connections
- Wye field constructed should have three pipe connections
- Wye manufactured should have three pipe connections

FLOWSPLIT_ALL.AML – The AML program calculates the number of incoming and outgoing pipes at each node. The calculations are then used to determine if the WASTE coverage pipes have been captured in the correct direction of flow.

ERROR_CHECK.AML – The ERROR_CHECK.AML performs a series of logic checks against all available coverages based on the rules defined for the project. It is a "catch all" program for potential error conditions as they are identified. The following describes the current inventory of checks:

- Node valence error
- Active connected to Abandoned
- Raw System connected to Distribution or Transmission Systems
- Sanitary System connected to Stormwater System
- Flow direction to an invalid node
- Inlet node not connected to Inlet line
- Offset Manhole not connected to Offset Access line
- Downstream Elevation greater than Upstream Elevation
- To node Invert Elevation greater than pipe Down Stream Elevation
- From node Invert Elevation less than pipe Upstream Elevation
- Material Reinforced Concrete Pipe prior to 1905
- Date installed of pipe and node does not match
- Pipes connecting at Collar with same diameter

- Hydrant or Hydrant Valve not connected to Hydrant line
- Domestic Service Valve or Domestic Service End not connected to Service line
- Fire Service Valve or Fire Service End not connected to Fire line
- Pipes connecting at Reducer with same diameter
- Invalid node type connected to Blow off
- Cross Connection Valve not connected to Cross Connection line
- Vertical Offset not connected to Vertical Offset line
- Intake not connected to Pump Intake line
- Round Connection Valve not connected to Round Connection line
- Material Ductile Iron prior to 1955
- HPFS Hydrant not connected to Hydrant line
- Invalid node type connected to HPFS Blow off
- Vertical Offset not connected to HPFS Vertical Offset line
- Potential Transmission/Distribution System connectivity/coding error

DOWN_CHECKS.AML **–** This program creates two temporary coverages representing the SAN/COMB system and STORM/COMB system to validate that downstream pipe diameters are equal to or larger than upstream pipe diameters.

REMOVE_DUPES.AML – Because of the many to many and many to one relationships evaluated in the DOWN_CHECKS.AML program, there will be duplicate records in the output text files. This AML processes each text file and eliminates duplicate records.

HYDDIST.AML – Checks for hydrants in carriageways.

INLETDIST.AML **–**Checks for inlets in carriageways that are greater than 5 feet from curb lines.

CHECK_BENDS.AML – Checks for areas where STORM pipes should bend around SAN manholes.

MANHOLE_CHECK.AML –Checks for duplicate manhole IDs.

If the dataset has passed all automated and AML tests and has met the acceptance criteria for visual inspection and random sampling, the dataset is considered PASSED and can be processed for delivery to PWD. If the dataset fails any test and does not meet acceptance criteria for random sampling, the data is considered FAILED and will be returned with error reports for editing. Once edits are completed or exceptions are documented, the dataset will be returned for an additional sequence of QC procedures. This process will be repeated until all tests have received a PASS status.

Based on acceptable results of the combined test categories, the delivery area will then be packaged and delivered to PWD in the following format:

Water.e00 ARC/INFO GIS export file of the water distribution system

Waste.e00 ARC/INFO GIS export file of the collection system

HPFS.e00 ARC/INFO GIS export file of the high-pressure fire system (where applicable)
DelvXXX.mdb Microsoft Access Database of all database files for the three GIS coverages.

Legacy Conversion

One of the first uses of the converted data was to link the GIS data to existing maintenance information that is stored in Legacy databases. There was a significant amount of historical maintenance information stored on the hydrant, valves, and inlet systems that is updated on a daily basis. All of the information is stored on a HP mainframe which is not currently linked to any of the GIS information. The information in the Legacy database will be utilized in PWD's future maintenance management systems.

The problem with linking the databases is that the geographic information in the legacy databases was in a free text format based on non-specific GIS information such as street corners, non-geodetic standardized directions, and distance offsets. Also each dataset had its own set of locational data which had to be parsed into information that could be geocoded to relative positions so that it could be matched to the existing GIS data. Developing the Legacy to GIS linkage was done in the following two steps which are described below, and also shown in Figure 5:

1. Convert Legacy database address and feature type information (Hydrant, Inlet, and Valve) into a legacy data point geodatabase based on the address information presently stored in the legacy databases.

