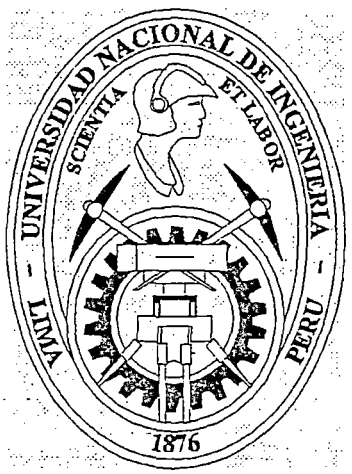
Universidad Nacional de Ingeniería

FACULTAD DE INGENIERIA INDUSTRIAL Y DE SISTEMAS



HERRAMIENTAS PARA LA ADMINISTRACIÓN MECANIZADA DE NOMBRES

TESIS

Para Optar el Título Profesional de:

INGENIERO DE SISTEMAS

Víctor Mondragón Chuquisengo

Lima – Perú

2005

DEDICATORIA

A Dios y mi familia que son la gran motivación de mi vida.

AGRADECIMIENTO

A mi gran amigo Daniel Hoshi con quien compartí este proyecto.

INDICE

RESUMEN
INTRODUCCIÓN....1
CAPÍTULO I: DIAGNÓSTICO ACTUAL
1.1 Identificación del problema....4
1.2 Herramientas existentes....4
CAPÍTULO II: SISTEMAS PROPUESTOS
2.1 Marco teórico....7
2.2 Formulación del problema....15
2.3 Recolección y procesamiento de datos reales....17
2.4 Formulación del modelo....18
2.5 Estimación de las reglas de comportamiento.... 21
2.6 Evaluación del modelo y reglas aplicadas....32
2.7 Formulación del soporte mecanizado....39
2.8 Validaciones....42
2.9 Aplicaciones prácticas....54
CAPÍTULO III: ANÁLISIS ECONÓMICO FINANCIERO....63
CAPÍTULO IV: ANÁLISIS COMPARATIVO ENTRE SISTEMA EXISTENTE Y SISTEMA PROPUESTO
4.1 Ventajas y desventajas de las soluciones actuales....67
4.2 Ventajas y desventajas de la solución propuesta....68
4.3 Resumen de la comparación....69
CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES
5.1 Conclusiones....70
5.2 Recomendaciones....70
GLOSARIO DE TERMINOS....72
BIBLIOGRAFÍA....73
ANEXOS....74

DESCRIPTORES TEMÁTICOS

- AUTOMATIZACIÓN
- ACOTAMIENTO
- BÚSQUEDA
- CLAVE
- CODIGO
- COMPARAR
- DIRECCIÓN
- NORMALIZACIÓN
- MNEMOTÉCNICA
- PALABRA

RESUMEN

El autor presenta el análisis, diseño y puesta en ejecución de un modelo analizador de datos alfanuméricos, tomándose como aplicación los "nombres" tanto de persona naturales como personas jurídicas.

En general la lógica del diseño propuesto abre las puertas a múltiples aplicaciones. A modo demostrativo se presenta también su aplicación a las direcciones.

En el caso de nombres, el concepto "Clave-nombre" será el elemento final de un trabajo de análisis, diseño y construcción de un juego de elementos normalizados.

El concepto "normalizado" representa un tratamiento de homologación o estandarización que en el límite se pueden considerar seudo-códigos.

La representación de nombres/apellidos como seudo-códigos permite finalmente su explotación en múltiples usos, destacando principalmente las búsquedas de listas que pudiesen ser: búsqueda fonética, búsqueda parcial, búsqueda invertida, búsqueda con alias y en general aplicaciones diversas pues dicho concepto permite ajustar la precisión de búsqueda a un número variable de caracteres. Para procesos mecanizados de grandes cantidades de información por lotes, ésta herramienta retorna altísima performance en comparaciones (matching). Adicionalmente esta herramienta se explota para normalizar data (por ejemplo direcciones).

El prototipo inicial de esta herramienta fue desarrollado en una institución líder del Sistema Financiero peruano y una empresa pública peruana de contribuciones la viene aplicando.

INTRODUCCIÓN

DEFINICIÓN Y PLANTEAMIENTO DE LOS PROBLEMAS

Al no existir reglas para el registro de nombres, al existir diversos alfabetos, diversos sonidos y la libertad de elegir por nombre cualquier combinación antojadiza de letras el problema de registro y comparación de nombres/apellidos tiende a ser difuso.

El general para evitar estos problemas en los sistemas administrativos y de información se tiende a CODIFICAR elementos. Por ejemplo "estado civil" se puede representar en una tabla codificada:

's': soltero

'c': casado

'v': viudo

'd': divorciado

Lamentablemente el universo de nombres/apellidos no es un conjunto finito por lo tanto estos elementos no se pueden tratar directamente como códigos lo cual induce al autor a crear seudo-códigos de referencia.

IMPORTANCIA DEL TEMA

En las Instituciones que realizan transacciones relativas a nombres/apellidos existen diversos problemas que provocan ineficiente uso de recursos en horas/hombre, materiales, dinero y tiempo.

Desde pequeños hasta grandes problemas surgen como consecuencia de no contar con una herramienta de apoyo en el manejo de nombres/apellidos.

Por citar algunos casos:

Por citar algunos casos:

- Registro ineficiente de nombres/apellidos: Por ejemplo registrar apellido/nombre ó nombre/apellido, escribir mas de un blanco entre elemento y elemento, escribir caracteres extraños o acentos que no son reconocidos en el intercambio de información entre plataformas (ASCII ó EBCDIC)
- Búsquedas ineficientes: Al no contar con una herramienta potente se requiere realizar 'n' consultas e iteraciones quedando la posibilidad de obviar alguna que coincidiese con lo buscado.
- Imposibilidad de comparaciones batch: Al no estar "normalizados" los nombres las comparaciones son limitadas debido al alto porcentaje de error.
- Pérdidas por negligencia o dolo: Desde que el documento oficial de identidad no explota el dígito de chequeo, expone su registro a un margen de error.

En el caso de una entidad privada del sector financiero se puede citar:

- Búsqueda parcial, fonética, con alias e invertida: Para apertura de clientes, investigación crediticia y consultas en general.
- Unificación ó desunificación de cuentas de clientes: propuesta automática para evitar la tediosa labor manual.
- Optimización de las búsquedas por orden judicial o de Contribuciones. Diariamente los bancos reciben órdenes de levantamiento de secreto bancario para diversas personas o empresas. Mayormente como dato se recibe nombres sin códigos de identidad. La mecanización de esta labor reduce ostensiblemente el uso de recursos.
- Depuración de nombres/apellidos antes de su integración a las Bases de Datos de Marketing.
- Normalización de nombres/apellidos a tiempo de ingreso de información

- Agilización en la preparación del reporte de Clientes deudores (RCD) para la Superintendencia de Banca y Seguros:

Para el caso de empresas públicas su uso es innumerable, por ejemplo:

- Para la RENIEC para validar y filtrar probables registros indebidos
- Para el Banco de la Nación para identificar los probables fallecidos que siguen cobrando pensión.
- Para el CNI por los trabajos de inteligencia(cruce de información)
- Para la SUNAT por cruces de información (identificar posible defraudación)
- Para la Superintendencia de Banca y Seguros a tiempo de confección del Reporte consolidado de deudores que suele tener un retraso de 60 ó más días.
- Para la Policía Nacional, ESSALUD, ONP y otras para identificación y cruces de información.

OBJETIVO DEL ESTUDIO

El autor muestra la metodología para desarrollar un soporte analizador de datos alfanuméricos tomando como aplicación práctica la problemática de nombres.

A modo práctico se mostrará el análisis, diseño y puesta en marcha de un software encapsulado comercialmente.

ALCANCES

El alcance es mostrar el desarrollo y aplicación práctica de herramientas para la administración mecanizada de nombres.

Si bien el título es refiere "nombres", para demostrar la potencia conceptual, se mostrará su uso orientado a "personas jurídicas" y también para atender la problemática de las direcciones en las instituciones.

Cabe notar que las herramientas propuestas están diseñadas para la realidad peruana pero la metodología es aplicable a otros idiomas.

CAPÍTULO I
DIAGNÓSTICO ACTUAL

1.1 IDENTIFICACIÓN DEL PROBLEMA

La problemática actual se resume en:

- No-explotación de dígitos de chequeo en los documentos de identidad motivo por el cual se expone a riesgos en la manipulación de nombres/apellidos.
- Necesidad de una herramienta potente para el registro, normalización y comparaciones de nombres/apellidos.
- Alto costo en las instituciones en la administración de nombres.
- Posibilidad de negligencia o dolo en el manejo de la información de nombres.
- Necesidad de una metodología objetiva para la estructuración del difuso problema de nombres.
- Problemas en el almacenamiento electrónico de nombres debido a conversiones en el intercambio entre plataformas tecnológicas (ejemplo ASCII y EBCDIC), mayúsculas, minúsculas, acentos y caracteres especiales.

1.2 HERRAMIENTAS EXISTENTES

Muchas instituciones han desarrollado herramientas de "búsquedas" habiendo conseguido relativo éxito en búsquedas fonéticas y búsquedas parciales pero serias limitaciones en búsquedas con "alias" y búsquedas con "clave invertida".

El autor validó diversas lógicas basadas en barridos de listas que terminan consumiendo recursos del ordenador o también invocaciones a software producto desarrollados para una realidad distinta a la peruana.
La mayoría de software como los productos Microsoft poseen querys que atienden parte de la problemática de búsquedas y comparaciones con las limitaciones antes mencionadas aunado a altos tiempos de respuesta.
En los buscadores WEB también se aprecia parte de estas limitaciones si bien destaca el uso de key-words o alias pero se limita en el manejo de sonidos y claves invertidas.

CAPÍTULO II
SISTEMAS PROPUESTOS

El autor propone establecer un modelo compuesto por:

- La realidad, el universo de nombres
- Un subsistema de base de datos de nombres calificados
- Un subsistema de criterios de comportamiento
- Un subsistema de retroalimentación de información

UNIVERSO DE NOMBRES

El autor trata la realidad de nombres de una institución cualquiera en el Perú donde el idioma castellano es oficial. Cabe acotar este alcance pues las reglas de comportamiento difieren entre idioma e idioma.

Cabe destacar que el punto de partida en la solución de un problema es la realidad y no el modelo. El autor ha apreciado algunos casos que pretenden que la realidad se acomode a modelos pre-concebidos cuando el fundamento es al revez.

BASE DE DATOS

El autor extraerá del universo real una muestra representativa de nombres.

Luego analizará la información y tras un análisis de frecuencia de mayor a menor calificará cada elemento.

Esta base de datos analizada y calificada busca ser una base de datos de conocimientos pues poseerá atributos que permitirán una sencilla aplicación de los criterios inferidos.

Si bien esta base de nombres no cubrirá el 100% de casos, se aproximará a este objetivo mediante validaciones estadísticas.

CRITERIOS DE COMPORTAMIENTO

El autor realiza un análisis de entidades y relaciones a fin de establecer la forma usual en que se relacionan las palabras. Es un arte diferenciar aquellos comportamientos genéricos de aquellos específicos.

En la validación se ajusta criterios indebidamente generalizados mediante prueba y error.

Cabe recalcar que lo más valioso de esta propuesta es la metodología que abre las puertas para nuevas aplicaciones.

SISTEMA DE RETROALIMENTACIÓN DE INFORMACIÓN

Como la mayoría de sistemas, el modelo propuesto es dinámico y siempre susceptible de ajustar, mejorar o evolucionar. Por lo tanto el autor incluye un subsistema de administración para las herramientas propuestas.

2.1 MARCO TEORICO

2.1.1 SISTEMAS

Se entiende por sistemas 'conjunto de entes relacionados con características o atributos en cada ente y que en su conjunto buscan un fin común'

Por ejemplo un sistema de lenguaje posee:

ENTIDADES	ATRIBUTOS
Palabras	Fonología, ortografía, morfología
Individuos	Conocimientos, culturas, costumbres.
Reglas	Gramática, Sintaxis etc.

Instituciones rectoras	Preservar reglas, aceptar cambios.

Así se tiene la definición de lenguaje en la enciclopedia Lexipedia:

"Conjunto de sonidos articulados mediante los cuales el hombre manifiesta lo que piensa o siente"

Existen varias relaciones entre las entidades de un sistema. Para el caso del sistema de lenguaje podemos citar:

1. Las palabras se clasifican en sustantivos, verbos, artículos, adjetivos etc.
2. Los individuos emplean las palabras.
3. Las instituciones rectoras (ejemplo: Real Academia de la lengua) preservan las reglas del lenguaje.
4. Los individuos proponen nuevas palabras.

La relación 1 es de Clase o clasificación y es estática.

La relación 4 es dinámica (anualmente se aprueban nuevas palabras)

En conjunto; los atributos de una entidad definen su estado y los estados de las entidades más importantes definen el estado del sistema.

Los objetivos que se persiguen al estudiar uno o varios fenómenos en función de un sistema son aprender como cambian los estados, predecir el cambio y controlarlo. La mayor parte de los estudios combinan estos objetivos en mayor o menor grado. A una combinación específica de estos objetivos le concierne la relación entre las entradas y las salidas de un sistema según se presenta en la figura 1.

Las entradas se refieren a los estímulos externos de un sistema que producen cambios en el estado del sistema. Las salidas se refieren a mediciones de estos cambios de estado. Suelen ocurrir tres variantes de la evaluación de alternativas:

1. Se puede llevar a cabo un análisis directo donde se especifica las entradas al sistema y después se miden las salidas.

2. Se conoce la entrada y se especifican ciertas características pertinentes para la salida.
3. Se especifica el sistema y se desea determinar la entrada que produce una salida deseada.

Para el presente trabajo el autor aplica la alternativa 1.

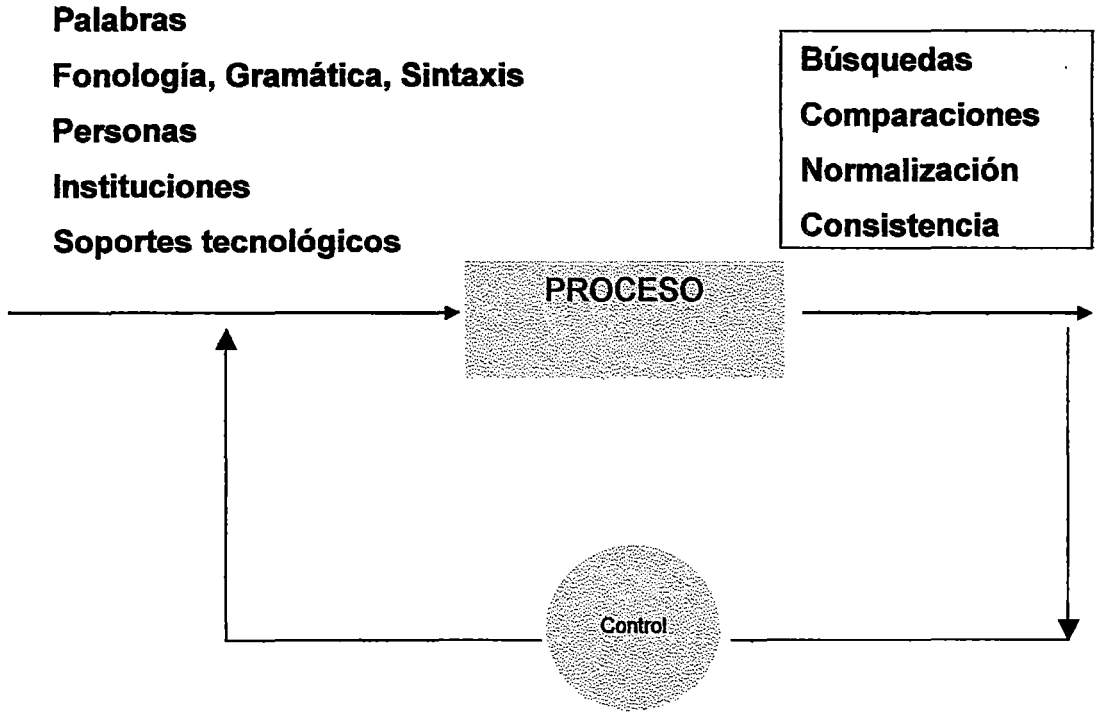


FIGURA 1: ENTRADAS Y SALIDAS DEL SISTEMA

2.1.2 CLASIFICACIÓN DE LOS SISTEMAS

Hay varios patrones útiles de clasificación de sistemas, así tenemos:

Según naturaleza:

- Naturales: Ejemplo los sistemas físicos y biológicos
- Artificiales: Ejemplo los sistemas sociales, económicos y políticos. El sistema de lenguaje es un sistema artificial.

Según interacción con el entorno:

- Sistema cerrado: No intercambian energía o información con su entorno, (son casi ideales)
- Sistema abierto: La mayoría de sistemas son abiertos, ejemplo el cuerpo humano. El sistema de lenguaje es un sistema abierto pues realiza intercambios con el entorno.

Según grado de adaptabilidad:

- Sistema adaptable: Reacciona ante cambios del entorno. El sistema de lenguaje es adaptable pues evoluciona.
- Sistema inadaptable: No reacciona en forma conveniente ante cambios del entorno según su finalidad original. Ejemplo el cuerpo de un difunto

Según el estado:

En cualquier momento se puede describir el estado de un sistema observando el valor actual de sus atributos.

- Sistema estable: Si los valores de sus atributos permanecen constantes o se encuentran dentro de límites definidos.
- Sistema inestable: Si los valores de los atributos fluctúan mucho.

2.1.3 RENDIMIENTO DE LOS SISTEMAS

Los sistemas se analizan para comprender como ocurre un cambio, así como predecirlos y controlarlos. La ejecución de un sistema es la secuencia de estados que un sistema adopta en un intervalo de tiempo. Estos estados suelen medirse de alguna forma para proporcionar una medida del rendimiento.

El concepto medida de rendimiento varía según los sistemas. En una empresa comercial las utilidades pueden ser la medida. En un sistema de lenguaje la medida puede ser el grado en que satisface la comunicación entre individuos.

2.1.4 ANÁLISIS DE SISTEMAS.

Según el diccionario Lexipedia "análisis es la separación y distinción de las partes de un todo hasta llegar a conocer los principios o elementos de éste"

Por lo tanto los objetivos del análisis de sistemas son la descripción y explicación de su comportamiento. En el caso de una situación no estructurada, por ejemplo analizador de palabras de un sistema de lenguaje, lo primordial es su descripción. ¿Qué parece incluirse? ¿Cómo parecen interactuar los componentes? ¿Cambia el sistema con el tiempo? Los científicos conductuales esperan entender algún día poder entender el aprendizaje y la percepción en los seres humanos.

La segunda finalidad del análisis de sistemas es la explicación de la conducta del sistema. Se responde a las preguntas sobre "cómo" y quizás el "por qué" de las conductas. Este aspecto del análisis de sistemas plantea cierto número de preguntas de índole filosófica. ¿Por ejemplo, qué se entiende mediante la palabra EXPLICACIÓN?

Si una explicación justifica el comportamiento de un sistema en 99% de los casos, ¿es suficiente? ¿Hay medidas objetivas que permitan evaluar explicaciones de comportamientos de sistemas?.

En cada sistema debemos regresar a las cuestiones de representación, suficiencia, validación y utilidad. Las críticas al análisis de sistemas giran en torno a la comprensión o la falta de comprensión de estos conceptos.

2.1.5 MODELOS

Modelo se define como "abstracción selectiva de la realidad" "representación de algo en pequeño" (diccionario Lexipedia).

Los propósitos de usar modelos son (1):

- Hacen posible que un investigador organice sus conocimientos teóricos y sus observaciones empíricas sobre un sistema y deduzca las consecuencias lógicas de esta organización.

(1) Tomado de Conceptos y Métodos de Simulación – George Fishman

- Facilita la comprensión del sistema.
- Acelera el análisis
- Constituye un modelo de referencia para probar la aceptación de las modificaciones del sistema.
- Es más fácil de manipular que el sistema mismo
- Hace posible controlar mas fuentes de variación que lo que permitiría el estudio directo de un sistema
- Suele ser menos costoso

PRINCIPIOS UTILIZADOS EN EL MODELADO.

FORMACIÓN DE BLOQUES.

La descripción del sistema se debe organizar en una serie de bloques o subsistemas con el propósito de simplificar la especificación de las interacciones dentro del sistema. Cada bloque describe parte del sistema que depende de pocas, preferiblemente una, variables de entrada y produce unas pocas variables de salida.

Luego puede describirse el sistema como un todo en términos de las interconexiones entre los bloques.

El presente trabajo respecto a un "sistema analizador de palabras" modela el subsistema nombres de personas naturales y/o jurídicas.

RELEVANCIA:

El modelo solo debe de incluir los aspectos del sistema relevantes a los objetivos del estudio. En el presente trabajo, se enfatizará el estudio de elementos y relaciones entre nombres de personas naturales y jurídicas.

EXACTITUD:

Debe de tenerse en cuenta la exactitud de la información que se recibe. El caso de "nombres" es un conjunto de datos poco estructurado que se puede representar de diversas formas, por lo tanto se requerirá procesos básicos de normalización de la información.

AGREGACIÓN:

Un factor adicional que debe de considerarse es el grado con que pueden agruparse las distintas entidades individuales en entidades más grandes. La presente tesis mostrará que el modelamiento aplicado a nombres es extensible a otros subsistemas por ejemplo direcciones.

2.1.6 MARCO METODOLÓGICO A APLICAR

En el presente trabajo el autor emplea la siguiente metodología:

- Formulación del problema
- Recolección y procesamiento de datos tomados de la realidad.
- Formulación de un modelo
- Estimación las reglas de comportamiento a partir de los datos reales.
- Evaluación del modelo y de las reglas aplicadas
- Formulación del soporte mecanizado
- Validación
- Diseño de experimentos de simulación
- Análisis de los datos simulados.

La figura 2 muestra una un problema difuso de palabras al cual se pretende acotar y estructurar a fin de llevarlo a límites manejables

Aunque el orden metodológico propuesto queda abierto a la discusión, en la figura 3 se muestra una ordenación tal que en las experiencias pasadas condujo a resultados aceptables.

2.2 FORMULACIÓN DEL PROBLEMA

Muchas veces el principal problema es precisamente su formulación, es decir un problema mal formulado o mal planteado nos puede llevar a resultados inesperados.

