Section 8 – Technical Documentation of DataStore

GEOCODING AND MAPPING DATABASE DESIGN

The Project design calls for the construction of relational database management system to house the case records and census demographics, support tasks necessary to geocode and analyze cases and facilitate the mapping of the data. Rather than building an ad-hoc database for one-time use with the specific Grantees, OIG's instructions are to use this Project to create a working prototype of a permanent geocoding and mapping support database that would be used for similar projects with other Grantees.

Technology

In keeping with the rules of the Project, the information system uses only widely adopted, commercially available software technology.

- *Microsoft SQL Server 2000, Standard Edition*, provides the relational database environment for all databases. For the purposes of geocoding the case records and preparing them for mapping, the database system was hosted on a *Dell PowerEdge* server running *Windows 2000 Server*.
- MapInfo's *MapMarker ESP* gives SQL Server the needed geocoding capabilities.
- MapInfo's *SpatialWare for SQL Server*, an extension to Microsoft SQL Server, enables spatial or geographic operations within SQL Server. It provides support for some of the geocoding tasks such as assigning records with un-geocodable addresses to Census Block groups (which is discussed in the Enhanced Geocoding Strategies document found in Section 6.) SpatialWare also handles the tasks needed to assign case records to Congressional Districts and Places.
- *Microsoft Access, Office 2003*, provides the relational database environment for desktop application databases. Principally used to house the census and base data geometry for mapping and for distribution of the case data to end-users.
- *ESRI's ArcGIS for Windows* used to geographically analyze, map, and present case and census data.

Case Database Design

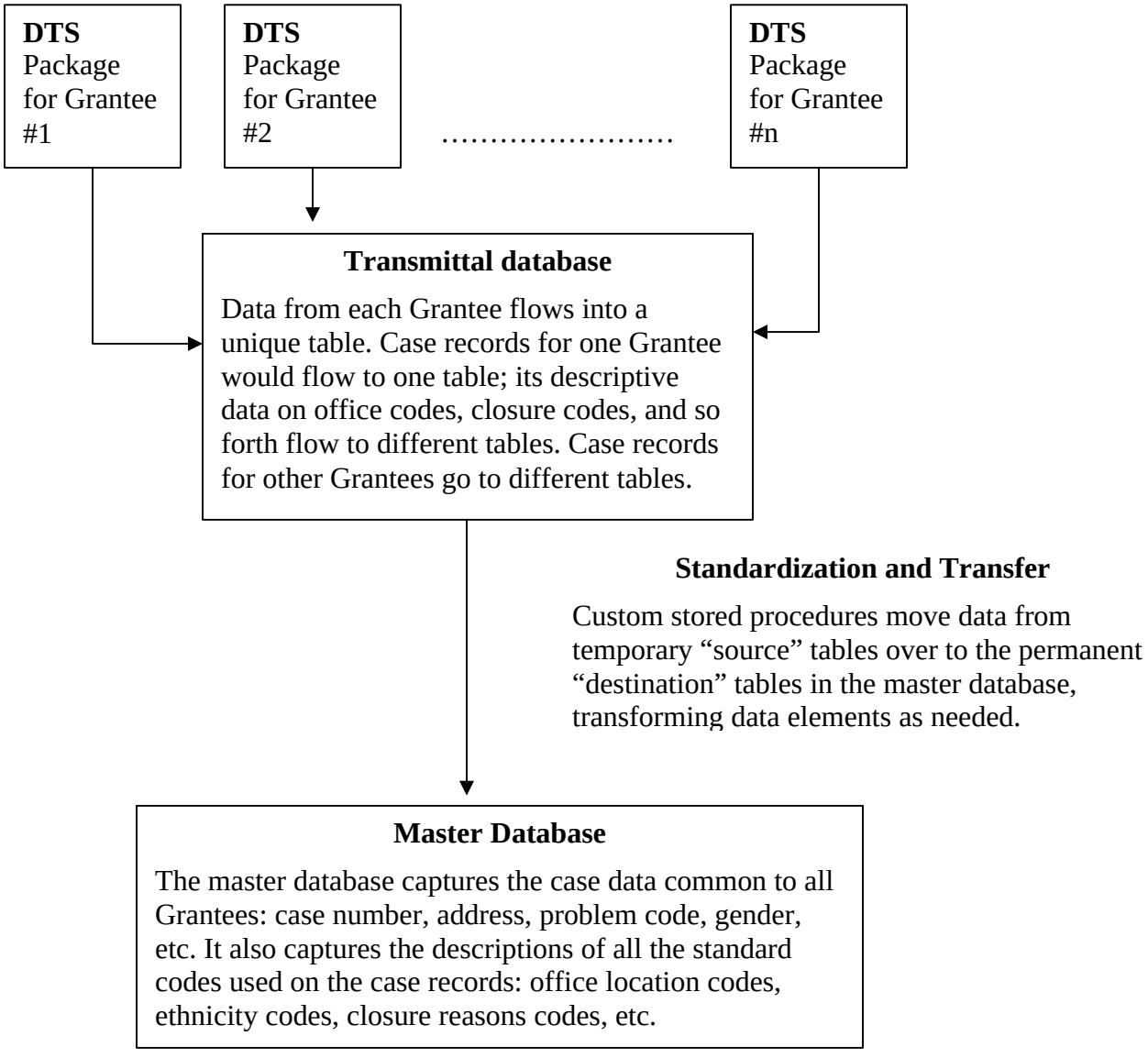
The case mapping system actually consists of two independent databases.

