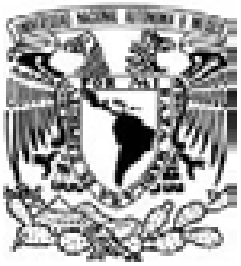


UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE INGENIERÍA

MODELO DE GESTIÓN DE LA INFORMACIÓN DE SISTEMAS REGISTRALES: UN ENFOQUE PRÁCTICO

T E S I S

QUE PARA OBTENER EL TITULO DE
INGENIERO EN COMPUTACIÓN

P R E S E N T A :

SERGIO GERARDO ZAVALA MENDOZA

DIRECTOR DE TESIS
ING. FRANCISCO JOSÉ RODRÍGUEZ RAMÍREZ

2008

Dedico este trabajo a
Ana Lia y Cris

Índice

I. Introducción.

Los sistemas de registro son tan antiguos como la misma historia de la humanidad y nacen por la necesidad de registrar hechos históricos, preservar la cultura y tradiciones religiosas, normar las conductas civiles, controlar el comercio o simplemente para comunicar algo.

Es innegable que la evolución de los sistemas registrales ha ido de la mano con la evolución de la escritura y de tecnología existente para dejar constancia escrita de diversos hechos. La especialización de los mismos también evoluciona conforme crece la necesidad de dejar cualquier tipo de constancia escrita. Así, se evoluciona desde los antiguos escribas que utilizaban rollos y códices –sin olvidar el registro que se realizaba tallado en piedras o pinturas- pasando por el desarrollo de la imprenta, el nacimiento de oficias registrales, principalmente de la propiedad, las bibliotecas que desarrollaron en el concepto de catálogos, hasta el nacimiento de las tecnologías de la información.

Hace cerca de 5 mil años el hombre comenzó a dejar registro de las cosas en forma de escritura, para ello utilizó diversos materiales y técnicas para crear los registros. Los pueblos de Mesopotamia fueron los primeros que escribieron sobre tabletas de barro a las que hacían incisiones con un punzón; la escritura también se manifestó en tabletas de cera, pedazos de cerámica, papiros, pieles de animales y más tarde se escribió sobre papel. Hasta el siglo XV se revolucionó la producción de libros en Europa, gracias a dos innovaciones: los europeos aprendieron de los musulmanes (que a su vez aprendieron de los chinos) a confeccionar papel y el alemán Gutenberg inventó la imprenta basada en tipos móviles de metal. Pero antes de que se inventaran los libros como actualmente los conocemos, existieron los rollos, códices y manuscritos.

Desde la antigüedad el uso de sistema registrales ha sido una tarea que ha realizado el hombre para poder tener el control sobre diversos aspectos de la vida. Así por ejemplo, los antiguos fenicios idearon mecanismos de registro para que cada comerciante pudiera anotar con cierta facilidad información acerca de sus transacciones. Así, la intensa actividad económica de los fenicios los llevó a crear un sistema de registro que al final se convirtió en un alfabeto fonético, es decir símbolos que representaban sonidos –Los fenicios no “inventaron” el alfabeto, pero sí contribuyeron a desarrollarlo, tanto que el alfabeto fenicio se considera el antecedente antiguo del alfabeto moderno.

En México, una vez concluida la conquista militar, los españoles iniciaron la evangelización, tarea que los reyes de España pusieron en manos de diferentes órdenes religiosas, al considerar que tenían la autoridad moral y el empeño para realizar dicha obra. Entre ellos estaban franciscanos, dominicos, agustinos y jesuitas. La obra de las congregaciones mencionadas dependía de las limosnas, donaciones y de su trabajo. Por esta razón, a su arribo a la Nueva España celebraban reuniones con las autoridades civiles (en este caso el virrey, quien era el representante del rey), las autoridades eclesiásticas (obispo, arzobispo) y los miembros más adinerados de la sociedad, a fin de solicitar el apoyo económico que requerían para emprender la tarea. Las limosnas y las dotes de por vida, que cubrían un número determinado de colegiaturas, fueron la primera fuente de ingresos económicos para la obra educativa de los jesuitas; sin embargo, no serían suficientes para dar la seguridad financiera que dicha tarea requería.

Lo que indudablemente redituó mayores recursos fueron las donaciones de casas y haciendas. Todas ellas eran otorgadas después de la muerte de ricos terratenientes, o bien de familias que no tenían descendencia. Los hermanos de la compañía de Jesús se dieron a la tarea de lograr que estos bienes fueran productivos. Para ello implementaron un detallado y cuidadoso sistema de registro con el fin de evitar pérdidas. El éxito que obtuvieron los llevó a ser reconocidos como los mejores administradores agrícolas[1].

En la actualidad los sistemas registrales son la columna vertebral del funcionamiento de muchas instituciones, el desarrollo de la tecnología ha permitido la incorporación, cada vez más frecuente, de mecanismos de administración de la información incorporando diversos componentes como los son: la gestión de la calidad de la información, la administración de riesgos, el desarrollo de modelos de conocimiento y de comportamiento.

Los sistemas registrales revisten tal importancia que son fuente de conocimiento en el cual se resguarda no sólo información, sino la certeza jurídica de la posesión de algún bien. Así, desde el punto de vista legal y con base en las obligaciones, derechos y/o beneficios que se adquieren por pertenecer a algún registro, podríamos catalogar a los registros de la siguiente manera:

1 Vid, González Marín, Silvia: Historia de la Hacienda de Chapingo, México, Universidad Autónoma Chapingo, 1996, p.36.

Sistema que dan personalidad Jurídica

Registros civiles (nacimientos, defunciones, matrimonios, etc.)
Registros mercantiles (constitución de sociedades, cooperativas, asociaciones, etc.)

Sistemas de posesión de bienes

Registro público de la propiedad (posesión sobre bienes inmuebles)
Registro de patentes (uso y goce de patentes, marcas, etc.)
Registros bancarios (cuentas de inversión, ahorro, préstamos hipotecarios, etc.)

Sistemas para otorgar servicios

Registros de telefonías
Registros de luz
Registros de aguas
Registros de beneficiarios de seguridad social

Sistemas de filiación

Registros de clubes deportivos
Registros de tiendas departamentales
Registros de clubes de compras
Registros de clientes
Registros de alumnos

Sistemas de Control

Registro federal de contribuyentes
Registro aduanal
Registro de automóviles
Registro predial
Registro de Población (CURP)
Registro de licencias de funcionamiento
Registro de buroes de crédito

Como puede observarse, debido a los millones de usuarios que forman parte de los sistemas registrales, es casi imposible imaginar el funcionamiento de nuestra vida sin la existencia de estos. Sin embargo, y no obstante el avance tecnológico en materia de sistemas de

información, los sistemas registrales enfrentan grandes retos ya que la falta de calidad de estos y una mala administración de los mismos acarrea grandes costos a las empresas e instituciones y a sus propios usuarios, que en muchas ocasiones no son perceptibles y peor aún son difíciles de identificar o de medir.

Si nos centramos en tres grandes apartados la calidad de la información: 1) Calidad del dato, 2) que éste corresponda a la realidad y 3) que se tenga al total del universo de la información; encontraremos de forma muy recurrente fallas en alguno de estos temas. Como podrá suponerse, las deficiencias en la calidad de la información conllevan a problemas en la interpretación y uso que se le dé a la misma, de forma general podríamos mencionar que afecta la generación de conocimiento. Entendiendo al conocimiento, sin tratar de realizar una definición, como la asociación del cúmulo de información, su interpretación y la experiencias de las personas que se conjugan para la operación de una institución.

No obstante, el aplicar esfuerzos para mejorar la calidad de la información, por ese solo hecho, no se mejora el conocimiento, por lo que un modelo de gestión de la calidad de la información, debe acompañarse con un modelo de gestión del conocimiento, conjugando las capacidades y experiencias del personal de la institución, las estrategias de la institución en materia de información y el empleo de las nuevas tecnologías de la información que puedan llevar al establecimiento de un modelo genérico de Gestión de la Información en los sistemas registrales.

