Application of Lessons Learned in the Illinois EDC Project Toward Emerging Election Data Standards and System Guidelines.

Scott Hilkert, Managing Principal of Catalyst Consulting Group, Inc.

Oct 23, 2009; Vers.1.1.
Sept. 8, 2009; Vers.1.0.

Introduction

This brief paper seeks to convey lessons learned from a recent election data collection project conducted in Illinois that may be of interest to those interested in designing universal election data and system standards. The Illinois Election Data Collection (EDC) project was sponsored by an Election Assistance Commission (EAC) Grant which, among other objectives, sought to spark innovation in the automatic collection of precinct level election results data. The conclusions drawn from the Illinois EDC project apply to a narrow but critical subset of election data. The first section of this paper attempts to identify the full range of distinct systems commonly used to execute elections and the data paths in between each system so that the focus of the Illinois EDC project can be put into a broader context. This section can also serve as a primer for those attempting to understand the processes affected by election data formats and standards.

The second section describes specific challenges faced during the Illinois EDC project and techniques devised for overcoming those challenges. These same challenges would be faced by anyone attempting to produce statewide or nationwide election data repositories by merging and converting data from locally managed election systems. The most prominent lesson learned was that the lack of consistency in data and system use between election jurisdictions prevented true automatic data collection without manual intervention. This section also outlines some potential remedies for the challenges identified. This paper does not seek to promote one single data standard, but instead identifies new election data management processes that will complement whatever data standard emerges.

Election Systems and Data Flow.

Figure 1 below depicts the broad range of individual election subsystems and functions that are used in the management and execution of elections in numerous states. The black arrows represent flows of data between systems. In smaller local jurisdictions or in states with a central statewide "top-down" election management system, many of the subsystems shown are combined into a few all-inclusive software packages. In large jurisdictions and in states with integrated "bottom-up" election management systems, these subsystems are often provided by separate products or custom built solutions tailored to unique local election management processes. The dotted line down the middle (labeled "9") demarks a system boundary that exists almost universally in all election jurisdictions. The systems on the left typically have an open architecture and involve data relating to individual voters. The systems on the right have a closed architecture and deliberately avoid tracking individual voters storing only anonymous vote counts and aggregated data. The unavoidable existence of system boundaries and the need for common data to exist one each side provides an opportunity to expand product interoperability and integration options through the emergence of a uniform national election data standard. In addition to the system boundaries shown, new functions and subsystems such as provisional ballot tracking, on-line voter registration, and election auditing controls are constantly being invented as election laws and technologies change.

