



# VIEW500

IDENTITY MANAGEMENT AND XML  
DIRECTORY SERVICES SOLUTIONS

## WHITE PAPER: APPROXIMATE MATCHING

Written By: Andrew Sciberras  
and Dr Steven Legg

Published: 2008

© eB2B.com Pty. Ltd



## Introduction

Human users won't always be precise in searching a directory: names can be misheard, transcribed incorrectly or shortened; a user may not be familiar with the conventional function or service keywords; an acronym or abbreviation could have been used rather than the full title, etc.

This document describes the approximate matching strategy implemented in the View500 Directory Server. It briefly overviews Search Filters and describes how the View500 WebDUA builds simple filters with approximate matching always requested.

Approximate matching is one of the major contributors to View500's ease of use and user friendliness, and is a major differentiator between View500 and other directory systems.

The algorithms implemented in View500 were developed especially for View500, and are not available on other systems. Their specification forms a significant portion of the intellectual property contained in the View500 software. The majority of other directory vendors implement an approximate matching strategy called Soundex. A comparison between Soundex and View500's Approximate Matching is provided within this document.

## Background to search filters

A search request specifies a "filter" against which all prospective entries are tested. Only those entries having attribute values satisfying the filter are returned. The filter has full generality, including and, or, and not operators. A simple example of a search filter (in plain English) may be:

FirstName EQUALS Bob

Filters, like the one above often contain an Attribute (FirstName), the Type of Match (Equals), and a Value (Bob) that corresponds to type of match being performed.

The LDAP Standards require a search filter to specify the type of match to be performed for a supplied filter value. Commonly used matching types include:

Equality Match    FirstName = Bob    Firstname equals Bob

Substring Match    FirstName = B\*    Firstname Starts With Bob

Ordering Match    FirstName <= Bob    Firstname that is before Bob

Presence Match    FirstName = \*    Firstname is Present

Approximate Match    FirstName ~= Bob    Firstname approximately equals Bob

The semantics of approximate match are not specified in the directory standards; each implementation can do it differently, or in fact ignore it by treating approximate matching to be simple equality matching. Most support at best a simple Soundex algorithm. View500 implements a very rich form of approximate match that is far superior to the Soundex algorithm that was developed pre-1920.

For simplicity and usability reasons, the View500 WebDUA will by default build a simple search filter that is an and of all items specified, and an approximate match on

each item. However, this default behaviour may be overridden allowing a rich set of capabilities for the more advanced user.

## View500 approximate matching techniques

View500 implements 12 approximate match types, which it can perform when evaluating an approximate match type search operation. View500's indexes can be configured on an attribute by attribute basis to select the appropriate combination of approximate match types desired.

The approximate match capabilities can be separated into two distinct groups; Whole Value Matching and Keyword Matching. The value of an attribute may contain one or more words; for example Organizational Unit Name = New South Wales Sales Division. Whole Value matching techniques will apply to the entire value ("New South Wales Sales Division"), whilst keyword matching rules will apply to the individual Keywords of an entire value.

Keywords are formed by breaking values into individual words (by discarding punctuation) and rejecting "noise" words (in, at, the, of etc).

### Whole value equality matching

The supplied value must match exactly, except that upper and lower case differences, extraneous spaces, and punctuation are ignored.

yahoo will match Yahoo!

The values are equivalent when case and punctuation are ignored.

### Whole value synonym matching

The supplied value will match if it is a synonym of a value in an entry.

Synonyms are typically used for given names, but can be set up for other text based attributes. Synonyms are held within the View500 Directory and are currently maintained through a command line interface. An enhancement that has been implemented for release in late 2008 will allow synonyms to be easily updated through the Administration Console.

Bob will match Robert

Bob and Robert are synonyms and considered to be equivalent.

### Whole value abbreviation matching

The supplied value will match if it is an abbreviation of a value in an entry.

Abbreviations are automatically generated by View500 according to built in rules and taking into consideration noise word settings.

NSW will match New South Wales

NSW is an abbreviation of New South Wales.

## Whole value phonetic matching

The supplied value will match if it sounds like a value in an entry.

Rules to determine phonetic matching are built into View500 using proprietary methods are based on the English language. Phonetic matching allows View500 to accommodate for human users who may type a word that sounds like a given word, whilst not being accurately spelt.

fizzeotherap will match physiotherapy

fizzeotherap sounds like physiotherapy and will therefore match.

## Whole value typing correction matching

The supplied value will match if it is spelt like a value in the entry.