Finalmente, es importante resaltar que hoy en día uno de los activos más importantes de las instituciones lo constituye la información y los modelos diseñados para hacer uso de la misma y generar el conocimiento necesario para el óptimo funcionamiento de éstas, por lo que los sistemas registrales constituyen la principal herramienta para la operación de las mismas. Ya sean compañías privadas o instituciones de gobierno, ambas llevan sistemas de registro para ofrecer sus servicios a sus clientes o ciudadanos. Es por esto, y a fin de mejorar la eficacia de las mismas, que es necesario incorporar en las estrategias operativas la gestión de la información y con ello la del conocimiento. Para lograr esto, se plantea el presente trabajo de tesis bajo el título de: Modelo de Gestión de la Información de Sistemas Registrales: Un enfoque práctico.

La hipótesis que se plantea en este trabajo de tesis parte de la idea de que en los sistemas registrales se cuenta con una gran cantidad de información, en lo que una mala calidad de la misma se convierte en un

factor desfavorable para la operación de las instituciones y para la generación o acumulación de conocimiento. Por lo que se pretende desarrollar una propuesta de Gestión de los sistemas registrales atendiendo principalmente el enfoque a la calidad de la información y a la generación del conocimiento.

A lo largo del trabajo se hará referencia a diversos modelos registrales, pero se pondrá especial énfasis en el Registro Federal de Contribuyentes (RFC), el cual hoy reviste gran importancia por ser el instrumento de control del Gobierno Federal para el pago de impuestos y que es el principal insumo para garantizar la recaudación tributaria.

Se escogió el RFC ya que por su complejidad, nos permite desarrollar ampliamente un modelo de gestión de la información en los sistemas registrales. Debido a la confidencialidad de la información, se presentará sólo aquella que sea de dominio público, en el caso de las diversas propuestas y ejercicios que se planteen se mostrarán siempre datos ficticios o bien datos que hayan sido obtenidos de alguna fuente de información de dominio público, como la que es obtenida en Internet.

II. Planteamiento del Problema.

Cuántas veces hemos escuchado la frase “tengo bomberazo”, y en cuántas ocasiones esta frase está asociada a la necesidad de producir información para que la gerencia de una institución tome decisiones. En cuántas ocasiones la producción de información por bomberazo conlleva a la generación de datos erróneos y más allá, cuántas veces ha sido necesario responder que no se dispone de información.

Si bien, en muchas ocasiones esto obedece a una mala planeación de las actividades a realizar dentro de las instituciones, también se debe, en un gran número de casos a que los sistemas de información no producen los datos necesarios y útiles que permitan crear el conocimiento que apoye la toma de decisiones y también en gran medida a la falta de calidad en la información.

Bryan Aucoin, Arquitecto de Datos en la Central de Inteligencia del gobierno de los Estados Unidos durante el año 2003, identifica como uno de los factores más críticos en la calidad de los datos que enfrentan muchas de las organizaciones, tanto en el sector público como en el privado, a la imposibilidad de que los directivos consigan respuestas confiables y oportunas a preguntas aparentemente simples, respuestas que necesitan para tomar decisiones informadas.

Como ejemplo, en el caso del Servicio de Administración Tributaria (SAT) de México, podríamos plantear la siguiente pregunta: Si se envía una carta invitación a todos los contribuyentes para que se acerquen al SAT a realizar su declaración y pagar impuestos, ¿Cuántos responderían?

Es una pregunta realmente simple, pero si comenzamos a examinar todos los factores por los que un contribuyente podría no atender al llamado del SAT, encontraríamos una gran cantidad de ellos; sin embargo, sólo unos cuantos de ellos son medibles, y pocos son controlables por la institución.

Así, tratando de acotar a los factores aparentemente controlables por la institución, restringiéndonos al universo de estudio de este trabajo (calidad de los datos) y, partiendo del supuesto de que al ser una autoridad con carácter fiscalizador, todos los contribuyentes atenderían al llamado y suponiendo que el Servicio Postal Mexicano puede cubrir todo el país, la respuesta se centraría sólo en conocer a cuántos

contribuyentes, con la información de la que dispone el SAT, realmente les llegará la carta. Lo que dependerá básicamente de los siguientes aspectos:

1. Qué tan exacta es la información del domicilio de los contribuyentes, es decir: ¿la información que se encuentra en la base de datos corresponde a la realidad?, ¿se sabe cuál es el nivel de desactualización de los domicilios?

2. Si queremos hacer llegar la carta a todos los contribuyentes, la pregunta a responder sería: ¿se cuenta con la totalidad de la información del domicilio en cada uno de los registros para que la carta llegue? y una pregunta aún más importante ¿están en la base de datos la totalidad de los contribuyentes?

3. Otra pregunta importante a contestar sería: ¿se cuenta ya con la información actualizada en la base de datos de los contribuyentes que realizaron la actualización de su domicilio en los últimos días?

4. ¿El domicilio está lo suficientemente detallado para que el servicio postal pueda realizar su trabajo?

A las preguntas antes realizadas, se agregan otras de igual sencillez, pero no menos complejas de responder, por ejemplo: Consientes de la dinámica poblacional y sobre todo del dinamismo que tiene el comportamiento de la economía, la variabilidad de la información de los contribuyentes puede ser muy grande, en este sentido surgen las siguientes preguntas: ¿cuál sería la probabilidad de localizar a un contribuyente en su domicilio fiscal? o ¿es factible encontrar a un contribuyente cuando éste no fue localizado en el domicilio que se tiene registrado?.

Dar respuesta a estas preguntas no es posible con el solo hecho de garantizar la calidad de la información de la base de datos, sobre todo si conocemos que existen contribuyentes que no quieren ser localizados por el SAT o que simplemente, por los horarios de servicio, no pueden ser encontrados en sus domicilios ya que estos laboran en los mismos horarios de oficina que el personal del SAT. Así, para responder a este tipo de preguntas, se debe trabajar en métodos para generar conocimiento, con base en la información que tanto el SAT posea como de aquella que pueda gestionar. Por ejemplo, si en una base de datos del pago del impuesto sobre automóviles nuevos se tiene un domicilio más reciente que el del RFC, seguramente será más factible localizar al contribuyente en este nuevo domicilio.

II.1 La calidad y la cobertura.

El SAT ha realizado esfuerzos por conocer el grado de calidad de su padrón del RFC, lo cual se pone de manifiesto en el documento publicado el 28 de septiembre del 2005 en su portal, en él menciona las cifras de calidad y cobertura del Padrón de Contribuyentes y reconoce que "El RFC (Registro Federal de Contribuyentes) es un sistema dinámico en el que de manera constante ocurren movimientos de inscripciones, cambios de denominación o razón social, cambios de domicilio, aumento o disminución de obligaciones, apertura y cierre de establecimientos, conversión de asalariado a trabajador por cuenta propia, fallecimientos, entre otros, por lo que es necesario mantenerlo actualizado, de acuerdo a los cambios que ocurren en la realidad económica del país". (Fig. 1).

Fig. 1: Imagen de la página del SAT
http://www.sat.gob.mx/sitio_internet/servicios/campanas/cumplimiento_voluntario/111_4515.html)

El complemento a 100% de los datos encontrados en el portal nos lleva a una cobertura del 56.5% y una calidad de la información de estos contribuyentes del 60.5%, así podríamos aventurarnos a hacer el producto de estos números para calcular cual es la probabilidad de que un contribuyente esté registrado y que sus datos estén bien en el RFC,

en otras palabras podríamos obtener un indicador sobre el índice de calidad global del RFC:

Índice de calidad del RFC = (Exactitud) (Totalidad)

Exactitud = Calidad = 60.5%

Totalidad = Cobertura =56.5%

Índice de calidad del RFC = (60.5%) (56.5%)

Índice de calidad del RFC = 34.18%

Es decir, el RFC para los fines del cumplimiento de la Misión del SAT, "Recaudar las contribuciones federales y controlar la entrada y salida de mercancías del territorio nacional, garantizando la correcta aplicación de la legislación y promoviendo el cumplimiento voluntario y oportuno", sólo está en posibilidades de fiscalizar y exigir el cumplimiento de poco más de una tercera parte de los contribuyentes obligados a estar inscritos y pagar impuestos.