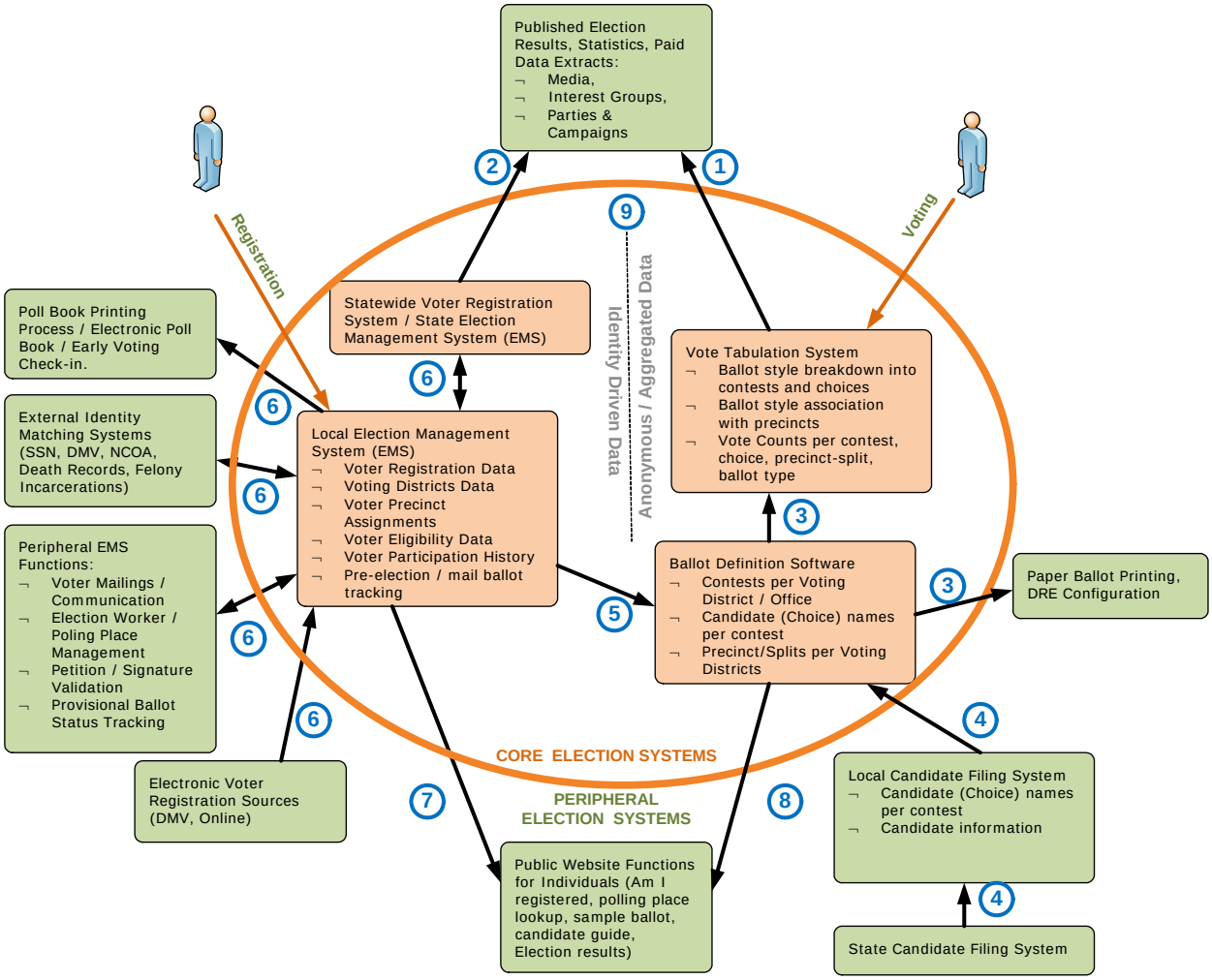


Figure 1: Common Election Subsystems and Flows of Data

The following list describes the numbered data paths depicted as black arrows in Figure 1 above:

1 Data exports from vote tabulation systems. This data is principally used to describe the detailed results of an election. It consists of elemental counts of votes organized by election, state, local jurisdiction, party (only relevant in primaries,) contest, choice (includes candidate, yes or no on issues and judicial retention, special non-choices such as overvotes and undervotes) precinct, and ballot type (includes polling place, early, grace, by mail, provisional) Commercial vote tabulation systems typically offer a wide range of standard and configurable export data file formats to cover this function.

2 Data exported from voter registration or election management systems. This data is often purchased from the election authority and serves a range of purposes such as campaign and interest group contact management, redistricting analysis, and jury selection. The data typically includes records on individual voters, their demographic information, their contact information, their election participation history, their assigned voting districts and precincts, and more. Export file formats vary widely and are often proprietary

to the election management system or tailored to specific election management processes in use at state and local levels.

3 Data used to configure vote tabulation systems for each election. This data covers the precincts in the jurisdiction, the contests and choices assigned to each ballot position of each ballot style, the ballot styles used in each precinct, the rules for each contest such as "vote for 1", or "vote for up to 3." This data typically must conform to input file formats prescribed by the tabulation system vendors . In some cases it is manually keyed into the system by a system administrators. This same data is also an input to ballots printing or to configuring DRE (touch screen) voting terminals

4 Data from candidate filings at either the state or local level. It identifies the contests entered and the candidate names as they will appear on the ballot. It often includes the order that candidates will appear on ballots based on statutory rules and rulings. This data must be passed from states to each local jurisdiction for candidates filing for federal and state level contests. States and localities often store filing information in home grown databases and spreadsheets. This data is often exchanged between state and local levels in the form of printed notices and is hand entered into ballot design software with no data standard at all.

5 Data used to configure vote tabulation systems for each election. Unlike item 3, this data generally remains constant between elections and defines voting districts such as congressional, judicial, educational, county, city, township, etc. It defines which precincts and precinct splits intersect these districts and it defines which offices and contests relate to each district. These are among the rare data elements that exist on both sides of the boundary between vote tabulation systems and voter registration systems and is used for different purposes on each side. In many instances, the ballot design function is integrated into the election management system with a shared database so that the more significant system boundary must be crossed by data item 3 above. When maintained in separate systems, data is often copied using improvised export and import files or is hand entered in one or both systems.

