Multi-level Information Modeling and Preservation of eGOV Data*

Richard Marciano[1], Bertram Ludäscher[1], Ilya Zaslavsky[1],
Reagan Moore[1], and Keith Pezzoli[2]

University of California San Diego, 9500 Gilman Drive, San Diego, CA 92093 USA
[1] San Diego Supercomputer Center, MC 0505
{marciano,ludaesch,zaslavsk,moore}@sdsc.edu,
http://www.sdsc.edu
[2] Urban Studies and Planning Program, MC 0517
kpezzoli@sdsc.edu, http://regionalworkbench.org

Abstract. This paper addresses the issue of long-term preservation of and access to digital government information. We show how the preservation process is enhanced by storing an infrastructure-independent representation of the raw data, together with a model dependency graph (an executable graph of database view mappings). This allows for the design of decision-support tools and services for improving government transparency and promoting citizen access to eGOV data. A case-study, the *Florida Ballots Project*, is used to illustrate the approach.

1 Introduction and Approach

A common demand is that e-Government services promote citizen access to government information [1], such as official records kept at an archival institution [3]. Today, thanks to the ubiquitous Web, access to digital data is often less of a problem than actual information content. We argue that a multi-level information or "deep" modeling approach combined with an appropriate infrastructure independent representation mechanism can greatly enhance the value of eGOV data to the interested public, future researchers, and "digital archeologists/historians". We use the 2000 U.S. Presidential Election as an example of the deep modeling approach.

"On behalf of the State Elections Canvassing Commission and in accordance with the laws of the State of Florida, I hereby declare Governor George W. Bush the winner of Florida's 25 Electoral Votes," said Florida's Secretary of State, Katherine Harris, as she certified George W. Bush the winner over Al Gore, on November 26, 2000. The National Archives and Records Administration (NARA), went on to record this 25-Vote result by collecting two documents for permanent retention:

- *Certificate of Ascertainment,* containing the proposed Electors: http://www.nara.gov/fedreg/elctcoll/2000/certafl.html

* Work partially supported by NSF/NPACI ACI-9619020 award (National Archives and Records Administration / NARA supplement) and National Historical Publications and Records Commission / NHPRC award ("Methodologies for Preservation and Access of Software-dependent Electronic Records").

R. Traunmüller and K. Lenk (Eds.): EGOV 2002, LNCS 2456, pp. 93–100, 2002.

- *Certificate of Vote*, capturing the winning Electors: http://www.nara.gov/fedreg/elctcoll/2000/certvfl.html

More recently, two election media studies, started rethinking the entire process:

(1) *USA Today / the Miami Herald* on April 4, 2001, http://www.cnn.com/2001/ALLPOLITICS/04/04/florida.recount.01/

(2) The NORC *Florida Ballots Project*[1], on November 12, 2001, http://www.norc.org/fl, the results of which we use in our case study.

These studies present parameters under which either candidate could have won.[2] They suggest that with the 25 Votes, one should consider the retention of a parameter space that captures greater context. Fig. 1 depicts a model dependency graph we derived from examining NORC, and tries to formally define such a suitable parameter space as an example of our "deep modeling" approach.