A CH2M HILL developed Legacy data geocoding/migration application was developed and used to process the Citywide legacy data. The legacy data conversion processes includes parsing the address, geocoding, and migrating the legacy data points to correct street locations. The data conversion was done primarily by automated processes with limited amount of manual checking and update. The final results of this task were a Citywide legacy data point geodatabase. Only address and feature type information from the legacy database was transferred to the converted legacy geodatabase. Legacy data records that did not pass the automatic conversion process were forwarded to PWD for further review and correction.

2. Match Legacy data points to GIS data points using a CH2M HILL developed Legacy-GIS matching program.

The matching entailed using certain matching algorithms which utilized geographic proximity and the attribute information in both databases to determine automated matches or non-matches between the data sets. The matching results were stored in an ArcGIS feature class indicating the links between the legacy database and the GIS database. Matched features in question went through a manual matching process. An application was developed to make the manual matching process more efficient and to ensure quality of the data.

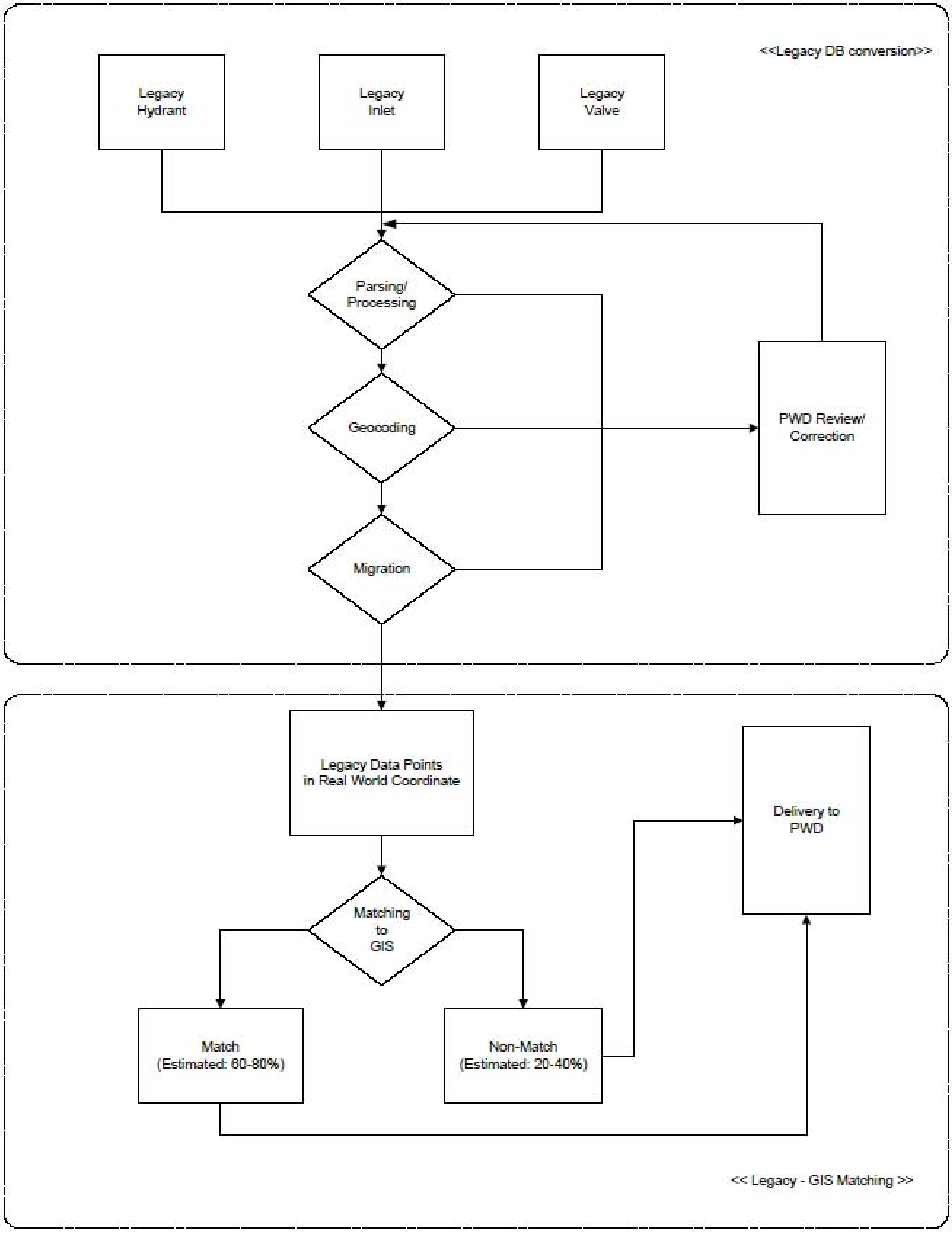


Figure 5: Steps to Develop the Legacy to GIS linkage

The steps involved in the legacy database conversion and matching process are defined in more detail below.