6 Data Shared Between Election Management Subsystems. . Many of these subsystems are all part of a single integrated EMS product and all share a common database which means that this data is not exchanged between subsystems at all. When subsystems are separated into separate products, they often communicate between each other using proprietary file exports or more modern service oriented architecture components such as XML web services. One of the most crucial election functions served by this data is the printing of poll books or eligibility lists which are used by election judges to sign in voters and issue ballots.

7 Data Shared Between Public Election Management Subsystems This refers to a subset of data used to serve public voter self-service website functions such as "am I registered?" or "where is my polling place?" or "which government districts do I live in?" For security reasons, publicly accessible websites often use data copies that exclude personal identifying information such as last-4 social security digits. The data copies are kept in a completely separate database from that of the election management system. Data is exchanged between databases using common database transfer techniques or is accessed by the website using XML webservices.

8 Data Shared Between Public Election Management Subsystems Very similar to number 7 except that this data serves a different set of self-service website functions including "what candidates are running in the

districts I am voting in?" When ballot definition and candidate filing functions are managed in separate subsystems, a separate data flow is needed to serve this website function.

9 EMS / VTS System Boundary. This is intended to demonstrate the deliberate absence of a data flow between two core election systems (Election Management and Vote Tabulation) which are often provided by separate products from separate vendors in a single jurisdiction. The Vote Tabulation Systems have intentionally closed architectures and encrypted databases which prevent visibility and manipulation of all data except through the proprietary vote tabulation software. Voter Registration systems (frequently called election management systems) have open architectures and shared databases which are integrated with numerous other systems.

The Illinois EDC Project was entirely focused on data flows 1 & 2 as listed above and as called out in Figure 1. Many of the challenges presented in the following section stem from the fact that data flow #1 had to be merged from exports files taken from the tabulation system in each of Illinois 110 local election jurisdictions. Data flow #2 came from the Illinois Statewide Voter registration system which is completely separate from the 110 tabulation systems. From a technical perspective, the goal of this EAC funded project was to automate and merge data exports from the 111 data sources and produce a single electronic data file that represented relevant election data for the whole state.

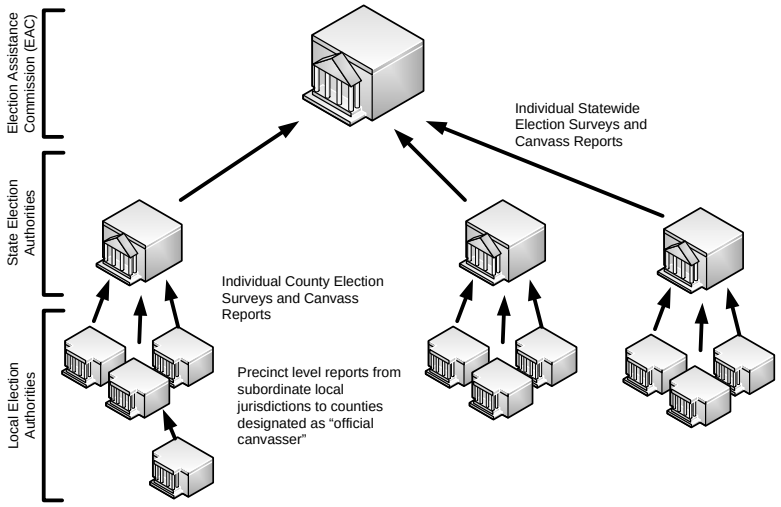


Figure 2: Election Results Reporting Structure

Figure 2 above depicts the reporting hierarchy which the EAC relies on to collect nation-wide election data every two years. The flow of information through this hierarchy is often achieved using improvised spread sheets and mailed paper surveys Illinois and other states have long sought a means of automating this process. The EAC grant provided a means of advancing this goal. Reporting had always been restricted to detail at the county-level. A further goal of the EAC's sponsorship of the Illinois EDC project was for this data to be broken down more granularly by precinct.

Challenges Encountered and Addressed in the Illinois EDC project.