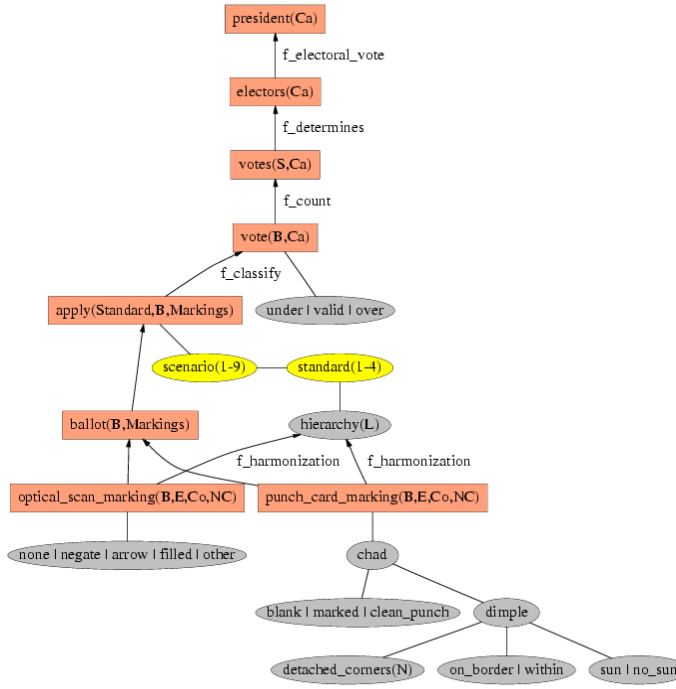


Fig. 1. 2000 Presidential Florida Election Model Dependency Graph, derived from NORC.

[1] At NORC, the National Opinion Research Center, at the University of Chicago, a consortium of news organizations including: the New York Times, the Wall Street Journal, the Washington Post Co. (The Washington Post, Newsweek), Tribune Publishing (LA Times, Chicago Tribune, etc.), CNN, Associated Press, and others. While the first study only looked at the ***Undervote*** (ballots rejected because no vote was recorded for the president), the second study looked at all 180,000 uncertified ballots in Florida's 67 counties, including the ***Overvote*** (ballots rejected because more than one presidential candidate was selected).

[2] Interactive analysis of the USA Today study at: http://usatoday.com/graphics/news/gra/gvote/frame.htm.

2 Multi-level Model Dependency Graphs

For several reasons, the depth of information modeling that corresponds to news reports and even official archival records is quite limited (e.g., the top three nodes in Fig.1), with NORC being a notable exception. One of the findings was that, depending on the specific scenario or applied standards, different outcomes can result.

The multiple possible outcomes can be made precise through "deep modeling" using a *graph of database mappings* as follows. The graph in Fig. 1 is an abstraction of such a graph (i.e., a network of "views" in the database sense). A view is a relational table that is defined by a query expression. Note that views can be layered and defined on top of other views, resulting in a graph of mappings. The overall graph itself defines a (complex) view, mapping the ("raw") input data to the final result (the President Elect in Fig.1). In general, a deep modeling approach using database views comprises:

1. ***relational schemas*** for all relevant entities and relationships (in the figure: parameterized entities and attributes)
2. ***view definitions*** (= database queries) precisely defining the mappings from one schema to another
3. ***constraints*** (logic formulas) over the relational schemas (to express, e.g., which standards can be applied to which ballot type)

Thus, in the actual graph, nodes correspond to relational views defined on top of other views or base tables. In our abstraction in Fig.1, nodes stand for *parameterized entities* (boxes) and *attributes* (ovals), while directed edges denote *database mappings*, i.e., functions between relational schemas.

Together with the raw data, the graph of database mappings can then be *executed* as a (complex) database query with a *verifiable* and non-controversial output. Of course this does not prevent a political controversy from happening, but it can be dealt with at a less superficial and more informed level: In Fig. 1, the *scenario* (see *Appendix E*) and/or the specific *standard* (see *Appendix D*) being applied to specific sets of ballots, uniquely determine the database tuples in the views above; in particular, the topmost tuple, i.e., which president should be named president elect. Thus, the only degree of freedom and non-determinism that such a graph of mappings allows is in the input data, in this case, the raw ballot data and the scenario/standard to apply.

NORC did everything to guarantee that the raw data was as objective as possible – in particular the coders did *not* compute the function *f_classify* themselves, i.e., they did not determine the votes. Instead every coder just described the markings seen and the *f_classify* determines the vote (under/valid/over) as a function of the standard and the markings on the ballot (see *Appendix B*). The crux is that those functions can be expressed and implemented as *database queries.* For example, the edge:

f_electoral_vote: *electors(S,Ca)* $\rightarrow$ *president(Ca)*

means that whether candidate *Ca* is elected president is a function (called "electoral vote") of the electors (of all States *S*) of *Ca.* The latter is itself a function of ***votes****(S,Ca)*, i.e., the votes that candidate *Ca* received in state *S.* Clearly, given the corresponding relational tables, the result of *president(Ca)* or *electors(Ca)* can be represented as a database query on a table representing the votes per state and candidate (=*votes(S,Ca)*).