Legacy Database Conversion

The purpose of the legacy database conversion was to develop coordinates for each of the Legacy data points. As discussed previously, the issue with determining a coordinate for each point is that the geographic information in the legacy databases was in a free text format based on non-specific GIS information such as street corners, non-geodetic standardized directions, and distance offsets. Therefore the first step in the process was to determine coordinates for each legacy feature. The following describes the tasks that were used to determine the coordinates.

1. Legacy Data Collection and Preparation

At the beginning of the legacy database conversion process, Citywide legacy database were acquired and prepared for the parsing, geocoding, and migration process,

- Collect Hydrant, Inlet, and Valve Legacy database
- Covert the Legacy databases to Access database
- Selected data fields were extracted from the legacy databases
- Examine the location and location type information
- Systematically clean and normalize the location field as necessary

2. Automatic Legacy Data Parsing

After the legacy databases were prepared for the conversion program, an automated parsing program was run to parse the legacy data location information. The automatic conversion program was updated to recognize some special address cases found in the pilot study, such as "HP-" suffix for high pressure fire hydrant address, a missing comma (",") between two street names, using address number on the opposite site of street, and direction flags (S/B, N/B, etc) to improve the automatic parsing performance and subsequent data migration results. If legacy data records did not pass the enhanced automatic parsing process, they were forwarded for further review and/or manual correction.

3. Automatic Geocoding

After the legacy data addresses were parsed, an automatic geocoding program was run to geocode the legacy data points. Legacy data records that did not pass this process were forwarded for further review and/or correction.

4. Automatic Migration

After the Legacy data points were geocoded, an automatic migration program was run to migrate the data points to the correct street locations. Legacy data records that did not pass this process were forwarded for further review and/or manual migration.

5. Manual Check/Update Migration

A QC of the Geocoding/Migration results was performed on a limited amount of manual check and corrections. Legacy data points often could not be geocoded/migrated correctly along double-line or multiple line streets or highways or at the intersection/exchange with them. It was necessary for this data to be QCed and/or geocoded and migrated manually. T intersections often caused errors in the automatic geocoding/migration results where "On Street" and "Intersecting Street" are difficult to match with the legacy database records. A program was developed to automatically detect T intersections and create points to mark their locations for manual checking.

6. Create Converted Legacy Geodatabase

The geocdoed and migrated legacy data was imported into a geodatabase:

- Import geocoded and migrated legacy data points into geodatabase
- Join legacy data point with its feature type attribute data
- Build citywide Legacy data point geodatabase

Legacy-GIS Data Matching

Once a geographic coordinate was derived for each of the Legacy data points, the next step was to determine if the Legacy point had a corresponding GIS point to which to match. This was a very difficult algorithm to develop due to the complexities of the attribute information and the differences in the format of the GIS and Legacy databases. The following is the process that we determined was the most efficient and produced the highest matching results.

1. Prepare Matching Geodatabase

   Prior to performing the legacy to GIS matching program, a Matching geodatabase was created for the delivery which hosted all the related Legacy and GIS data,

   - Create matching geodatabase
   - Select Legacy data from converted legacy geodatabase by delivery tiles and import in to matching geodatabase
   - Convert GIS coverages and attribute data and import them into matching geodatabase

2. Automatic Matching

Automatic matching programs were run to perform the legacy to GIS matching. The automatic matching program comes with a default search buffer radius. However, the program needed to run multiple times with varying search buffers which can be set at runtime to achieve the best matching results. The scoring of the matching process is discussed further in the matching algorithm section below.

3. Manually Checking

The Legacy Team manually checked the matching results to remove multiple matches and confirm or reject questionable matches.