Among the 110 local tabulation systems deployed in Illinois, all four US tabulation system vendors are represented: *Hart Intercivic, ES&S, Premier (recently acquired by ES&S from Diebold,) and Sequoia*. The 4 different tabulation system products each were able to export election results data in a variety of configurable data formats. Hart Intercivic offered a hierarchical XML file format while the others offered delimited flat file formats. Data from all of these 110 system installations had to be merged along with data from the Statewide Voter Registration System to satisfy EAC objectives.

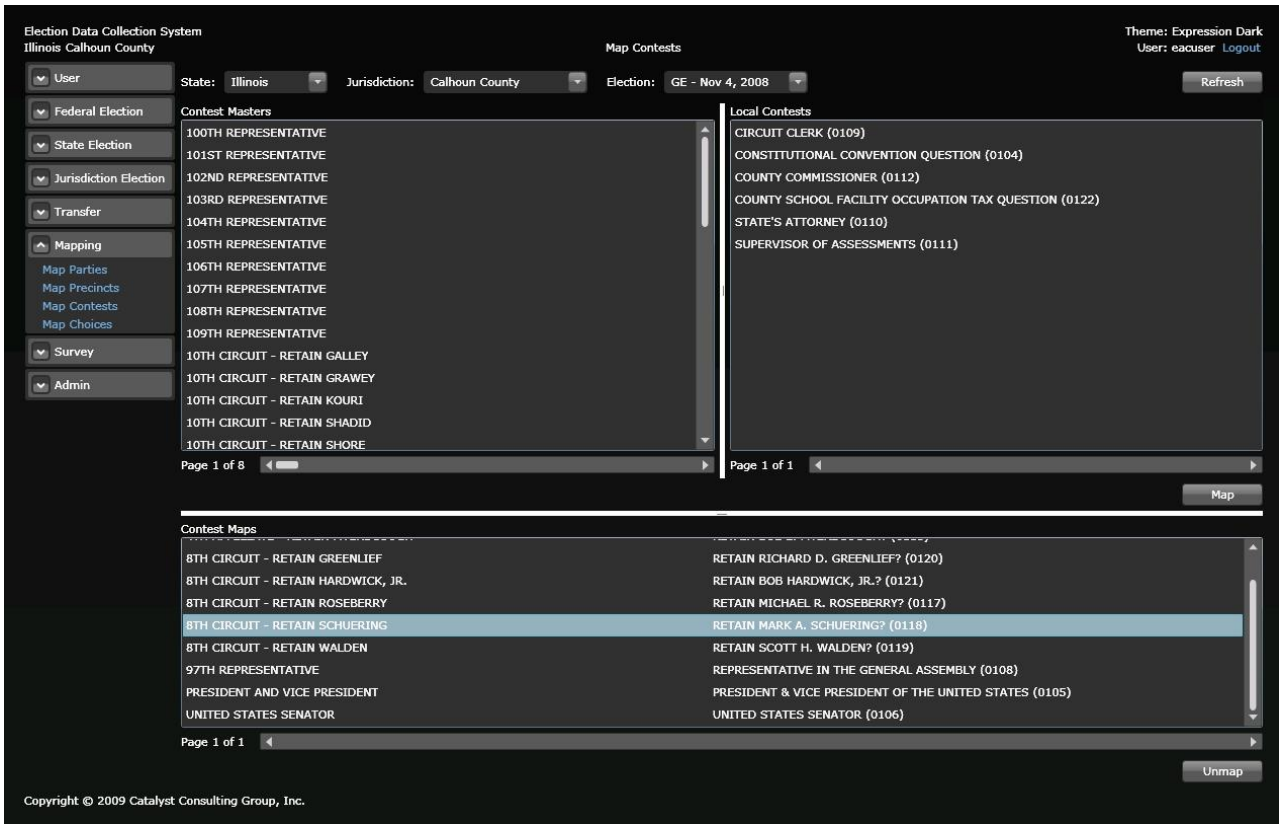
The EAC invited Illinois to invent its own data format for the export of its statewide election data. The outcome was a new XML Schema that most closely captured the hierarchical relationship between vote count attributes (as listed under data flow #1 in the previous section.) For brevity, this data format is not presented here but is available upon request from this author. Again, this paper does not seek to promote a particular data format but instead seeks to present lessons learned that will apply to any election data standard that emerges.