El presente trabajo se centra en la necesidad de desarrollar herramientas potentes para el registro, normalización y comparación de nombres y apellidos.

Este problema tiende a ser difuso en cuanto a su extensión y dinamismo. Por lo tanto gran parte del problema radica en su ESTRUCTURACIÓN a fin de acotarlo a límites manejables.

Dicha problemática requiere un enfoque multidisciplinario pues enfrenta aspectos lingüísticos, aspectos técnicos y aspectos administrativos.

Cabe notar que más que atender un problema de nombres y apellidos, el autor busca sentar las bases para un sistema de administración de palabras en general.

FIGURA 2: PROBLEMÁTICA A ESTRUCTURAR

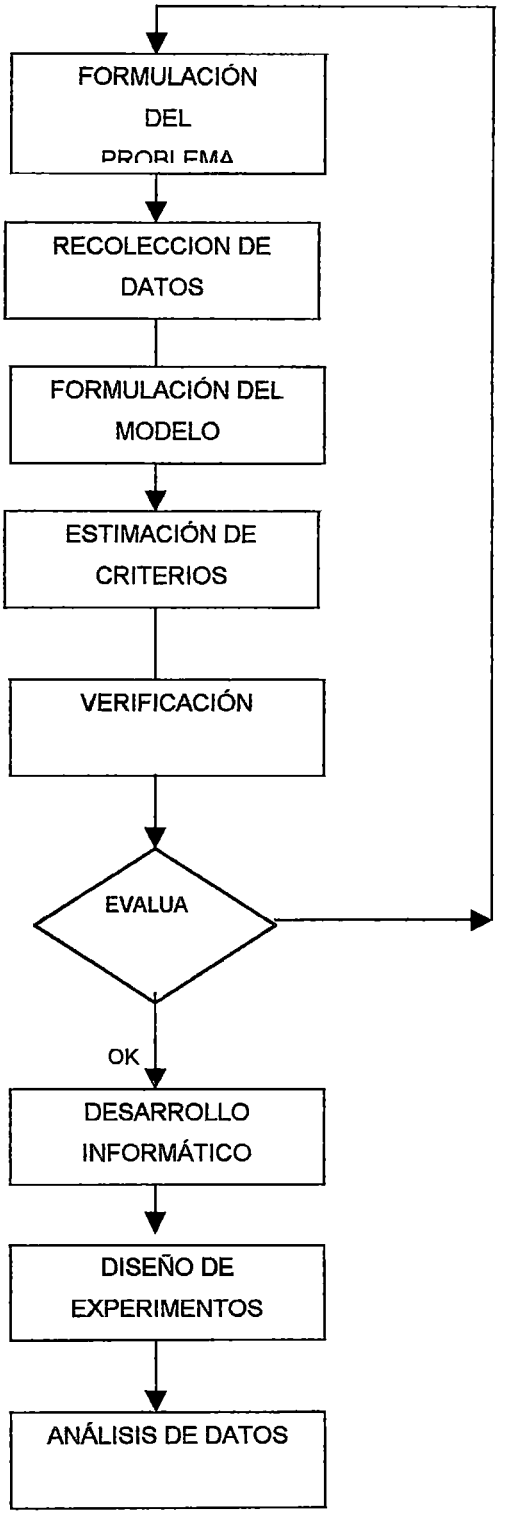


FIGURA 3: METODOLOGÍA A APLICAR

2.3 RECOLECCIÓN Y PROCESAMIENTO DE DATOS REALES

Algunos modelos fracasan cuando se busca llevar la realidad hacia modelos pre-concebidos. Cada realidad y cada problema tiene matices que hacen distinto un problema de otro, es decir, el punto de partida siempre es la realidad.

El caso en estudio se plantea en la realidad peruana, es decir:

- Idioma castellano que implica la aplicación de sus lineamientos tales como sintaxis, morfología y fonética entre otros.
- Nombres y apellidos mayormente castellanos con gran porcentaje de nombres ingleses, italianos, franceses y quechuas entre otros.
- Diversidad de plataformas tecnológicas de registro y almacenamiento de datos tal el caso que algunas plataformas usan el código EDCDIC y otras el código ASCII donde la conversión de algunos caracteres como la letra "Ñ" o los acentos no siempre se ejecutan debidamente.
- Limitaciones a nivel institucional en la administración de nombres, desde pequeñas instituciones, empresas y organismos públicos que no poseen adecuados procedimientos para el registro y control de nombres
- Archivos históricos con registros inconsistentes provenientes de años pasados

El modelo en estudio requiere recolectar un universo de nombre de diversas fuentes de información como pudiese ser la guía telefónica, contribuyentes de las municipalidades, clientes de empresas, la RENIEC entre otros.

Una vez obtenida la información se procede al análisis, es decir:

- Separar cada nombre en elementos y registrarlos
- Establecer reglas de formación o comportamiento
- Esbozar agrupaciones o atributos en común
- Calificar los elementos según pre-agrupaciones

CONTROL

Para determinar si seguimos acopiando mas bases de datos de nombres se experimenta con nuevos nombres. Por ejemplo conseguimos una nueva base de datos y validamos qué porcentaje de nombres ya se encuentran registrados en la base original.

Finalmente se hace un análisis de frecuencia para analizar que elementos son los mas referenciados y validar si se alcanzó un nivel de acopio o satisfacción razonable.

La figura 4 muestra un experimento de acopiamiento de nombres de diversas bases de datos donde se aprecia que añadiendo mayor cantidad de elementos a la base de datos de nombres en el límite el rendimiento asemeja una curva asintótica.

Número de elementos	Rendimiento
5,000	58.22%
10,000	71.85%
20,000	90.87%
30,000	96.12%
45,209	99.17%

FIGURA 4: ACOPIAMIENTO (VS) NIVELES DE RENDIMIENTO

Esta base de datos servirá para validar los experimentos y una vez calificado cada elemento-nombre dicha base evolucionará hacia base de datos de conocimientos.

Cabe notar que el acopio no tiene fin y el modelo contemplará mecanismos para seguir añadiendo elementos-nombre conforme se presenten y no los encuentre en dicha base de datos.

2.4 FORMULACIÓN DEL MODELO

El modelo planteado consiste en diseñar un sistema que ante el ingreso de nombres diversos permita estructurar dichos elementos tal que facilite su explotación.

Se entiende por 'estructurar' a un conjunto de procedimientos de administración de nombres, tales como registrar, organizar, normalizar y controlar de modo tal que facilite la atención de la problemática de nombres.

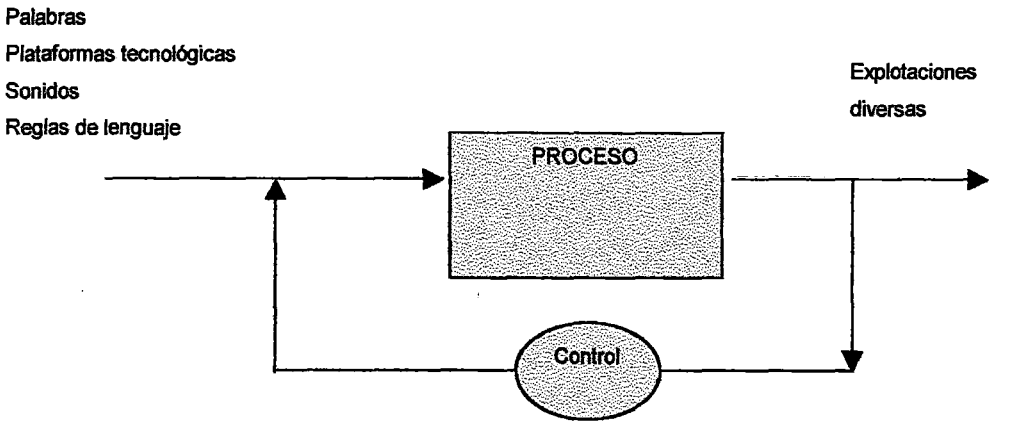


FIGURA 5: PROCESO DE ESTRUCTURACIÓN

El presente modelo se desarrolla sobre 4 subsistemas:

- La realidad o universo de nombres
- Una base de datos que propugna ser una base de conocimientos
- Un conjunto de criterios para normalizar los elementos-nombre
- Una relación básica de rutinas estándar para explotación diversa.

UNIVERSO DE NOMBRES

Es un subsistema no finito pues cada día aparecen nuevos nombres tales como "H2O" conque un conocido psicólogo inscribió a su hija. Los nombres no tienen reglas, puede escogerse una combinación antojadiza de caracteres.

BASE DE DATOS

Este subsistema se construye para almacenar nombres y agruparlos para su calificación. Sirve de consulta en caso de nombres ya almacenados y en caso de no encontrarse el nombre, éste se añade.

Esta base de datos propugna ser una base de conocimiento pues se retroalimenta, almacena criterios de agrupación y elementos calificados.

CRITERIOS Y NORMALIZACIÓN

Este subsistema también es dinámico y susceptible de ajustes y mejoras. Contiene lógicas que se comprueban antes de su aplicación y sirven para estructurar los nombres convirtiéndolos en elementos normalizados y aptos para facilitar su explotación.

RUTINAS DE EXPLOTACIÓN

Es un conjunto básico de procedimientos que permite explotar las bondades de haber estructurado la data (nombres/apellidos). Estas rutinas serán invocadas por diversos procesos que requieran hacer búsquedas, comparaciones, normalizaciones de nombres entre otros.

La figura 6 muestra la relación entre los subsistemas mencionados.

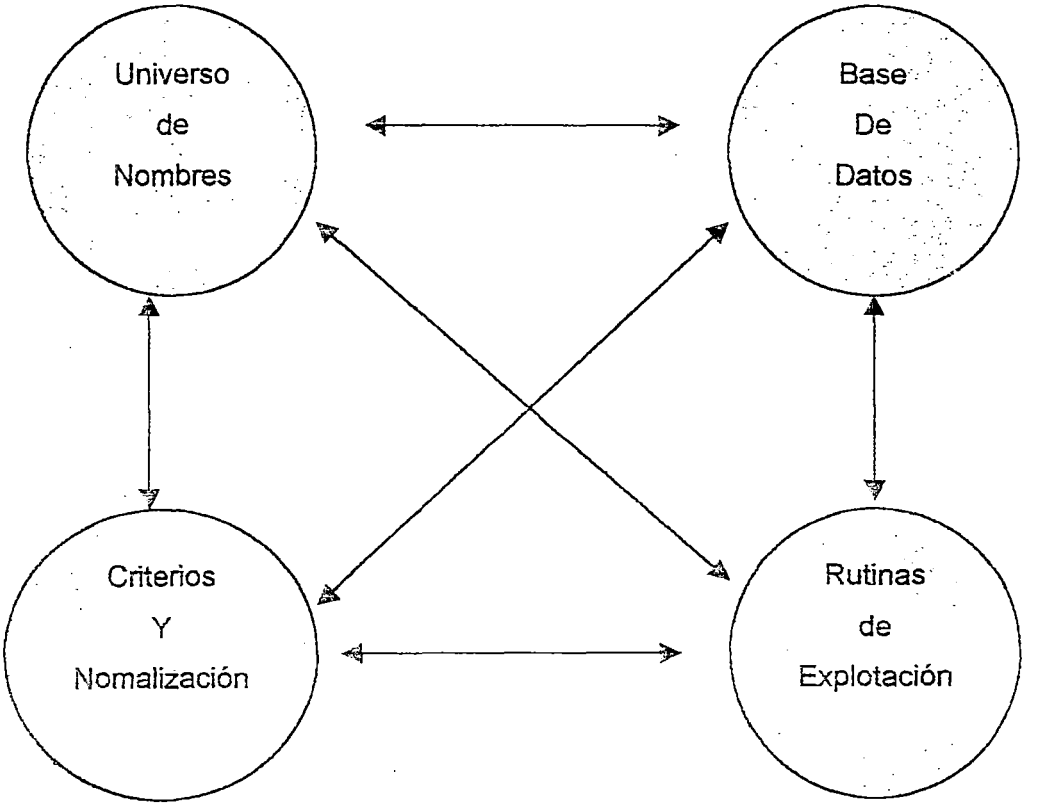


Figura 6: SUBSISTEMAS DEL MODELO A APLICAR

2.5 ESTIMACIÓN DE LAS REGLAS DE COMPORTAMIENTO

El autor estima las siguientes reglas básicas de comportamiento:

- Se puede extraer una muestra representativa del universo de nombres y almacenarla.
- Las cadenas de datos nombre son susceptibles de dimensionar (elementos y longitud de la cadena)
- Los caracteres de las cadenas de nombres son susceptibles de homologar.
- Las cadenas de nombres poseen caracteres innecesarios que se pueden eliminar.
- Los nombres tienen factores comunes de comportamiento por lo cual son susceptibles de agruparse.
- Los elementos nombre poseen características adicionales que deben almacenarse.
- Un elemento nombre pueden referirse mediante diversos nombres sustitutos o alias.
- Por sus características se puede discriminar si el nombre es de persona natural o de persona jurídica
- Los elementos nombre pueden registrarse en su mayoría en una base de datos de conocimiento y de ahí obtener mayores datos.
- Las letras de los nombres poseen sonidos susceptibles de normalizar hacia un menor número de caracteres.
- Los elementos nombre son susceptibles de representar como seudo-código o claves nombre.

2.5.1 Almacenamiento de una muestra representativa del universo de nombres.

Si bien el universo de nombres tiende a infinito podemos alcanzar un alto grado de aproximación tal como se trató en el punto 2.3 la recolección de elementos-nombre desde el mundo real. Esta recolección de elementos se

somete a análisis de agrupación y calificación. El objetivo es disponer de una muestra representativa que contenga el mayor porcentaje del universo de nombres.

Dicha base de datos propugna ser una base de datos de conocimientos dinámica pues evoluciona al ajustar o recalificar sus elementos o al agregar nuevos elementos.

La calificación de cada elemento es exhaustivamente contrastada con la realidad, en especial aquellos elementos que se presentan con mayor frecuencia en las datas muestrales (ejemplo, Juan, María, García, Fernández etc.).

2.5.2 Dimensionado de cadenas de datos nombre

Analizar implica separar un todo en sus partes. De este modo una cadena de nombres primero debe ser almacenada en un área de trabajo para:

- Dimensionar la longitud de cada elemento y su longitud total
- Separar cada elemento nombre para su respectivo análisis

En la figura 7 se muestra algunos ejemplos de cadenas nombre.

José Luis de la Riva Agüero y Sánchez Montalvo
Lavalle Acuña José María
María José Jorge Fernández de Córdova
Wilson Francisco Javier Llerena Gonzales
Comercial importadora La sirena s.a.
Agregados calcáreos La perfecta Ltda.
Distribuidora de materiales de construcción 2 de mayo SAC

Figura 7: EJEMPLOS DE CADENAS-NOMBRE

José	Luis	De	La	Riva	Agüero	Y	Sánchez	Montalvo
Lavalle	Acuña	José	María					
María	José	Jorge	Fernández	De	Córdova			

Wilson	Francisco	Javier	Llerena	Gonzales				
Comercial	Importadora	La	sirena	s.a.				
Agregados	Calcáreos	La	perfecta	Ltda..				
Distribuidora	De	Materiales	De	Construcción	2	De	Mayo	SAC

Figura 8: CADENAS-NOMBRE DESAGREGADAS

Cabe notar que el empleo de matrices es de gran utilidad cuando los procedimientos se trasladan a programas de computadora.

2.5.3 Homologación de caracteres

Las cadenas de nombres tienen diversas proveniencias:

- La palabra hablada
- Los escritos manuales
- Los archivos magnéticos de distintas plataformas tecnológicas.

El autor propugna acotar el universo de elementos hacia un conjunto mas estructurado para facilitar su tratamiento.

Por lo tanto se registran los nombres aplicando consideraciones básicas tales como:

- Convertir formatos EBCDIC y ASCII u otros hacia un solo formato.
- Homologar letras mayúsculas y minúsculas hacia mayúsculas. (en lenguaje de programación Cobol sería:
  INSPECT DATOS CONVERTING 'abcdefghijklmnopqrstuvwxyzñ' TO 'ABCDEFGHIJKLMNOPQRSTUVWXYZÑ').
- Hacer conversiones para eliminar acentos (en lenguaje de programación Cobol sería:
  INSPECT DATOS CONVERTING X'9FA082A1A2A3C1C9CDD3DA' TO 'AAEIOUAEIOU').
- Convertir caracteres extraños en blancos
- Convertir cadenas de caracteres blancos hacia un solo blanco

2.5.4 Eliminación de caracteres innecesarios

Hay caracteres que son poco relevantes para el registro de nombres. Por lo tanto el autor postula su eliminación convirtiéndolos en blanco.

En lenguaje de programación Cobol podría ser:

INSPECT DATOS1 CONVERTING '?¿+-<>{}[]Vj!,;.:_"%*=|°^¬#$~ '?¨$~+.<>,!_-V{};:%()*ª_á|=›³"

TO ' '.

2.5.5 Agrupación de nombres

En la práctica al analizar las cadenas de nombres se aprecia características comunes bajo ciertas agrupaciones.

El autor postula las siguientes agrupaciones:

- Nombres
- Apellidos
- Nombres que también son apellidos
- Actividades en general
- Siglas de empresas jurídicas
- Artículos y pronombres
- Palabras relativas a santos
- Oficios, profesiones o títulos
- Palabras relativas a instituciones diversas
- Palabras relativas a actividades económicas
- Meses
- Números expresados en palabras
- Nombre reducido
- Apellido reducido

A las agrupaciones se las puede calificar de modo tal que en un proceso mecanizado sea más simple su identificación.

En la figura 9 se muestra algunas agrupaciones y sus calificaciones

AGRUPACIÓN	CALIFICACIÓN	EJEMPLOS
Nombres	A	Juan, Pedro, María, Graciela, Susana
Apellidos	B	González, Fernández, García, Mendoza
Actividades en general	C	Carrocería, canal, aéreo, natura etc.

Siglas de empresas jurídicas	D	SA, SAC, EIRL etc.
Artículos y pronombres (conectores)	F	El, la, los, las, una, un etc.
Palabras relativas a santos	G	San, santa, santísimo, Sto. etc.
Oficios, profesiones o títulos	H	Guardia, duque, fray, general, licenciado etc.
Palabras relativas a instituciones diversas	I	Cuartel, embajada, episcopado, ejército, consulado etc.
Palabras relativas a actividades económicas	K	Corporativo, servicios, aerolíneas, constructores, contabilidad etc.
Nombres que también son apellidos	L	Washington, Reyna, Domingo, Jorge etc.
Meses	R	Enero, marzo, abril, etc.
Números expresados en palabras	U	Catorce, cuarta, III, vigésimo, etc.
Nombre reducido (truncado)	V	Aniba, ampar, anselm, juanit, leopold etc.
Apellido reducido (truncado)	W	Alvare, Andrad, Gonzag, Padill etc.
Palabras no encontradas en la tabla de elementos nombre calificados	Z	Ejm. Nombres coreanos, pakistaníes, hindúes etc.

Figura 9: AGRUPACIÓN DE NOMBRES Y SU CALIFICACIÓN

2.5.6 Adición de características adicionales a los nombres

A tiempo de acopiar o almacenar el universo de nombres se puede aprovechar para incluir otros datos que sean de posible utilidad. Por el ejemplo el sexo de un nombre.

En el análisis posterior se afinarán los criterios para discernir por ejemplo:

José María → Masculino

María José → Femenino

Pueden existir elementos que provienen de bases de datos anteriores y son copiados con sus bondades o defectos. Tal el caso de nombres

truncados o mal escritos que en un momento posterior se pueden calificar como errores de registro.

Este tipo de calificaciones también son apoyos a la lógica cuando se quiera separar nombres de apellidos o discriminar personas naturales de jurídicas.

En la figura 10 se aprecia algunos ejemplos de calificación adicional.

Nombre	CALIFICACIÓN	SEXO	ERROR
Jesús	A	M	
Juan	A	M	
Rosa	A	F	
Fernández	B		
Eysaguir	W		X
Eysaguirr	W		X
Eysaguirre	B		

Figura 10: CALIFICACIÓN ADICIONAL

Se puede apreciar que el nombre Jesús esta calificado con sexo M, es decir masculino. Sin embargo a veces se usa como nombre de sexo femenino. En el registro de arriba se optó por calificarlo según mayoría de casos.

Por otra parte, la calificación de sexo no es pertinente en los apellidos (nombres con calificación B).

Finalmente se aprecia dos elementos truncados (calificación W) que después de contrastar con el mundo real se opta por registrarlos con X en la columna error como recomendando no tomar en cuenta. Sin embargo estos elementos quedan almacenados a modo histórico.

2.5.7 Uso de nombres sustitutos o alias

Diversos elementos nombre pueden ser referidos por nombres alternos o 'alias'.

Este tipo de calificaciones es sumamente delicado y debe probarse exhaustivamente contra la data real pues si bien pudiéramos satisfacer algún caso específico, se podría estar generalizando algún supuesto indebido.

El uso de alias permite acceder a un elemento nombre desde diversos nombres.

En la figura 11 se puede apreciar un segmento de la base de datos calificados.

ELEMENTO NOMBRE	ALIAS	CALIFICACIÓN
COMANDANCIA		K
COMANDARY		B
COMBE		B
COMBI		B
COMBIAN		B
COMBINA		B
COMBUST		K
COMBUSTIBLE	COMBUSTIBLE	K
COMBUSTIBLES	COMBUSTIBLE	K
COMECA	COMBUSTIBLE	B
COMEDOR		K
COMEDORES		C
COMENA		B
COMER		K
COMERC	COMERCI	K
COMERCI	COMERCI	K
COMERCIA	COMERCI	K
COMERCIAL	COMERCI	K
COMERCIALE	COMERCI	K
COMERCIALES	COMERCI	K
COMERCIALIZ	COMERCI	K
COMERCIALIZA	COMERCI	K
COMERCIALIZAC	COMERCI	K
COMERCIALIZACAO	COMERCI	K
COMERCIALIZACI	COMERCI	K

COMERCIALIZACIO	COMERCI	K
COMERCIALIZACION	COMERCI	K
COMERCIALIZACIONES	COMERCI	K
COMERCIALIZAD	COMERCI	K
COMERCIALIZADO	COMERCI	K
COMERCIALIZADOR	COMERCI	K
COMERCIALIZADORA	COMERCI	K
COMERCIALIZADORE	COMERCI	K
COMERCIALIZADORES	COMERCI	K
COMERCIALIZORA	COMERCI	K

Figura 11: SEGMENTO DE LA BASE DE DATOS DE NOMBRES CALIFICADOS

2.5.8 Discriminación de personas naturales de jurídicas.

Para discriminar si una cadena de datos nombre referencia un nombre de persona natural o persona jurídica se requiere un apoyo inicial en la tabla de nombres-calificados y la aplicación de un conjunto de criterios de discernimiento. Este proceso sigue la siguiente secuencia:

- Homologar caracteres y eliminar acentos
- Eliminar caracteres especiales y blancos innecesarios
- Desagregar la cadena nombre en un arreglo de elementos
- Obtener de la base de datos de conocimiento la calificación de cada elemento
- Si el elemento no existe en dicha base de datos se procede a inferir su calificación
- Se asigna pesos o ponderaciones a cada elemento
- Se evalúa las calificaciones y ponderaciones de cada elemento de la cadena y en su conjunto
- Se evalúa el sentido respecto a la ubicación de las palabras.