- Master Case Database: The "master" database houses the case data, the geocodes, and all the routines needed to organize and process the data. It houses case data from all Grantees in a single master table, thereby enabling analysis across either a group of Grantees or a single Grantee. More importantly, it allows for the same set of procedures to be used with case data from any Grantee.
- Transmittal Database: This database acts as a sort of way station for new case data. New case records first move into it from the transmittal files. Here the data is verified and standardized to meet the structure used in the master database. If all is well, it is copied

into the master database. In a permanent mapping database system, use of the transmittal database would provide an extra layer of security. Connections from external sources would be made to the transfer database only, thereby blocking access to the case records.

Grantee Data Capture Procedure

Grantees typically provide a unique collection of transmittal files, each differing on file format, field names, data types, and even field collections. Because data for all Grantees is housed in one common database, individual data files from each Grantee need to be arranged and standardized to fit the database design. For each Grantee, a custom Data Transformation Services (DTS) package moves the data from the original source files (e.g., Microsoft Excel) into a transmittal database. From there, a second custom process standardizes the data as it moves it into the master database tables.



Case Master Database Structure

The master database uses a conventional relational and normalized design. Case data from all Grantees is housed in one table. Geocoding data is housed in a separate but parallel table. That is, each case record always appears in both tables, with one table holding the original address and descriptive data, and the other holding the geocoding results. Separate "look-up" or "legend" tables house descriptions of problem codes, office locations, ethnicity codes, and so forth.

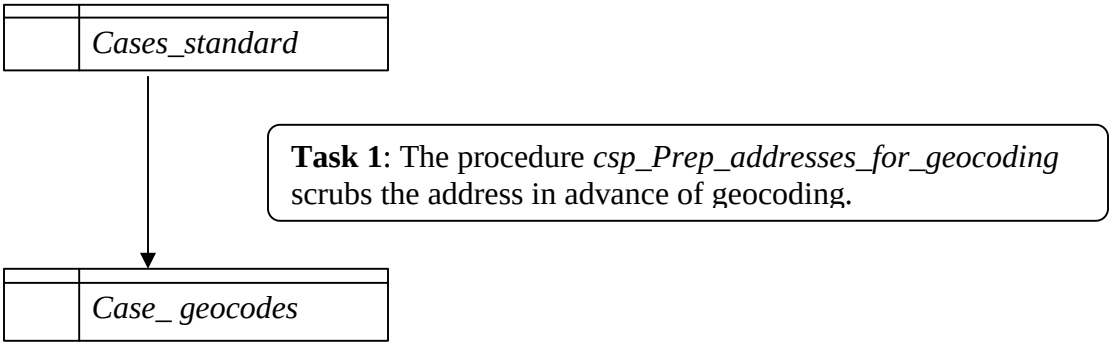
Table Name[1]	Description
Case Data	
Cases standard	All "standard" data from the Grantees is housed here, including the Grantee, case number, address, problem and closure codes, demographics, and open and closed dates. It also carries an identifier to link a case record back to the transmittal file provided by the Grantee.
Case geocodes	The results of the geocoding process are stored here, including standardized addresses, and Census, Congressional, and Place ids.
Related "look-up" Tables	
Grantees	This table contains a list of all Grantees represented in case tables.
Sources	Original transmittal files provided by the Grantees are documented here.
Office codes	Some Grantees use branch offices. Their addresses are listed here.
Intake codes	Each Grantee uses a unique set of "in-take" codes. All are standardized and stored in this table.
Problem codes	All Grantees use the same set of "problem codes" to describe the cases. These are captured here.
Closure codes	Reasons for closing cases are captured here. Each Grantee uses it own set of codes. This table captures each Grantee's set of codes.
Ethnicity codes	Standard race and ethnicity codes for the client.
Gender codes	Standard gender codes for the clients.