A partir de estos resultados, queda de manifiesto la carencia de un sistema de gestión de la calidad de la información en el SAT para la operación del Registro Federal de Contribuyentes. Cabe mencionar que la medición de la calidad se basa en sólo una variable y no especifica la calidad de otras variables como son: datos de identificación, actividad preponderante, obligaciones fiscales, régimen de tributación, que son variables de sumo interés para el desarrollo de las funciones de esa Institución, por lo que se presume que los resultados de la calidad, atendiendo a las variables críticas para el desempeño de sus funciones, pueden ser altamente preocupantes.

II.2 Acceso a otras fuentes de información.

Ante el problema que enfrentan los sistemas de registro para tener información de calidad y completa, en muchas ocasiones se ven en la necesidad de recurrir a fuentes de información externa para completar la información o bien subsanar las deficiencias de calidad que sufren.

Por ejemplo, el Registro Nacional de la Pesca, dependiente de la Comisión Nacional de Pesca, es un Registro Público, en el cual se lleva el

control de las Embarcaciones Pesqueras. Este registro, enfrenta el grave problema de que no tiene registrado al total de embarcaciones que practican la actividad pesquera, sobre todo las embarcaciones menores ya que un gran número de ellas no posee permiso para realizar ésta, por lo que no se acercan a dicha autoridad. Por lo anterior, y en el supuesto de querer emprender un proceso de regularización, tendría que conocerse cuál es el universo de embarcaciones que están en factibilidad de dedicarse a la actividad pesquera, esto lo podrían lograr mediante la aplicación de un censo, lo cual requeriría una gran inversión de recursos, o bien se podría recurrir a otros sistemas de registro, como es el caso de Registro Público Marítimo, en el que se debe inscribir cualquier tipo de embarcación o instrumento flotante que sirva para la navegación a fin de poder hacerse a la mar.

Los sistemas de registro de clientes de los bancos, sólo tienen en sus bases de datos al total de clientes del banco, pero estos sistemas no cuentan con la información que les pueda identificar a los clientes potenciales de ciertos productos bancarios, por lo que recurren a empresas de mercadeo, que tienen bases de datos especializadas de personas y categorizadas de acuerdo al nivel de ingreso o probabilidad de compra de éstas, para realizar campañas sobre sus productos. Al respecto, la Asociación Mexicana de Agencias de Investigación de Mercados y Opinión Pública, tiene desarrollado una batería de preguntas para calificar a las personas de acuerdo a su nivel socioeconómico.

El SAT, enfrenta el gran problema de no poder localizar a cerca del 40% de los contribuyentes por deficiencias en la calidad del domicilio almacenado en el RFC, lo que lo enfrenta a la imperiosa necesidad de obtener información de fuentes externas para hacerse de nuevos domicilios de los contribuyentes y ejercer sus acciones de fiscalización o cobro.

II.3 Transportabilidad.

Una vez obtenida la información de fuentes externas, el gran problema radica en la transportabilidad, es decir en la facilidad de llevar los datos de un individuo de una base de datos a otra.

En la introducción a este trabajo, se dio una categorización muy genérica de los sistemas registrales, más como un ejercicio para mostrar la gran cantidad y diversidad de sistemas registrales que existen, que el tratar de realizar una clasificación ortodoxa.

Esa categorización nos permite imaginar en cuantos sistemas de registro estarán nuestros nombres, además pensemos que no se consideraron los sistemas de registro que utilizan las empresas para mercadeo, ni encuestas que hemos contestado en algún momento de nuestra vida, las listas de correo electrónicos en las que podemos estar, o las bases de datos de claves y passwords de diversos servicios en Internet. En fin, es imposible tener una idea exacta de en cuántos sistemas de registro dejamos huella a lo largo de nuestras vidas.

El lector podrá imaginar el gran cúmulo de información que existe sobre cada individuo en las diferentes bases de datos, ¿podría algún psicólogo desarrollar una nueva teoría del comportamiento, con base en nuestro andar por los diversos sistemas registrales?, ¿las compañías de mercadeo podrían predecir nuestros gustos e intenciones de compra?, ¿las empresas podrían decidir sobre que servicios ofrecer sus clientes? o instituciones como el Servicio de Administración Tributaría, ¿podría establecer patrones de comportamiento de los contribuyentes que pretenden eludir o evadir al Fisco? Quizá algunas de estas preguntas podríamos responderlas, sin temor a equivocarnos, y una vez que nos damos cuenta de que nuestra vida está registrada en muchos lugares, con un Sí.

Lo anterior nos abre la pauta a uno de los grandes problemas que enfrentan los sistemas de registro en nuestro país: La Transportabilidad. Definiendo a ésta como la capacidad de transportar datos de un registro a otro, o sea, la facilidad para intercambiar datos entre diversos sistemas registrales a través de la existencia de una clave en común.

En México el problema no ha sido resuelto y no se ve solución, al menos en el corto o mediano plazo, aun con los esfuerzos de la Secretaría de Gobernación, a través de la instauración del RENAPO (Registro Nacional de Población) para impulsar el uso de la CURP (Clave Única de Registro de Población). Éste es un instrumento que sirve para registrar en forma individual a todos los habitantes de México, nacionales y extranjeros, así como aquellos mexicanos que radican en otros países.

No obstante la existencia de la CURP, y siendo de uso obligatorio en los registros de personas de las dependencias y entidades de la Administración Pública Federal, su uso como clave de identificación única aún no se ha difundido.

La CURP se integra con dieciocho elementos (Fig., 2), representados por letras y números, que se generan a partir de los datos contenidos en el

documento probatorio de tu identidad (acta de nacimiento, carta de naturalización o documento migratorio), y que se refieren a:

- El primero y segundo apellidos, así como al nombre de pila.
- La fecha de nacimiento.
- El sexo.
- La entidad federativa de nacimiento.

Los dos últimos elementos de la CURP evitan la duplicidad de la Clave y garantizan su correcta integración.

EJEMPLO:

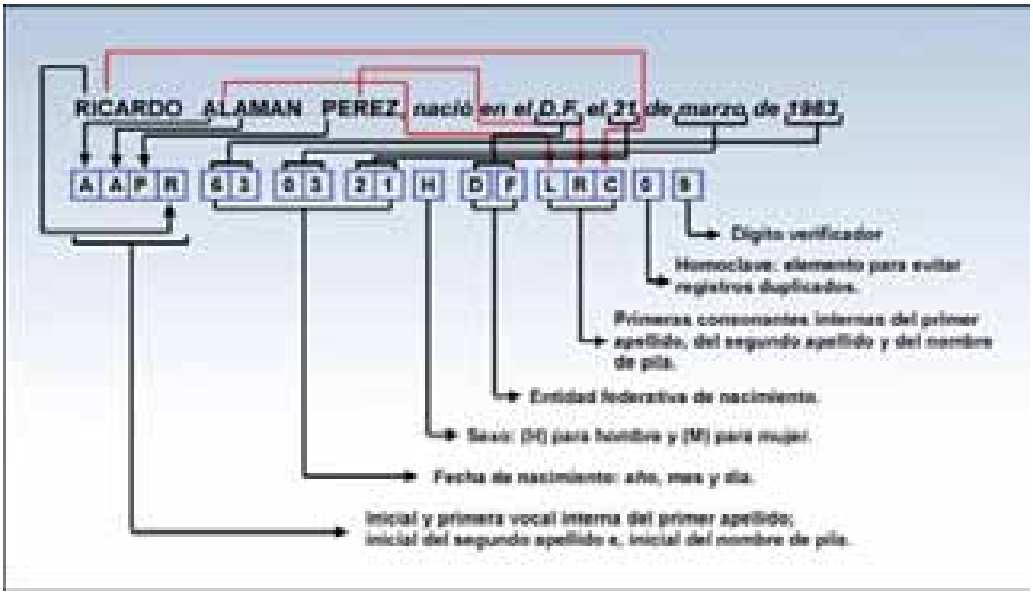


Fig. 2: Imagen de la página de SEGOB: http://www.segob.gob.mx

Según los resultados de la auditoria realizada al sistema de Registro de la CURP, cuyos resultados generales fueron dados a conocer por el Universal a finales del año 2002 (Fig. 3), se encontraron inconsistencias en 17 millones de los 90 millones de registros de esa base de datos.
Si se realiza, al igual que para el RFC, un cálculo del índice de calidad se obtiene:

Índice de Calidad de la CURP = (90 – 17 / 90)

Índice de Calidad de la CURP = 81.1% de calidad


EL UNIVERSAL.com.mx
México

Hallan en clave única 17 millones de fallas

Fig. 3: Imagen de la página del periódico El Universal
http://www2.eluniversal.com.mx/pls/impreso/noticia.html?id_nota=105925&tabla=notas

Para la fecha en que el Universal realiza la publicación de la nota, se menciona cerca de 300 trámites se realizan con la CURP –al día de hoy deben ser mucho más–. Estos problemas con el tiempo debieron incrementarse, debido a que no existe, o al menos no es palpable en la experiencia de los usuarios del registro, un mecanismo ágil de actualización. Existe la posibilidad de que se generen errores en el registro que quieran ser corregidos por las personas, al hacer esto, sólo se corrigen los datos en la institución que realizó el trámite, por ejemplo el IMSS (Instituto Mexicano del Seguro Social), y en los registros de gobernación, pero los cambios no se ven reflejados en las demás instancias del gobierno federal, y mucho menos, en las instituciones de los gobiernos estatales o privadas.

Analicemos ahora a otros dos grandes sistemas de registros: IMSS y el SAT. La clave del RFC del Servicio de Administración Tributaria se construye de forma muy similar a la de la CURP, ésta es una clave de tipo alfanumérico a trece posiciones, los primeros 10 son similares a los de la CURP, de hecho la CURP se basó en el RFC para generar su propia

clave. La gran pregunta es: ¿por qué no la compartieron? Los problemas del RFC ya fueron discutidos en un punto anterior.

Se sabe que el SAT ha hecho esfuerzos por incorporar la CURP a sus sistemas registrales, incluso tiene un enlace de red para comunicación directa con RENAPO para validar la CURP. El problema que se enfrenta es que RENAPO sólo distingue si la CURP fue o no emitida por esa instancia, y su capacidad para determinar si un registro es duplicado o no, se limita a que el usuario utilice el mismo documento probatorio para su incorporación a la CURP.

El Instituto Mexicano del Seguro Social, es una de las instituciones que un mayor número de personas tiene en sus sistemas de registro. Se basa principalmente en 2 grandes sistemas registrales: El Patronal y el de Beneficiarios de la Seguridad Social con más de 15 millones de asegurados permanentes y 25 millones de familiares beneficiados (Fig.4). En ambos registros se utilizan claves numéricas para la identificación de las personas, se sabe que ha realizado un esfuerzo por incorporar a sus sistemas el uso de la CURP y del RFC. Las claves que el IMSS utiliza son conocidas con los nombres de; NSS, Número de Seguridad Social y NRP, Número de Registro Patronal.

Población Derechohabiente (Asegurada)
2006

Delegación	Asegurados					
	Por temporalidad		Por ámbito		Por condición laboral	
	Permanentes	Eventuales	Urbano	Campo	Trabajadores	No trabajadores
Total Nacional	12638480	[illegible]	17981000	324000	14080367	4204833
Aguascalientes	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Baja California	428902	104536	724384	8054	661374	71864
Baja Calif. Sur	99085	38428	128151	8944	118758	13283
Campeche	115158	38480	145251	8557	119230	28418
Coahuila	508485	78791	873852	11423	526236	58040
Colima	117180	28548	150834	5054	96542	48788

Fig. 4. imagen obtenida del portal del IMSS: http://www.imss.gob.mx/dpm/dties/

En resumen, no existe una clave de acceso en común confiable entre los diferentes sistemas de registro, por lo que cualquier iniciativa para transportar datos de un registro a otro y hacer uso de la información requiere de un gran esfuerzo por las instituciones para realizar los procesos de identificación.

Con la metodología planteada más adelante para el procesamiento de cruce de bases de datos a fin de realizar la identificación de personas, podría iniciarse un camino para la solución de este problema. Podría en un futuro con los esfuerzos pertinentes, que las bases de datos del SAT o del IMSS incorporen en sus registros las tres grandes claves (CURP,

RFC y NSS) servir de base para la construcción de catálogos puentes entre las claves de esos sistemas de registro y así impulsar de manera decidida la constitución de una clave única de población, tanto para la instancias federales, como las locales y las privadas.

Países como Estados Unidos, España y otros desarrollados tienen claves únicas mediante la cual se identifican de forma única las personas, es de uso generalizado el número de seguridad social, que tal cual, o con algunas modificaciones es utilizado para el registro de personas en los sistemas registrales.

No obstante, el problema persiste y en el desarrollo de la tesis se abordará una posible solución.

II.4 Sistematización del intercambio de información.

Aunado al problema de la transportabilidad, se suma la falta de sistematización para el intercambio de información con agentes externos.

Aun teniendo a disposición la información, se pierde la oportunidad de la misma debido a que se requiere invertir mucho tiempo para lograr transportar los datos de una a otra base de datos y lograr usar la información.

Es necesario identificar con qué frecuencia se requerirá obtener información de fuentes externas tanto para enriquecer la información de los sistemas de registro y mejorar la calidad de los mismo, como para contar con información que nos permita ampliar el conocimiento que se tiene sobre el universo de personas al que corresponde el sistema de registro.

En el caso del Servicio de Administración Tributaria, el uso de fuentes externas de información puede ser de gran relevancia, ya que mediante éstas podría:

- Obtener domicilios de contribuyentes a los que no localiza,
- Contar con Información para identificar bienes de los contribuyentes morosos y ejecutar procedimientos de embargo,
- Planificar mejor sus procesos de auditorías a contribuyentes mediante una identificación del comportamiento del sector

económico en comparación con la estructura financiera de las empresas y

- Disminuir el volumen de cartas o notificaciones que no son entregadas por no localizar al contribuyente.

Cuando las condiciones de necesidad de información son permanentes, se requerirá establecer los mecanismos de intercambio de información y el desarrollo de lo que se conoce como “Data Factory”, a fin de crear de forma institucional procesos permanentes de acopio, procesamiento y gestión de la información.

II.5 Uso de la información para generar conocimiento.

El uso de fuentes externas de información como medio para generar conocimiento es una acción más del modelo de gestión de la información que se plantea. En el caso del SAT, conocer al contribuyente y usar la información como un monitor de la dinámica que estos tienen en su “Ciclo de Vida” facilitaría las tareas de la Administración Tributaria.

Para el SAT es imperativo contar con modelos de conocimiento acerca del contribuyente basados en fuentes externas, que apoyen los mecanismos de: actualización de datos, localización e incorporación de contribuyentes, fiscalización y cobranza, y segmentación que permitan realizar análisis sobre éstos y su actividad económica.