En resumen, si la cadena tiene elementos calificados como relativos a datos jurídicos, es un nombre jurídico. Caso contrario se evalúa que sus elementos tengan calificación única de nombres y apellidos.

En caso de identificarse una cadena de persona natural se realiza un análisis complementario:

- Si el nombre o apellido es truncado se reemplaza por su alias.
- Califica nombres que también son apellidos según su ubicación en la cadena quedando recalificados o como nombre o como apellido.
- De acuerdo los nombres se identifican su sexo
- De acuerdo a la ubicación y combinación se identifica el primer nombre y el primer apellido
- Finalmente se elimina letras dobles y se homologa letras según sonido

2.5.9 Normalización de sonidos.

En el idioma castellano existe letras o sílabas que tienen sonido similar. En algunos casos se identificó que el registro de algunos nombres históricos se realizó según su sonido.

Adicionalmente, una forma de acotar el universo de caracteres es homologar el registro de algunos caracteres que tienen sonido similar.

Una propuesta inicial se aprecia en la figura 12.

Por ejemplo, algunas conversiones en lenguaje de programación Cobol para la letra Ñ serian INSPECT DATOS CONVERTING '?"$~+.<>,!_-V{};:%()*ª_á|=›³" TO 'NNNNNNN'.

INSPECT WA-WRD-AUX CONVERTING X'0FA4' TO 'NN'.

Se deben considerar todos los casos posibles tales como proveniencias de otras plataformas que en la transmisión se convierten en caracteres especiales o se trasladan a código hexadecimal.

Como en cada paso, se contrasta contra el mundo real.

Caracteres iniciales	Propuesta
CE	SE
CI	SI
SH	CH
Y	i
&	i

Caracteres iniciales	Propuesta
Blancos	Un blanco
Ñ	N
RR	R
Z	S
Q	K
G	J
LL	L
W	V
V	B

Figura 12: NORMALIZACIÓN DE SONIDOS

En la figura 13 se aprecia algunos nombres normalizados en sonido.

Nombre original	Ejemplos normalización de sonidos
Víctor David Ramírez Llanos	BICTOR DABID RAMIRES LANOS
Diseños gráficos & servicios SAC	DISENOS GRAFICOS I SERBISIOS SAC
Zenón Gonzalo Vega La Torre	SENON GONSALO BEGA LA TORE
Yanet Cecilia Díaz Rivera	IANET SESILIA DIAS RIBERA
Nelly Zoila Carranza Núñez	NELI SOILA CARANSA NUNES
Private Sources Systems s.a.	PRIBATE SOURSES SISTEMS S.A.

Figura 13: EJEMPLOS DE NORMALIZACIÓN DE SONIDOS

2.5.10 Ponderación de palabras relevantes.

Palabras-relevantes son los elementos más representativos de una cadena de elementos nombre.

El orden de relevancia es un trabajo especial que se contrasta contra el mundo real. En la figura 14 se postula el orden de importancia de las agrupaciones descritas en el punto 2.5.5

TIPO DE AGRUPACIÓN	CALIFICACIÓN	PONDERACIÓN
Palabras no encontradas	Z	A
Apellidos	B	A
Apellidos recortados	W	A
Nombres que son apellidos	L	B
Nombres	A	B
Nombres recortados	V	B
Números expresados en palabras	U	C
Actividades en general	C	D
Instituciones diversas	I	D
Actividades económicas	K	D
Palabras relativas a santos	G	E
Títulos y oficios	H	F
Identificadores jurídicos	D	G
Conectores	F	H
Palabras de una sola letra	Y	I

Figura 14: PONDERACIÓN DE AGRUPACIONES DE PALABRAS

En la columna PONDERACIÓN se ha usado letras similares a la columna CALIFICACIÓN pero debe notarse que son atributos distintos.

Como se aprecia, el autor asume que las palabras no encontradas en la tabla de nombres calificados son las palabras más relevantes, seguido de apellidos y nombres.

Los números en todas sus expresiones también son relevantes en especial para el uso de cadenas-dirección como se muestra en los ejemplos prácticos del anexo I.

Las actividades económicas, instituciones diversas y actividades en general tienen una misma ponderación.

Demás agrupaciones tienen ponderación menor como por ejemplo los conectores.

En resumen se estableció las siguientes reglas de comportamiento:

PROBLEMÁTICA	ESTIMACIÓN DE SOLUCIÓN
El universo de palabras es grande	Es susceptible de extraer una muestra representativa y almacenarla
Las dimensiones de las cadenas-nombre son difusas	Es susceptible de estructurar en elementos
La representación de caracteres entre plataformas tecnológicas es diversa	Son susceptibles de homologar los caracteres provenientes de diversas plataformas
Las cadenas-nombre en ciertos casos poseen caracteres innecesarios	Es susceptible de identificar y eliminar los caracteres innecesarios.
Hay variados tipos de palabras relativas a nombres	Las palabras relativas a nombre son susceptibles de agrupar según características comunes.
Las palabras relativas a nombres tienen diversos atributos.	Los atributos de las palabras relativas a nombres son susceptibles de identificar, calificar y almacenar.
Ciertas palabras relativas a nombres tienen alias	Los alias o nombres-sustitutos son susceptibles de identificar y almacenar
Discriminar nombres de personas naturales y personas jurídicas.	Por análisis es susceptible identificar sin una cadena-nombre es persona natural o persona jurídica.
En algunos casos a letras distintas corresponden sonidos similares	Los sonidos son susceptibles de normalizar hacia un número menor de caracteres
En algunos casos los nombres se escriben de diversas maneras.	Los elementos-nombre son susceptibles de representar como seudo-códigos.

FIGURA 15: PROBLEMÁTICA (VS) SOLUCIONES ESTIMADAS

2.6 EVALUACIÓN DEL MODELO Y REGLAS APLICADAS

El proceso de estructuración de los elementos-nombre busca acotar el universo hacia un subconjunto más pequeño y por ende más susceptible de manejar.

La evaluación de las reglas de comportamiento descritas entre los puntos 2.5.1 al 2.5.10 se realiza al generar las claves-nombre que consiste en desarrollar elementos normalizados de referencia cual si fuesen seudo-

códigos. Dichas claves-nombre se almacenan paralelamente a la cadena-nombre original.

La hipótesis es analizar cada cadena-nombre, calificando y ponderando cada uno de sus elementos a fin de identificar las palabras más relevantes y almacenarlas a modo de clave de acceso.

Es discutible la cantidad de elementos de la clave-nombre y su longitud.

Previamente y a modo de repaso se ha procedido a:

ANÁLISIS DE ELEMENTOS

Los elementos en general son el universo de palabras o números que interrelacionados conforman una idea.

Para el caso en estudio dicho universo esta acotado a nombres de personas naturales y jurídicas de la realidad peruana.

Al interior de cada elemento se encuentra:

- Letras
- Números (escrito en letras y/o números)
- Signos especiales
- En el caso específico de nombres consideramos dos grandes bloques:
- Nombres de personas naturales
- Nombres de personas jurídicas
- Los nombres de personas naturales se componen de:
- Nombres
- Apellidos
- Ambos a su vez pueden contener conectores tales como:
- Artículos (ejemplo EL, LA, LOS LAS)
- Pronombres (ejemplo EL, LA, LOS, LAS)
- Interjecciones. (ejemplo Y).

ANÁLISIS DE AGRUPACIONES

El autor definió agrupaciones como subconjuntos de elementos que poseen una característica común en su forma de significado y/o relación con otros elementos. Este concepto es más notorio en el análisis de nombres de personas jurídicas.

Algunas agrupaciones o tipos de palabras son:

- Meses (ejemplo enero, febrero, marzo...diciembre)
- Nombres (ejemplo Juan, María, Pedro, etc.)
- Apellidos (ejemplo Vega, Pérez, Fernández, etc.)
- Tipo empresa jurídica (ejemplo SRL, SA, SAC)
- Títulos (ejemplo Capitán, Coronel, Conde, San, Don etc.)

En general, según el problema a atender, las agrupaciones se infieren analizando visualmente el universo de elementos en cuestión.

ANÁLISIS DE RELACIONES.

Los elementos se escriben a continuación de otros elementos y poseen reglas de formación o relaciones.

Por ejemplo para el nombre de persona jurídica 'Comercial distribuidora 2 de mayo'

Lo usual en lenguaje castellano es que antes de 'mayo' debe estar escrito el día (en este caso '2'), asimismo, después del día se escribe comúnmente la palabra 'de'.

Lo usual es que después de 'tipo negocio' está escrito el nombre del establecimiento y así sucesivamente.

El análisis de relaciones permite inferir las reglas de comportamiento y por ende identificar y extraer las palabras 'relevantes'.

Algunas relaciones son:

Después de un título viene generalmente un nombre, algunos ejemplos:

- Coronel Bolognesi.
- Conde de Superunda

- Santa Rosa de Lima

Los nombres se registran juntos y los apellidos juntos. Para el caso de 'Juan Martín Padilla Espinoza' suele escribirse:

- Juan Martín Padilla Espinoza (nombres y apellidos)
- Padilla Espinoza Juan Martín (apellidos y nombres)

Es poco probable que encontremos registrado como Padilla Juan Espinoza Martín.

En la figura 16 el autor muestra algunas combinaciones..

Apell. Paterno	Apell. Materno	1er nombre	2do nombre	USUAL
1er nombre	2do nombre	Apell. Paterno	Apell. Materno	USUAL
Apell. Paterno	1er nombre	Apell. Materno	2do nombre	DESUSUAL
1er nombre	Apell. materno	Apell. Paterno	2do nombre	DESUSUAL

Figura 16: ORDEN DE NOMBRE-APELLIDOS

Las reglas de formación son fundamentales para analizar una cadena-nombre antes de construir los elementos normalizados.

La identificación de las reglas de formación se infiere del universo de datos a tratar. Inicialmente se plantea una hipótesis. En los criterios se desarrolla una tesis (lógica del programa) y se aplica a un universo de datos de prueba (experimentación). Si el resultado cumple para un nivel de deseado (ejm 99.7%) se confirma su aplicación (comprobación). Caso contrario se desecha.

Se aplica las reglas de comportamiento descritas en los puntos 2.5.1 al 2.5.10.

A modo de ejemplo analizaremos las cadenas-nombre de la figura 17 (una cadena-nombre de persona natural y otra de persona jurídica):

CADENAS-NOMBRE
José Luis de la Riva Agüero y Sánchez Montalvo
Compañía depósito de licores y cerveza el universo srl.

Figura 17: EJEMPLO DE CADENA NOMBRE PERSONA NATURAL Y JURÍDCA

Se procede a:

Dimensionar la longitud de la cadena a analizar.

Se elimina blancos a la izquierda (si hubiesen).

Se convierte la cadena a mayúsculas EBCDIC

Se eliminan los acentos.

Si hay mas de un blanco juntos, se convierten en uno solo.

Se elimina caracteres especiales.

Se convierte la Ñ en N

Se explosiona las palabras

Se analiza la eliminación de apóstrofes.

Como éste, hay varios puntos de análisis especial. Si fuese el caso que la cadena-nombre dijese: Compañía 'El Universo' srl., la eliminación de apóstrofes procede y no afecta, sin embargo si la cadena-nombre dijese: Juan Carlos O'brien Mendoza se requiere una análisis mas refinado (revisar si el apóstrofe es impar, revisar contra la base de datos si esta registrado como nombre o apellido, revisar la posición de la palabra dentro de la cadena etc.)

Hasta aquí tenemos la figura 18:

JOSE	LUIS	DE	LA	RIVA	AGUERO	Y	SANCHEZ	MONTALVO
COMPAÑIA	DISTRIBUIDORA	DE	LICORES	Y	CERVEZA	EL	UNIVERSO	SRL

Figura 18: DESAGREGACIÓN INICIAL DE ELEMENTOS

Seguidamente se accede a la base de datos de conocimientos y se extrae la calificación de cada elemento. En la figura 19 se aprecia los datos calificados según los agrupamientos mostrados en la figura número 8.

Cadena-nombre	JOSE	LUIS	DE	LA	RIVA	AGUERO	Y	SANCHEZ	MONTALVO
Calificación	A	A	F	F	B	B	F	B	B

Cadena-nombre	COMPANÍA	DISTRIBUIDORA	DE	LICORES	Y	CERVEZA	EL	UNiVERSO	SRL
Calificación	K	K	F	C	F	C	F	Z	D

Figura 19: ELEMENTOS DESAGREGADOS Y CALIFICADOS

Cabe notar que la base de datos de conocimientos contiene mas de un 99% de los elementos de la realidad de donde se tomaron los datos muestrales. En el ejemplo se aprecia que la palabra UNIVERSO no fue encontrada en dicha base de datos. Para efectos de la clave-nombre el elemento UNIVERSO tendrá una ponderación mayor y debe notarse que la base de datos es susceptible de enriquecer posteriormente con nuevos elementos previa validación.

Luego de la calificación se procede a analizar si la cadena-nombre corresponde a una persona natural o jurídica.

En los ejemplos arriba señalados se aprecia claramente que la primera cadena-nombre tiene elementos calificados como:

A : nombres

B : apellidos

F : conectores

Por lo tanto se infiere que es un nombre de persona natural.

La segunda cadena-nombre tiene:

C: actividades diversas

D: identificador de nombre jurídico

F: conectores

K: actividades económicas

El peso del elemento D determina que es un nombre jurídico pues aun si el nombre tuviese apellidos ejemplo JUAN PEREZ MENDOZA SRL, se concluyese que es jurídico.

Luego se analiza si alguna palabra tiene la calificación de santo (por ejemplo San Martín, Santa Gadea etc.), esto ayuda a discriminar si una cadena es jurídica o natural, también si es nombre o apellido. Por ejemplo: Juan Pedro San Martín Mendoza, inicialmente SAN sería un tercer nombre pero en el análisis se determina que es parte del apellido.

Seguidamente se procede a ponderar las palabras según sus agrupaciones tal como se explicó en el punto 2.5.10. En la figura 20 se muestra las ponderaciones del ejemplo dado.

Cadena-nombre	JOSE	LUIS	DE	LA	RIVA	AGUERO	Y	SANCHEZ	MONTALVO
Calificación	A	A	F	F	B	B	F	B	B
Ponderación	B	B	H	H	A	A	H	A	A

Cadena-nombre	COMPANIA	DISTRIBUIDORA	DE	LICORES	Y	CERVEZA	EL	UNIVERSO	SRL
Calificación	K	K	F	C	F	C	F	Z	D
Ponderación	D	D	H	D	H	D	H	A	G

FIGURA 20: ELEMENTOS NOMBRE CALIFICADOS Y PONDERADOS

En resumen, las palabras 'DE', 'Y', 'LA' poseen peso inferior, los tipos de negocio y tipos de empresa ('SRL') tienen un peso medio- bajo. Las palabras 'DEPÓSITO', 'LICORES' Y 'CERVEZA' tienen un peso medio y la palabra 'UNIVERSO' tiene el peso mayor.

Para el caso de una persona natural lo más relevante es el 'Primer apellido', seguido del 'Primer nombre', seguido del 'segundo apellido' y finalmente el 'segundo nombre'.

En la figura 21 se muestra como se procede a ordenar las palabras según ponderación o importancia.

Cadena-nombre	RIVA	AGÜERO	SÁNCHEZ	MONTALVO	JOSE	LUIS	DE	LA	Y
Calificación	B	B	B	B	A	A	F	F	F
Ponderación	A	A	A	A	B	B	H	H	H

Cadena-nombre	UNIVERSO	COMPANIA	DISTRIBUIDORA	LICORES	CERVEZA	SRL	DE	Y	EL
Calificación	B	K	K	C	C	D	F	F	F
Ponderación	A	D	D	D	D	G	H	H	H

FIGURA 21. ELEMENTOS NOMBRE ORDENADOS SEGÚN PONDERACIÓN

Finalmente en la figura número 22 se elimina letras dobles, se procede a cambiar letras según normalización de sonidos explicado en el punto 2.5.9 y se presenta según ponderación o relevancia. Para el caso de nombre de persona natural se aplica la secuencia 1er apellido, 1er nombre, 2do apellido, 2do nombre.

RIBA	JOSE	AGUERO	LUIS
UNIBERSO	COMPANIA	DISTRIBUIDORA	LICORES

Figura 22: ELEMENTOS-NOMBRE MAS RELEVANTES

2.7 FORMULACIÓN DEL SOPORTE MECANIZADO.

Aprovechando las bondades de la tecnología, las propuestas realizadas se plasman en un soporte mecanizado.

Previamente el autor define:

- Elección de una plataforma de cómputo.

- Definir el alcance a desarrollar
- Dotar de soportes de retroalimentación

2.7.1 Plataforma de cómputo

El autor elige una plataforma que satisfaga la mayor amplitud de requerimientos del mercado. Se plantea un lenguaje C de programación para redes en general o lenguaje COBOL para mainframes.

Con un diseño en tres capas se superan las variedades de bases de datos, servidores y comunicaciones.

2.7.2 Alcance a desarrollar

El autor plantea desarrollar rutinas mínimas a modo encapsulado para que puedan ser invocadas desde diversos programas aplicativos.

Las aplicaciones son de diversa índole y más que atender un problema de nombres el autor pretende atender la problemática de palabras en general. En el anexo I se aprecia su uso para el problema de direcciones.

Debe notarse que los criterios dependen de la realidad muestral que se tome, en el presente caso se toma la realidad del idioma castellano en el Perú.

2.7.3 Soportes de retroalimentación.

Un sistema estático tiende a desaparecer. La propuesta del autor presenta un conjunto de procesos para realimentar el soporte informático.

Debe notarse que cada criterio o propuesta se evalúa exhaustivamente contra la realidad por lo cual los programas deben estar aptos para ajustarse o afinar en forma paramétrica evitando modificar los códigos fuente en la medida de lo posible.

2.7.4 Diseño informático.

En el anexo II se muestra el diseño informático que contiene:

- Diseño de tablas y llaves de acceso
- Relación de rutinas básicas
- Relación de programas de administración

Cabe notar que la presente propuesta se ofrece a instalaciones informáticas que ya vienen operando, es decir, el autor propone un juego mínimo de entidades (archivos y programas) que permitan su explotación inmediata desde cualquier instalación existente.

En tal sentido la implantación sigue los siguientes pasos:

1. Cargar en la base de datos las tablas de base de nombres calificados.
2. Reservar 48 bytes en el archivo central de nombres de la institución.
3. Mediante un proceso batch generar las claves-nombre de cada nombre de la institución llenando los 48 bytes reservados.
4. Registrar en la base de datos las claves de acceso para explotación de la clave- nombre.
5. Incluir la rutina de generación de clave-nombre en los programas aplicativos de la institución que actualizan nombres.
6. Finalmente, se incluye la rutina básica de búsqueda o comparación en tantos programas aplicativos como sean menester.

Una muestra práctica de estos pasos se muestra a continuación:

1. Sea la estructura inicial de la tabla de nombres de la instalación:

```
01 NOMBRES
   02 NOMBRE                          CHAR(120),
   02 ULTIMA_MODIFICACIÓN,
                  03 FECHA            CHAR(08),
                  03 HORA             CHAR(08),
                  03 USUARIO          CHAR(10).
```

2. Se procede a reservar 48 bytes para la clave-nombre:

```
01 NOMBRES
   02 NOMBRE                          CHAR(120),
   02 CLAVE_PARCIAL1                  CHAR(08),
   02 CLAVE_NOMBRE                    CHAR(32),
   02 CLAVE_PARCIAL2                  CHAR(08),
02 ULTIMA_MODIFICACIÓN,
                  03 FECHA            CHAR(08),
                  03 HORA             CHAR(08),
                  03 USUARIO          CHAR(10).
```

3. Se registra en la base de datos las claves alternas:

```
01 NOMBRES
   02 NOMBRE                          CHAR(120),
   02 CLAVE_PARCIAL1                  CHAR(16),
   02 FILLER                          CHAR(16),
   02 CLAVE_PARCIAL2                  CHAR(16),
  02 ULTIMA_MODIFICACIÓN,
                  03 FECHA            CHAR(08),
                  03 HORA             CHAR(08),
                  03 USUARIO          CHAR(10).
```

De este modo se puede explotar la clave-invertida para los nombres de personas naturales, es decir por ejemplo JUAN PEREZ sería equivalente a PEREZ JUAN o también MORALES BERMUDEZ es equivalente a BERMUDEZ MORALES.

La rutina básica de generación de clave-nombre se incluye en las rutinas en línea o batch que registran o modifican los nombre. De este modo se asegura que cada nombre tenga su respectiva clave-nombre. Obviamente, se estima que hay una sola tabla central de nombres para evitar las anomalías de inserción y de actualización.

Finalmente se incluye la rutina básica de búsqueda o comparación en los programas que así lo requieran tales como consulta alfabética de clientes, cruce para identificar nombres duplicados, programas para separar nombres y apellidos (normalizar) etc.

2.8 VALIDACIONES.

Como se mostró en la figura número 3 la metodología exige la constante validación de los criterios contra la realidad.

Considerando que el universo de nombres es infinito, el autor sugiere un límite de aceptación o confianza superior a 99%. Es decir cada hipótesis se valida que cumpla o satisfaga mas del 99% de casos reales.