Please see Appendix C for a diagram of the table relationships, including foreign keys.

Geocoding Procedures

The geocoding process involves multiple tasks. The initial task prepares the case addresses for geocoding by scrubbing the original address of extraneous data like apartment numbers and commas. To preserve the original address information, the scrubbed address is copied into the table *Case_geocodes* and original address left as is in the *Cases_standard* table.

1 The actual table name in the database uses the "_" character instead of a space (e.g., "cases_standard" instead of "cases standard.")



The remaining geocoding tasks are executed against the Case_geocodes table have been captured in "stored procedures" in Microsoft SQL Server.[2]

Task 2:

Some addresses are not geocodable or require extra processing (e.g., homeless and rural routes). The procedure *csp_Determine_Address_Status* classifies the addresses.[3] For cases addressed to a Rural Route, an attempt is made to convert the Rural Route to a regular residential address using the US Postal Service's Locatable Address Conversion System (LACS).[4]

Task 3:

The procedure *csp_Geocode_with_MapMarker* passes the case records though MapInfo's MapMarker ESP product to geocode each to the best possible Census geography available for an address (e.g., Census Block, Block Group, Tract, or County). The FIPS Code of the Census geography is written into the *Case_geocodes* table.

Task 4:

Some cases have addresses that are not recognizable to the MapMarker geocoder and therefore do not get assigned to a Census geography. The procedure *csp_Apply_Alternative_Geocoding_Method* assigns these cases to a Block Group based on the ZIP Code.[5] The new Census ID is written to the *Case_geocodes* table.

Task 5:

After the case records have been assigned to a Census geography, two procedures make additional assignments: *csp_Assign_cases_to_Places_spw* assigns cases to Census Places (e.g., cities and towns); and *csp_Assign_cases_to_Congress108_spw* assigns them to Congressional Districts (for the 108th Congress). The values also are stored in the *Case_geocodes* table.

Reporting Procedures

Geocoding statistics for each Grantee is presented using a standard report template (created with Microsoft Excel), a sample of which is shown in Appendix B. Data for each section of the spreadsheet is generated using a specific stored procedure in SQL Server.

---

2 The document *Standard Geocoding Approach* explains in detail the geocoding procedures.

3 The document *Standard Geocoding Approach* explains the different address categories.

4 Please go to http://www.usps.com/ncsc/addressservices/addressqualityservices/lacsystem.htm for more information about this service.

5 Please see the document *Enhanced Geocoding Approach* for a narrative describing the process.

SQL Procedure	Report Section
csp_Report_Year_closed	*Cases Closed by Year*
csp_Report_Address_status	*Case Address Categories* breaks down the cases by the type of address (e.g., the address status).
csp_Report_Census_type	*Census 2000 Geography Assignment* reports the number of cases assigned to each level of Census geography: block, block group, etc.
csp_Report_ZIP_code_assigned	*ZIP Code Assignment*, which counts the number of cases with ZIP Codes.
csp_Report_Coordinate_type	*Map Coordinate Assignment (Lat/Long)*, which counts the number of cases by the accuracy level of the geocodes (e.g., street addresses or US Postal).
csp_Report_GeoResult_by_Other	*Supplemental Geocoding Results* records the number of cases that fall inside a Place or can be accurately assigned to a Congressional District.