Some citizens may be interested in the top-most node only: who is the president elect? Others may choose to study the reports from the news agencies and study how many electors each candidate could win or how many certified votes per state each candidate had. The extremely close outcome of the presidential race (the differences in votes between candidates were below the statistical error margin in the state of Florida – the state which ultimately determined the election, 271 to 266 Electoral Votes for Gore) sparked an enormous controversy about the official outcome of the election and almost led to a constitutional crisis.

As a result NORC conducted a thorough study aimed at resolving the issues. Translated in our framework, this means that one can resolve the controversial issues in a precise and for the interested citizen, verifiable way (depending on the available raw data of course). The model dependency graph shows that at the lowest level, the raw data consists, e.g., of *optical_scan_markings* and *punch_card_markings*. For example, *B,E,Co,NC* means that on the ballot with identifier *B*, the coder *Co* has described the element *E* (e.g., a specific chad or specific box for a candidate) to have a marking NC (=NORC Code, e.g., *dimpled chad with two detached corners*). In the graph, the node *ballot(B,Markings)* then provides a convenient way to represent the information: the ballot *B* with all of its markings (including, for each element *E* and each coder *Co*, the observed markings encoded as *NC*). One point of the controversy was which *standard* should be applied to determine the intent of the voter.[3] Depending on the county or even precinct, and the type of ballot (see *Appendix A*) different standards could be applied. For convenience, NORC created *scenarios*, where each scenario explicitly states which standard is applied to which set of ballots. The markings coming from different ballot types (optical and punch card) were "harmonized" (see *Appendix C*) so that one could easily express standards even if applied *across* different ballot types. This "harmonization mapping" was itself documented but was added only as another convenience: one can still apply standards directly to the markings of a ballot (but without harmonization one needs to do this for each ballot type individually).

Thus, under the assumption that the raw data is uncontroversial, by using a graph of mappings, the dispute can be localized to the specific choice of scenarios/standards being applied. In this way, transparency and verifiability of the process can be guaranteed for every interested citizen.

3 Preservation Issues

The modeling of the study as a network of database transformations also has advantages for the preservation information. The NORC study provided all raw data online and precise descriptions of the mappings as part of the accompanying documentation. Moreover, a "Scenario Manager" (see Fig. 2) has been developed that allows the user to inspect the outcome of applying different scenarios. From an archival point of you, however, the specific choice of system (Microsoft Access)

[3] Of course if other more *reliable* technologies were available that would lead to unambiguous voting results and avoid the discussion about which standard to apply – however, this is not the point here: even if this specific controversy was caused by an anachronistic voting system, many other eGOV data issues (e.g., redistricting) will always present a "deep modeling challenge".

introduces an infrastructure dependency: *"Will a researcher be able to evaluate and experiment with the study 5, 10, or 50 years from now?"*[4]