The process determined an average of 20% to 40% of the legacy data points that can not be matched to GIS features by the automatic matching process, which were reported for further review. Significant manual effort was required to cleanup these data.

4. Matching results and Summary report

The matching results were exported from the match processing geodatabase to be used as the link between the legacy database and the GIS data by PWD.

Matching Process

As stated above the matching process entailed developing a matching algorithm to determine if the Legacy to GIS match resulted in an automated match, a match that had to be manually reviewed, or a non-match. Matching criteria was developed for each of the data sets (hydrants, valves and inlets). The purpose of the matching program was to develop scores based on attribute information in both datasets as well as geographic proximity. The following describes one of the matching algorithms for the hydrant data set:

Matching Algorithm

In the Legacy Database, each feature, i.e., hydrant, inlet, or valve, has Location Type information. The proposed methodology was based on the each Location Type since each requires a different resolution.

Hydrant MA – Address

For the Address location type, the Legacy (geocoded) features are located at the approximate street address. There are five major processing steps:

1. Read the Legacy ***Location*** value

    - Determine the ***on street*** name
    - Determine the ***address number***

- Adjust the ***Street Type*** of the “on street” name to be consistent with the Street Type value in the Street Centerline and Water map layers

If the on street and street number can not be read correctly, the process terminates and continues with the next record. This may happen when the format of the Location information is not in accordance with the Location Type information.

2. Find ***nearby (200ft) candidate features*** with the same feature type in the Water map layer

    - Create a ***buffer*** (circle like polygon) using the migrated curb corner as the center with 200 ft radius
    - Find ***a collection of features*** with the same feature type in the Water map layer within the buffer

If no nearby candidate features are found within the buffer, the process terminates and continues with the next record. The may happen because of data error in either the Legacy or the Water data set.

3. Determine the ***on Street Name*** and ***side*** of each nearby candidate features

    - Find the ***nearest street segment*** of a nearby candidate feature
    - Determine whether the ***address number on the left-hand side*** (based on the digitized direction) of a street segment should be ***even or odd*** number (L_F_ADD field in the Street Centerline map)

4. Apply a set of ***criteria*** to these nearby candidate features

    - ***Proximity Criteria*** – all nearby candidate features found with a buffer met the proximity criteria and were given a score value of ***35***
    - ***Nearest Distance criteria*** – there are three possible conditions:
        - ***25*** - if there is only one nearby candidate feature found, it was assigned a score value of 25
        - ***20*** – if more than one nearby candidate features found, the nearest feature of the nearby candidate features is given a score value of 20
        - ***10*** – if more than one nearby candidate features found, all features, except the nearest one, have a score value of 10
    - ***Address Number and Street Name Criteria*** – there are four possible conditions:
        - ***25*** - if the nearby candidate features have the same street name (in the Street field) as the on street name and their address numbers are located at the correct street side, they will have a score value of 25;
        - ***15*** - if only one criteria is correct (either on street name or address number), they will have a score value of 15
        - ***15*** - if both conditions failed but their ***from_street*** or ***to_street*** value matches with the Legacy on street name, the score value is 15
        - ***0*** - otherwise (none matched), their score value is 0
    - ***Water System Criteria*** – If the main_size value in the Legacy data is < 16 inches, their water main system should be distribution system, i.e., the value of the System field in the Water data set should be DIST; otherwise, they data belongs to the transmission system and the system field in the Water data set should be

TRANS. . Any nearby candidate features that met this criteria has a score value of ***15***; otherwise, their score value is ***0***

5. Establish the ***association*** between Legacy and Water data sets

   - The results of the association is ***a line shapefile*** with a set of attributes - a line is drawn from the Legacy geocoded (or migrated curb corner) street intersection point to individual nearby candidate point feature.
   - ***ID Information*** – the ID values from both Legacy and Water data sets are recorded for each line.
   - ***Miscellaneous Information*** – map sheet, feature, and location type are recorded.
   - ***Distance*** – the distance of the connected line is stored.
   - ***Criteria Scores*** – the score of each criterion is also recorded in separate fields in order, i.e., criteria 1 to criteria 4.
   - ***Probability*** – the sum of all 4 scores is appended.

