Architecture Article: AA001 / VoteCal Search & Matching

1 Executive Summary

The VoteCal Search Service is an independent service application that is used for voter searching and matching. The service is structured in a manner that supports data collection and search for nearly any type of business entity. This includes voters, death records, and felon records.

Since it is exposed as a stand-alone service, search can be made available to any application that needs to use the service and does not have to be limited to applications within the VoteCal system.

Requirements Covered

Req. Num.	Requirement Text
S3.1	VoteCal must allow an authorized user to query and locate an existing registered voter in the system for update using a variety or combination of criteria, including: ▪ Full or partial first name; ▪ "Smart name" variances on first name; ▪ Full or partial middle name; ▪ Full or partial last name; ▪ Soundex variations on last name; ▪ Full or partial residence address; ▪ Full or partial mailing address; ▪ Telephone number; ▪ VoteCal assigned UID; ▪ DL/ID #; ▪ Registration affidavit number; ▪ SSN4; ▪ Date of birth (DOB); ▪ Place of birth; ▪ Political party affiliation; ▪ Precinct; and ▪ Political district.
S3.3	VoteCal must provide the ability for an authorized user to search and retrieve one or more registered voters by using wildcard characters in one or more fields. For example: "Saun*" in the last name field to find all last names starting with "Saun".

	"*aun* in the last name field to find all last names containing "aun". "*ders" in the last name field to find all last names ending in "ders". "Saun*" in the last name field and "Har*" in the first name field to find records with last names starting with "Saun" and first name starting with "Har".
S10.1	VoteCal must include a user-configurable method for authorized SOS administrators to: ▪ Establish sets of registration record matching criteria; ▪ Configure which criteria apply to each type of matching function (e.g., new registration matching, death record matching, NCOA matching, etc.); ▪ Assign “confidence” levels to each criteria set as it applies to each matching function; and ▪ Establish threshold confidence levels required for manual or automatic application of matches for each matching function.
S10.2	VoteCal must provide the ability for SOS administrators to establish one or more basis for matching data in a registration record field, including (where applicable): ▪ Exact character match; ▪ First “X” characters of the field (where “X” is user configurable); ▪ Same characters and order in string, but with spaces and punctuation removed; ▪ Soundex match (or alternative method based on phonetic pronunciation); ▪ “Smartnames” match based on common variations of First Name established by SOS administrators (e.g., Robert = Bob, Bobby, Rob); ▪ “X” matching characters within string; and ▪ Same month and year.
S10.3	VoteCal must provide the ability for SOS administrators to identify a set of matching criteria based on combinations of individual field match settings, such as: ▪ First Name- with “Smartnames”; Last Name- first 4 characters; and Date of Birth- same day and month or ▪ DL/ID#- exact match; First Name- with “Smartnames”; Last Name- with Soundex.
S10.4	VoteCal must provide the ability for SOS administrators to configure and update whether or not an established matching criteria set is applied to each matching function, including: ▪ New & updated voter registration; ▪ Duplicate registration checks; ▪ DMV Motor Voter processing; ▪ Death record matching; and ▪ Felon record matching.
S10.5	VoteCal must provide the ability for SOS administrators to individually establish “confidence” values to each established matching criteria set as it applies to each potential matching function.

2 Description

There are several aspects of the search service that will be addressed below. These include: the manner in which data is collected (indexed) so that it may be searched, how a search request is sent to the search service, and how the search results are retrieved and returned to the caller.

2.1 Indexing Data

Data is gathered for searching by one of two processes called Full Indexing and Incremental Indexing. Full Indexing leverages the VoteCal Job Processing Service to periodically do a complete refresh of data available for search. Incremental Indexing takes place whenever a record is added or modified in the VoteCal system resulting in a single entity getting added or modified in the underlying search database.

Each type of entity for which data is collected for search has its data segmented into what is called a search scope. This search scope limits the type of records that are returned in the results. For example voter records, death records, and felon records are all contained within their respective search scopes so that a search scoped on death records for “John Smith” only looks at death records and does not return a voter record or felon record as a result.

Each search scope has its own full indexing job defined for the Job Processing Service so that they can be independently managed. It also has an adapter used by the search service so that real-time modifications to individual instances of the respective entity are reflected in the search data.

A scope definition also includes the fields that are available to be searched on for its respective entity type. Each field is defined as one of three types: Standard, Identifier, and Filter. These types are discussed in greater detail later in this document in the section titled “Gathering Results”.