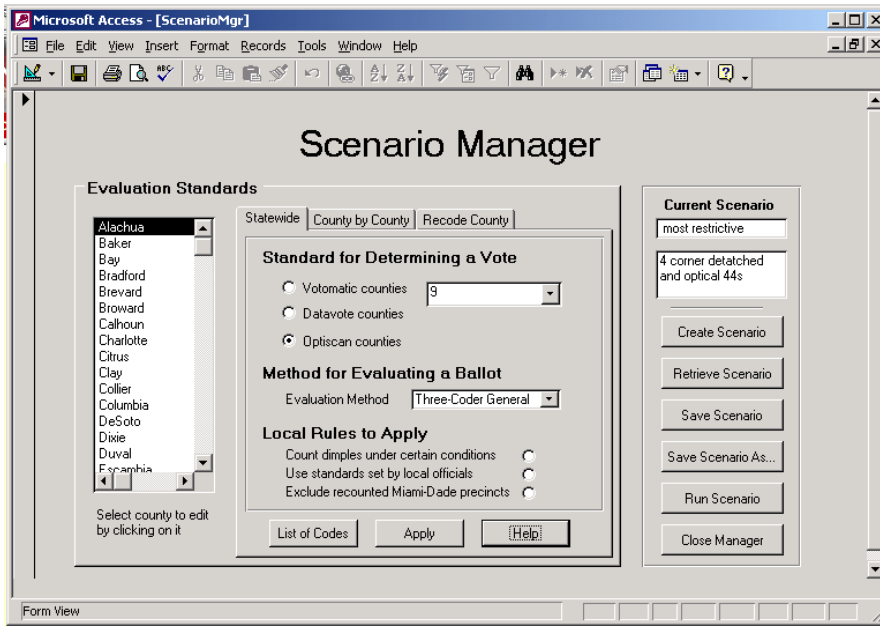


Fig. 2. Snapshot of a NORC tool created by Elliot Jaspin (Cox Newspapers)

A more robust and generic solution is to explicitly model the functionality of the Scenario Manager in an infrastructure independent way. First, mappings in a standard such as SQL. In this way, an "archival package" can contain the raw data together with all mappings and can be run on any future SQL engine, provided the standard is still supported.

One can go even one step further and create a *self-instantiating, self-validating archive* [2]. As in the case of SQL, we store an *executable* version of the constraints and mappings of a scenario. Moreover, we also package the *execution engine itself.* If the engine is SQL, this may not be a viable solution. Instead, for our running example, one could express all mappings as logic programs (Datalog/Prolog queries) and archive the complete execution engine (some complete Prolog system are smaller than the raw data of the NORC study) as part of the archive. Assuming that the logic engine is implemented in an infrastructure independent way (e.g., a Prolog engine in Java Byte Code), the complete analysis can be unrolled in the future by instantiating the graph of database mappings, and validating its integrity constraints. If the mappings satisfy certain properties, the analysis can in fact be *reversed*, i.e., one could try to solve the inverse problem and ask under which scenarios/standards a specific outcome is obtained.

[4] In fact, the problem occurs today: the Scenario Manager crashed several times during our experiments.

4 Conclusion

Multi-level (or "deep") information modeling provides a mechanism for capturing process information in a formal and unambiguous way as a network of database transformations. The characterization of the modeling process itself leads to the notion of self-instantiating, self-validating archives [2].

References

1. Cowell, E., Jacobs, J., Peterson, K.: Government Documents at the Crossroads. American Libraries. Infotrac. (Sep. 2001), 52–55
2. Ludäscher, B., Marciano, R., Moore, R.: Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-based Archives, ACM SIGMOD Record, Vol. 30, No. 3 (2001) 54–63.
3. Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., Gupta, A.: Collection-based Persistent Digital Archives, D-Lib Magazine Vol. 6, No. 3 & 4, (Mar. 2000, Apr. 2000)