Challenge #1: Inconsistency in Contest Naming.

For each contest in the statewide output file, data had to be combined from the corresponding contest in each local input file (item 1 from Figure 1). In all cases, the contests could only be distinguished by either the contest name or a contest identifier. Because each local jurisdiction programmed its tabulation system independently using data (item 3 from Figure 1) without a standard naming convention, the contest names and identifiers were all inconsistent between counties.

For example, Illinois 17th Congressional District spans 12 whole counties and portions of 2 more counties. Example names for this contest from among theses counties: "US House of Representative," "Representative in the United States Congress," or simply "Congress." These were acceptable and unambiguous for voters in counties which were entirely contained within the congressional district. But without being identified as the 17th district, it was impossible to automatically merge this data into a statewide report without cross-referencing an outside data source containing the jurisdiction-to-congressional district associations. . Even in counties that included the numerical district number, the difference in naming styles ("US House of Representatives, Illinois 17th District" vs. "Representative in the US Congress Dist 17") made it difficult to automatically merge the data without manually applying match/ search rules. In addition to the full contest name, all the data file formats encountered also offered a contest identifier code to help distinguish between contests. But regardless of the vendor system, this code was assigned differently and arbitrarily in each jurisdiction based on the sequence of contests programmed into the local tabulation system. Thus the contest identifier was also unusable as a means for automatically merging the data.

The software developed for the EDC project therefore included a user interface for allowing a data technician to easily create mappings between each county's contest names and a statewide master list of contest names in order to merge the data. A sample screen shot of the software is provided below:



This same approach was also required to map different codes for the same political party (such as "Reform," "Ref," "F") to the same value and for different forms of the same Candidate Names (such as "John Joseph Polachek" and "John J. Polacheck".) These data mappings required over two-hundred analyst hours to complete because fully automatic mapping was not possible.

Potential Remedy: Establish a nation-wide contest naming convention and set of codes or short names which could be used as contest identifiers which would be used by local election authorities when configuring their tabulation systems. This would apply to items 1,3,4,5 in Figure 1. Encourage Tabulation System Vendors to emphasize this list as part of system training. Enforce the use of these conventions through education, regulation, or both. With consistent identifies, the manual data mapping described above would not be required. At the very least, this naming convention could cover Federal offices, but it would have value for State legislative, judicial, or any districts that routinely cross county borders. Some "Federal Information Processing Standards." (FIPS) Codes were published for this purpose but were not commonly used by local election jurisdictions. These codes and were withdrawn in 2008 deferring to the Census Bureau's Geographic Information System standards initiative. Through internet searches, it was not immediately clear what identifier convention will emerge from this body and if it will be suitable for election contest naming.

Challenge #2: Inconsistency in Precinct Naming.

One ancillary objective of the Illinois EDC project was to compare voter participation record counts (part of item 2 in figure 1) from the Statewide Voter Registration System with ballot counts from the local tabulation systems for each precinct. These two counts are derived from separate and independent processes: bar-code scanning ballot requests / poll book pages, vs. tabulation system counts of ballots. When these two counts match for each precinct exactly, it is a strong indicator of high data accuracy. There are legitimate reasons for slight inconsistencies, but when the two counts of a precinct have a deviation significantly larger than the typical precinct deviation, it is an indicator of an error or some abnormality that warrants further investigation.

The challenge encountered in Illinois was that the naming/identifying scheme for precincts in the Voter Registration System was inconsistent with that of the tabulation systems. To solve this, a precinct name mapping interface similar to the one used to handle contest naming inconsistencies was created. Again, this required manual intervention by analysts and prevented automatic data gathering. Example: A precinct name in the tabulation system output file was "Northfield 7" while the VR System used "5007" to represent the same precinct. Without using an external list to know that the county used the prefix digits "50" to represent Northfield, it was impossible to automatically reconcile the names and perform the two-source comparison.

Potential Remedy: Prescribe a national precinct naming identification standard and educate election authorities on how to implement it. Encourage system vendors to include a "National Precinct Code" field in their database schema that can be mapped to the local jurisdiction precinct names. Encourage vendors to promote this feature as part of their standard training programs.

Potential Benefit: Voter participation records and other data from voter registration systems should be included in the scope of a national standard as this provides a separately derived point of comparison for data gathered through the tabulation systems.