En el caso particular del SAT, el desarrollo de un modelo de gestión de la información del Registro Federal de Contribuyentes apoyará a hacer más eficiente la administración tributaria, y combatir la evasión, el contrabando y la informalidad e institucionalmente se podrá:

- Mejorar la efectividad en las actividades de control de obligaciones, fiscalización y cobranza. Apoyando en la planeación de las auditorías fiscales se espera incidir en los procesos de selección de los contribuyentes a auditar, lo que traerá una recuperación en la recaudación por posible evasión de impuestos,

- Disponer en el área de cobranza de información para localizar a contribuyentes que estén no localizados y se puedan ejercer acciones de cobro. Además, contará con información de bienes para iniciar los embargos necesarios que garanticen el interés fiscal y

- Apoyar en los procesos de diligenciación, puesto que los esfuerzos de verificación de domicilios podrán ser orientados hacia contribuyentes con mayor riesgo de no localización y se contará con elementos para la localización efectiva de estos.

Fortalecer el Registro Federal de Contribuyentes

- El área de Geografía se beneficiará al poder geo-referenciar a los contribuyentes del RFC, de manera más eficiente, con lo que se podrán realizar análisis estadísticos, en sentido espacial, sobre las diferentes variables socioeconómicas y fiscales.

- Se dispondrá de información necesaria sobre los contribuyentes que servirán de apoyo en las estrategias y en la planeación para la actualización del padrón mediante procesos masivos o de campo.

- Se contará con información para la incorporación de contribuyentes activos y potenciales, así como para la actualización de contribuyentes no localizados.

Transformar los datos en conocimiento para la toma de decisiones

- Con los modelos de conocimiento del contribuyente se podrán establecer relaciones entre contribuyentes, que van desde empleado-patrón hasta posibles relaciones de parentesco entre personas. Se podrán definir indicadores de riesgo con base al grado de formalidad (cumplimiento) del contribuyente e indicadores de signos de riqueza.

- Se podrá apoyar en la creación de modelos de segmentación y de riesgo y en general realizar análisis sobre el comportamiento del contribuyente basado en la aplicación de técnicas estadísticas y de minería de datos.

- Para generar un verdadero conocimiento del contribuyente se requiere una cultura de calidad de la información, la cual promueva el uso de estándares de nombres y domicilios, así como catálogos cuyo uso en las diferentes Entidades Federativas e Instituciones facilite en el futuro el intercambio de información entre estas.

II.6 Impacto en las instituciones.

Con el gran desarrollo de las tecnologías de la información, el crecimiento exponencial que el Internet ha tenido y con él al avance en los sistemas computacionales y de almacenamiento, el ser humano se enfrenta como nunca antes a un mundo en el que la información se convierte en el motor de la sociedad y de su desarrollo, hoy en día vivimos en la edad de la sociedad de la información.

Un ejemplo de este gran cúmulo de información y de los esfuerzos por que ésta se comparta y que a partir de ella se genere conocimiento lo es el portal de Internet WIKIPEDIA. En este portal se encontró la siguiente definición:

> "Una sociedad de la información[2] es aquella en la que la creación, distribución y manipulación de la información forman parte importante de las actividades culturales y económicas. La sociedad de la información es vista como la sucesora de la sociedad industrial […] y no está limitada a Internet, aunque esta ha desempeñado un papel muy importante como un medio que facilita el acceso e intercambio de información y datos.
>
> El reto para los individuos que se desarrollan en todas las áreas de conocimiento es vivir de acuerdo a las exigencias de este nuevo tipo de sociedad, estar informados y actualizados, innovar, pero sobre todo generar propuestas y generar conocimiento, el cual surge de los millones de datos que circulan en la red."

El gran cúmulo de información al cual pueden acceder las instituciones y por consiguiente con la cual pueden ser integrados, actualizados o enriquecidos los sistemas de registro es enorme y plantea como gran reto el poder administrar la información de tal forma que se pueda generar el conocimiento institucional requerido para que los procesos y servicios institucionales sean eficaces y eficientes.

La hipótesis que se plantea es que poniendo foco en los temas de calidad de la información se puede mejorar la eficiencia y eficacia de las instituciones, ya que los problemas de calidad la información se convierte en un factor desfavorable para que las éstas cumplan los objetivos que persiguen.

2 http://es.wikipedia.org/wiki/Sociedad_de_la_informaci%C3%B3n

Como podrá suponerse, el problema de la calidad de la información, conlleva a problemas en la interpretación y uso que se le dé a la misma, de forma general podríamos mencionar que afecta la generación de conocimiento. Entendiendo al conocimiento, sin tratar de realizar una definición en este momento, como la asociación del cúmulo de información, su interpretación y las experiencias de las personas que se conjugan para la operación de una institución.

Como se mostró específicamente para los casos del SAT y del RENAPO, y en general es un problema que existe en los diferentes sistemas de registro, es común encontrar problemas de calidad en tres grandes apartados: 1) que los datos están escritos de forma correcta, 2) que la información registrada corresponda a la realidad y 3) que se tenga al universo de las personas que deben conformar el Registro.

La falta de calidad en la información en alguna de esas vertientes, necesariamente traen consigo costos indirectos a las instituciones que muchas veces son muy difíciles de medir e incluso de percibir; sin embargo, es un tema en el que tradicionalmente las instituciones no centran sus estrategias para reducir costos y mejorar la calidad de sus servicios.

Con base en la información publicada por el SAT en el resumen ejecutivo de los libros blancos correspondientes a la entrega de la administración federal durante los años 2000-2006[3] y tomando como referencia costos de mercado de diversos servicios se realizó un ejercicio de costos para tratar de demostrar el impacto en términos monetarios en los que pueden incurrirse por una mala calidad de la información.

Para mostrar el impacto de la calidad de la información se costeará una de las etapas de la campaña de actualización del RFC que el SAT realizó en agosto del 2005. En la campaña realizada se trató de enviar cartas a contribuyente en San Luis Potosí en los municipios de Soledad Graciano Sánchez y de San Luis Potosí a fin de invitarlos a actualizar su información.

El envió de cartas del SAT fue por un total de 153,346 cartas, a domicilios fiscales y otros domicilios alternos de los contribuyentes. Según las cifras de calidad, el máximo esperado de entrega de cartas sería del 60.5%, suponiendo un margen de error del 5%, que no es

3 Portal del SAT, módulo de transparencia, sección de libros blancos:
http://www.sat.gob.mx/sitio_internet/transparencia/51_8956.html

especificado en las cifras del SAT, el total de cartas que se esperaría entregar estaría en el rango del 55.5% y 65.5%.

Según los resultados que se mencionan, no se logró entregar el 40% de las cartas, por lo que el porcentaje de cartas entregadas queda dentro del rango estimado para la calidad del padrón del RFC.

Mediante estudios de mercado se obtuvieron algunos precios unitarios para estimar los costos en los que se incurrió en este universo de contribuyentes, la siguiente tabla (Tab. 1) muestra los resultados:

Operativo de Correspondencia en San Luis Potosí y Soledad Graciano Sánchez 15 agosto del 2005 Costos estimados según invetigación de mercado de precios unitarios	Unidades	Precio Unitario	Costo Total
Servicios de Impresión y Correspondencia			
Impresión de cartas, ensobretado y personalización	153,346	2.30	352,695.80
Envío de Correspondencia por SEPOMEX con acuse de recibo	153,346	22.25	3,411,948.50
Costo total estimado por la producción y envío de cartas		24.55	3,764,644.30
IVA			564,696.65
Costo total con IVA			**4,329,340.95**
No se logró entregar el 40% de las cartas	61,338		
Precio Unitario Con IVA		28.23	
Costo estimado de las cartas que no lograron entregarse	**61,338**	**28.35**	**1,738,943.64**

Tab. 1: Costo estimado de las cartas que no pudieran entregarse por deficiencias en la calidad de los datos.

Como puede inferirse, derivado de los datos de calidad, eran de esperarse estos resultados; sin embargo, sirve este ejercicio para corroborar la hipótesis de que los problemas de la calidad de la información en los sistemas de registro traen consigo costos indirectos y no permiten a las instituciones lograr el cumplimiento de sus objetivos.