Slight mistakes in the spelling due to typographical errors like missing characters, additional characters or the transposition of characters are accommodated through the Typing Correction matching.

Enbiromnent will match Environment

Mistyped and transposed characters errors are accounted for.

## Whole value prefix matching

The supplied value will match if it is a prefix (i.e. the initial characters match) of a value in the entry. Matching is case insensitive and punctuation and extraneous spaces are ignored.

Gov will match Government

Gov is a prefix of Government

## Keyword equality matching

This is exactly the same as whole value equality matching, except that a match occurs if any keywords in the supplied value equality match any keyword in a value in an entry.

eB2Bcom will match eB2B.com Pty Ltd

eB2Bcom matches a keyword of "eB2B.com Pty Ltd" when punctuation is removed.

## Keyword synonym matching

This is exactly the same as whole value synonym matching, except that a match occurs if any keywords in the supplied value synonym match any keyword in a value in an entry.

Soccer will match Football Team

Soccer is a synonym of a keyword within "Football Team".

## Keyword phonetic matching

This is exactly the same as whole value phonetic matching, except that a match occurs if any keywords in the supplied value phonetic match any keyword in a value in an entry.

Hockey will match Australian Jockey Association

Hockey sounds like a keyword within "Australian Jockey Association".

## Keyword typing correction matching

This is exactly the same as whole value typing correction matching, except that a match occurs if any keywords in the supplied value typing correction, match any keyword in a value in an entry.

Dircetor will match Executive Director

Dircetor is a misspelt keyword within "Executive Director".

## Keyword stem matching

The supplied value will match if it and a keyword value within an entry share the same stem.

In linguistics, a stem is the part of the word that is common to all of its inflected variants.

For example, the stem "wait" will match all of its inflected variants, such as "waits", "waited" and "waiting".

Optics will match Optical Services

Optics and a keyword within "Optical Services" share the same stem (optic).

## Keyword Mandarin phonetic matching using Pinyin

The supplied value will match if it sounds like the stored Mandarin value. This form of approximate matching also works in reverse, whereby a supplied Mandarin character will match Pinyin values that sound the same.

Ni Hao will match 你好

Ni Hao is the Pinyin spelling of what 你好 sounds like when it is spoken.

## Comparison to Soundex

Soundex is a phonetic algorithm for indexing names by their sound when pronounced in English. The basic aim is for names with the same pronunciation to be encoded to the same string so that matching can occur despite minor differences in spelling. Soundex is the most widely known of all phonetic algorithms and is often used (incorrectly) as a synonym for "phonetic algorithm".

Soundex was developed by Robert Russell and Margaret Odell and patented in 1918 and 1922 (U.S. Patent 1,261,167 and U.S. Patent 1,435,663 ). A variation called American Soundex was used in the 1930s for a retrospective analysis of the US censuses from 1890 through 1920. The Soundex code came to prominence in the 1960s when it was the subject of several articles in the Communications and Journal

of the Association for Computing Machinery (CACM and JACM), and especially when described in Donald Knuth's magnum opus, *The Art of Computer Programming*.

The Soundex code consists of a letter and three numbers. The letter simply preserves the initial letter of the name. The numbers are obtained from the remaining letters according to rules: letters that sound similar (such as S and F, P and V, or M and N) are converted to a single number, vowels and some other letters are ignored, repeated letters are ignored, and the code truncated or padded to be four characters long.

For example, Rosewell is encoded as R240, where the 2 represents the S and the 4 represents the L; all other letters apart from the initial R are ignored and a 0 added to pad the code to four characters. The same code represents other similar names such as Rosewall, Roswell, Rowswell and even Russell.

This will tend to produce quite a number of false positives. The other major issue with it is that words will only match when the first letter of the word is the same. For example while View500 would match the names "kris" and "chris", Soundex would not match these values. (K620 and C620 respectively) but "chris" does match "Church", "Craig", "Chairs", "Cargo", "Charge", "Cork", "Corky", "Courage", "Crash", "Creek", "Cruise", "Crook" and "Croak".

## Conclusion

Where Soundex is only a single method of searching for a particular search field, View500 can offer any number of approximate search methods and combine the results to produce a concise list of results. The choice of which search methods to apply to each searchable attribute in View500 is completely configurable and specified in the directory schema.

View500 therefore facilitates a far richer and satisfying user experience when searching for information when compared to other directory technologies. View500 isn't simply a network application used to hold user credentials, nor is it a relational database with an LDAP front end, it is a purpose built directory server - designed and optimised for efficient and accurate searching, whilst accommodating a human user base.