

# The Voter Registry Dilemma Whitepaper



# The Voter Registry Dilemma

## Five Ten-Thousandths of a Percent

In the year 2000 a national presidential election of the world's most powerful nation was won by .0005% of the national vote despite the fact that more citizens voted for the other candidate.<sup>1</sup> Whether the world would be better or worse if the election result had been different, we will never know. We do know however, if history was rewritten based on a different elections outcome of that year, that history made in the decade prior to 2010 would be unrecognizable from today's account.

**National and International Legitimacy, Perception, Sovereignty based on Free, Fair and Credible elections.** At its root, your democracy is the direct relationship people have with your government. The population grants the new government the mandate to wield its power within the constraints of constitutions and jurisprudence. In a democracy, the greatest tool the public has is the vote, with eligible individuals exercising their power on Election Day. One legal person, in one legal place, casting one legal vote and that vote is counted once. The basis of political legitimacy is the *perception* by the population that their vote matters, that all votes have been fairly and accurately recorded, that there has been no external influence, that there is no corruption of results, and that the new government has lawfully come to power.

As an example, let's look at the challenges of the infamous 2000 US election in Florida *without* talking about hanging chads, butterfly ballots and perceived inability to count reliably. This whitepaper talks to the *real* issue at hand, an issue that was faced in Florida and more importantly faces every elections organization and democratic government in the world today. The real issue is the integrity of the voter registry.

There are more critical and unanswered questions than the efficacy of butterfly ballots.

- How many citizens of foreign countries voted in Florida? With a utility bill<sup>2</sup> you can get your name on the voters list in Florida if you are under 65. If you are over 65, you don't even need that. How many Canadians voted in the 2000 election in a country where they are not even a citizen?
- How many people were full time residents of Florida or how many are "snowbirds" that have a vacation home in Florida and voted by Absentee ballot in their home state?
- How many people were on the voter role, but got married, changed their last name and then found that they could not vote on Election Day?

---

<sup>1</sup> In the 2000 Presidential Elections 105,405,100 votes were cast and counted in the United States of America. Vice President Al Gore received 543,895 more votes nationally than Candidate George W. Bush but lost the election on an Electoral College count when the State of Florida pledged their electoral college votes to George W. Bush as he had 537 more votes in the state than Vice President Gore.(0.009% of the votes cast and counted in the state).

<sup>2</sup> <http://election.dos.state.fl.us/pdf/webappform.pdf>

- How many people were removed from the voter's list because they happen to share a name a birthday with a convicted felon who was ineligible to vote?
- How many people moved their home from one electoral district to another and stayed on the list in both districts, some with polling locations just a few minutes away?
- How many students who generally move annually found that they could be on none, one, two, three or four lists simultaneously?
- How many people named Bob or Robert, Susan or Suzie found themselves on the list more than once?

The answer is not certain, but what is certain is that the margin of error is **larger than 0.0005%** and may have altered the outcome of the election. History depends on the elections institutions charged with the safekeeping of democracy doing an exemplary job of the ensuring the integrity of the voting registry.

## What's in a Name?

Every election organization in the world is tasked with the challenge to ensure that every citizen who is eligible to vote gets the opportunity to vote. To find eligible voters the elections organization will likely use multiple sources of data with names and demographic information of probable or possible voters. Once aggregated, the sources will no doubt produce duplicate listings for a voter potentially in different electoral geographies. The resolution to this problem is more challenging than most would expect.

If you were in any international airport in the world and had a Mr. Li paged to the "nearest courtesy phone for a message", how many people would call back? With more than 150 million Mr. Li's in the world there is a very good chance of getting more than one. Much less than 1% of the world's population can be identified uniquely by last name. In some electoral areas it approaches near 0%. For example Singh and Kaur are very common last names in the Sikh community. In a tradition that began more than 300 years ago, the name Singh is given to every Amrit (baptized) male and Kaur to every female Sikh. There is a population of over 30 million people worldwide that have these names.