Challenge #3: Small Discrepancies Between Machine Counted Data and Official Canvass Reports.

Normally, local jurisdictions certify their official "canvass" of election results in the form of a readable report of total vote counts per choice, per contest. These canvass reports are not produced in a prescribed format and most do not break down results by ballot type. States collect these local reports and manually combine results for federal and statewide offices to produce and certify the official state canvass.

An ancillary goal of the Illinois EDC project was to determine if the Tabulation System export files collected from each local election jurisdiction could serve as the basis for automatic electronic canvassing. The finding was that at least a third of the jurisdictions had small (On order of 1 vote per precinct) deviations between the counts from their tabulation system export files and the counts published in their official canvass report. It was beyond the scope of the project to investigate every discrepancy, but informal inquiries revealed 2 general causes:

1. When provisional or other disputed ballots were counted, their results were not programmed into the tabulation system but instead were manually added to the official canvass report afterwards;

2. The date/time of export of the tabulation system file was not exactly known and the export was apparently performed before disputed, absentee, or provisional ballots were input into the tabulation system.

Furthermore, when the tabulation system export files were sought from local jurisdictions 6 weeks after the election, it was no longer available in some cases. Three jurisdictions cited hard drive failures or other technical difficulties which occurred after they had reported their official canvass. These data losses would be especially troubling should an audit or recount be requested due to a close race or suspicion of fraud.

In order to have a completed statewide data file that matched the official canvass reports from the local jurisdictions, manual adjustments had to be inserted into the underlying database used to generate the EDC's statewide output file. A user interface was developed to allow analysts to enter offsets at the precinct or contest level. The original machine count was not overwritten, instead an offset record was created that included the analyst's identity, the date/time of the adjustment, a reason code, and a comment field.

Correction of inconsistencies is not the only use for an adjustment component to the data standard. Although rare, there is precedence for some courts to rule that official vote counts must be modified for a candidate. (Of interest in developing a universal election data standard: these court-ordered adjustments are sometimes based on non-integer constants multiplied by numbers of registered voters and can result in non-integer official vote totals.)

Potential Remedy: Any national data standard should include a component for capturing adjustments to vote counts without overwriting the original machine count. Implement national or state level repositories to which jurisdictions can upload their tabulation system output files in one or more pre-determined formats. Through regulation and/or education, encourage all jurisdictions to use this at the time of results certification. The repository would include a user interface for merging the files into a national or state level election results database and include an interface for entering official offsets to the machine counts.

Not a Challenge for Illinois EDC: Lack of Standard Data Format

In comparison to the above, it was not a significant challenge that the four tabulation system types encountered in Illinois required handling four different export data file formats. Analysis of these file formats and the creation of four data conversion modules to translate them was a trivial matter due to the wide availability of middleware tools and application frameworks designed to manage data from diverse sources. This was also possible because the key data elements required were present in all four systems.

Conclusion

Additional Challenges were encountered and addressed during the Illinois EDC project but have been omitted from this paper for the sake of brevity. The common theme emerging from these additional challenges and the challenges described in this report was an underlying inconsistent usage of tabulation system features by local election jurisdictions.

Recap of Potential Remedies to Challenges Identified in the Illinois EDC Project:

- Beyond a national standard data format, national standard naming conventions and identifiers for election contests, precincts, and other data elements are needed.
- Voter participation records and other data from voter registration systems should be included in the scope of a national standard as this provides a separately derived point of comparison for data gathered through the tabulation systems.
- Regulation and/or education of local election officials should be promoted to standardize the way tabulation systems are used.
- Any national data standard should allow for the tracking of manual adjustments made to the original machine counted data.
- Regulation and/or education of local election officials should promote the preservation of data exported from these systems. A national repository or separate state repositories designed to receive uploaded tabulation system exports from local election authorities would serve this goal.

Any of these remedies or merely the consideration of lessons learned from the Illinois EDC project in developing a national data standard would advance the ultimate objective of accurate, timely, detailed, and multijurisdictional reporting of results and validation of election processes.

About the author:
Scott Hilkert is a Managing Principal of Catalyst Consulting Group and served as the solution architect for both the Illinois Statewide Voter Registration System and the Illinois Election Data Capture Project sponsored by the EAC. He is currently serving as Catalyst's project director for the California Statewide Voter Registration System.