Comments on the CA SoS' Precinct Level Data Pilot Project

Joseph Lorenzo Hall[1,2], Philip B. Stark[3], Henry E. Brady[4,5], and Jasjeet S. Sekhon[4]

[1]University of California, Berkeley; School of Information
[2]Princeton University; Center for Information Technology Policy
[3]University of California, Berkeley; Department of Statistics
[4]University of California, Berkeley; Department of Political Science
[5]University of California, Berkeley; Goldman School of Public Policy

November 16, 2009

Thank you for inviting comments on your Precinct Level Data Pilot Project, which seeks to provide precinct-level vote tabulation data to the public. We applaud Secretary Bowen's pilot program. Timely precinct-level election results from California counties are crucial for establishing the integrity of California's elections, for supporting analyses of election results and for designing and conducting post-election vote-tabulation audits.

We have examined the sample data from the four counties—Orange County, Sacramento County, San Francisco County and San Luis Obispo County—that provided data for the Pilot Project.[1] We submit these comments in the hope that you find them helpful as the Pilot Project goes forward.

Summary The data would be far more useful if they were reported in a uniform, structured, standard format. The sample formats for the four counties are all decipherable, with varying degrees of effort. However, each has its own format, its own labeling of precincts, contests, candidates, and type of vote. It is not possible to parse the formats automatically to combine results across counties. Because the formats differ, dealing with the data from each county is a separate programming task.

Consider what would be involved in getting the results for a statewide contest, or a contest that spanned even a few counties. One would have to figure out which counties had precincts in the contest, how each of those counties named the contest, how each of those counties identified the candidates, how those counties had formatted the data, etc., to find the relevant subtotals. A uniform, structured, standard format would eliminate the need for each of these steps.

The focus of this initial period of the pilot project should be to ensure that counties are able to report all the correct data elements, and at the appropriate level of granularity, desirable for precinct-level data. Then, the Secretary's technical staff, in conjunction with interested parties like ourselves, can work to create standardized structured data formats from what the counties can provide.

Uniform Formats There should be a common understanding of the definition and naming for each data field. For example, it is difficult to tell if Orange County's `Precint ID` or `Precinct Name` correspond to Sacramento's `Precinct Code` or the unnamed precinct numberings used by San Francisco and San Luis Obispo. Uniformly named fields whose contents are uniformly defined are essential to any state-wide or cross-county analyses.

Data uniformity requirements should include:

[1]See: `http://www.sos.ca.gov/elections/sov/2009-special/precinct-data/index.htm`

1. uniform labeling of contests that cross county lines;
2. uniform naming/identification of candidates in those contests;
3. use of unique precinct identifiers statewide, with a separate statewide table to relate those identifiers to geography;
4. uniform naming of various vote types (for a candidate, yes, no, undervotes, overvotes, etc.); and,
5. if the data are to be reported as `.txt` or `.csv` files, uniform column ordering and labeling (the use of OASIS/EML would obviate the need for this requirement).

Structured Formats Only two counties, Sacramento and San Luis Obispo, provided a data dictionary or data description with their data. In addition to the data uniformity we discuss in the last section, it is essential that the data is structured. That is to say, human involvement in processing the data should not be necessary; the data should be structured in such a way that computer programs and scripts can process the data.

None of the four formats is self-documenting and none of the formats is structured. That said, there are important differences among each county's sample format.

Orange County's format is easy to decipher; uniformly naming precincts, contests, candidates, and vote types are perhaps the biggest issues. Sacramento County's fixed-width format is hard to parse. Output such that the character position in a line determines the meaning of the character was useful in the past—e.g., for mainframe computers—where data storage was at a premium; it would help immensely if Sacramento used commas to delimit fields.[2] The San Francisco County format is the worst of the four. The file has extra lines, summaries, etc., that would require hand editing to make it possible to extract and process the data electronically. The format would be more robust if quotes or other delimiters were used for text strings (but that is not the most serious problem). These data are not really machine readable for analytical purposes. San Luis Obispo County's format is comparable to Orange County's. Again, standardizing the naming of precincts, contests, candidates, and vote types is a serious issue.

Standardized Formats The pilot project data files from these four counties appears in four different formats: three versions of Comma-Separated-Value (CSV) formats (Orange, San Francisco and San Luis Obispo) and one in a fixed-width format (Sacramento). The data files should be available both in the version received from the counties and in a standard file format that is the same for each county.

While CSV files with consistent, structured rows and columns along with a separate data dictionary might solve each of these problems, we'd like to suggest that XML data, possibly in the form of Election Markup Language (EML), would serve all these needs simultaneously.

We hope that these shortcomings can be addressed to produce useful, uniform statewide data for California elections.

[2] In fact, in our experience, Sacramento's use of this "fixed-width format" caused at least one error in parsing their data into fields. That is, if one opens their sample file in a spreadsheet program like OpenOffice or MS Excel and uses the provided data dictionary's table of field lengths to parse each line into separate data fields, the resulting data is mis-parsed. Specifically, we suspect that one of the `Party Code` or `District Type ID` field is actually 2 characters wide instead of the reported 3 characters. If both of those fields *are* 3 characters wide, the fields that follow these two fields are not parsed correctly. For example, the `Contest Title` field which appears to map to `REGISTERED VOTERS - TOTAL` becomes "`R`" at the end of the previous field and "`EGISTERED VOTERS - TOTAL`" as the contents of the `Contest Title` field. For a version of the Sacramento data with the `Party Code` field parsed as a 2-character field instead of a 3-character field, see the files here: `http://josephhall.org/casos/pldpp/`.