Observations and Comments on Case Data

Duplicate Case Records

All Grantees inadvertently transmitted a small number of duplicate case records.[6] These were identified and removed from the master tables, and are not counted in the statistical tables.

Street Addresses

From the geocoding perspective, there is nothing overly remarkable about the street addresses. While it is true that a large proportion of the addresses are for sub-units of a building (e.g., an apartment), sub-units do not pose a particular problem for geocoders. All units in a building necessarily have the same map coordinates because they are all in the same building.[7]

Many case records have ½ unit addresses, such as 326 ½ S 24th Street. To a geocoder, a ½ unit is equivalent to an apartment like address. Geocoders have no particular problem with this type of address.

Finally, addresses in the case records frequently contained extraneous commas. For example, "1 E, Main St." contains an extra comma. Since the source file (ASCII formatted) was delimited using commas, we scrubbed these out before importing the records to prevent the street address

6 A case record is considered to be duplicated when two or more case records agree on the address, city, state, ZIP Code, problem code, year case closed, gender, and ethnicity, and the ages are within 3 years.

7 Case record addresses contain the phrases "Apt," "Bldg," "Space," "Unit", or "Floor" or the # symbol.

from being split between two fields in the database. (A better delimiter would be vertical bar like the "|".)

City and ZIP Code Coverage

A sizeable proportion of the city names on the case records do not match with names recognized by the Postal Service or the Census Bureau (as either a municipality or Census Designated Place). Also, city names are often abbreviated. For example, records may list the city name as "Los A." or "Santa." for Los Angeles and Santa Ana respectively.

It is not necessarily true that unrecognized names will lead to un-geocodable addresses. Commercial geocoders are designed to handle spelling mistakes in city names and incorrect ZIP Codes. As long as a ZIP Code is present or the city name is recognizable, the geocoder should be able to resolve the addresses. In fact, virtually all case records have a valid in-state ZIP Code. Still, because it seemed obvious from the ZIP Codes that "Los A" is short for "Los Angeles," case records with "Los A" as the city were edited.

USERS GUIDE TO THE DATA STORE

Components of the Data Store

Data for this project were stored in five separate containers. One SQL database contained all of the case records and look-up tables for case codes, as well as the comprehensive list of ZIP Codes for California, and additional data submitted by OIG. Another SQL database stored the relevant Census tables for the 1990 Census, and a separate SQL database stored the tables for the 2000 Census.

In each of these SQL databases, there are map specific views which dynamically access data from one or more of these SQL databases for a particular grantee or area, according to individual map specifications. The result of a SQL view is a table that includes the calculated data that will be used for mapping. There are also separate Access Databases which stored the geometry for the Census boundaries and base map data using ESRI's geodatabase data model. The primary components of the Data Store are shown in Figure 8.1.