Como se ha mencionado, los costos asociados a los problemas de calidad en los sistemas de registro son difíciles de medir; sin embargo, y a efectos de tener un valor estimado del costo asociado a los problemas de calidad que enfrenta el padrón de contribuyentes del SAT, se utilizan los datos anteriores para hacer una aproximación, los resultados se muestran en la siguiente tabla (Tab.2):

Costos de un Operativo de Correspondencia para todo el padrón del RFC Se excluyen asalariados Costos estimados según investigación de mercado de precios unitarios	Unidades	Precio Unitario	Costo Total
Servicios de Impresión y Correspondencia			
Impresión de cartas, sobretodos y personalización	12,418,797	2.90	36,027,530.13
Envío de Correspondencia por SEPOMEX con acuse de recibo	12,418,797	22.25	288,181,230.25
Costo total estimado por la producción y envío de cartas		25.15	324,208,760.38
IVA			51,873,401.66
Costo total con IVA			376,082,162.04
Se espera no lograr entregar el 59.5% (según datos de censo)	5,317,217		
Precio Unitario Con IVA		36.23	
Costo en el que se incurre por la calidad del padrón	5,317,217	36.24	192,708,427.68

Nota: Datos del Padrón del RFC + agosto del 2006 obtenidos de las Cifras Oficiales de la entrega de la Administración del ejercicio 2000-2006 publicadas en el portal del SAT

Tab. 2. Costo estimado en el que se incurre por la calidad de los datos del Registro Federal de Contribuyentes

En este ejercicio se excluyó el universo de contribuyentes asalariados ya que el pago de impuestos de estos se realiza a través del patrón el cual tiene la obligación de realizar las retenciones de impuestos correspondientes y enterarlas al SAT, además el proceso de inscripción al RFC lo realiza el patrón, por lo cual en ese registro queda asentado como domicilio fiscal el domicilio del empleador.

Por otra parte, como quedó de manifiesto por la calidad del padrón, no es posible localizar a cerca del 40% de los contribuyentes, por lo que, si aplicamos esta proporción a los contribuyentes que el SAT tiene en su cartera de deudores del fisco (fig. 5), la cual asciende a más de 536 mil millones de pesos, se obtendrán resultados realmente impresionantes.

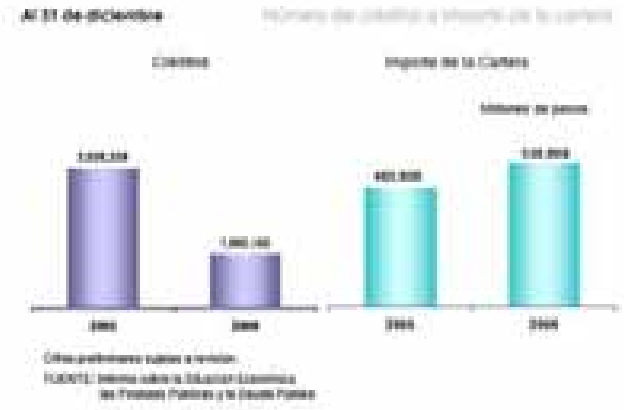


Fig. 5 Importe total en la cartera de créditos
Fuente: Portal del SAT, Informe Tributario de Gestión 2006, cuarta trimestre
http://www.sat.gob.mx/sitio_internet/informe_tributario/reporte_anual_2006_74/

Para realizar los cálculos se parte del hecho de que el proceso de medición de la calidad del RFC se realizó mediante un proceso muestral con selección aleatoria, en el que están representados, tanto los contribuyentes con créditos fiscales como los que no tienen créditos, y que el sentido común nos indicaría que será menos factible localizar a un contribuyente deudor del fisco que a uno cumplido, por lo que la estimación podría se inferior a la realidad; y aplicando nuevamente el factor de no localización de 39.5% que es el complemento a 100% se tienen los resultados que se muestran en la siguiente tabla (Tab. 3):

Costo en el que se incurre por los problemas de calidad en el padrón del RFC al no poder localizar a contribuyentes con créditos		
Total de Créditos	1,965,160	
% de Contribuyentes no factibles de localizar en su domicilio	39.50%	
Total de créditos asociados a no localización		**776,238**
Valor de la Cartera en millones de pesos	536,669	
Monto promedio de cada adeudo en millones de pesos	0.27	
Valor no recuperable de la cartera por no localización en millones de pesos		**211,984**

Tab. 3: Estimación del valor de los créditos no cobrables por deficiencias de la calidad de la información del RFC

Se reconoce que estos son grandes números y que al no disponer de la base de datos para hacer un análisis estadístico formal, se puede estar cometiendo errores, sobre todo por el tratamiento de los “Fuera de Rango”, ya que puede existir un grupo muy pequeño de contribuyentes con créditos multimillonarios; sin embargo, por sí mismo, el solo hecho de saber que no se podrá localizar a el 39.5% de los contribuyentes con créditos para ejercer acciones cobro es un dato realmente preocupante.

Con este par de ejercicios, acotados al quehacer diario del SAT, queda de manifiesto que esa institución requiere realizar grandes esfuerzos para mantener la calidad de la información de su padrón de contribuyentes, este trabajo de tesis, mostrará en los capítulos siguientes algunas iniciativas para incorporar en las instituciones procesos de gestión de la información que les permita disminuir los problemas asociados a la calidad de la información por una parte y por otra a la generación de conocimiento que a partir de ella se materializa.

Es claro, al menos en este primer ejemplo, que el conocimiento que se tiene sobre el contribuyente, estará directamente afectado por la calidad del padrón. Si no es posible localizarlo físicamente y éste no tiene la necesidad de acercarse a la autoridad fiscal, no podemos esperar que el resto de su información sea correcta.

En conclusión, es importante que las instituciones definan mecanismos orientados a medir la calidad de la información y realizar los esfuerzos necesarios para tratar de costear los impactos, aunque sea de forma indirecta. Así, se podrán tomar con mayor facilidad la decisión de incorporar una estrategia institucional para medir permanentemente la calidad de la información, y más aún, para establecer una política para una adecuada administración de la información de la empresa que abarque tanto los temas de calidad, como de generación del conocimiento, lo que en este trabajo estamos denominado “La Gestión de la Información en los Sistemas de Registro”.

No obstante, el realizar esfuerzos para mejorar la calidad de la información, por ese solo hecho, no se mejora el conocimiento, por lo que un modelo de gestión de la calidad de la información, debe acompañarse con un modelo de gestión del conocimiento, conjugando las capacidades y experiencias del personal de la institución, las estrategias de la institución en materia de información y el empleo de las nuevas tecnologías de la información se puede llegar al establecimiento de un modelo genérico de Gestión de la Información en los sistemas registrales.

Finalmente, es importante resaltar que hoy en día uno de los activos más importantes de las instituciones lo constituye la información y los modelos diseñados para hacer uso de la misma y generar el conocimiento necesario para el óptimo funcionamiento de éstas, por lo que los sistemas registrales constituyen la principal herramienta para la operación de las mismas. Ya sean compañías privadas o instituciones de gobierno, ambas llevan sistemas de registro para ofrecer sus servicios a sus clientes o ciudadanos. Es por esto y a fin de mejorar la eficacia de las mismas que es necesario incorporar en las estrategias operativas la gestión de la información y con ello la del conocimiento. Para lograr esto, se plantea el presente trabajo de tesis bajo el título de: Modelo de Gestión de la información de Sistemas Registrales: Un enfoque práctico.

III. El Modelo de Gestión de la Información.

III.1 Marco conceptual

Los ***datos*** son sucesos u observaciones acerca de fenómenos, son hechos objetivos sobre acontecimientos que son representados mediante códigos, palabras o imágenes, etc.

Los datos describen únicamente una parte de lo que pasa en la realidad y no proporcionan juicios de valor o interpretaciones, y por lo tanto no son orientativos para la acción. La toma de decisiones se basará en datos, pero estos nunca dirán lo que se debe hacer. Los datos no dicen algo acerca de lo que es importante o no. A pesar de todo, los datos son importantes para las organizaciones, ya que son la base para la creación de información[4].