2.8.1 BASE DE DATOS DE NOMBRES

El punto 2.3 trata del acopiamiento de nombres de personas naturales y jurídicas. En la práctica el autor realizó una recolección de nombres de una

institución financiera peruana con 700,000 clientes y posteriormente realizó un análisis de frecuencia para ordenar de mayor a menor los elementos mas repetidos.

Seguidamente, procedió a una calificación manual de agrupamiento y sexo.

Cabe notar que con 45,209 elementos-nombre ingresados se alcanzó un nivel de rendimiento de 99.17%, es decir el 99.17% de cada nombre elegido al azar de ésta u otra institución peruana ya encontraba sus elementos registrados con un nivel de confianza de 99.17%.

Se comprobó acopiando nombres de otras instituciones de distinto giro comercial y sometiéndolos a un proceso mecanizado donde se corroboró el nivel de confianza alcanzado.

Debe notarse que la tabla de nombres con 45,209 elementos calificados ocupa tan solo 1.87 megabytes lo cual lo hace asequible para su explotación en instalaciones que posean limitada infraestructura de cómputo.

2.8.2 DIMENSIONAMIENTO DE CADENAS-NOMBRE

Se validó que para analizar los nombres, éstos deben estar compuestos por una tira continua de palabras separadas al menos por un blanco. Es indistinto el orden, ya sea nombre y apellido o viceversa pues la lógica diseñada permite distinguir plenamente un nombre de un apellido y almacenarlos en arreglos.

El autor sugiere una longitud óptima de 120 posiciones para una cadena-nombre.

2.8.3 TRATAMIENTO DE CARACTERES ESPECIALES.

Como los nombres estaban almacenados en diversas plataformas, rápidamente se apreció el problema de conversión de caracteres entre plataformas y finalmente se concluyó estandarizar hacia código EBCDIC de mainframe.

Sin embargo, en dicho archivo de nombres el autor registró los nombres tal cual fueron recibidos, esto para respetar las fuentes origen donde los nombres aun permanecían registrados de tal modo. En la figura 22 se aprecia un segmento del archivo donde se aprecia la problemática de la letra 'Ñ.

En general se corroboró las hipótesis planteadas:

Conveniencia de estandarizar el uso de mayúsculas.

Conveniencia de hacer las conversiones necesarias para eliminar los acentos.

Conveniencia de eliminar los caracteres especiales tales como: '?¿+-<>{}[]V¡!,;.:_"%*=|°^¬#$~' '?¨$~+.<>,!_-V{};:%()*ª_á|=›³"'

Dichas conversiones no afectaron los nombres en si, mas bien acotaron el universo a un número menor de caracteres con lo cual disminuye el universo de combinaciones.

NOMBRE	CALIFICACIÓN
_AHUINRIPA	B
_AHUIS	B
_ARQUEZ	B
_ATO	B
_AUPA	B
_AUPARI	B
_AUPAS	B
_AURI	B

Figura 23: LETRAS 'Ñ' MAL CONVERTIDAS

2.8.4 AGRUPACIONES Y PONDERACIONES.

El autor desarrolló las agrupaciones en función de atributos comunes y su utilidad se validó a tiempo de la generación de claves-nombre donde facilitó los criterios de programación.

Cabe notar que las agrupaciones deben responder a la realidad. En el trabajo realizado los registros de nombres en muchos casos contenían nombres reducidos por lo cual se hizo necesario crear una agrupamiento para estos casos que posteriormente se validó contra tarjetas manuales de registro de nombres y de ser el caso se registró la marca de error como se aprecia en la figura 24.

ELEMENTO	CALIFICACIÓN	ERROR
AYQUIP	W	
AYQUIPA	B	
AZABACH	W	X
AZABACHE	B	
BACILI	V	X
BACILIA	A	
BACILIO	A	
BEATRI	V	X
BEATRIZ	A	

Figura 24: ELEMENTOS-NOMBRE MARCADOS CON EXCEPCIÓN

En el caso de nombres que son apellidos, se calificó según lo mas frecuente, como nombre (calificación A) o como apellido (calificación B). En los casos en que no había una mayoría contundente se calificó como 'L' como se aprecia en la figura 25.

ELEMENTO	CALIFICACIÓN
BALERIO	L
BARTOLO	L
BARY	L
BARRY	A
BASILIO	L
BAUDELIO	A
BAUTISTA	B
BEATO	A
BEBELU	B

FIGURA 25 CALIFICACIÓN DE NOMBRES-APELLIDOS

Se validó que las agrupaciones no solo apoyan en el tratamiento de nombres sino palabras en general, tal el caso de elementos-dirección que se trata en el anexo I y cuyos agrupamientos de aprecian en la figura 26.

AGRUPACIÓN	CALIFICACIÓN	EJEMPLOS
TIPOS DE VIA	M	Jirón, calle, óvalo, avenida, pasaje etc.
TIPOS DE DIRECCIÓN	N	Apartado, block, cuadra, manzana, etc.
ZONAS	O	Complejo, fundo, habitacional etc.
DISTRITOS	S	Lince, Breña, Ate, La Victoria etc.

FIGURA 26: AGRUPAMIENTOS DE ELEMENTOS-DIRECCIÓN

3.8.5 EMPLEO DE ALIAS

El autor corroboró lo delicado que es generalizar un criterio con el fin de satisfacer un caso específico. En la práctica un exhaustivo análisis contra data real rechazó varios supuestos que si bien solucionaban ciertos casos provocaban confusiones colaterales. Tal es el caso de algunos alias que merecieron pruebas variadas y finalmente fueron rechazadas como por ejemplo se aprecia en la figura 27.

ELEMENTO	ALIAS	VALIDACIÓN
GMO	GUILLERMO	
GUILLERMO		
JHONI	JHONNY	
JHONIL	JHONNY	RECHAZADO
JHONN	JHONNY	RECHAZADO
JHONNE	JHONNY	RECHAZADO
JHONNI	JHONNY	
JHONNY		

JHONY	JHONNY	

FIGURA 27: ALIAS DE NOMBRES RECHAZADOS

3.8.6 NORMALIZACIÓN DE SONIDOS.

El autor realizó diversas pruebas para corroborar o rechazar ciertas homologaciones de sonido. En la práctica algunos casos particulares provocaban problemas colaterales por lo cual se terminó denegando su generalización.

El autor destaca que esta reducción acorta el número de letras, por ende el tamaño de posibilidades del universo. Por ejemplo las permutaciones de las palabras con 4 letras son: 28! / (28 – 4)! = 28 x 27 x 26 x 25 = 491,400 pero si disminuyésemos 6 letras del alfabeto se disminuye a: 22! / (22 – 4)! = 22 x 21 x 20 x 19 = 175,560 posibilidades lo cual acorta el universo.

Caracteres iniciales	Normalizado	Contraste mundo real
CE	SE	Ok
CI	SI	Ok
SH	CH	rechazado
Y	i	Ok
&	i	Ok
Blancos	Un blanco	Ok
Ñ	N	Ok
RR	R	Ok
Z	S	Ok
Q	K	rechazado
G	J	rechazado
LL	L	Ok
W	V	rechazado
V	B	ok

FIGURA 28: HOMOLOGACIONES DE SONIDOS RECHAZADOS

3.8.7 GENERACIÓN DE CLAVE-NOMBRE.

Experimentalmente el autor corroboró los atributos de la clave nombre:

Número de elementos:

La conveniencia de manejar un juego de pocos elementos facilita su administración siempre y cuando no se pierda efectividad. Para el caso de nombres de personas naturales se concluyó que con 4 elementos se alcanza un óptimo. Esta cantidad sirve para incluir dos nombres y dos apellidos que en la realidad coinciden con mas del 90% de los casos de nombres.

Esta comprobación es también aplicable a nombres de personas jurídicas.

Longitud de cada elemento:

La experimentación realizada por el autor concluye que con 8 caracteres se alcanza un óptimo entre menor tamaño y representatividad, es decir así ampliásemos a 10, 12 o 14 caracteres la longitud de cada elemento normalizado el grado de eficiencia sería el mismo.

En al figura 29 se aprecia algunos segmentos de nombres de personas naturales y su clave-nombre y en la figura 30 se aprecia de personas jurídicas.

CADENA-NOMBRE	1ER. APELLIDO	1ER. NOMBRE	2DO. APELLDIO	2DO NOMBRE
NAZARIA TERESA YIMURA YIMURA DE	IIMURA	NASARIA	IIMURA	TERESA
SABINE PIA LUMBRERAS HORNUNG	LUMBRERA	SABINE	HORNUNG	PIA
ANA INES REATEGUI VELA	REATEGUI	ANA	BELA	INES
CARLOS ALBERTO ZAPATER CATERIANO	SAPATER	CARLOS	CATERIAN	ALBERTO
ARMANDO JESUS MORENO VELAZCO	MORENO	ARMANDO	BELASCO	JESUS
JOSE ANTONIO ROBLES FLORES	ROBLES	JOSE	FLORES	ANTONIO
ENRIQUE JAVIER PAZ ESCRIBENS PAS	PAS	ENRIQUE	ESCRIBEN	JABIER
WILLIAM JULIO OBREGON MENDOZA	OBREGON	WILIAM	MENDOSA	JULIO
MARIELA ISABEL CAMARGO ROMAN	CAMARGO	MARIELA	ROMAN	ISABEL
ALEX ROBERTO ALBUJAR CRUZ	ALBUJAR	ALEX	CRUS	ROBERTO

JORGE MARTIN SANTANA ORMEÑO	SANTANA	JORGE	ORMENO	MARTIN
ADA GEORGINA AMPUERO CARDENAS	AMPUERO	ADA	CARDENAS	GEORGINA
ELENA BEATRIZ GAZZO VERAND	GASO	ELENA	BERAND	BEATRIS
DAVID RITCHIE BALLENAS	RITCHIE	DABID	BALENAS	
TOMAS ALBERTO MINAURO LA TORRE	MINAURO	TOMAS	TORE	ALBERTO
MARGARITA MARIA PONCE DE LEON ROS	LEON	MARGARIT	PONSE	MARIA

FIGURA 29: EJEMPLOS CLAVE-NOMBRE DE PERSONAS NATURALES

CADENA-NOMBRE	1RA. PALABRA	2DA. PALABRA	3RA. PALABRA	4TA. PALABRA
INST PERUANO DE ADMINISTRACION DE EMPRESAS IPAE	IPAE	INSTITUT	PERUAN	ADMINIST
INST SUPERIOR DE ESTUDIOS TEOLOGICOS JUAN XXIII	XXIII	JUAN	INSTITUT	SUPERIOR
INST. PERUANO DE PATERNIDAD RESPONSABLE INPPARES	PATERNID	RESPONSA	INPARES	INSTITUT
AGENCIA DE VIAJES Y TURISMO TACNA TRAVEL INTERNACIONAL SERVICE S.R.L.	TACNA	TRABEL	AGENSIA	BIAJE
CAMINO AGENCIA DE VIAJES Y SERVICIOS	CAMINO	AGENSIA	BIAJE	SERVISIO
CAMINOS DEL INKA IMPORT & EXPORT E.I.R.L.	INKA	CAMINOS	IMPORTA	EXPORTA
CENTRO RESERVAS AREQUIPA S.A.C.	AREQUIPA	SENTRO	RESERBA	SAC
CESAR`S TRAVEL SERVICE S.R. LTDA.	SESARS	TRABEL	SERBISE	SRL
ILUSIONES AGENCIA DE VIAJES Y TURISMO S.A.	ILUSIONE	AGENSIA	BIAJE	TURIS
IMPERIO SERVICIOS Y	IMPERIO	SERBISIO	REPRESEN	EIRL

REPRESENTACIONES E.I.R.L.				

FIGURA 30: EJEMPLOS DE CLAVE-NOMBRE DE PERSONAS JURÍDICAS

Se debe notar que en la segunda fila de la figura 30 la palabra PERUANO se estandarizó a su alias de la tabla de nombres calificados que es PERUAN así como otras palabras que aparecen con menos de 8 caracteres pues su alias así lo indica. Este tipo de a consideración ayuda a acortar el universo de combinaciones.

3.8.8 LONGITUD A TIEMPO DE COMPARACIONES

A tiempo de comparaciones de nombres de personas naturales los procesos de validación concluyeron que entre 20 y 18 caracteres se obtiene óptimos resultados.

El autor realizó diversas simulaciones combinatorias como por ejemplo:

Considerar 8 bytes de los cuatro elementos de la clave

Considerar 7 bytes de los cuatro elementos de la clave

Considerar 8 bytes de los tres primeros elementos de la clave

Considerar 7 bytes de los tres primeros elementos de la clave

Considerar 6 bytes de los cuatro elementos de la clave y diversas combinaciones más.

En la práctica tomando 6 caracteres de los tres primeros elementos de la clave son suficientes para alcanzar altos rendimientos en comparaciones de nombres de personas naturales (es decir no se toma en cuenta los dos últimos bytes de los tres primeros elementos y el 4to elemento no se considera).

Esta experimentación también se validó en los nombres de personas jurídicas obteniéndose igualmente resultados óptimos.

El autor remarca que la flexibilidad de la herramienta propuesta permite manejar mejor el recurso según el problema a atender.

Imaginemos que recibimos una lista de nombres que requerimos comparar contra otra lista. En primer lugar ambos archivos de nombres deben tener

generadas sus claves-nombre según lo explicado. En la figura número 31 se aprecia dichas claves:

CADENA-NOMBRE	CLAVE PARTE 1	CLAVE PARTE 2	CLAVE PARTE 3	CLAVE PARTE 4
JOSE LUIS BUSTAMANTE Y RIVERO	BUSTAMAN	JOSE	RIBERO	LUIS
IMPERIO SERVICIOS Y REPRESENTACIONES E.I.R.L	IMPERIO	SERBISIO	REPRESEN	EIRL

FIGURA 31: ELEMENTOS DE CLAVES-NOMBRE

Finalmente para obtener mejores resultados en la comparación se emplea solo 18 caracteres (los 6 primeros caracteres de los tres primeros elementos). En la figura 32 se aprecia las claves-nombre a comparar.

CADENA-NOMBRE	CLAVE PARTE 1	CLAVE PARTE 2	CLAVE PARTE 3	CLAVE PARTE 4
JOSE LUIS BUSTAMANTE Y RIVERO	BUSTAM	JOSE	RIBERO	
IMPERIO SERVICIOS Y REPRESENTACIONES E.I.R.L.	IMPERI	SERBIS	REPRES	

FIGURA 32: ELEMEMTOS DE LA CLAVE-NOMBRE EN 6 CARACTERES

La justificación es clara pues al haber acortado los nombres se acorta la posibilidad de diferencias en los caracteres finales sin perder exactitud. Por ejemplo si un apellido dijese BUSTAMANTERO en vez de BUSTAMANTE afectaría poco pues para una mala comparación también tendrían que coincidir los otros elementos JOSE y RIVERO lo cual es poco probable.

3.9 APLICACIONES PRACTICAS.

En la introducción se señaló diversas aplicaciones prácticas. A modo demostrativo el autor explica algunos casos.

CASO 1 – NORMALIZACIÓN DE NOMBRES.

En una institución los nombres registrados no respetan necesariamente el orden apellido-nombre, tampoco se dispone de nombres desagregados en el ingreso inicial y en el almacenamiento.

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-nombre

Se emplea 48 bytes del archivo de nombres de la institución para generar la clave-nombre de cada registro.

Se invoca la rutina generadora de clave-nombre desde los programas de la institución que tratan el ingreso inicial y el almacenamiento de nombres.

Inicialmente se tenía registros como se muestra en la figura 33.

LUIS ALBERTO FERNÁNDEZ DE CORDOVA LLERENA
TENORIO GARCIA JUAN JOSE
JUAN PABLO SANTA CRUZ FERNÁNDEZ
MARIA DEL PILAR NÚÑEZ MARTINEZ
JUAN DE DIOS PIZARRO DIAZ
CARMEN LUCY ALZA MOSTACERO

FIGURA 33: NOMBRES A NORMALIZAR

Finalmente los nombres quedaron reordenados como apellido, nombre y separados en columnas distintas como se aprecia en la figura 34. Nótese que esta desagregación no es la clave-nombre.

NOMBRE REORDENADO	PRIMER APELLIDO	SEGUNDO APELLIDO	PRIMER NOMBRE	SEGUNDO NOMBRE
FERNANDEZ DE CORDOVA LLERENA LUIS ALBERTO	FERNANDEZ DE CORDOVA	LLERENA	LUIS	ALBERTO
TENORIO GARCIA	TENORIO	GARCIA	JUAN	JOSE

JUAN JOSE				
SANTA CRUZ FERNANDEZ JUAN PABLO	SANTA CRUZ	FERNANDEZ	JUAN	PABLO
NUÑEZ MARTINEZ MARIA CECILIA	NUÑEZ	MARTINEZ	MARIA	CECILIA
PIZARRO DIAZ JUAN DE DIOS	PIZARRO	DIAZ	JUAN	DIOS
ALZA MOSTACERO CARMEN LUCY	ALZA	MOSTACERO	CARMEN	LUCY

FIGURA 34: NOMBRES NORMALIZADOS

CASO 2 – NOMBRES CORTOS DE PERSONAS JURÍDICAS.

En una institución se requiere tener un nombre corto u abreviado para imprimir en los vouchers que se entregan en las ventanillas de atención al publico o en los cajeros automáticos o en general en ciertos reportes.

Las reglas de abreviación en 20 caracteres están normadas mediante una directiva de Organización y Métodos pero en la práctica algunos empleados la incumplen ocasionando registros disímiles.

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-nombre

Se carga el archivo de palabras abreviadas normadas (ejemplo cia, corp, ind, serv, etc.)

Se añade al programa de ingreso inicial de nombres cortos la invocación de la rutina generadora de clave-nombre y otra rutina que accede al archivo de palabras abreviadas normadas.

Inicialmente se tenia registros tales como los de la figura 35.

NOMBRE LARGO	NOMBRE CORTO	OBSERVACIÓN
COMPAÑÍA INDUSTRIAL LA ENSENADA S.A.	COMP IND LA ENSENAD	El abreviado de COMPAÑÍA debe ser CIA.
INVERSIONES Y SERVICIOS NUEVA ERA SAC	INVER Y SERV NUEVA E	El abreviado de INVERSIONES debe ser INV
CORPORACIÓN NACIONAL LA VIDA SRL	CORP. NAC LA VIDA SR	Se truncó última palabra.

FIGURA 35: NOMBRES CORTOS INICIALES

Finalmente los nombres cortos son propuestos por el software disminuyendo las diferencias de registro como se muestra en la figura 36.

NOMBRE LARGO	NOMBRE CORTO
COMPAÑÍA INDUSTRIAL LA ENSENADA S.A.	CIA IND LA ENSENADA
INVERSIONES Y SERVICIOS NUEVA ERA SAC	INV Y SERV NUEVA ERA
CORPORACIÓN NACIONAL LA VIDA SRL	CORP NAC LA VIDA SRL

FIGURA 36: NOMBRES CORTOS PROPUESTOS

CASO 3 – BÚSQUEDA DE NOMBRE INVERTIDO.

En las instituciones se dispone de búsquedas de listas convencionales y en algunos casos se requiere mas iteraciones para ubicar determinado nombre, tal el caso de nombre invertidos.

Por ejemplo, en una institución financiera el cliente requerido se llama ROBERTO EMILIO MARTINEZ MOROSINI y se busca el informe crediticio de ROBERTO MARTINEZ MORISINI pero se encuentra registrado como EMILIO MARTINEZ o como MOROSINI MARTINEZ.

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-nombre

Se emplea 48 bytes del archivo de nombres de la institución para generar la clave-nombre de cada registro del archivo de nombres.
Como el autor explicó en el punto 2.8.7 la clave-nombre mantiene el orden que se muestra en la figura 37:

1ER. APELLIDO	1ER. NOMBRE	2DO. APELLIDO	2DO. NOMBRE
MARTINEZ	ROBERTO	MOROSINI	EMILIO

FIGURA 37: NOMBRE DE EJEMPLO

Por lo tanto, se añadió dos claves alternantes al principio y al final de la clave-nombre y facilitan las búsquedas invertidas como se aprecia en la figura 38.

2DO. APELLIDO	1ER. APELLIDO	1ER. NOMBRE	2DO. APELLIDO	2DO. NOMBRE	1ER APELLIDO
	MARTINEZ	ROBERTO	MOROSINI	EMILIO	

CLAVE ALTERNANTE1				CLAVE ALTERNANTE2	
2DO. APELLIDO	1ER. APELLIDO	1ER. NOMBRE	2DO. APELLIDO	2DO. NOMBRE	1ER APELLIDO
MOROSINI	MARTINEZ	ROBERTO	MOROSINI	EMILIO	MARTINEZ

FIGURA 38: CLAVES PARA BÚSQUEDA DE NOMBRE INVERTIDO

De este modo al ingresar ROBERTO MARTINEZ MOROSINI el ordenador accesa por clave en forma rápida casos variados tales como se muestran en la figura número 39.

MARTINEZ MOROSINI ROBERTO EMILIO
MOROSINI MARTINEZ ROBERTO EMILIO
MARTINEZ MOROSINI ROBERTO
ROBERTO EMILIO MARTINEZ MOROSINI

FIGURA 39: NOMBRES INVERTIDOS ENCONTRADOS

CASO 4 – BÚSQUEDA CON ALIAS.

En las instituciones financieras los funcionarios de créditos algunas veces no tienen en mente el nombre exacto de un cliente, por lo tanto realizan varias iteraciones en el ordenador buscando el nombre deseado.

Por ejemplo, un funcionario de crédito requiere información de un cliente y recuerda las palabras COMERCIAL IMPORTADORA CUZCO.

Cabe notar que mayormente las instituciones tienen desarrolladas búsquedas convencionales tales como:

Búsqueda exacta (buscará exactamente la tira COMERCIAL IMPORTADORA CUZCO).

Búsqueda aproximada (buscará toda cadena que contenga COMERCIAL IMPORTADORA CUZCO y algo mas).

Búsqueda por sonido similar (buscará CUZCO escrito con 's' y con 'z').

A la solución propuesta el autor le denomina 'búsqueda variada' y combina las anteriores búsquedas mas el empleo de alias que se definieron en la base de datos de nombres calificados (ver punto 2.5.7).

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-nombre

Se emplea 48 bytes del archivo de nombres de la institución para generar la clave-nombre de cada registro del archivo de nombres.

Se modifica el programa de búsqueda de nombre de la institución haciendo que invoque la rutina de clave-nombre.