Clearly more data points are required to determine uniqueness. Combined with a first name, the probability of uniquely identifying a person rises dramatically except if the person's first name is widely given such as Muhammad. Now the world's most popular first name, it is estimated that more than 25 million men today have this first name. However even with a first name/last name combination, the problem becomes acute for organizations managing voters lists within elections geographies such as a city. In two similar sized Canadian cities (population approximately 500,000) the probability of having a duplicate name is **100 times more likely** in one city than the other.<sup>3</sup> This clearly illustrates that standard name-driven technology to "de-

---

<sup>3</sup> The examples were the cities of London and Brampton Ontario. Both cities are approximately 500,000 people in population. The name A Singh appears 100 times more often in Brampton than in London as an example.

duplicate” voters lists will not be effective in its own as even the probabilities of a match vary dramatically from area to area.

## Spellbound and Transformed

In addition to inherent problems of using textual names for the voters’ registry, there are always problems with data quality and data completeness.

- A. Smith
- Andy Smith
- Andi Smith
- Andrew Smith
- A.J. Smith
- A. John Smith
- J. Andrew Smith
- Andrew J. Smith
- Andrew Smith III
- Andrew Smith Jr.
- Andrew Smith Sr.
- Andrew Smith Junior
- Andy Smith (key transposition)
- Andi Smythe (Phonetic error)
- Etc.

All of the name permutations’ above could be the same elector or each could be a unique legitimate elector. It may depend on what source or sources provided a feed to the voters registry to establish the level of confidence in the name veracity. The source could come from various methods: enumeration, a form keyed in by the elections agency from an in person or web based authenticated registration, on line form, a feed from government property or tax rolls, a feed from another government agency or perhaps a feed from a third party trying to identify potential voters that have never been on the list.

The types of algorithms applied include phonetics (soundex), typo-confidence, New York State Identification and Intelligence System<sup>4</sup> (NYSIIS), Jaro-Winkler<sup>5</sup>, Scramble (multi-field), root name, alias transfer etc. All of these techniques can be applied to the data source to increase the level of confidence that a name provided is or is not unique in the system. These and many others are necessary in improving match confidence however still insufficient to improve the accuracy to acceptable levels.

## Extending the Demographic Dataset

Legislation throughout the world varies but continues to extend the privacy rights of citizens to not have extended demographic information in governmental data sources. This includes demographic attributes such as date of birth, sex, race, biometric data and other identification numbers that could assist the elections organizations in determining the uniqueness of a voter on a registry. The more data points that could be a probable match, the more certainty can be obtained about the potential uniqueness of that record. The number of data points continues to decline as concerns for privacy and the use of the data expands. Today with 98% accuracy<sup>6</sup>

---

<sup>4</sup> <http://en.wikipedia.org/wiki/NYSIIS>

<sup>5</sup> [http://en.wikipedia.org/wiki/Jaro-Winkler\\_distance](http://en.wikipedia.org/wiki/Jaro-Winkler_distance)

<sup>6</sup> Word/Coleman Demographic Aspects of Surnames (2003)

you can determine the race of Mr. Krueger, Mr. Jefferson, Mrs. Zhang, Mr. Khan, Ms. Patel and Miss Velazquez, all by their last names. Is that in danger also of disappearing? What is important to note is that names and age alone will never be enough dataset to ensure integrity.

One of the critical elements for not only determining where and in what district a voter will cast a vote, is the importance of attaching a geographic reference point to a voter. It is equally important to note that a geographic reference point cannot reliably be a textual address.

## The Trouble with Voter Addressing

If you placed your car on Yonge Street<sup>7</sup> in Toronto, Canada heading north, you would drive across Ontario some 1,896 Kilometers without turning off it until you reached the United States Minnesota border on the far western shore of Lake Superior. It was recognized for many years by the Guinness Book of World Records as the “Longest Street in the World”. However, Yonge Street changes its name over 50 times between Toronto and its termination point. You can have an address at 12345 Yonge Street that is also 12345 York Region #1. It is the same address with multiple street names.