La ***Información*** se genera mediante la transformación y el análisis de los datos relevantes y un propósito determinado, son la interpretación de los fenómenos observados que describen una situación a partir de los datos a los que se ha aportado un significado. Los datos se convierten en información cuando se les añade significado. Los datos son transformados en información al añadirles valor, para esto existen varios métodos:

Contextualizando: sabemos para qué propósito se generaron los datos.

Categorizando: conocemos las unidades de análisis de los componentes principales de los datos.

Calculando: los datos pueden haber sido analizados matemática o estadísticamente.

Corrigiendo: los errores se han eliminado de los datos.

Condensando: los datos se han podido resumir de forma más concisa.

El ***conocimiento*** tiene un alcance superior: implica la acción de la mente humana sobre la información. Se constituye a partir de la

4 http://www.gestiondelconocimiento.com/conceptos_diferenciaentredato.htm

información; representa una compresión del contexto, las relaciones dentro de un sistema, la capacidad para identificar puntos críticos y debilidades, aprender de experiencias del pasado y comprender las implicaciones futuras de las acciones llevadas a cabo para resolver problemas[5]. El conocimiento en más que información: se trata de información aplicada, ésta se transforma en conocimiento cuando se introduce en un modelo mental.

Para Davenport y Prusak[6] el conocimiento es una mezcla de experiencia, valores, información y "saber hacer" que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción. Como se mencionó, se origina por diversos procesos mentales.

El conocimiento se deriva de la información, así como la información se deriva de los datos. Para que la información se convierta en conocimiento, debe existir un proceso mental con la cual se produzca gracias a:

- Comparación.
- Consecuencias.
- Conexiones.
- Conversación.

Los datos se perciben mediante los sentidos, éstos los integran y generan la información necesaria para producir el conocimiento que es el que finalmente permite tomar decisiones para realizar las acciones cotidianas que aseguran la existencia social.

La ***acción*** es el ejercicio de la posibilidad de hacer, y es resultado del conocimiento que se genera en la mente de los individuos al procesar la información. La acción se deriva del proceso de toma de decisiones sobre el conocimiento generado, es lo que nos permite obrar y realizar actos libres y conscientes.

En los sistemas de información, los datos habitan en el mundo de los bits y bytes, en ellos, mediante el empleo de códigos se puede almacenar y representar la información que es de utilidad para la empresa.

---

5 "Investigación de mercados: Métodos de recogida y análisis de la información para la toma de decisiones", Juan Antonio Trespalacios Gutiérrez, Laurentino Bello Acebrón, Rodolfo Vázquez Casielles

6 "Working Knowledge: How Organizations Manage What They Know", Laurence Prusak, Thomas H. Davenport

La información no es una simple colección de datos, es necesario que exista una interrelación entre ellos para que estos se constituyan en información o bien que exista una asociación con objeto o fenómeno. La información habita en el mundo de los sistemas y programas, los cuales estableces las asociaciones necesarias para que los datos puedan ser interpretados.

Mientras que la información establece las asociaciones necesarias para entender los datos, el conocimiento vive en el mundo de la mente humana, se adquiere el nivel de conocimiento cuando se somete la información a un modelo mental.

La acción radica en el mundo de los hechos y de las consecuencias, es una correspondencia entre el conocimiento y las conductas que se asumen.

El nivel jerárquico de esto conceptos se puede observar en la pirámide informacional (Fig. 6).

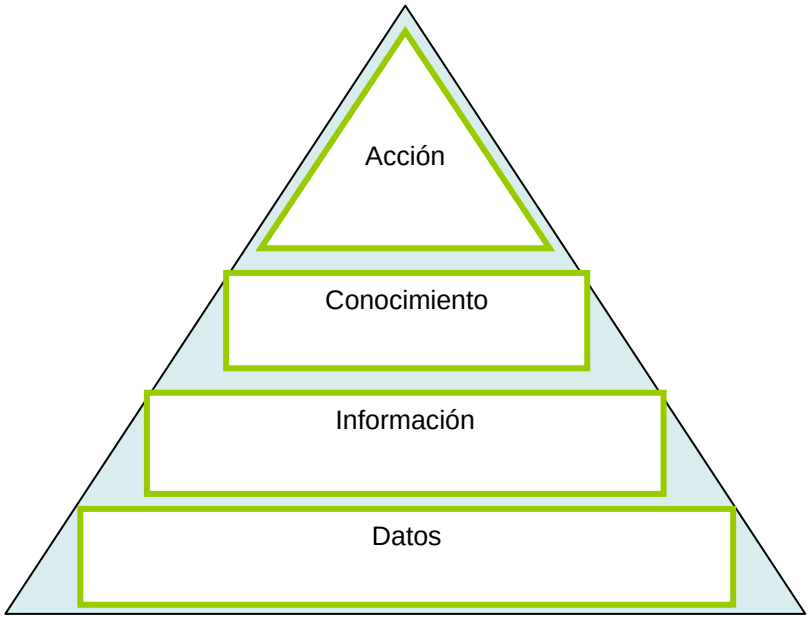


Fig. 6. Pirámide Informacional

Por fortuna, algunos modelos mentales, ya sean simples o complejos, pueden llevarse al campo de los sistemas informáticos mediante el desarrollo de algoritmos que permiten que el conocimiento sea aprovechado de forma eficiente por las empresas y en consecuencia automatizar la toma de decisiones sobre las acciones a seguir en el caso de que se presenten tales o cuales variables.

El concepto de ***registro***, desde el punto de vista informático, se trata de un conjunto de datos relacionados entre sí, que constituyen una unidad de información almacenado en una base de datos. El concepto, también nos remite a la acción y efecto de registrar, que quiere decir, grabar, inscribir, anotar, dejar constancia de los hechos; o bien, guardar datos en un sistema de información.

Ahora bien, entendemos como un sistema de registro al conjunto de datos estructurados y ordenados que pueden ser consultados para extraer de ellos información.

Para los fines del modelo de gestión de información de sistemas registrales acotaremos los sistemas de registros a aquellos cuyo fin se centra en el registro de personas y de los datos asociados a las mismas. En este sentido y para los fines de establecer una definición, nos basaremos en el concepto de Sistema de Datos Personales que el Instituto Federal de Acceso a la Información da en los lineamientos para la protección de datos personales.

Entenderemos como ***Sistema Registral***, a cualquier conjunto ordenando de datos que para su uso sean sometidos a un tratamiento informático y que por ende requieran de una herramienta tecnológica específica para su acceso, recuperación o procesamiento.

Gestión, el diccionario de la Real Academia Española la define, como la acción de gestionar, que es: “Hacer diligencias conducentes al logro de un negocio o de un deseo cualquiera”. En la acepción de diligencia, cuando se refiere al “hacer alguien sus diligencias” define el término como: “poner todos los medios para conseguir un fin”. En este sentido, el término “Gestión” se definiría como: “Poner todos los medios conducentes para conseguir el logro de un negocio o deseo”.

Quizá la palabra en inglés Management, sea un poco más representativa del enfoque que necesitamos para la definición del término Gestión, en su traducción literaria, quiere decir “gerencia”, es decir, la acción de dirigir y controlar un grupo de una o más personas o entidades con el fin de coordinar y de armonizar a ese grupo para lograr una meta.
Buscando un enfoque de procesos, Harold Koontz lo define como “el proceso mediante el cual se obtiene, despliega o utiliza una variedad de recursos básicos para apoyar los objetivos de la organización.”[7]

7 Harold Koontz & Heinz Weihrich, "Administración: Una perspectiva global", McGraw Hill, España, 1995

Con base en estas definiciones y para el desarrollo de este trabajo entendemos por gestión: *"Conjunto de acciones o procesos mediante las cuales se asignan, dirigen y controlan un grupo de recursos necesarios a fin de coordinarlos y armonizarlos para el logro de un objetivo"*.

III.2 Desarrollo del modelo de gestión de la información.