En la figura número 40 se muestra un posible resultado de la búsqueda mencionada.

COMERCIAL IMPORT. CUZCO SRL.
COMERCIAL IMPORTADORA CUZCO OMBLIGO DEL MUNDO SAC
COMERCIALIZACION E IMPORTACIÓN DEL CUZCO SA

COMERCIALIZADORA IMPORTACION EXPORTACION CUZCO
COMERCIANTES DE IMPORTACIONES CUZCO SAC
ASOCIACION COMERCIALIZADORA E IMPORTADORA CUZCO
GALERIA COMERCIAL EL CUZCO IMPORTACIONES
COMERCIANTES REUNIDOS CUZCO IMPORT LTD

FIGURA 40: RESULTADO EN BÚSQUEDA POR ALIAS.

Debe notarse que con el buen diseño de claves establecido la búsqueda es directa y no es como otros procesos que realizan diversos recorridos en la base de datos consumiendo tiempo y recursos de máquina.

CASO 5 – COMPARACIONES DE LISTAS POR LOTES (BATCH)

Realizar comparaciones de nombres a modo convencional alcanza resultados muy limitados debido a las combinaciones variadas que se presentan, es mas, requeriría amplios recursos de máquina y grandes lógicas combinatorias. Sin embargo si tenemos los nombres normalizados mediante la clave-nombre la tarea se torna sencilla con pocos recursos de máquina y alto rendimiento.

Un ejemplo común es la desunificación de nombres, por ejemplo, un cliente tiene diversas cuentas y éstas apuntan al mismo código de cliente. Sin embargo cabe la posibilidad que alguna cuenta fue relacionada indebidamente y exige un proceso de validación.

Otro ejemplo puede ser que se recibe una base externa de clientes potenciales y se quiere validar si éstos ya están registrados en la base de nombres de la institución y no se confía del registro de los documentos de identidad.

Mensualmente las entidades financieras envían a la Superintendencia de Banca y Seguros su relación de clientes y cuentas (Registro crediticio de deudores) y la Superintendencia consolida las colocaciones por cliente a

nivel del sistema financiero y devuelve la información 60 o más días después desconociendo el autor los motivos de esta demora.

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-nombre.

Se instala la rutina de comparación batch entre dos listas.

Se emplea 48 bytes del archivo de nombres de la institución para generar la clave-nombre de cada registro de los archivos a comparar.

Se crea un proceso que invoca la rutina de comparación batch para los archivos deseados.

El autor realizó comparaciones de un archivo central de nombres de 500,000 registros contra listas de 5,000 registros obteniendo resultados en 3 minutos en promedio en un mainframe S/390 de IBM.

Adicionalmente la constatación manual confirmó en diversas comparaciones de archivos una eficacia superior al 99% de certeza.

Como el autor expuso en el punto 2.8.8, el uso de la clave-nombre permite ajustar la herramienta a niveles de precisión y exigencia convenientes empleando de las 32 posiciones de la clave-nombre solo 20, 18, 16 o menos caracteres según se requiera.

CASO 6 – COMPARACIONES DE LOTES APLICANDO CLAVE-DIRECCIÓN (BATCH).

El propósito del autor mas que mostrar herramientas para el tratamiento de nombres es proponer una metodología para el tratamiento de palabras en general.

En tal sentido en el anexo I se muestra la aplicación de la misma metodología para la atención de la problemática de direcciones.

En una institución financiera del Perú se tenia registros de 100,000 clientes jurídicos donde se quería validar si el código de contribuyente registrado correspondía realmente a los registros de la SUNAT.

La alternativa era recibir el archivo magnético de la SUNAT y proceder a una revisión visual de reportes comparativos. Se estimó que demandaría un equipo de 8 personas por una semana.

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-dirección.

Se instala la rutina generadora de clave-nombre.

Se instala la rutina de comparación entre dos listas siendo la misma para clave dirección y para clave-nombre.

Se emplea 48 bytes de una copia del archivo de nombres de la institución para generar las claves-nombre y clave-nombre de cada registro de los archivos a comparar.

Se crea un proceso que invoca la rutina de comparación batch para los archivos deseados.

Debe notarse que la clave-dirección también ocupa 48 bytes de almacenamiento pero en este caso solo se generó en memoria para la comparación.

El autor realizó comparaciones de un archivo central de contribuyentes de 800,000 registros aproximadamente contra una lista de 100,000 registros.

El proceso demoró 7 minutos y emitió tres reportes propuesta

- Clientes con el mismo RUC y nombre y dirección similar.
- Clientes con el mismo RUC y nombre pero dirección distinta.
- Clientes con el mismo RUC y nombre distinto.

El equipo de 8 personas revisó manualmente los reportes y en un día ratificó la eficacia superior al 99% del proceso mecanizado.

Cabe mencionar que este proceso permitió validar la aplicación conjunta de dos herramientas que complementadas proveen mayor grado de confianza en los resultados.

CASO 7 – NORMALIZACIÓN DE DIRECCIONES.

Esta aplicación corresponde a la metodología descrita en el anexo I y consistió en atender entre otros la siguiente problemática:

El departamento de Calidad de la institución normó que la apertura de datos clientes no debía demorar mas de 11 minutos. La dirección de los clientes se registraba en casilleros normalizados de una pantalla lo cual consumía tiempo de registro. La pantalla de ingreso contenía entradas tal como se muestran en la figura 41.

TIPO-VIA	NOMBRE-VIA	NUMERACIÓN	TIPO-ZONA	NOMBRE-ZONA	DISTRITO

SELECCIONAR:

ALAMEDA		APARTADO	A.A.H.H.		ANCON
AV		BLOCK	CASERIO		ATE
CALLE		CHALET	CONJUNTO HAB.		BELLAVISTA
CARRET.		CUADRA	COOP. VIVIENDA		BREÑA
JR. .		DPTO.	LOTIZACION		CALLAO
MALECÓN		ESQUINA	PUEBLO JOVEN		CARMEN LEGUA
OVALO		GRUPO	RESIDENCIAL		CERCADO
PJE.		LOTE	UNIDAD VECINAL		LA PERLA
PLAZA		MANZANA	URB		LA PUNTA
PQUE.		TIENDA	ZONA INDUST.		VENTANILLA
ETC.		ETC.	ETC.		ETC.

FIGURA 41: CASILLEROS PARA REGISTRO DE DIRECCIONES

SOLUCIÓN:

Se instala los archivos de nombres calificados (menos de 2 megabytes de espacio).

Se instala la rutina generadora de clave-dirección.

Se añade la rutina generadora de clave-dirección al programa de ingreso de direcciones.

Finalmente se consigue:

- Una tira dirección ingresada es automáticamente explosionada en sus partes de forma tal que el operador se limita a corroborar visualmente.
- La máquina estandariza el uso de abreviaciones evitando la heterogeneidad de registros.
- Se acorta el tiempo de ingreso de información.

En la práctica el autor confirmó una exactitud superior al 99% en las desagregaciones.

En la figura número 42 se puede apreciar algunos ejemplos de desagregación automática de direcciones.

TIRAS DE DIRECCIONES
CALLE ANDREA DEL SARTO 285 URBANIZACON VIPEP SURQUILLO
URB. ALAMOS II AVENIDA CENTRAL MZ. I LOTE 7 SURCO
JIRON HUALLAGA 739 INTERIOR 112 LIMA CERCADO
ASENTAMIENTO HUMANO SAN JOSE CARRETERA CENTRAL S/N ATE
UNIDAD VECINAL MIRONES BLOCK 5 DPTO. 6 CERCADO

TIRAS DIRECCIÓN DESAGREGADAS					
TIPO-VIA	NOMBRE-VIA	NUMERACIÓN	TIPO-ZONA	NOMBRE-ZONA	DISTRITO
CALLE	ANDREA DEL	285	URB.	VIPEP	SURQUILLO

	SARTO				
AV.	CENTRAL	MZ. I LOTE 7	URB.	ALAMOS II	SANTIAGO DE SURCO
JR.	HUALLAGA	739 INT. 112			CERCADO
CARRET.	CENTRAL	S/N	A.A.H.H.	SAN JOSE	ATE
		BLOCK 5 DPTO. 6	UNIDAD VECINAL	MIRONES	CERCADO

FIGURA 42: DIRECCIONES DESAGREGADAS EN SUS ELEMENTOS

Cabe notar que el diseño de la clave-dirección ofrece bondades similares a la clave-nombre (normalización, alias, sonidos, clave invertida etc.).
Las aplicaciones son variadas pues una vez estructurada una dirección es susceptible de muchas aplicaciones.

CAPÍTULO III
ANÁLISIS ECONÓMICO FINANCIERO

El desarrollo a modo comercial de la solución planteada considera dos módulos:

- Desarrollo de los soportes básicos.
- Implantación de aplicaciones tipo.

El desarrollo de los soportes básicos incluye:

- Acopio de bases de datos diversas (mínimo 300,000 nombres)
- Análisis de frecuencia de los elementos mas repetidos
- Calificación y agrupamiento de palabras de mayor frecuencia
- Carga e indexación del archivo de nombres calificados
- Desarrollo de la rutina generadora de clave-nombre
- Desarrollo de la rutina generadora de clave-dirección
- Desarrollo de la rutina básica de comparación (nombres y/o direcciones).
- Desarrollo del sistema de administración de la tabla de nombres calificados.

El costo asociado se muestra en la figura número 43.

Desarrollo de los soportes básicos				
CANTIDAD	RECURSO	COSTO MENSUAL	MESES	COSTO TOTAL S/.
10	Digitadores	600	1	6,000
1	Jefe proyecto	6,000	2	12,000

2	Analistas-programadores	5,500	2	22,000
Total				40,000

FIGURA 43: COSTO DEL DESARROLLO DE SOPORTES BÁSICOS

En la práctica este costo es hundido pues ya se ha realizado y solo estaría pendiente adaptar la solución propuesta a las diversas instituciones que lo soliciten.

La implantación de las aplicaciones tipo incluye:

- Adicionar al programa de registro nombres de la institución la rutina de generación de la clave- nombre.
- Adicionar al programa de registro de direcciones de la institución la rutina de generación de clave-dirección.
- Cargar en el archivo de nombres de la institución las claves-nombre.
- Cargar en el archivo de direcciones de la institución las claves-dirección.
- Adicionar al programa de búsqueda de nombres de la institución la rutina de búsqueda mediante clave-nombre (búsqueda online opción variada).
- Adicionar a un programa de comparación de nombres de la institución la invocación de la rutina básica de comparación (búsqueda batch a precisión variada).
- Instalar el sistema de administración de la tabla de nombres calificados.
- Capacitación al personal de la institución para el mantenimiento de las tablas de nombres calificados y la explotación de las rutinas básicas.

El costo asociado se muestra en la figura 44.

Implantación de aplicaciones-tipo				
CANTIDAD	RECURSO	COSTO MENSUAL	MESES	COSTO TOTAL S/.
1	Jefe proyecto	6,000	1.5	9,000
2	Analistas-programadores	5,500	1.5	16,500
Total				25,500

FIGURA 44: COSTO DE IMPLANTACIÓN APLICACIONES TIPO

El autor plantea una oferta comercial al precio de S/. 25,500 mas un contrato de mantenimiento según se muestra en la figura 45:

DETALLE	COSTO S/.
Módulo básico nombres y direcciones	0
Implantación de aplicaciones tipo	25,500
Mantenimiento mensual 8 horas/hombre	1,000
Mantenimiento nuevos desarrollos hora/hombre	100

Figura 45: COSTO INCLUYENDO MANTENIMIENTO

El autor considera los siguientes aspectos complementarios:

- Cantidad de usuarios : Ilimitado.
- Derechos de propiedad definitiva sobre los objetos ejecutables.
- Ofrecimiento de entrega gratuita de los fuentes al tercer año.
- Los derechos otorgados al comprador no faculta comercialización posterior.

La institución que adquiera el software justificará la inversión desde el primer mes de implantación. Por ejemplo, en las instituciones financieras el envío de correspondencia debe tener máximo de 3% de devoluciones. En la figura 46 se muestra costos de correspondencia mensual devuelta asociados a cantidad de encartes.

CANTIDAD DE ENCARTES	DEVOLUCIONES (3%)	COSTO S/.
20,000	600	420
50,000	1,500	1,050
80,000	2,400	7,200
100,000	3,000	9,000
200,000	6,000	18,000

FIGURA 46: COSTO MENSUAL POR CORRESPONDENCIA DEVUELTA

El autor a través de la aplicación en una institución líder del sistema financiero peruano determinó entre otros resultados:

- Disminución de 15% de tiempo en el registro de nombres y direcciones de clientes.
- Disminución de problemas de unificación o desunificación de clientes en 50%.
- Reducción del número de intentos de búsqueda online de nombres de clientes en 70%.
- Reducción del porcentaje de devoluciones de correspondencia de 4% a 1.2%.
- Reducción de tiempo en la atención de solicitudes de información de juzgados en 80%.

Lo anterior cuantificado en horas hombre justifica económica y financieramente la aplicación propuesta.

CAPÍTULO IV

ANÁLISIS COMPARATIVO ENTRE EL SISTEMA EXISTENTE Y EL SISTEMA PROPUESTO

4.1 VENTAJAS Y DESVENTAJAS DE LAS SOLUCIONES ACTUALES.

Actualmente existen software producto que poseen querys incorporados y diversas instituciones han desarrollado rutinas propias de búsqueda.

VENTAJAS:

- Son de rápida aplicación.
- Para el manejo de listas de tamaño medio (ejemplo 50,000 elementos) son prácticas.
- Se apoyan en algunos casos en software producto de costo módico.

DESVENTAJAS:

- Cuando los archivos son grandes (por ejemplo 100,000 elementos) el tiempo de respuesta es lento.
- Poseen capacidad limitada de búsqueda por alias
- Poseen capacidad limitada para el manejo de acentos, caracteres especiales, minúsculas y conversiones en general.
- No poseen capacidad de búsqueda invertida.
- No poseen capacidad de búsqueda variable (ajustada a cierta precisión puntual deseada)
- Tienen demasiado margen de error en comparaciones batch.
- Tienen limitaciones para trabajar con otras herramientas (por ejemplo direcciones) con lo cual se haría más confiable la comparación.

- La mayoría de herramientas se apoyan en barridos por claves lo cual consume recursos de máquina y tiempo.

4.2 VENTAJAS Y DESVENTAJAS DE LA SOLUCIÓN PROPUESTA.

VENTAJAS:

- Es indistinto al tamaño de archivo pues en vez de accesos a porciones de archivo (barridos), se ubica directamente en los datos requeridos gracias a su buen diseño de claves.
- Posee amplia capacidad de búsqueda por alias
- Posee capacidad para el manejo de acentos, caracteres especiales, minúsculas y conversiones en general.
- Posee capacidad de búsqueda invertida.
- Posee capacidad de búsqueda variable (ajustada a cierta precisión puntual deseada)
- Tiene mínimo margen de error en comparaciones batch.
- Se puede complementar fácilmente con otras herramientas (por ejemplo direcciones) con lo cual se hace más confiable una comparación.
- Las búsquedas y comparaciones consumen pocos recursos de máquina.

DESVENTAJAS:

Como desventaja de la metodología propuesta el autor señala:

- El desarrollo está adaptado a la realidad peruana. Es decir, si se quiere usar por ejemplo en Corea requeriría rehacer todo el proceso de análisis y diseño por encontrarse ante otro alfabeto, otros sonidos y otras relaciones.
- El concepto de búsqueda fonética para el presente desarrollo contempla elementos prioritariamente castellanos.
- Parte de la certeza de la herramienta depende de la sapiencia con que los analistas registran los atributos en los archivos de nombres calificados.

4.3 RESUMEN DE LA COMPARACIÓN.

La solución propuesta dispone mayores ventajas y menores desventajas que las soluciones existentes en el mercado peruano.

Cabe notar que los buscadores WEB han desarrollado parte de las características de la solución propuesta pero la solución del autor esta mejor orientada a la realidad peruana.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES.

CONCLUSIONES:

1. Lo más significativo del presente trabajo es la metodología que abre las puertas a nuevas aplicaciones.
2. El autor muestra como un problema difuso es susceptible de estructurar y hasta normalizar a fin de facilitar su solución.
3. El concepto "clave-nombre" es de uso práctico e inmediato.
4. La lógica aplicada es extensiva a otras aplicaciones análogas tales como nombres jurídicos y direcciones.
5. Existe un amplio mercado de necesidades que serían atendidos por estas herramientas.
6. El diseño propuesto es dinámico, se retroalimenta pues posee un módulo de mantenimiento que ante nuevas palabras permite el enriquecimiento continuo de la base de nombres calificados.
7. Es recomendable el registro de patente de esta solución así como su incursión en otras realidades distintas al habla hispana.

RECOMENDACIONES

1. Para aplicar estos conceptos a idiomas distintos al castellano se recomienda analizar la problemática de dicho idioma pues los alfabetos pueden ser diferentes, así como los sonidos, las reglas y relaciones entre palabras en general.
2. Cada criterio que se añada a la lógica de las rutinas debe ser exhaustivamente validada con un amplio número de iteraciones o casos pues se puede dar el caso que una corrección de un caso

específico puede dañar la generalidad de un número mayor de casos.

3. Según la necesidad a atender, el usuario deberá ajustar el tamaño de la clave-nombre en la búsqueda de un óptimo de precisión.
4. No se debe perder de vista un adecuado diseño y combinación de claves de acceso a la base de datos a fin de optimizar los tiempos de respuesta.

GLOSARIO DE TERMINOS

AUTOMATIZACIÓN: Ejecución automática de trabajos industriales, administrativos o científicos sin intervención humana.

ACOTAMIENTO: Delimitación de algo material o inmaterial o restricción en el uso de algo.

BÚSQUEDA: Busca: acción de buscar. Buscar: hacer lo necesario para encontrar una cosa o persona.

CLAVE: Convención o conjunto de convenciones necesarias para efectuar las operaciones de cifrar o descifrar.

CODIGO: Sistema de signos o reglas que permiten formular y comprender un mensaje.

COMPARAR: Examinar dos o mas cosas para descubrir sus relaciones, diferencias o semejanzas.

DIRECCIÓN: indicación precisa del lugar donde alguien habita o se encuentra un establecimiento.

NORMALIZACIÓN: Tecnología: Ubicación de las medidas y calidades de los productos industriales o manufacturados para simplificar la fabricación y reducir el costo de los mismos.

MNEMOTECNIA: Técnica para desarrollar la memoria por medio de una serie de ejecuciones apropiadas

PALABRA: Conjunto de sonidos o de letras que representan un ser , una cosa o una idea o concepto.

BIBLIOGRAFÍA.

- Técnicas de Simulación en computadoras. Naylor-Balintfy-Burdick-Kong Chu
  Editorial LIMUSA 1ra. edición– México 1977
- Análisis y diseño de los sistemas de información. Jeffrey L. Whitten
  Editorial IRWIN 2da edición– Sydney Australia 1996
- Estudio de proyectos y soluciones de problemas. Fonythe, Keenan
  Editorial LIMUSA – México 1976
- Estructuración y procesamiento de datos. Ivan Flores
  Editorial PARANINFO 2da edición – Madrid 1984
- LISP el lenguaje de la inteligencia artificial. A. A. Beck
  Ediciones Amaya Multimedia 1ra. Edición – Madrid 1986
- Análisis y diseño de sistemas de información . James A. Senn
  Editorial Mc. Graw Hill 2da. Edición – Bogotá 1997
- Análisis y diseño de aplicaciones informáticas de Gestión – Mario G. Piattini
  Editorial Alfaomega 2da. Edición – México 2000
- Sistemas expertos – Enrique Castillo
  Editorial PARANINFO 1ra. Edición – Madrid 1989

ANEXO I

APLICACIÓN DE LA METODOLOGÍA EN DIRECCIONES

I.1 ESTIMACIÓN DE LAS REGLAS DE COMPORTAMIENTO.

Para las direcciones el autor infiere las siguientes reglas de comportamiento

- Se puede extraer una muestra representativa del universo de direcciones.
- Las cadenas de direcciones son susceptibles de dimensionar (elementos y longitud de la cadena)
- Los caracteres de las cadenas dirección son susceptibles de homologar.
- Las cadenas dirección poseen caracteres innecesarios que se pueden eliminar.
- Las direcciones tienen factores comunes de comportamiento por lo cual sus elementos son susceptibles de agruparse.
- Un elemento dirección puede referirse mediante diversos nombres sustitutos o alias.
- Los elementos dirección pueden registrarse en su mayoría en una base de datos de conocimiento y de ahí obtener mayores datos.
- Las letras de las direcciones poseen sonidos susceptibles de normalizar hacia un menor número de caracteres.
- Los elementos de una dirección son susceptibles de representar como seudo-código o claves dirección.

I.2 ALMACENAMIENTO DE UNA MUESTRA REPRESENTATIVA DEL UNIVERSO DE DIRECCIONES.

Si bien el universo de direcciones tiende a infinito, se puede alcanzar un alto grado de aproximación mediante la recolección de elementos-dirección desde el mundo real. Esta recolección de elementos se somete a análisis

de agrupación y calificación. El objetivo es disponer de una muestra representativa que contenga un gran porcentaje del universo de elementos-dirección.

El autor acopió 200,000 direcciones diversas de clientes de una institución financiera, realizó un análisis de frecuencia de los elementos más repetitivos y procedió a su registro y agrupamiento.

I.3 DIMENSIONAMIENTO DE CADENAS DE DATOS DIRECCIÓN

Similar a lo explicado en el punto 2.5.2, una cadena dirección primero debe ser almacenada en un área de trabajo para:

- Dimensionar la longitud de cada elemento y su longitud total
- Separar cada elemento dirección para su respectivo análisis

En la figura 47 se muestra algunos ejemplos de cadenas dirección.

Av. Central mz. I lote 7 urbanización Alamos II Monterrico Surco
Calle Andrea del Sarto 285 urbanización VIPEP Surquillo
Conjunto habitacional Los Próceres block 5 dpto 101 Surco
Av. 2 de mayo 566 Miraflores
Carretera central km 2.5 Ate

FIGURA 47: EJEMPLOS DE CADENAS-DIRECCIÓN

La desagregación se muestra en la figura 48

TIPO VIA	VIA	NÚMERO	TIPO ZONA	NOMBRE ZONA	DISTRITO
Av.	Central	Mz i lote 7	Urbanización	Alamos II Monterrico	Surco
Calle	Andrea del Sarto	285	Urbanización	VIPEP	Surquillo
	Los Proceres	Block 5 dpto 101			Surco
Av.	2 de mayo	566			Miraflores
Carretera	Central	Km 2.5			Ate

FIGURA 48: DESAGREGACIÓN DE CADENAS-DIRECCIÓN

I.4 HOMOLOGACIÓN DE CARACTERES.