If you lived or live in any of the following communities in Ontario:

- Yorkville
- Brockton
- Riverdale
- Parkdale
- Seaton Village
- Sunnyside
- East Toronto
- York
- Carlton
- Davenport
- West Toronto
- Wychwood
- Bracondale
- Midway
- Earlscourt
- Dovercourt
- Mimico
- Etobicoke
- Moore Park
- Leaside
- North York
- Forest Hill
- Swansea
- Long Branch

They are now through amalgamation and annex, the **City of Toronto**. Despite amalgamation, many organizations continue to prefer the names of the old municipalities instead of using Toronto. Including the post office as one often has to give the name of one of the original communities because amalgamation resulted in duplicate street names that are disambiguated only by referring to the former communities.

In short, textual addresses are dynamic and to that extent unreliable as they change over time and in the cases of new geographies may not physically exist yet be in address registries.

---

<sup>7</sup> [http://en.wikipedia.org/wiki/Yonge\\_Street#Yonge\\_Street\\_as\\_the\\_.22Longest\\_Street\\_in\\_the\\_World.22](http://en.wikipedia.org/wiki/Yonge_Street#Yonge_Street_as_the_.22Longest_Street_in_the_World.22)

## The Answer – Geo-Coding Voters

A specific reference point on a map that never changes can be clearly understood by electoral geography. By geo-enabling your voter list, you can import many different geographic mappings of land parcels called GIS fabrics, to help you discern a true physical location of a voter. There is likely land GIS registry fabric that can correlate parcels, concessions, lots and other identifiable areas of geography. Postal services may maintain fabric with postal codes by region and potentially post office areas (electors with post office box addresses, general delivery or a Rural Route delivery address). Tax roll and property assessment fabric maintained at various levels of government that may have higher currency of data (example: newly built homes or subdivisions) and also community-unique identifiers for properties or locations.

## How do you Start?

The accurate geo-coding of an elector starts with the layering of land parcels with all the variable address naming conventions (the GIS fabrics) to ensure that no matter what textual address is provided, that it will resolve and be fixed to a single geo-reference point (an X – Y coordinate on a map). Where no explicit address point exists such as a Post Office box, the address can at least be fixed to a single and common point in the area. The overall system accuracy is dependent on its ability to import and resolve addresses from as many fabrics as required to support the textual address format provided.

The concept of one legal voter at one legal address or more specifically a single geographic reference point is now a very powerful concept for maintaining a voter list.

For example, if you can place Mr. Andrew Smith at 12345 Yonge Street, Toronto and Mr. Andy Smith at 12345 Highway 11 North York and it is the same geographic reference point the probability of those two records being a duplicate increase substantially. By adding date of birth, sex and other available demographic points we can raise the probability to near certainty, *if the data quality is high enough.*

## Increased Confidence through Multi-phase Processing

If your data set provided a correct full legal and unique name, and a full legal address that was resolvable to a single geo-reference point and fully accurate demographics supported the uniqueness, then the processing logic would be relatively simple to determine a qualified voters list with a high degree of confidence that it did not contain duplicates or bad data (vital statistics errors). However, most data is not of that quality and the processing becomes heuristic in order to find the optimal data sets and improve the confidence in the data.

There are 2 steps required in the processing:

- Identification of what we suspect we know about the data and tagging the record(s)
- Processing the tags in an expert rules system that allows us to make subjective decisions about the data **and** create a plan and execute a plan for the remediation of the data.

Why are there two distinct steps? The answer is context. Until every record in the dataset is processed we do not know how many (if any) are possible duplicates. Each record is processed against every other record producing a tag that links records. Until the chain of tags is linked together, we don't know how to potentially merge or remediate records as the tags **only in combination with each other** can accurately reflect the confidence we have in the decision being made.

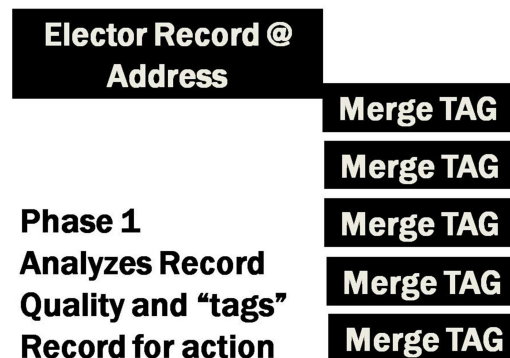


Figure 1 - First Phase Tagging and Rules

- Andrew Smith at 123 Main street Toronto, geo-code ID 987654321 – record 100
- A Smith at 123 Main street North York geo-code ID 987654321 – record 200
- Andy Smith at 123 Main street Unit 402 Toronto geo-code ID 987654321 – record 300