It should be noted that all punctuation and spaces are removed from field values when data is indexed.

2.2 Sending a Query

The VoteCal Search Service is exposed as a SOAP/XML web service. Fundamentally, a query consists of two parts: an indicator of the scope to search within, and a list of field names and corresponding values to search on.

2.3 Gathering Results

When the search service processes a search request it first evaluates each field for which a parameter was sent to identify the type of field. The field types are defined as follows:

Standard – a standard field is one that has a varying degree by which it can distinguish an instance of an entity. The degree to which these fields uniquely identify an entity generally grows exponentially when combined with each other. Examples of standard fields are first or last name and SSN4.

Identifier – an identifier field is one that can theoretically uniquely identify one and only one entity. Some examples of these fields are a driver’s license number or social security number.

Filter – a filter field is one with a low degree of uniqueness and generally has a finite set of values and applies to a large subset of entities. Examples of filter fields are county and political party.

Search uses the Standard and Identifier type fields to form an actual database query. In the interest of performance, the Filter fields are not included in the database query. The query is created in a manner

that will allow for the matching of each field individually. The results returned for each field are examined and scored (see Match Scoring).

An entity is included in the result set if it achieves a match on greater than XX % (configurable) of the fields used in the search request or if it achieves a match on at least 1 Identifier field.

Any Filter field used in the search request is then applied to the remaining result set. Any entity that does not match on every Filter field supplied is removed from the result set.

The remaining entities in the result set are then assigned a confidence rating (see Confidence Rating), sorted in descending order according to the rating and returned to the requestor.

2.3.1 Match Scoring

Each field used in a search is scored independently. Points are awarded to an entity for matching a field with any of the following matching methods.

2.3.1.1 Exact Matches

1 point is awarded to an entity when it matches the value supplied in a search for the respective field exactly.

2.3.1.2 Soundex Matches

A Soundex is a 4 character code used to represent the phonetic sound of a word. Search will award a partial point, as configured, for a field getting matched with this method.

2.3.1.3 Synonym Matches

Synonym matches, sometimes referred to as “Smart Names” are used to match variations of names. For instance Bob and Robert can be matched together with a synonym match. Search will award a partial point, as configured, for a field getting matched with this method.

2.3.1.4 Wildcards

A wildcard character (*) may be used in a search field to match on part of a value. For instance a search on a first name field for “Chris*” would match results for “Chris”, “Christopher”, “Christian”, etc. A full point is awarded for a field matched when a wildcard is used; however, all other match methods are disabled on the field when the wildcard is used.

2.3.2 Confidence Rating

A confidence rating is assigned to each item returned in a search result set. This rating is an indicator of how certain the system is that the item returned is the one that is being searched for. There are two factors involved in calculating an item’s confidence rating: Match Grade and Match Quality. Both of these factors are expressed as a percentage and each is discussed in detail below. The confidence rating is the product resulting from multiplying the 2 factors together.

2.3.2.1 Match Grade

The Match Grade is the total match score awarded to an item divided by the highest possible match score, expressed as a percentage. For instance, if a search is done on first name, last name and birth date, there are 3 possible match points (one for each field). If an item matches first and last name, but does not match the birth date, the item receives a Match Grade of 66.7%.

2.3.2.2 Match Quality

The Match Quality is an indicator of the statistical probability of the combination of fields that a search result item matches on identifying one and only one record in the data population. For example, if a search is performed using first name, last name, and birth date, and an item is returned that matches only first and last name, the Match Quality factor of that item's confidence rating is equal to the probability that a first and last name uniquely identify an individual.

3 Matching Considerations

The term "Matching" refers to the automated searching on one type of entity based on data extracted from another type of entity. For example, Death Record Matching works by performing a search for voter records systematically with data taken from a death record, effectively trying to "match" two independent pieces of data together. The following are additional considerations that are taken into account with respect to searches as they pertain to matching:

- A match cannot be automatically applied when more than one result is returned regardless of the confidence rating.
- A match cannot be automatically applied when a result matches on less than 2 fields regardless of the confidence rating. This consideration accounts for a scenario where an identifier type field matches a record, but another field used in the search did not match.
- A match cannot be automatically applied when an identifier field was used in the search and a result did not obtain a match on that field, regardless of confidence rating.

Revision History

Date	Document Version	Document Revision Description	Revision Author
03/01/2010	0.1	Initial Draft	Chad Hoffman
04/08/2010	0.2	Reformatted to move Requirements Covered section to the top	Victor Vergara