Similar al 2.5.3, se aplica las siguientes consideraciones de homologación:

- Convertir formatos EBCDIC y ASCII u otros hacia un solo formato.
- Homologar letras mayúsculas y minúsculas hacia mayúsculas.
- Hacer conversiones para eliminar acentos
- Convertir caracteres extraños en blancos
- Convertir cadenas de caracteres blancos hacia un solo blanco

I.5 ELIMINACIÓN DE CARACTERES INNECESARIOS

Similar al punto 2.5.4 los caracteres especiales que son poco relevantes para el registro de direcciones se convierten en blancos. Una excepción es el carácter # pues en la data muestral se encontró direcciones que lo contienen.

I.6 AGRUPACIÓN DE PALABRAS

A las agrupaciones descritas en el punto 2.5.5 el autor le añade cuatro nuevas agrupaciones propias de direcciones.

AGRUPACIÓN	CALIFICACIÓN	EJEMPLOS
Nombres	A	Juan, Pedro, María, Graciela, Susana
Apellidos	B	González, Fernández, García, Mendoza
Actividades en general	C	Carrocería, canal, aéreo, natura etc
Siglas de empresas jurídicas	D	S.A., S.A.C., S.R.L., E.I.R.L. etc.
Artículos y pronombres	F	El, la, los, las, una, un etc.
Palabras relativas a santos	G	San, santa, santísimo, Sto. etc.
Oficios, profesiones o títulos	H	Guardia, duque, fray, general, licenciado etc.
Palabras relativas a instituciones diversas	I	Cuartel, embajada, episcopado, ejército, consulado etc.
Palabras relativas a actividades	K	Corporativo, servicios, aerolíneas,

económicas		constructores, contabilidad etc.
Nombres que también son apellidos	L	Washington, Reyna, Domingo, Jorge etc.
Meses	R	Enero, marzo, abril, etc.
Números expresados en palabras	U	Catorce, cuarta, III, vigésimo, etc.
Nombre reducido (truncado)	V	Aniba, ampar, anselm, juanit, leopold etc.
Apellido reducido (truncado)	W	Alvare, Andrad, Gonzag, Padill etc.
TIPOS DE VIA	M	Jirón, calle, óvalo, avenida, pasaje etc.
TIPOS DE DIRECCIÓN	N	Apartado, block, cuadra, manzana, etc.
TIPOS DE ZONAS	O	Complejo, fundo, urbanización etc.
DISTRITOS	S	Lince, Breña, Ate, La Victoria etc.

FIGURA 49: NUEVAS AGRUPACIONES PARA CADENAS-DIRECCIÓN

Para estandarizar el autor emplea listas estandarizadas de elementos dirección tal como el caso de tipo-vía y tipo-zona:

TABLA TIPO-VIA
AV.
JR.
CALLE
PJE.
ALAMEDA
MALECÓN
OVALO
PQUE.
PLAZA
CARRET.
PROLONG.
VIA
AUTOPISTA

BOULEVARD

TABLA TIPO-ZONA
VILLA
URB.
PUEBLO JOVEN
UNIDAD VECINAL
CONJUNTO HAB.
A.A.H.H.
COOP. VIVIENDA
RESIDENCIAL
ZONA INDUST.
LOTIZACION
CASERIO
FUNDO
ASOC. VIVIENDA
CONJUNTO RES.
COMPLEJO HAB.
AGRUP. VECINAL
AMPLIACIÓN
BARRIO
CENTRO COMERC.
HACIENDA
URB. POPULAR
CONDOMINIO
UNIDAD VECINAL
COOP. VIVIENDA

FIGURA 50: TIPOS DE VÍA Y DE ZONA

I.7 USO DE NOMBRES SUSTITUTOS O ALIAS

Diversos elementos dirección pueden ser referenciados por nombres alternos o 'alias'.

En la figura 51 se puede apreciar un segmento de la base de elementos dirección calificados basado en la data real tal cual se encontró y validó.

ELEMENTO NOMBRE	ALIAS	CALIFICACIÓN
APARTAD	APARTADO	N
APARTADO	APARTADO	N
APDO	APARTADO	N
APTADO	APARTADO	N
APTDO	APARTADO	N
APARTAD	APARTADO	N
APARTADO	APARTADO	N
ASENT	ASENTAMIENTO	O
ASENTA	ASENTAMIENTO	O
ASENTAM	ASENTAMIENTO	O
ASENTAMIENTH	AAHH	O
ASENTAMIENTO	ASENTAMIENTO	O
ASH	AAHH	O
ASNT	ASENTAMIENTO	O
ASO	ASOCIACIÓN	O
ASOC	ASOCIACIÓN	O
ASOCIAC	ASOCIACIÓN	O
ASOCIACION	ASOCIACIÓN	O
ASOCV	ASOVIV	O
ASOCVIC	ASOVIV	O
AUT	AUTOPISTA	M
AUTOP	AUTOPISTA	M
AUTOPIST	AUTOPISTA	M
AUTOPISTA	AUTOPISTA	M
AV	AV.	M
AVD	AV.	M
AVDA	AV.	M
AVE	AV.	M
AVND	AV.	M
AVNDA	AV.	M

FIGURA 51: MUESTRA DE ELEMENTOS DIRECCIÓN CALIFICADOS

I.8 NORMALIZACIÓN DE SONIDOS.

Similar al punto 2.5.9 se propone homologar ciertas letras o sílabas que tienen sonido similar.

I.9 PONDERACIÓN DE PALABRAS RELEVANTES.

En la figura 52 se presenta el orden de importancia de las agrupaciones relativas a direcciones. Nótese la gran diferencia con la ponderación descrita en el punto 2.5.5 que el autor usó para nombres.

Se aprecia las siguientes diferencias:

- Nombres y apellidos se reúnen en una sola calificación (A).
- Las palabras no encontradas en el archivo de elementos-dirección reciben la mayor ponderación (calificación Z, ponderación A)
- Las agrupaciones de palabras C, I, K y D no aparecen y están contenidas en la calificación Z. Al disminuir el número de agrupaciones se simplifica la lógica.
- Los calificadores de tipo-vía, tipo-zona y dirección reciben poca ponderación.

TIPO DE AGRUPACIÓN	CALIFICACIÓN	PONDERACIÓN
No encontrado en tabla elementos-dirección	Z	A
Nombres y apellidos	A	B
Meses	R	B
Número y/o números expresados en palabras relativas al nombre de la vía	u	C
Palabras relativas a santos	G	D
Títulos y oficios	H	D
Palabras de una sola letra	Y	E
Número y/o números expresados en palabras relativas al número de la dirección	U	Y
Tipos de vía	M	Y
Tipos de zona	O	Y

Tipos de numeración	N	Y
Distritos de Lima	S	Z
Conectores	F	Z

FIGURA 52: PONDERACIÓN DE AGRUPACIONES DE ELEMENTOS-DIRECCIÓN

El autor remarca la diferencia entre las calificaciones u y U. Por ejemplo, sea la dirección: AV. 2 DE MAYO 836 MIRAFLORES.

La palabra 2 se refiere al nombre de la vía y tiene mas peso que 836 que se refiere a la numeración de la dirección.

I.10 ANÁLISIS DE RELACIONES.

Los elementos se escriben a continuación de otros elementos y poseen reglas de formación o relaciones.

Complementando lo mencionado en el punto 2.6 las cadenas direcciones en mayor o menor grado tienen las siguientes relaciones

Después de tipo-vía continua la el nombre de la vía:

TIPO-VÍA	NOMBRE DE LA VIA
JIRÓN	HUALLAGA
ALAMEDA	LOS CIPRESES
PASAJE	SAN JOAQUIN
CARRETERA	CENTRAL
PLAZA	LA BANDERA
PROLONGACIÓN	GAMARRA

FIGURA 53: EJEMPLOS DE TIPO-VIA Y VIA

Después de tipo-zona continua zona:

TIPO-ZONA	NOMBRE DE LA ZONA
URB.	LOS ALAMOS
ASENTAMIENTO HUMANO	ALBERTO FUJIMORI
RESIDENCIAL	MIRONES
COOP. VIVIENDA	HUANCAYO
LOTIZACIÓN	SAN GABRIEL

FIGURA 54: EJEMPLOS DE TIPO-ZONA Y ZONA

Después de tipo-numeración continúa la numeración:

TIPO-NUMERACION	NUMERACION
MANZANA.	M
LOTE	14
INTERIOR	210
BLOCK	H
APARTADO	16B

FIGURA 55: EJEMPLOS DE TIPO NUMERACIÓN Y NUMERACIÓN

Los datos que continúan a TIPO-VIA son más relevantes que los datos que continúan después de TIPO-ZONA.

Las reglas de formación son fundamentales para analizar una cadena-dirección antes de construir los elementos normalizados.

A modo de ejemplo analizaremos las cadenas-dirección de la figura 56

Av. Central mz. I lote 7 urbanización Alamos Surco
Calle Andrea del Sarto 285 urbanización VIPEP Surquillo
Avenida Los Próceres block 5 dpto 101 Surco

FIGURA 56: EJEMPLOS DEMOSTRATIVOS

Se procede a:

1. Dimensionar la longitud de la cadena a analizar.
2. Se elimina blancos a la izquierda (si hubiesen).
3. Se convierte la cadena a mayúsculas EBCDIC
4. Se eliminan los acentos.
5. Si hay mas de un blanco juntos, se convierten en uno solo.
6. Se elimina caracteres especiales.
7. Se convierte la Ñ en N
8. Se explosiona las palabras
9. Se analiza la eliminación de apóstrofes.

Hasta aquí se tiene la figura 57:

AV.	CENTRAL	MZ.	I	LOTE	7	URBANI ZACION	ALAMOS	SURCO
CALLE	ANDREA	DEL	SARTO	285	URBANIZACIÓN	VIPEP	SURQUILLO	
AVENIDA	LOS	PROCERES	BLOCK	5	DPTO.	101	SURCO	

FIGURA 57: DIRECCIONES DESAGREGADAS EN SUS ELEMENTOS

10. Se accede a la base de datos de elementos dirección, se extrae sus valores de alias y se califica cada elemento. En la figura 58 se aprecia los datos calificados según los agrupamientos mostrados en la figura anterior.

Cadena-nombre	AV.	CENTRAL	MZ.	I	LOTE	7	URB.	ALAMOS	SANTIAGO DE SURCO
Calificación	M	A	N	U	N	U	O	A	S
Significado Calificación	Tipo vía	Nombre	Tipo número	Nume-ración	Tipo número	Nume-ración	Tipo zona	Nombre	Distrito

Cadena-nombre	CALLE	ANDREA	DEL	SARTO	285	URB.	VIPEP	SURQUILLO
Calificación	M	A	F	A	U	O	Z	S
Significado	Tipo vía	Nombre	Conector	Nombre	Número	Tipo zona	No encontrado	Distrito

Cadena-nombre	AV.	LOS	PROCERES	BLOCK	5	DPTO.	101	SANTIAGO DE SURCO
Calificación	M	F	Z	N	U	N	U	S
Significado	Tipo vía	Conector	No encontrado	Tipo Numero	Nume-ración	Tipo Número	Nume-ración	Distrito

FIGURA 58: ELEMENTOS DIRECCIÓN CALIFICADOS

Cabe notar que la tabla de elementos dirección calificados contiene mas de un 90% de los elementos de la realidad de donde se tomaron los datos muestrales. En el ejemplo se aprecia que la palabra PROCERES no fue encontrada en dicha base de datos. Para efectos de la clave-dirección el elemento PROCERES tendrá una ponderación mayor y debe notarse que la base de datos es susceptible de enriquecer posteriormente con nuevos elementos previa validación.

11. Se analiza si alguna palabra tiene la calificación de santo.

12.Seguidamente se procede a ponderar las palabras según sus agrupaciones tal como el autor explicó en el punto 2.5.10.

Cadena-nombre	AV.	CENTRAL	MZ.	I	LOTE	7	URB.	ALAMOS	SANTIAGO DE SURCO
Calificación	M	A	N	U	N	U	O	A	S
Significado	Tipo vía	Nombre	Tipo número	Nume-ración	Tipo número	Nume-ración	Tipo zona	Nombre	Distrito
Ponderación	Y	B	Y	Y	Y	Y	Y	B	Z

Cadena-nombre	CALLE	ANDREA	DEL	SARTO	285	URB.	VIPEP	SURQUILLO
Calificación	M	A	F	A	U	O	A	S
Significado	Tipo vía	Nombre	Conector	Nombre	Número	Tipo zona	Nombre	Distrito
Ponderación	Y	B	Z	B	Y	Y	B	Z

Cadena-nombre	AV.	LOS	PROCERES	BLOCK	5	DPTO.	101	SANTIAGO DE SURCO
Calificación	M	F	Z	N	U	N	U	S
Significado	Tipo vía	Conec-tor	No encontrado	Tipo Numeración	Núme-ro	Tipo Numeración	Nume-ro	Distrito

Ponderación	Y	Z	A	Y	Y	Y	Y	Z

FIGURA 59: ELEMENTOS DIRECCIÓN PONDERADOS

En la figura 52 el autor mostró como se ordena las palabras según ponderación o importancia. Primera ordenación para Tipo-vía y nombre-vía y luego Tipo-zona y nombre-zona. Entre dos palabras con la misma ponderación, se ubicarán según el orden en que se encontraron originalmente.

Calificación	CENTRAL	AV.	MZ.	I	LOTE	7	ALAMOS	URB.	SANTIAGO DE SURCO
Ponderación	B	Y	Y	Y	Y	Y	B	Y	Z

Elementos	ANDREA	SARTO	CALLE	285	VIPEP	URB	DEL	SURQUILLO
Ponderación	B	B	Y	Y	B	Y	Z	Z

Elementos	PROCERES	AV.	BLOCK	5	DPTO.	101	LOS	SANTIAGO DE SURCO
Ponderación	A	Y	Y	Y	Y	Y	Z	Z

FIGURA 60: ELEMENTOS DIRECCIÓN ORDENADOS SEGÚN PONDERACIÓN

Finalmente en la figura 61 se muestra las palabras según importancia, se eliminan letras dobles y se procede a cambiar letras según normalización de sonidos explicado en el punto 2.5.9. Los ensayos validados contra la data real establecieron una estructura óptima de clave-dirección con 4 elementos de 8 caracteres cada uno donde los tres primeros elementos son los de mayor ponderación y el cuarto elemento es fijo y corresponde al tipo vía.

Clave-	Parte1	Parte2	Parte3	Parte4:

dirección				VIA
Elementos	SENTRAL	MS	I	AB

Clave-dirección	Parte1	Parte2	Parte3	Parte4: VIA
Elementos	ANDREA	SARTO	285	CALE

Clave-dirección	Parte1	Parte2	Parte3	Parte4: VIA
Elementos	PROSERES	BLOCK	5	AB

FIGURA 61: CLAVES-DIRECCIÓN GENERADAS

ANEXO II

- Diagramas de contexto y de flujo de datos
- Diseño de tablas y diseño de llaves de acceso
- Modelo de datos
- Relación de rutinas básicas
- Relación de programas de administración