What the tag tells us is the comparison with the other record(s) and what type of possible match has been determined. In the very simplistic example above there are only 2 tags in play

- Record 100 – Possible truncated first name with Record 200 while match on Geo-code reference and Last Name
- Record 100 – Possible common versus formal name with Record 300 while match on Geo-code reference and Last Name

The merge phase of the processing needs to be expert rules system driven because the optimal answers rarely fit a defined procedure. Expert systems are by nature non-procedural. In the very simple example above we have a number of choices to make.

- Is A Smith the same person as Andrew Smith or is it Arlene Smith at the same address?
- Is Andrew Smith really Andy Smith or is Andy Smith the younger son of Andrew?

- Should the address for Andrew Smith really include unit A in it or is Unit a separate address for someone else?

When we originally process the records in the first phase, we identify possible duplicates but we also elaborate and create data that will assist us in our processing. For example:

- We collect information on the number of persons in a specific location and;
- if they do or do not have the same last name and;
- whether the data source of the data is newer for some records than for others and;
- the confidence level associated with the data source

First Name	Last Name	Address	City	Geo--Code	People	Data Currency	Source Confidence
Andrew	Smith	123 Main	Toronto	987654321	64	2 months	1
A	Smith	123 Main	North York	987654321	64	12 months	4
Andy	Smith	123 Main #402	Toronto	987654321	64	2 years	2

The expert system rules are now leveraged in this phase and look at the problem heuristically.

64 people in the same geo-code likely implies a high density dwelling so the address that contains a unit code is likely more correct.

The source confidence for record 100 is a 1 – legal tax role and both record 200 and 300 have been tagged as possibly truncated or common name version of Andrew. Therefore Andrew is more likely correct.

First Name	Last Name	Address	City	Geo--Code	People	Data Currency	Source Confidence
Andrew	Smith	123 Main	Toronto	987654321	64	2 months	1
A	Smith	123 Main	North York	987654321	64	12 months	4
Andy	Smith	123 Main #402	Toronto	987654321	64	2 years	2



Figure 2 - Second Phase Decisions and Actions

The second phase of the processing acts on the guidance provided by the tagging and expert-system driven merge directives or rules.

The record in the registry now becomes the following.

First Name	Last Name	Address	City	Geo--Code	People	Data Currency	Source Confidence
Andrew	Smith	123 Main #402	Toronto	987654321	Active	Derived	0
Andrew	Smith	123 Main	Toronto	987654321	Archive <sup>8</sup>	2 months	1
A	Smith	123 Main	North York	987654321	Archive	12 months	4
Andy	Smith	123 Main #402	Toronto	987654321	Archive	2 years	2

## Where Do I Vote?

Another basic record processing challenge is multiple geographies.

First Name	Middle Name	Last Name	Date of Birth	Sex	Address	City	Source	Geo-Code	Currency
A	J	Smith	03.01.1960	M	123 Main Street	Toronto	1	987654321	1 month
A	J	Smith	03.01.1960	M	123 Main Street	Toronto	2	987654321	1 month
A	J	Smith	03.01.1960	M	324 Cottage Road	Cottageville	2	123456789	6 months

In the case above we have an exact match on name, date of birth and sex identifiers. There are 3 possibilities:

- AJ Smith owns 2 properties and is identified by the tax rolls on both (1 in the city and 1 for vacation)
- AJ Smith has recently moved from 1 address to the other and the records at either/both locations do not fully reflect this yet
- AJ Smith is in fact two legal voters and two legal locations

<sup>8</sup> We need to archive data not delete it as we may in fact be incorrect in our optimization and may need to return the data to a previous state.



**Figure 3 Andrew Jenkins and Allan James Smith**

As the matched names are not fully elaborated, our confidence in a duplicate match is not high on its own. When combined with a match on core demographic such as date of birth and sex identifiers the probability rises to a match standard.

