

The Prototype would not have been usable without specific data sets to support each application. Since creating new data sets was unrealistic within the scope of the project, data was extracted from production systems, scrubbed, and integrated in the Prototype so its functions and features could be used in ways similar to real world uses. The data sources used were not necessarily the best authoritative source for a production system. They were chosen to provide enough appropriate data to support the Prototype.

Since the scope of the Prototype could not support data for the entire state, each application was populated with data about county and municipal governments from the following 15 NYS Counties: Monroe, Niagara, Ontario, Broome, Cayuga, Cortland, Jefferson, Clinton, Essex, Albany, Saratoga, Schenectady, Washington, Ulster, Westchester. Data about three NYS agencies (NYS Department of Agriculture and Markets, NYS Office of the State Comptroller, and NYS Office of Real Property Services) was also included. All data in the Prototype was specific to one or more of the applications. Applications and corresponding data sources are shown in Table 3.

Table 3. Data Sources by Application

Data Sources and Limitations

Application	Data Source	Notes
Overall Gateway	<ul style="list-style-type: none"> • CTG • CGI • Keane 	<ul style="list-style-type: none"> • All user role information was provided and validated by CTG • Links for the resources section was gathered, categorized, and summarized by CTG • Frequently Asked Questions were developed by CTG • Help was written by CGI Information Systems & Management Consultants, Inc. and Keane, Inc.
Contact Repository Application	<ul style="list-style-type: none"> • NYS Department of Agriculture and Markets • NYS Office of the State Comptroller • NYS Office of Real Property Services 	<ul style="list-style-type: none"> • Contact information for local jurisdictions were obtained from the three state agencies. Not every official from each of the jurisdictions was populated in the Prototype. • Contact information for state government officials was obtained by the NYS Office of the State Comptroller. • All contact information is the most updated version.
Dog Licensing Application	<ul style="list-style-type: none"> • NYS Department of Agriculture and Markets 	<ul style="list-style-type: none"> • There are approximately 150 records for each municipality within 15 counties in NYS. • All records were randomly chosen from the years 1999 to October 2003.
Parcel Transfer Verification Check Application	<ul style="list-style-type: none"> • NYS Office of Real Property Services 	<ul style="list-style-type: none"> • Only Counties that use SalesNet were eligible to have data run through the Prototype. Of those Counties, four were chosen: Clinton, Niagara, Cortland, Broome. • The data was supplied by the NYS Office of Real Property Services for these four municipalities within the time range of March 1, 2003 and August 31, 2003. • There were approximately 300-500 records per County populated in the Prototype. • SalesNet extracts for the dates between September 1, 2003 and

		October 31, 2003 were sent to the Prototype from the counties during the field test.
--	--	--

To be usable by the Prototype, all the data sets needed to go through at least one of four transitions:

- **migration** – one-time move from one system to another,
- **integration** – of multiple data sources into a single set,
- **cleaned** – scrubbed for inconsistencies. or
- **re-creation** – new data set created with new business rules

As suggested in the transitions listed above, data sets are not neutral. They contain attributes and qualities that affect their validity and value. Therefore, in preparing the data sets for use in the Prototype, the development team needed to ask some fundamental questions of the data providers:

- How was the data collected?
- How was it managed?
- What do each of the data fields mean and how do they relate to one another?

Once the answers to these questions were understood, a new set of questions arose:

- How will the data be used in the prototype?
- How can the existing data fields be mapped into the new structure?

From here, solutions were developed that took the existing data sets and transformed them into a format and structure directly usable by the Prototype databases (migration, integration, improvement, re-creation).

As seen in the steps above, the Prototype Team and the Corporate Partners addressed all the traditional data issues such as:

- "dirty data," (e.g. inaccurate, duplicated, conflicting, or improperly defined),
- moving data from several sources into a centralized, relational structure,
- accounting for historical features and tracking over time, and
- incorporating new data fields that are not in the current sources but extend the usefulness of the data (e.g., email addresses for dog licenses).