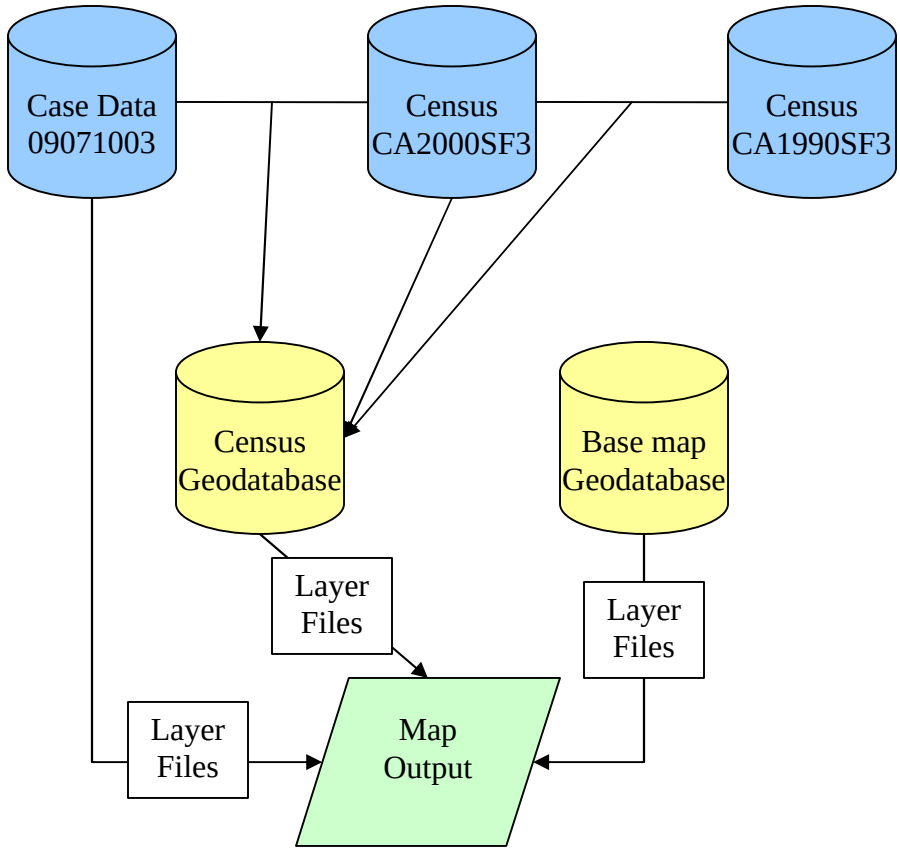


Figure 8.1

Linking Components of the Data Store

ArcGIS Layer files were used to link SQL data views (tables) and the MS Access geodatabase geography (census boundaries) together for mapping. Using ArcCatalog, a database connection was established to the required SQL database. This direct connection to SQL allowed map compilers to add SQL views to a map document as tabular data.

For case points, this tabular data was converted to points using the geocoded latitude and longitude coordinates. Those virtual points (created on the fly from in the specific map document) were then stored as a layer file (.lyr) which could be used in any map document without first loading the SQL view as a table. The layer file stores both the link to the SQL data view as well as the symbol properties (size, shape, and color) for the points.

Choropleth data is created in a similar way. The SQL view containing demographic or case densities, ratios, or values by Census Tract, block group, county, or ZIP code are added to a map document as a table containing the unique ID for each geographic entity. The corresponding geography is also added to the map document from the geodatabase, and also contains the same unique identifier for each entity.

The SQL data is joined to the geodatabase geography, and the entities (tracts, counties, etc) are given the proper legend symbols (ratio, density, etc). This is also saved as a layer file, which

stores the source of the geography, the SQL data, the join properties, and the symbol properties. All of these are retained and displayed in an ArcGIS map document which references the layer file. In this way, map specific layer files were created for each map produced.

The Case Data

The case data were submitted by each grantee to the geocoding contractor, who made efforts to standardize as much of the case data as possible (problem codes, ethnicity codes, closure codes, etc) and combine all case records into one table as described above. Tables to explain the codes (look-up tables) were also provided.

The tables and views created for the case record SQL database are shown in Appendix C.

The Census Data

Raw Census data segment files were downloaded from the U.S. Census Bureau's website. Only the segments that contained data required for mapping were obtained; these Census data repositories are not comprehensive stores of all data available. For example, the Census tables for California are so extensive that not only are the tables distributed by segment, but also the geographic extent is broken into segments.

There are two Census SQL repositories; one for 2000 Census data, and another for 1990 Census data. The 1990 Census data was only used for change maps. The tract relationship files that allow comparison from 1990 to 2000 were downloaded separately and stored in the 1990 Census data repository, as change was mapped based on the 1990 Census Tract geometry. All SQL views were created in the 2000 Census database, even if they referenced the 1990 Census data.

The tables and views created for the Census 1990 and 2000 SF3 data are shown in Appendix C.

The Geodatabase Mapping Data

Map geometries (county boundaries, roads, Census boundaries, etc) used for this project were stored in ESRI's Geodatabase format, which is a modified Microsoft Access database format (extension .mdb). These data were obtained in various formats from a number of different sources, including Geographic Data Technologies (GDT) and ESRI's Data and Maps publication.

Two geodatabases were created in this effort- one for general base map data and another for the geographies (Census and ZIP Codes) used in the geocoding process. See Appendix C for a listing of all feature data classes in those geodatabases.