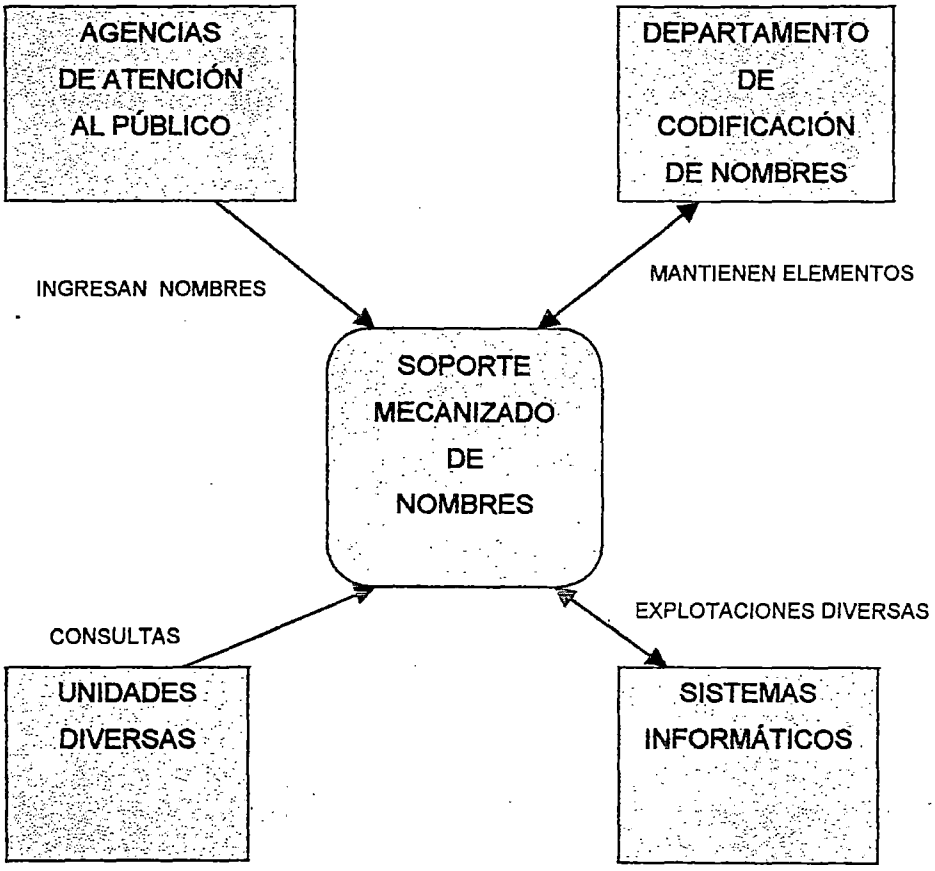


FIGURA 62: DIAGRAMA DE CONTEXTO

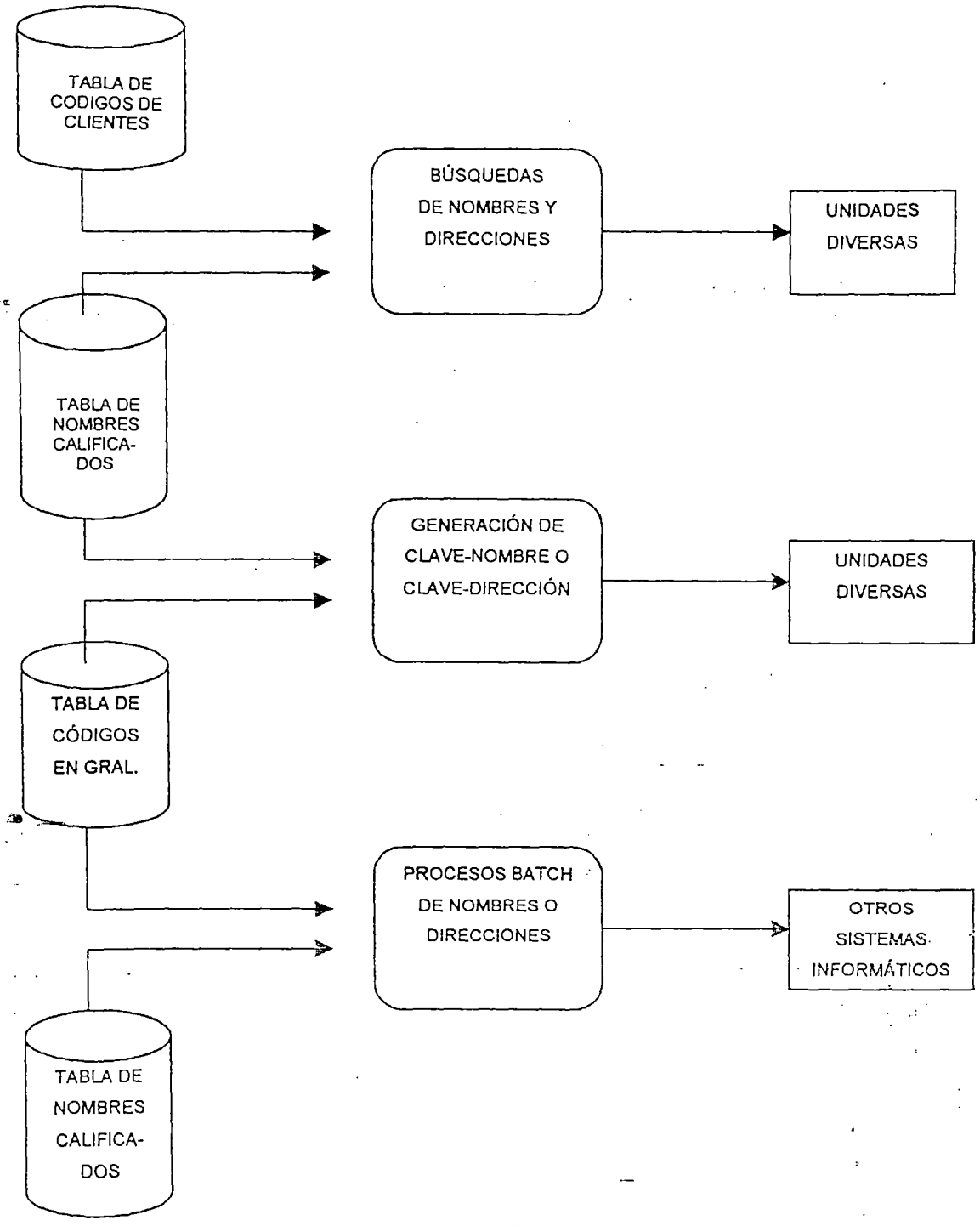


FIGURA 63: DIAGRAMA DE FLUJO DE DATOS

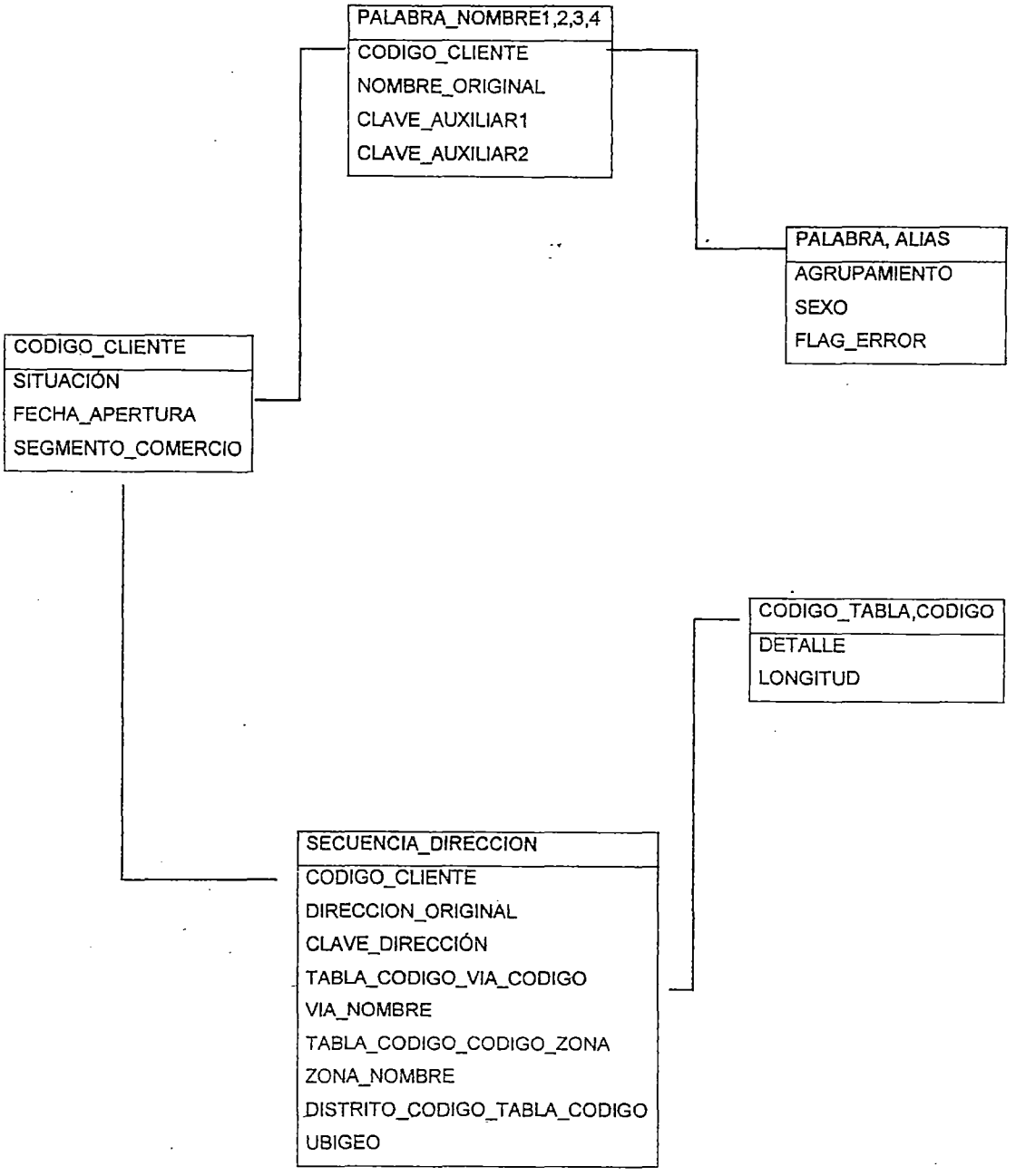


FIGURA 64: MODELO DE DATOS BÁSICO

II.1 DISEÑO DE TABLAS

La tabla de nombres calificados contiene las palabras de mayor frecuencia extraídas del universo real. Contiene nombres, apellidos, abreviaciones, en general las palabras en cuestión del problema a atender.

El autor recomienda crear una tabla análoga para las palabras del universo de direcciones a tratar que será explotada para la generación de la clave-dirección.

Tabla de nombres calificados				
Column name	Data type	Length	Obligatory	Comment
PALABRA	VARCHAR	20	Yes	Nombre, número, abreviación, palabra en general
ALIAS	VARCHAR	20		Nombre alias.
AGRUPAMIENTO	VARCHAR	1	yes	Calificación de agrupamiento
SEXO	VARCHAR	1		M,F, válido solo para nombre
FLAG_ERROR	VARCHAR	1		Valor 'X' cuando palabra es observada o error
MODIFICACIÓN_FECHA	TIME	8	YES	Fecha última modificación
MODIFICACIÓN_HORA	TIME	8	YES	Hora última modificación
MODIFICACIÓN_USUARIO	VARCHAR	20	YES	Usuario última modificación
MODIFICACIÓN_TERMINAL	VARCHAR	6	YES	Terminal ultima modificación

Figura 65: TABLA DE NOMBRES CALIFICADOS

Las sentencias para la definición de claves de acceso son

```
SQL> ALTER TABLE NOMBRES_CALIFICADOS
   > ADD CONSTRAINT NOMBRES_CALIFICADOS PK
   > PRIMARY KEY(PALABRA, ALIAS);
```

La tabla de nombres de la institución contiene los nombres de la institución, tanto nombres originales como sus respectivas claves normalizadas.

Tabla de nombres de la institución				
Column name	Data type	Length	Obligatory	Comment
CODIGO_CLIENTE	VARCHAR	12	YES	Código de cliente
NOMBRE_ORIGINAL	VARCHAR	120	YES	Cadena nombre
CLAVE_AUXILIAR1	VARCHAR	8		1er nombre en cadena nombre
PALABRA_NOMBRE1	VARCHAR	8	YES	1er Apellido en cadena nombre
PALABRA_NOMBRE2	VARCHAR	8	YES	1er nombre en cadena nombre
PALABRA_NOMBRE3	VARCHAR	8		2do apellido en cadena nombre
PALABRA_NOMBRE4	VARCHAR	8		2do nombre en cadena nombre
CLAVE_AUXILIAR2	VARCHAR	8		2do apellido en cadena nombre
MODIFICACIÓN_FECHA	TIME	8	YES	Fecha última modificación
MODIFICACIÓN_HORA	TIME	8	YES	Hora última modificación
MODIFICACIÓN_USUARIO	VARCHAR	20	YES	Usuario última modificación
MODIFICACIÓN_TERMINAL	VARCHAR	6	YES	Terminal ultima modificación

FIGURA 66: TABLA DE NOMBRES DE LA INSTITUCION

Las sentencias para la definición de claves de acceso son

```
SQL> ALTER TABLE NOMBRES_INSTITUCION
     > ADD CONSTRAINT NOMBRES_INSTITUCION1_PK
     > PRIMARY KEY(CODIGO_CLIENTE)
SQL> ALTER TABLE NOMBRES_INSTITUCION
     > ADD CONSTRAINT NOMBRES_INSTITUCION2_PK
```

```
    > ADD CONSTRAINT(CLAVE_AUXILIAR1, PALABRA_NOMBRE1);
SQL> ALTER TABLE NOMBRES_INSTITUCION
    > ADD CONSTRAINT NOMBRES_INSTITUCION3_PK
    > PRIMARY KEY(PALABRA_NOMBRE4, CLAVE_AUXILIAR2)
SQL> ALTER TABLE NOMBRES_INSTITUCION
    > ADD CONSTRAINT NOMBRES_INSTITUCION4_PK
    > PRIMARY KEY(PALABRA_NOMBRE1, PALABRA_NOMBRE2,
      PALABRA_NOMBRE3, PALABRA_NOMBRE4);
SQL> ALTER TABLE NOMBRES_INSTITUCION
    > ADD CONSTRAINT NOMBRES_INSTITUCION5_PK
    > PRIMARY KEY(PALABRA_NOMBRE1, PALABRA_NOMBRE2)
```

La tabla de direcciones de la institución contiene las direcciones tanto originales como sus respectivas claves-dirección:

Tabla de direcciones de la institución				
Column name	Data type	Length	Obligatory	Comment
SECUENCIA_DIRECCIÓN	VARCHAR	3	YES	Tipo de dirección
CODIGO_CLIENTE	VARCHAR	12	YES	Código de cliente
DIRECCIÓN_ORIGINAL	VARCHAR	120	YES	Cadena dirección
CLAVE_DIRECCION	VARCHAR	32		4 elementos de 8 caracteres cada uno
VIA	VARCHAR	20		Av., jr., calle, pasaje etc.
VIA_CODIGO	VARCHAR	9		Según tabla de códigos
DIRECCIÓN	VARCHAR	60	YES	Cuerpo de la dirección
NUMERO	VARCHAR	20	YES	Número en todas sus formas.
ZONA	VARCHAR	60		AAHH, Urb., conj. Habitac etc.
ZONA_CODIGO	TIME	9	YES	Según tabla de códigos
DISTRITO	VARCHAR	20		Valido para Lima
CODIGO_DISTRITO	VARCHAR	9		Válido para Lima
UBIGEO	NUMBER	6		Válido para Lima

MODIFICACIÓN_FECHA	TIME	8	YES	Fecha última modificación
MODIFICACIÓN_HORA	TIME	8	YES	Hora última modificación
MODIFICACIÓN_USUARIO	VARCHAR	20	YES	Usuario última modificación
MODIFICACIÓN_TERMINAL	VARCHAR	6	YES	Terminal ultima modificación

FIGURA 67: TABLA DE DIRECCIONES DE LA INSTITUCIÓN

Las sentencias para la definición de claves de acceso son

```
SQL> ALTER TABLE DIRECCIONES_INSTITUCIÓN
  > ADD CONSTRAINT DIRECCIONES_INSTITUCIÓN_PK
  > PRIMARY KEY(CLAVE_DIRECCIÓN);
```

La tabla de códigos es un archivo de uso compartido y contiene información diversa que sirve de apoyo a la normalización de direcciones entre otras necesidades.

Tabla de códigos de la institución				
Column name	Data type	Length	Obligatory	Comment
CODIGO_TABLA	VARCHAR	3	YES	Corresponde a tipo vía, tipo zona, CIIU, ubigeo etc.
CODIGO	VARCHAR	6	YES	Código en sí.
DETALLE	VARCHAR	50	YES	Descripción del elemento
LONGITUD	NUMBER	1	YES	Según cada tabla.
MODIFICACIÓN_FECHA	TIME	8	YES	Fecha última modificación
MODIFICACIÓN_HORA	TIME	8	YES	Hora última modificación
MODIFICACIÓN_USUARIO	VARCHAR	20	YES	Usuario última modificación
MODIFICACIÓN_TERMINAL	VARCHAR	6	YES	Terminal ultima modificación

FIGURA 68: TABLAS DE LA INSTITUCIÓN

Las sentencias para la definición de claves de acceso son

SQL> ALTER TABLE CODIGOS_INSTITUCION

> ADD CONSTRAINT CODIGOS_INSTITUCION_PK

> PRIMARY KEY(CODIGO_TABLA,CODIGO);

La tabla log de modificaciones contiene las modificaciones históricas realizadas en la tabla de nombres calificados u otra tabla. Es muy útil para identificar quien y bajo que sustento realizó modificaciones pues cada cambio debe ser exhaustivamente validado contra la realidad.

Log de modificaciones de tabla nombres calificados				
Column name	Data type	Length	Obligatory	Comment
TABLA	VARCHAR	15	YES	Nombre de la tabla
CODIGO_CLIENTE	VARCHAR	12	YES	Código de cliente
NUMERO_CAMPO	NUMBER	2	YES	Número de campo 01, 02 etc.
VALOR_ANTERIOR	VARCHAR	50		Dato antes del cambio
VALOR_NUEVO	VARCHAR	50	YES	Dato después del cambio
MODIFICACIÓN_FECHA	TIME	8	YES	Fecha última modificación
MODIFICACIÓN_HORA	TIME	8	YES	Hora última modificación
MODIFICACIÓN_USUARIO	VARCHAR	20	YES	Usuario última modificación
MODIFICACIÓN_TERMINAL	VARCHAR	6	YES	Terminal ultima modificación

FIGURA 69: TABLA LOG DE MODIFICACIONES

Las sentencias para la definición de claves de acceso son

SQL> ALTER TABLE LOG_MODIFICACIONES

> ADD CONSTRAINT LOG_MODIFICACIONES_PK

> PRIMARY KEY(TABLA,NUMERO_CAMPO);

II.2 RUTINAS BÁSICAS.

Las rutinas básicas están diseñadas para ser invocadas vía parámetros desde otros programas aplicativos facilitando de este modo la programación modular.

RUTINA DE GENERACIÓN DE CLAVE NOMBRE.

INPUT: recibe una cadena nombre de 120 posiciones. Es indistinto si la cadena esta ordenada nombre-apellido o apellido-nombre.

OUTPUT: devuelve la clave-nombre compuesta por 4 elementos de 8 posiciones cada uno.

PROCESO: Accesa la tabla de nombres calificados y aplica la lógica explicada en el 2.8.7 a la cadena nombre recibida.

RUTINA DE GENERACIÓN DE CLAVE-DIRECCIÓN.

INPUT: recibe una cadena nombre de 120 posiciones. Es indistinto si la cadena esta ordenada o desordenada.

OUTPUT: devuelve la clave-dirección que compuesta por 4 elementos de 8 posiciones cada uno.

PROCESO: Accesa la tabla de nombres y direcciones calificadas y aplica la lógica explicada en el anexo I.

RUTINA DE BÚSQUEDA VARIADA DE NOMBRES.

INPUT: recibe una cadena nombre de 120 posiciones. Es indistinto si la cadena esta ordenada nombre-apellido o apellido-nombre.

OUTPUT: devuelve las cadenas-nombre que coinciden con la clave-nombre dependiendo de la precisión deseada.

PROCESO: Accesa la tabla de nombres calificados y aplica la lógica explicada en el punto 2.6 a la cadena nombre recibida y del archivo de nombres de la institución ubica los nombres que coinciden. El autor le denomina a esta búsqueda variada pues considera nombres invertidos, sonidos similares y alias si hubiesen.

RUTINA DE COMPARACIÓN DE DOS LISTAS DE NOMBRES.

INPUT: recibe una lista de nombres que debe compararse contra otra lista de nombres.

OUTPUT: Por cada nombre de la primera lista devuelve los nombres de la segunda lista que coinciden en la clave-dirección.
PROCESO: Es un matching entre dos tablas según clave-nombre que en proceso previo debió ser generada para ambas tablas.

RUTINA DE COMPARACIÓN DE DIRECCIONES.

INPUT: recibe dos cadenas de direcciones de 120 posiciones cada una..
OUTPUT: Devuelve las claves-dirección de las direcciones recibidas.
PROCESO: A ambas cadenas-dirección recibidas les genera su clave-dirección y devuelve dicha información en 32 bytes. Los procesos que invocan esta rutina pueden calibrar el número de caracteres o elementos hacia un óptimo tal cual se explicó en el punto 2.8.8.

RUTINA DE NORMALIZACIÓN DE DIRECCIONES.

INPUT: recibe una cadena-dirección de 120 posiciones.
OUTPUT: Devuelve la dirección desagregada en campos separados tales como vía, dirección, número, zona, distrito y ubigeo si fuese el caso.
PROCESO: A la cadena-dirección recibida le aplica la rutina de generación de clave-dirección y una lógica complementaria para desagregar la cadena en campos independientes.

II.3 RELACIÓN DE PROGRAMAS DE ADMINISTRACIÓN

A fin de mantener el sistema el autor refiere la necesidad de mantener las tablas de nombres calificados y direcciones calificadas.
Cada cambio debe ser celosamente revisado por un nivel de control y validado exhaustivamente contra la data real pues se puede dar el caso que algún cambio ayude a solucionar algunos casos específicos pero dañe considerablemente una generalidad de casos.
Las rutinas básicas para este mantenimiento se muestran en la figura 65

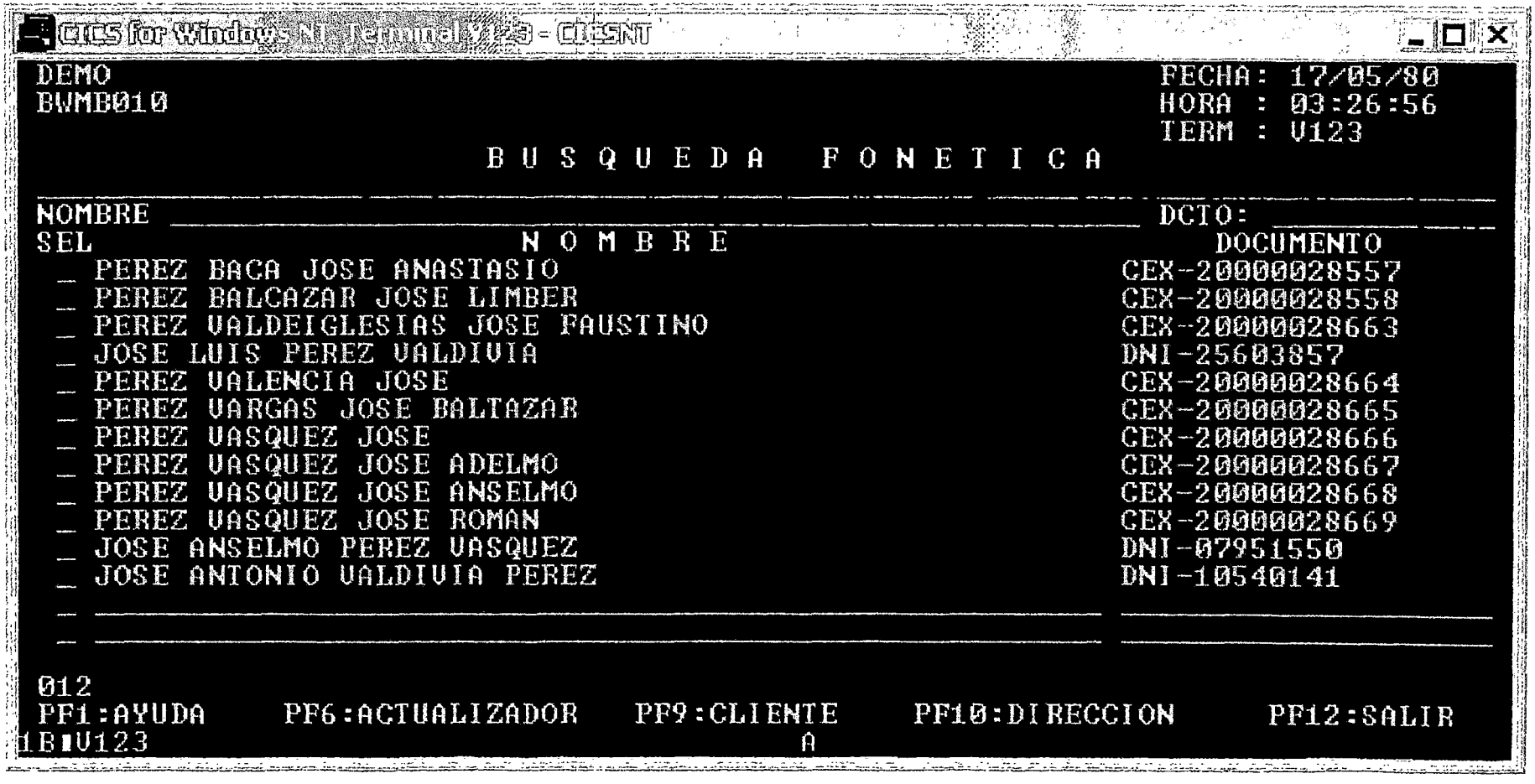
PROGRAMA	DESCRIPCIÓN
PROG001	Realiza altas, bajas y cambios en las palabras de la tabla nombres-calificados.

PROG002	Realiza altas, bajas y cambios en las palabras de la tabla direcciones-calificadas.
PROG003	Pantalla de consulta a las tablas de nombres y direcciones calificadas
PROG004	Reporte de últimos cambios realizados a las tablas de nombres y direcciones calificadas.