Manual Matching Process

As described above, once the automated matching process was completed a certain percentage of the features were determined to need a manual review. The manual review group was determined to be any automated match that did not have a matching result of 50 to 75 using the algorithm developed for each feature in addition to any feature that had multiple matches to legacy data. An application was developed to enable the manual review process to be completed in an efficient manner. The following tools were developed as part of the manual review application (Figure 6):

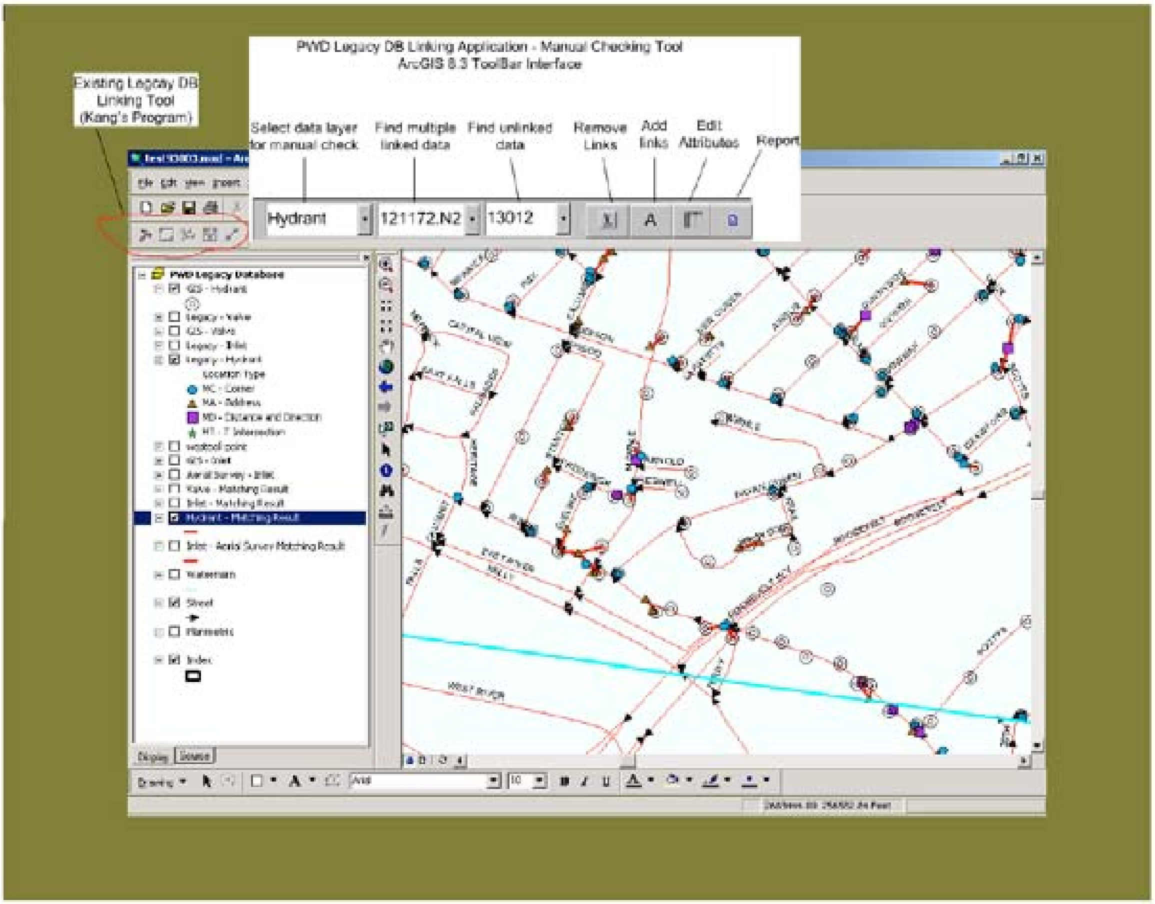

PWD Legacy DB Linking Application - Manual Checking Tool
ArcGIS 8.3 ToolBar Interface
Existing Legacy DB Linking Tool (Kang's Program)
Select data layer for manual check
Find multiple linked data
Find unlinked data
Remove Links
Add links
Edit Attributes
Report
Hydrant
121172.N2
13012
PWD Legacy Database

Figure 6: Legacy Matching Application

1. **Select data layer** – This tool utilizes a drop down list (combo box) to display a list of checking data layer names. Clicking on the down arrow will show the list of layer names. Users may then select a data layer to check on by its name. This tool will automatically locate the corresponding data layers from the map.