In this case the data source can assist our decision. The record sourced from income tax information is high confidence and current. We can also dynamically derive “helper” data to assist us with our decision like determining if AJ has previous records at either location.

In this case we use the data provided in an expert system rule to decide that when we process records of this nature what decisions do we make? The alternatives:

- Use 1 address as the primary (voting) address and archive the other based on ancillary data
- Do the above and advise AJ by letter of the decision and provide the ability for correction

In both of the above examples, the expert system rule engine allows you to create and act on policies you have developed for improving the data. Over time you will continue to optimize your policies so that the *accuracy of the voter registry will continually improve over time*. This approach also let's Elections organizations test policies to determine the overall impact on the registry and assess the quality level achieved by that policy.

## If It Were Only That Simple

While the above example describes the mechanics of the Voter Registry processing approach, it fails to communicate the complexity of the processing.

In the example above we illustrated a very simple scenario. In reality the voter name and demographic datasets are much larger and the comparators are much more extensive.

Simply matching **F**irst Name, **M**iddle Name, **L**ast Name generates 6 test combinations

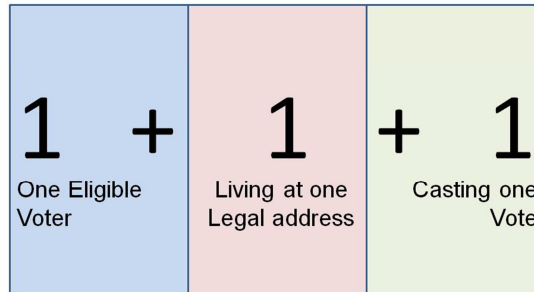
1. F and F,
2. M and M,
3. L and L
4. F and M,
5. F and L,
6. F and M and L

Each of these combinations is processed with more than 20 criteria (=, truncated, missing etc.). An example would be:

- First possible match with truncated name, middle does not match, last name exact match

This first test alone generates over 120 unique tests. To this we add further complexity such as named change analysis where we accommodate the fact that people with legally changed names (ie. marriage, divorce etc.). The processing continues to get more complex.

The combinations and permutations of test conditions quickly escalate to **tens of thousands** of viable tests for each record.



The Navantis Voter Registry solution automates the process of achieving 1 eligible voter at 1 legal address being placed on the registry and ensuring via voting strike-offs that a single vote is cast and counted.

The solution intakes an unlimited number of data feeds from any source including voter names and as much demographic information as can be provided and vital statistics information. It also intakes an unlimited number of GIS mapping fabrics which provide the mechanism to turn low-confidence address data into high-confidence geographic reference data.

The solution **automatically** generates the thousands of rules that are used to process and tag the data and then optimizes the rule sets for high performance and large data quantities.

The solution is initially configured with a standard set of expert rules and policies to assist in decision making, but allows the flexibility for each elections organization to develop and implement their own policies for attaining the highest quality possible in their registry. The rules are fully sensitive to the type and quantity of the available voter, vital statistics and demographic data.

### **Enterprise Technology**

The solution is built on state-of-the-art technologies from Microsoft and ESRI including advanced Microsoft SQL Server database technology, ESRI GIS, Microsoft BizTalk integration and orchestration engine and the BizTalk Business Rules Composer and Execution engine, a forward chaining expert systems engine based on rules (a set of if-then statements known as a Rete algorithm).

### **How it Works**

Each data feed either textual or GIS fabric is loaded and updated regularly into the Registry database. With each feed the engine reprocesses the data records tagging and updating tags to the registry records. The policies are then applied to make decisions based on the tags found and the registry is processed again where records are merged, archived, identified as optimized or marked as pristine.

It is a process of continuous improvement for your registry. As more data is made available to the system, the confidence level in every record will rise with the goal of having the highest quality data available for the election.

## **Conclusion**

Elections organizations understand their challenges in executing an election with the utmost integrity.

A concern for every government, elections organization and democracy in the world is:

**“Will everyone who is eligible to vote, be allowed to vote and have their vote count?”**

By the implementation of the ElectionReady Voter Register Management system the foundation for answering yes is established and perhaps equally important, each day using the system can propel that democracy closer and closer to an unequivocal **yes**, remembering that **0.0005%** error can change the world.