FIGURA 70: RUTINAS BÁSICAS DE MANTENIMIENTO

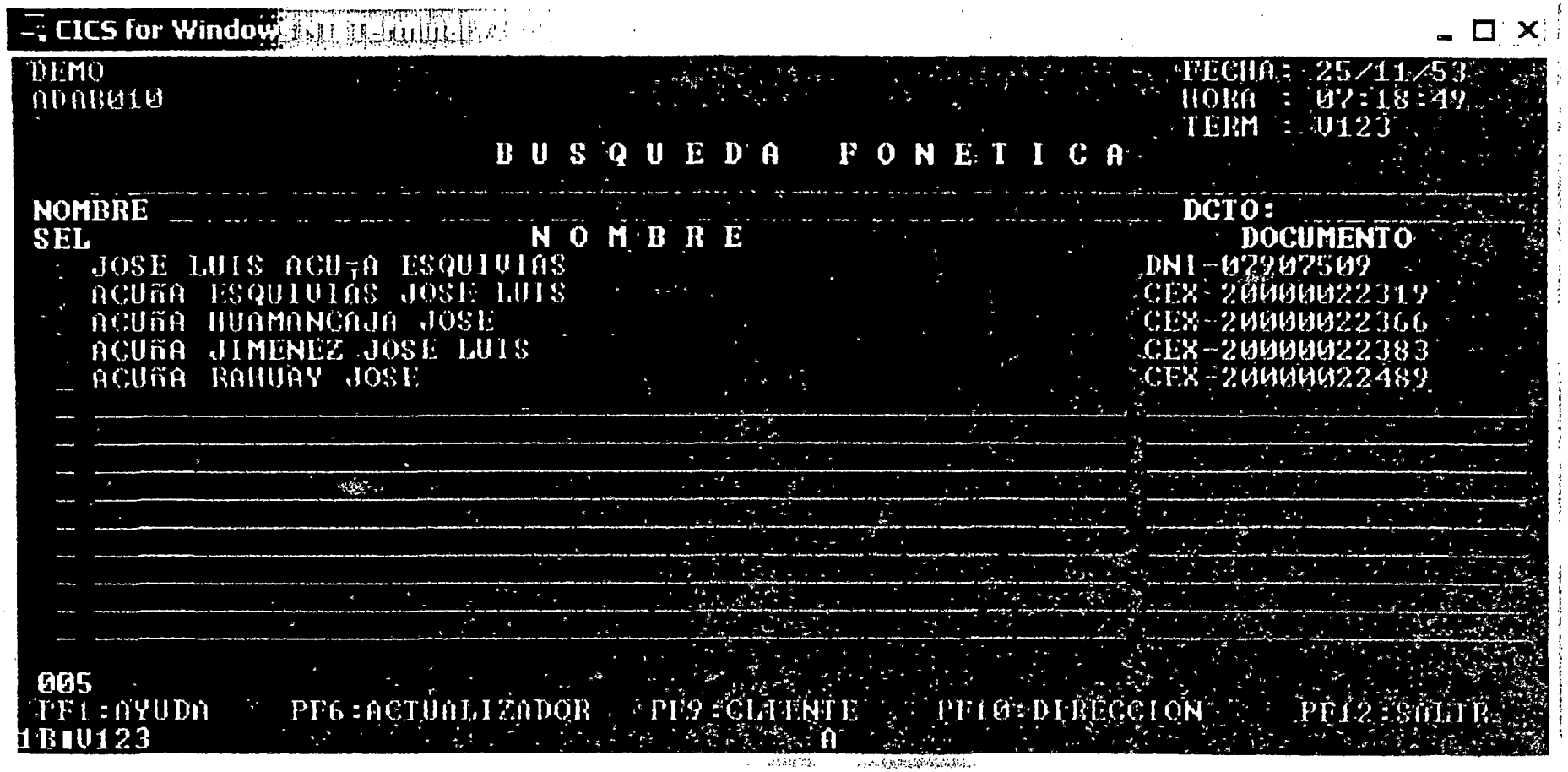
Casos de búsqueda - Orden de los elementos

- Ej.JOSE PEREZ VA



Casos de búsqueda – Caracteres de diversa plataforma (Ñ)

- Ej.JOSE ACUNA

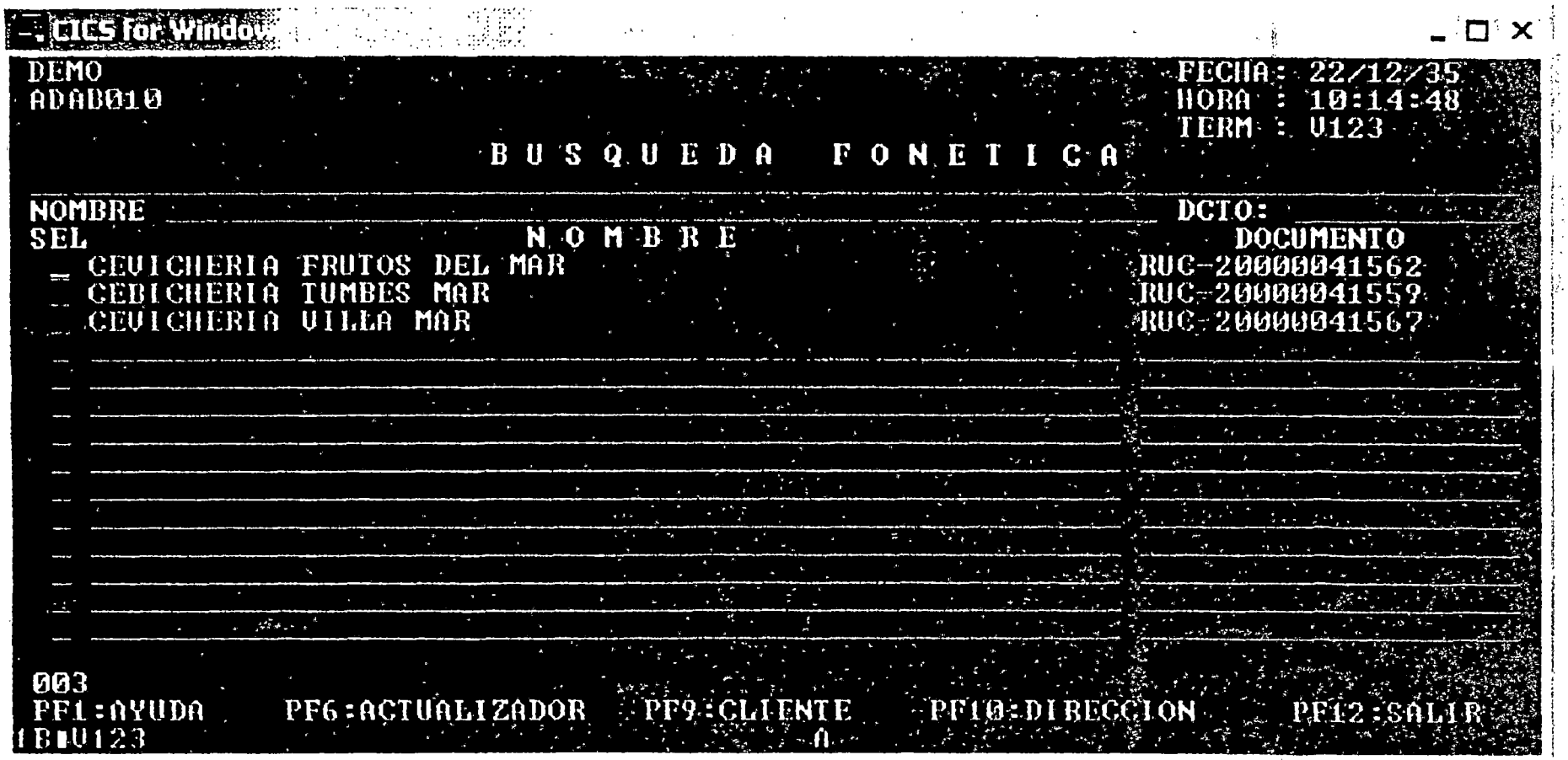


Casos de búsqueda – Errores comunes de escritura (letras dobles)

- Ej.MALLQUI

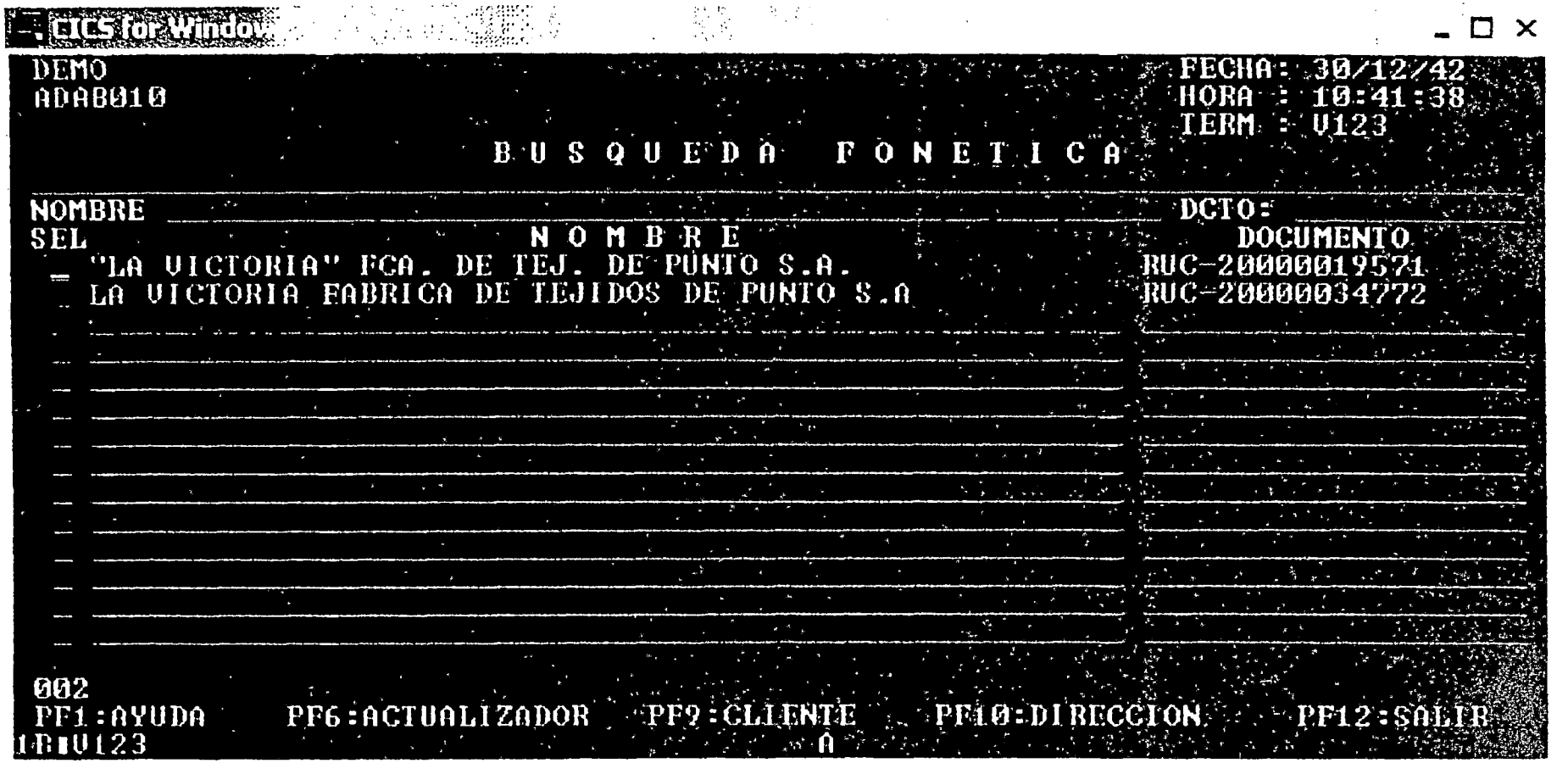
Casos de búsqueda – Sólo se conoce algún elemento

- Ej. CEVICHE MAR



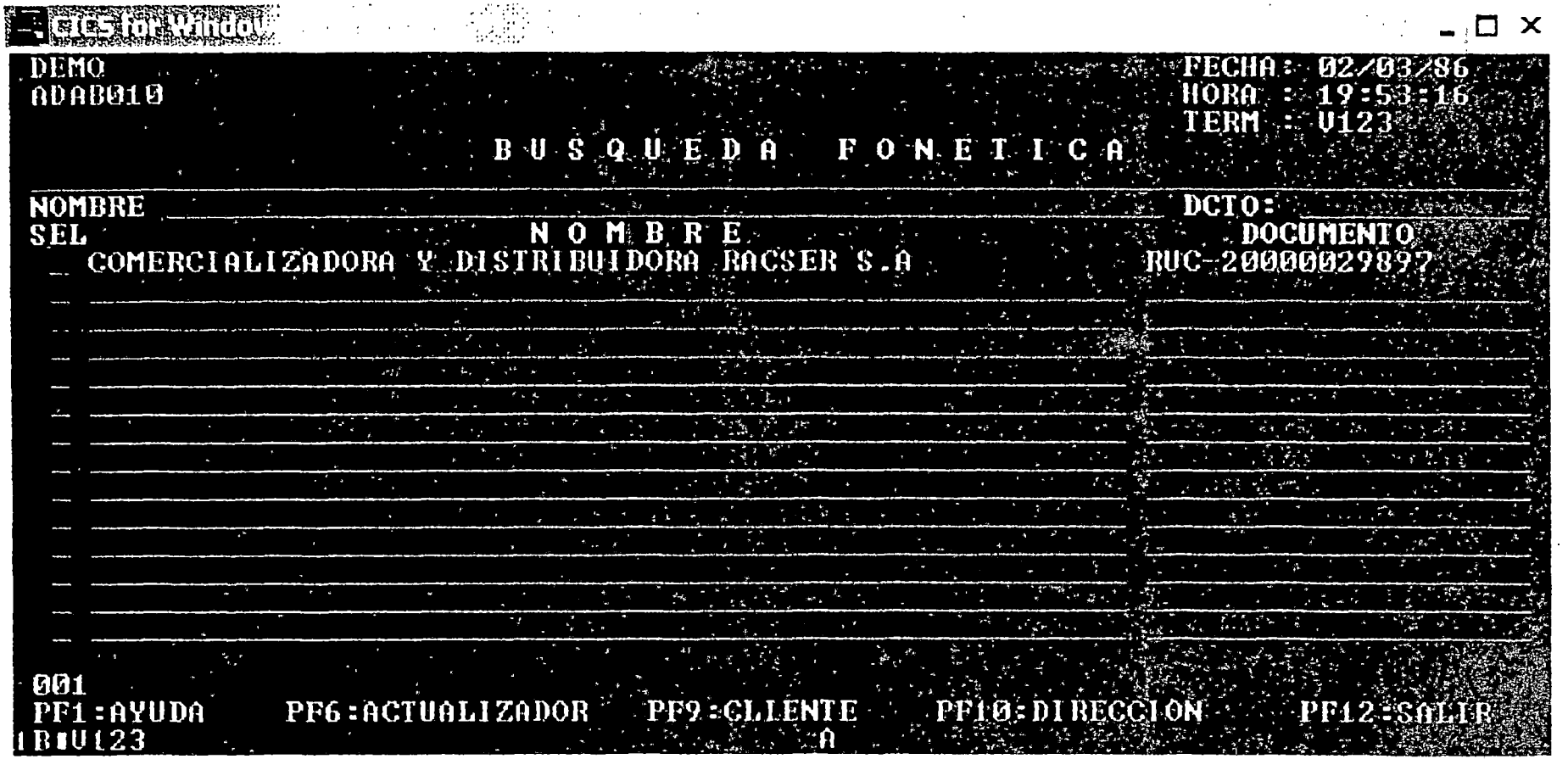
Casos de búsqueda – Uso de caracteres especiales / Uso de abreviaturas

- Ej. FABRICA LA VICTORIA



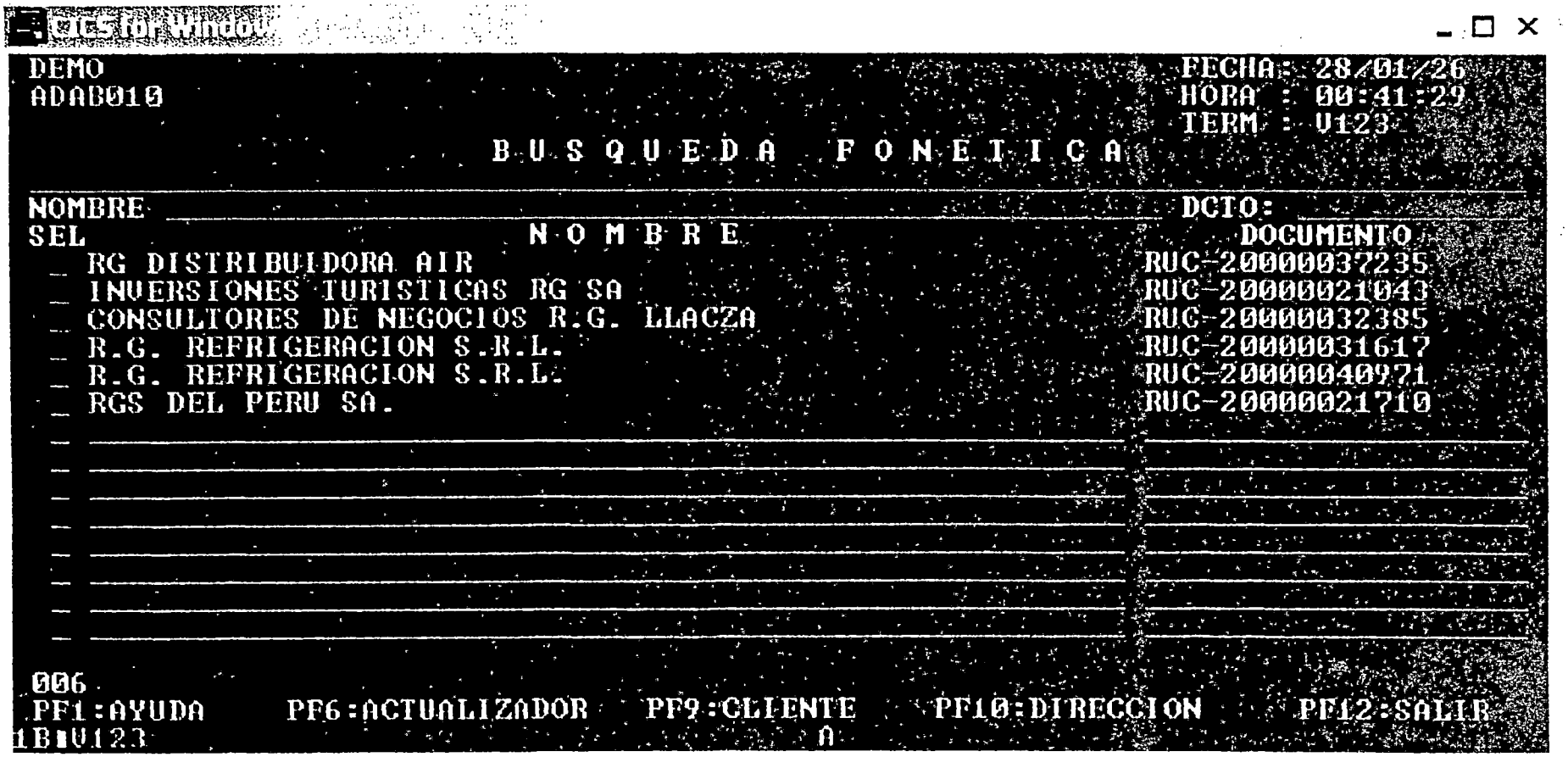
Casos de búsqueda – Uso de sinónimos

- Ej. COMERCIAL RACSER



Casos de búsqueda – Uso de siglas

Ej. RG



Reporte de unificación (Casos interesantes)

EDIT.ESCUELA ACTIVA SA	0 20000035965	AV. ARGENTINA 1325 - EL C
EDITORIAL ESCUELA ACTIVA S.A.	0 20000038779	AV. ARGENTINA 1325 - EL C
ALEJANDRO ACU¥A GONZALES	2 20000022352	MZ.N LT.28 URB. N,STOR GA
ACU¤A GONZALES ALEJANDRO	2 20000022353	MZ.N LT.28 URB. N,STOR GA
ANA MARIA ACUÑA PRINCIPE	1 08068049	CA SHELL 310 PISO 7 - MIR
ACU¤A PRINCIPE ANA MARIA	2 20000022483	MZ.B LT.20 URB. JUAN X X
ANGEL D. ACU¥A NEYRA	2 20000022433	AV BRASIL 367 INT.357 - E
ANGEL DAVID ACU¥A NEYRA	2 20000022434	AV BRASIL 367 INT.357 - E
ACU¤A NEYRA ANGEL DAVID	2 20000022432	AV RICARDO HERRERA 962 -
ACU¤A MORENO SANTOS F.	2 20000022593	AV GRAU 469 DP.4 - EL CER
SANTOS F. ACU¥A MORENO	2 20000022423	AV GRAU 469 DP.4 - EL CER
SANTOS FELIPE ACU¥A MORENO	2 20000022422	AV GRAU 469 DP.4 - EL CER
ACU¤A MORENO SANTOS FELIPE	2 20000022592	AV GRAU 469 DP.4 - EL CER

Reporte de unificación (Casos interesantes)

IMPORTADORES A & M S.R.L.	0 20000040689	JR. MANUEL VILLAR 211 - U
A & M IMPORTADORES S.R.L.	0 20000040690	JR. MANUEL VILLAR 211 URB
ARMAS CLARA G. FIGUEROA DE	2 20000022725	P.MASCAGNI MZ.FLT.24 - SU
ARMAS CLARA GABINA FIGUEROA DE	2 20000022726	PIETRO MASCAGNI MZ.FLT.24
BARRIOS FUENTES URQUIAGA ABOGADOS	0 20000030024	ARIAS ARAG□EZ 250 - MIRAF
ESTUDIO BARRIOS FUENTES URQUIAGA ABOG	0 20000032409	ARIAS ARAG□EZ 250 - MIRAF
DEL BOSQUE RONCAL D'ANGELO Y ASOC	2 20000020394	DONATELLO 206 - SAN BORJA
DEL BOSQUE, RONCAL, D`ANGELO Y ASOCIA	2 20000032621	DONATELLO 206 - SAN BORJA
CH Y V GRAFICOS SAC	0 20000020033	AV NICOLAS DUEÑAS 279 MIR
CH & V GRAFICOS S.A.C.	0 20000038725	AV. NICOL S DUE□AS 279 -
COMPA¥IA VILLA GAS	0 20000037129	AV. AVIACI¢N 3485 - SAN B
COMPA¥IA DISTRIBUIDORA VILLA GAS	0 20000037126	AV. AVIACI¢N 3485 - SIN D
CIA DE SERVICIOS ELECTRICOS INDUSTRIA	0 20000020050	AV AVIACION 1191 - LA VIC
COMPA¥IA DE SERVICIOS ELECTRICOS INDU	0 20000039562	AV. AVIACI¢N 1191 - LA VI

Reporte de unificación (Casos interesantes)

DALL`ORSO MERTZ ASOCIADOS AUDITORES	0	20000032392	AV. ARENALES 395 OF. 405
DALLORSO MERTZ ASOCIADOS CONTADORES P	2	20000032619	AV. ARENALES 395 OF. 495
GRUPO D & V EVENTOS SOCIALES	0	20000042735	PJE. UNI¢N 102 - URB. HUA
GRUPO D&V EVENTOS SOCIALES S.R.L.	0	20000042736	PJE. UNI¢N 102 - R¡MAC
EXCEL PRODUCTS SA	0	20000020614	JR AGUARICO 133 CHACRA CO
A.EXCEL PRODUCTS S.A.	0	20000040504	JR. AGUARICO 133 - BRE¤A
FIBERGLASS IMPORT Y FABRICACIONES	0	20000036512	- SIN DISTRITO
FIBERGLASS IMPORTACIONES Y FABRICACIO	0	20000036513	- EL CERCADO
FIBERGLASS IMPORTACIONES Y FABRICACIO	0	20000036514	AV. AVIACI¢N 3358 OF. 401
COLEGIO ISABEL FLORES DE OLIVA	0	20000034091	JUAN DELLEPIANE 530 - SAN
CEP ISABEL FLORES DE OLIVA	0	20000020012	JR JUAN DELLEPIANE 530 OR
IWASHITA, NUE Y ASOCIADOS SOCIEDAD CI	2	20000021058	REPUBLICA DE PANAMA 3030
IWASHITA, NUE Y ASOCIADOS S.C.	2	20000032644	AV. REP£BLICA DE PANAM 3
KU HORTENCIA YONG VDA. DE	2	20000026054	UCAYALI 636 - EL CERCADO
KU YONG HORTENCIA FELICITA	2	20000026061	BATERIA MAYPU 534 - LA VI