2. **Find multiple linked data** – This tool utilizes a drop down list to display a list of Ids of the multiple linked GIS and legacy data. Selecting one of them from the list will result in the sub-sequential selection of links on the map. Given the predefined map unit and projection, the map will automatically zoom to the scale that can display the selected feature and surrounding GIS, legacy, and base map data in detail.

3. **Find unlinked data** – This tool utilizes a drop down list to display a list of Ids for unmatched legacy data. Users may select any legacy ID from the list, the map will automatically zoom to the scale that can show the selected legacy data feature and surrounding GIS and base map data in detail.

4. **Remove links** – This tool allows users to click on a link and delete it from the map layer. The tool first selects the link on or very close to the mouse point at the point of click by querying feature layer according to spatial proximity rules. The selected record is then be deleted.

5. **Add links** – This tool is first used to select a legacy data point that has no links, and then it is used to select GIS data points to be linked to. After the pair of data points has been selected, the tool creates a new link between them and adds the record to the matching results data layer. Necessary attributions from both the GIS and legacy data points are written to the new link record.
6. **Edit links** – This tool allows users to open the attribute table for a matching result feature (link) to update the linking information.
7. **Report** – This tool allows users to summarize the print matching results and statistics used in the report.

Results

Once the Legacy datasets were run through the conversion and matching program including the manual matching the amount of matched features were in the range of 70% to 81% for the three data sets. Table 1 displays the summary of the results for approximately half of the geographic area of the City.

TABLE 1
Legacy Matching Results

	Hydrants	Inlets	Valves
Number of GIS records	5,197	17,939	16,730
Number of Legacy Records	4,548	15,098	13,214
Automatic Matched	2,991 (66%)	8,314 (55%)	8,723 (66%)
Manually Matched	195 (4%)	3,146 (21%)	1,981 (15%)
Total Matched	**3,188 (70%)**	**11,460 (76%)**	**10,719 (81%)**
No Match	1,105 (24%)	2,200 (14%)	1,712 (13%)
Forwarded to PWD for Review	257 (6%)	1,444 (10%)	805 (6%)
Total Not Matched	**1,362 (30%)**	**3,644 (24%)**	**2,517 (19%)**

Future Applications

Being able to match to such a high degree (70-80%) the newly acquired GIS coverage data to existing Legacy data for valves, inlets and hydrants allows PWD a whole new window into important system data. While the GIS coverages consist of 1.35 million features with an average of 40 attributes per feature, the history information of that data is not associated in the GIS dataset. Having this final deliverable of a linking table of the legacy ID number and the GIS Facility ID number allows a whole other layer of data to be accessible to GIS users.

There are several applications which currently exist at PWD which will become more integrated with the GIS dataset over time. But the linkage to the existing Legacy data, as was accomplished under this task of the project, was by far the most difficult to accomplish. The Hydraulic Model and SAP-DSS Applications already are tightly integrated with the current PWD GIS data.

Following is a list of the Planning and Inventory Management Applications which will either develop an initial integration or a stronger integration with GIS over the next few years.

Planning Applications

- Leak Detection System
- CAPIT (Capital Management System)
- SAP-DSS (Sewer Assessment Project-Decision Support System)
- Defective Lateral System
- Hydraulic Water Model

Inventory Management

- Legacy Systems for Valves, Inlets and Hydrants

Until now, these applications (Leak Detection, CAPIT, and Defective Lateral System) have used the Citywide Streets centerline data to provide a graphical link in each application if any graphical link exists at all. However, a street centerline is no replacement for the actual PWD feature that is being reported on in an application. While personnel need to locate their projects via a street centerline, i.e., along street, from street, and to street location, they are not reporting on the streets' characteristics themselves but on the actual water and sewer features involved in their work. It follows that with the existence of the PWD GIS dataset, their applications will need to be shifted to connect to the actual water and sewer assets via a unique GIS FacilityID.

This constitutes a large paradigm shift within the PWD organization. For over a hundred years the actual PWD facilities have almost been masked by their along, from, and to street locations. Now this positional description can be joined with actual PWD asset attribute and geospatial information to provide a more robust application environment for these planning and inventory management applications. While the Legacy Systems for Valves, Inlets and Hydrants, the Hydraulic Model and SAP are currently linked to the new GIS data and are the pioneer applications of this paradigm shift, the remaining applications will require some re-work to get them to link directly into the PWD GIS datasets. Also it should be noted that the Legacy Systems are planned to be migrated into a new Work Order Management System (WMS) which will link more efficiently to the GIS datasets. At the time of the transfer, the linking tables between the current legacy data and the PWD GIS data can be utilized to prime the WMS with the volume of asset histories that has been collected for more than 20 years.