Appendices

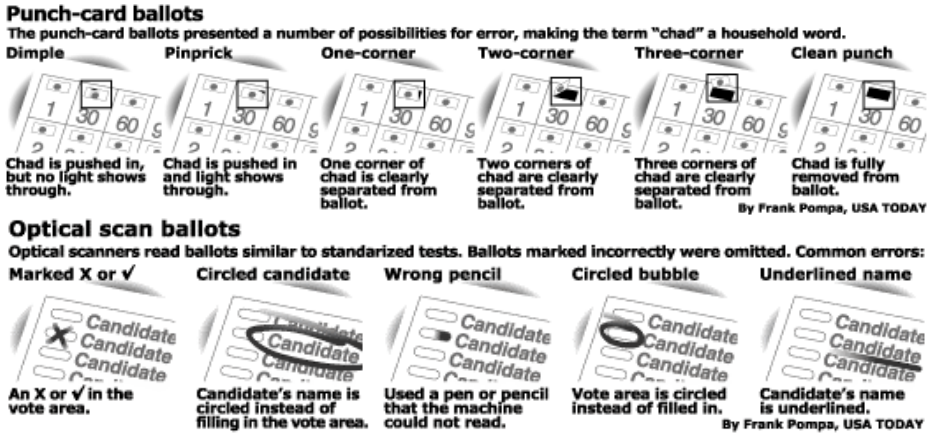
Appendix A: Types of Ballots

Logically, we distinguish between two types of ballots: (1) **punch-card** ballots (*Votamatic* and *Datavote*), and (2) **optical-scan** ballots. However, there were really 5 types of voting systems in use in Florida. **Votomatic**, where a hand-held stylus was used to punch the pre-scored paper or *chad*, **Datavote**, where voters use a mechanical punching machine, **Optical Scan**, where ovals are filled in, or arrows connected, **Lever**, where the *Datavote* process was followed, and **Paper**, where the *Optical Scan* process was followed.

Voting system	***Number of counties***	***Undervotes***	***Overvotes***	***Total number of uncertified ballots***
Votomatic	15	53,215	84,822	138,037
Datavote	9	771	4,427	5,198
Lever	1			
Optical Scan	41	7,204	24,571	31,775
Paper	1			
Total	**67**	**61,190**	**113,820**	**175,010**

Appendix B: Visual Markings and NORC Codes Used

Most of the uncertified ballots were due to *overvotes*, and most of the problems came from the *punch-card* ballots. It is easy to see why, by looking at the following animation of a *Votomatic* voting machine (Doug Jones, University of Iowa):
http://www.cs.uiowa.edu/~jones/voting/votomat/animate.html



Examples of Undervote ballots for punch-card and optical-scan from the USA Today study.

NORC Codes used to classify each mark on **punch-card** and **optical-scan** ballots.

Punch-card ballots		**Optical-scan ballots**	
Code	**Meaning**	**Code**	**Meaning**
0	blank, no mark seen	00	blank, no mark seen
1	1-corner of chad detached	11	circled party name
2	2-corners of chad detached	12	other mark on or near party name
3	3-corders of chad detached	21	circled candidate name
4	4-corners of chad detached, clean punch	22	other mark on or near candidate name
5	dimpled chad, no sunlight	31	arrow/oval mark other than fill: circle, x, /, check, scribble
6	dimpled chad, with sunlight	32	other mark near oval/arrow
7	dimple within chad area, off chad, with or without sunlight	44	arrow/oval filled
8	dimple on border of chad area, with or without sunlight	88	arrow/oval filled or marked other than fill, then erased or partially erased
9	chad marked with pencil or pen	99	negated mark: scribble-through, cross-out, "NO", and similar

Appendix C: Harmonized Codes

Equivalence Classes	**NORC Codes**	
	Punch-Card	**Optical-Scan**
0	0	00 / 99 / 88
1	8	
2	7	
3	5	11 / 12 / 21 / 22 / 31 / 32
4	6	
5	1	
6	2	
7	3	
8	4	44
9	9	

Appendix D: Standards

Standards to specify evidence of voter intent.

Standards	Equivalence Class Codes	
	Punch-Card	Optical-Scan
1. Dimple or better	>= 3	>= 3
2. One-corner detached	>= 5	>= 3
3. Two-corner detached	>= 6	>= 3
4. Dimple (if rest of ballot is dimpled)	(>= 6) && (3, 4, 5)	>= 3

Appendix E: Scenarios

Scenarios
1. Prevailing statewide standard
2. Supreme Court "simple"
3. Supreme Court "complex"
4. 67-county custom standards
5. Two-corners-detached statewide
6. "Most inclusive" statewide
7. "Most restrictive" statewide
8. The Gore 4-county recount strategy
9. "Dimples when other dimples present"

For example, *Scenario 5*., *Two-corners-detached statewide*, is based on arguments made by George W. Bush's attorneys during the 36-day period following Election Day, where Standard 3. is applied statewide.
Also, *Scenario 8*., *The Gore 4-county recount strategy*, is based on early post-election results, where the Gore camp requested hand counts in 4 heavily Democratic counties: Miami-Dade, Broward, Palm Beach and Volusia. Standard 2 is applied to some of Miami-Dade precincts.