Maintenance of the PWD GIS data alone and also its linkage to the Legacy data is of utmost importance to the PWD Data Conversion team. An investment as significant as this Data Conversion Project simply cannot be allowed to go out of synch with "what is happening on the street." The linkage between the GIS data and the Legacy data will be kept up to date by using a Microsoft Access project which links directly through ODBC connections to the HP Mainframe in which is stored the Legacy data. Queries are run periodically which show additions to the Hydrants, Inlets and Valves legacy database. Approximately 3000 valves, 80 inlets and 80 hydrants are added to the entire system annually. While the newly added inlets and hydrants can be manually linked as the number of them is manageable, PWD in-house staff will utilize Ch2MHILL's matching tool in Arcmap 8.2 to add linkages for the newly added valve data.

With these tools and a dedicated in-house GIS staff, it is assured that the GIS dataset's linkage to the legacy data will not become obsolete due to lack of maintenance.

Conclusion

The PWD GIS Dataset is a highly sophisticated dataset in that it was drawn piece by piece from the as-built drawings that describe, in a block by block fashion, the entire water, storm, hpfs, and waste water system for Philadelphia. For the first time in the history of PWD there is a seamless dataset available that represents the entire city's water and wastewater system. First phase usage of the dataset revolves around maintenance and basic reporting on the features and their attributes. All maintenance occurs within the ESRI environment. Basic reporting functionality on the dataset will be accomplished through a thin client web application. This will allow a large number of managerial and field personnel to review pre-set reports on the data and to visually inspect the data by navigating via the streets of Philadelphia. This will satisfy the immediate needs of the current GIS user population. Then the needs of the current applications which link or will link to PWD GIS data must be met. However, the most important integration of the GIS dataset with an application must be with the new Work Order Management System of which this task of linking PWD GIS data to the Legacy valves, inlets, and hydrants was a first step. Linking the GIS Spatial and attribute data with the work order histories of the data, even more than just the valves, inlets and hydrants, will provide an extensive set of data upon which many types of reporting can happen to assist in proactive maintenance of the Philadelphia Water and Wastewater System.

As time goes on, more and more complex spatial and attribute queries will be devised to gain insight into the system, its behavior, and possible future paths of change. And continuous checking of the data against real world assets by field personnel will result in an increasingly accurate dataset. It is envisioned that monthly meetings of the GIS Steering Committee will continue to assist in prioritizing all projects that involve usage of or linkage to this new and sophisticated GIS dataset. The GIS dataset should become a central hub in the wheel of organization of the PWD in order to maintain its main mission of providing clear, clean, quality drinking water to the residents of Philadelphia.

References

"The City of Philadelphia Water Department GIS Data Conversion Services Data Conversion WorkPlan." CH2MHILL Data Conversion Team: CH2MHILL, AGRA Baymont, Inc., Schoor DePalma, Inc., CSA North America, Inc., LAM Associates, Inc. March 27, 2002.

The Drinking Water Dictionary. Water Amer, American Water Works Association, Lee Jr. Bradley, James M. Symons

Author Information:

Roxanne Gregorio , GIS Manager, Philadelphia Water Department. Aramark Tower 1101 Market Street, Philadelphia, PA 19107 p. 215 685 6333 f. 215 685 6207 roxanne.gregorio@phila.gov

Andrew Hu, Senior GIS Analyst, CH2M HILL 139261 Cedar Road Suite 600, Herdon, VA. 20171 p. 703 471 1441 f. 703 471 0231 Andrew.Hu@ch2m.com

Christopher Peluso, Water Resources Engineer, CH2M HILL 1700 Market St. Suite 1600, Philadelphia, PA. 19103-3916 p. 215 563 4220 f. 215 563 3828 cpeluso@ch2m.com

Rami Raad – Project Engineer, CH2M HILL 1700 Market St. Suite 1600, Philadelphia, PA. 19103-3916 p. 215 563 4220 f. 215 563 3828 rraad@ch2m.com