El modelo de gestión de la información que aquí se presenta, parte del hecho de que la información representa por un lado, el eje por el cual los datos pueden llegar a generar conocimiento, y por el otro, el mecanismo para la transmisión del conocimiento, en cuyo caso los datos representan la posibilidad de que esa información, mediante la cual se transmite el conocimiento, pueda ser codificada y registrada.

El enfoque planteado se divide en cinco grandes apartados: el primero es la adquisición de la información, seguido de los modelos de información necesarios para su uso, que son característicos en cualquier sistema de registro, y se introducen a fin de resolver la problemática planteada, la gestión del conocimiento, de la calidad de la información y la mejora continua.

Es importante aclarar en este momento que la gestión del conocimiento se refiere a la forma de que crear y utilizar el conocimiento mediante el usos de herramientas de software, mas que a las prácticas orientadas a la formación de competencias laborables y la administración del conocimiento en las instituciones.

Para iniciar la construcción del Modelo de Gestión de la Información, partamos del análisis del Ciclo Tributario de Gestión del Servicio de Administración Tributaria[8] (Fig. 7). "El Ciclo Tributario es el conjunto de esfuerzos que constituyen el quehacer fundamental de la Administración Tributaria. Se desenvuelve alrededor de un marco jurídico establecido, desde el mandato de Ley que el SAT recibe respecto al cobro de un impuesto e incluye el registro e identificación de los contribuyentes, el ejercicio de sus obligaciones y derechos, las acciones para asegurar su cumplimiento y la administración interna de los recursos de la organización para el soporte de sus actividades sustantivas" (Fi. 7).

8 Expedientes del Cambio Solución Integral de a Administración Tributaria

ftp://ftp2.sat.gob.mx/asistencia_servicio_ftp/publicaciones/folletos2006/FolletoSI.pdf

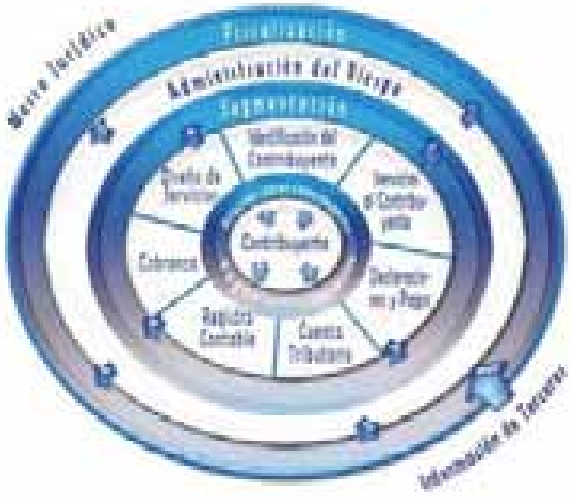


Fig. 7. Modelo Referencial del Ciclo Tributario
para la Solución Integral del SAT

Se elige este modelo para análisis debido a que en él, el eje de todas las operaciones es el contribuyente, según se manifiesta en el documento fuente de la información. Es decir, el Registro Federal de Contribuyentes se constituye en la columna vertebral de todos los procesos de la Administración Tributaria.

Fig. 8: Reorganización del Ciclo tributario.

Iniciamos por reorganizar el diagrama del ciclo tributario (Fig. 8), los diversos componentes del ciclo tributario se agrupan en 3 grandes ejes: Marketing (Adquisición de la información), Servicios (Modelos de Información) y Control (Conocimiento)

Como comentamos, el RFC es el eje de la Administración Tributaria, este sistema de registro está presente a lo largo del ciclo tributario y se presentará en cada una de las vertientes de la siguiente forma:

Marketing:

Este ciclo comienza con la acción de captar contribuyentes, para ello se deben emprender acciones orientadas a transmitir el mensaje a los contribuyentes de su obligación de estar inscrito en el RFC y de los beneficios que el país tiene por el pago de los impuestos. Esta campaña de comunicación debe acompañarse también de acciones que promuevan la constante actualización del RFC, así como el cumplimiento voluntario de las obligaciones fiscales.

Diversas estrategias de comunicación deben utilizarse para difundir los mensajes y crear una cultura fiscal (Civismo Fiscal) en la población, así los medios tradicionales de comunicación, como son la radio y la televisión se deben reforzar con acciones cara a cara, visitas a escuelas, cursos y seminarios, divulgación masiva en medios escritos o gráficos y con la incorporación del uso de tecnologías como son los portales en Internet, el correo electrónico.

La estrategia del Marketing se debe orientar principalmente a captar contribuyentes obligados y potenciales. Por eso, también es de gran importancia los acuerdos que se realicen con diversas instituciones, como bancos y universidades, para que sean un canal más de comunicación y de incorporación de contribuyentes.

Finalmente el marketing debe apoyar en la consolidación de la imagen institucional, si bien se debe transmitir la imagen de que se brindan servicios eficientes y de alta calidad, no debe dejarse a un lado la percepción de riesgo, es decir, fomentar el cumplimiento.

Servicios:

La estrategia de Marketing, debe estar muy bien soportada por el diseño de los servicios que permitan a la institución alcanzar el objetivo de captar al mayor número de contribuyentes obligados y potenciales. Los servicios que se otorguen deben tener una amplia gama de opciones para facilitar a la institución la captación de contribuyentes y conjugar estas acciones con las estrategias de control que sobre ellos se realizan.
Desde la perspectiva del contribuyente, los servicios deben diseñarse de tal forma que les facilite su incorporación al RFC y posteriormente el cumplimiento de sus obligaciones fiscales.

Como parte de los servicios que se otorgan, están también aquellas acciones de capacitación y fortalecimiento de la cultura fiscal, las acciones para facilitar el entendimiento de las leyes y aquellos tendientes a facilitar el cumplimiento de las obligaciones.

Uno de los retos de la Administración Tributaria es el llevar los servicios a cada rincón del país, para lo cual debe utilizar diversas estrategias como son el empleo de Internet, el uso del teléfono, unidades de atención móviles, entre otras. Su gran nicho de oportunidad es el poder incorporar al total de contribuyentes obligados y potenciales al Padrón del RFC.

Control:

El pago de un impuesto, como lo define la Real Academia Española, es la “Carga continua u obligación que impone el uso o disfrute de algo” o bien, “La obligación dineraria establecida por la ley, cuyo importe se destina al sostenimiento de las cargas públicas” o el “Tributo que se exige en función de la capacidad económica de los obligados a su pago”.
Como se puede ver, no tiene nada de voluntario, es por esto que se tienen que ejercer acciones de control y fiscalización para determinar el grado de cumplimiento de las obligaciones que son impuestas por la legislación.

Las acciones de control no se logran sin una adecuada identificación de los contribuyentes y de sus obligaciones fiscales, proceso que se da durante la identificación de los contribuyentes que constituyen el potencial recaudatorio. Éstas son producto de la necesidad de asegurar el cumplimiento de la Ley, por lo que a las

acciones de control siguen actividades tendientes a garantizar el interés fiscal.

Para llevar a cabo estas acciones es necesario, casi imperante, el contar con información de calidad y oportuna sobre todas aquellas variables que ayudan a conformar el potencial recaudatorio, ya sea del país, de un sector de la economía o de un contribuyente en específico. Para esto la acciones de identificación de contribuyentes, segmentación de los mismos, el uso de fuentes de información externa, en conjunción con técnicas de administración de riesgo, fiscalización y cobranza son muy necesarias.

El modelo de gestión de la información que se propone deriva del ciclo de gestión tributario, para ello incorporaremos las vertientes de calidad de la información y de mejora continua, que no habían sido contemplados y asimilaremos los términos de negocio correspondientes a Marketing, Servicios y Control con los de Adquisición de Información, Modelos de Información y Conocimiento, respectivamente. La siguiente figura (Fig. 9) muestra esos cambios y se constituye en nuestro modelo de gestión de la información:

Fig. 9 Modelo de Gestión de la Información en los sistemas de registro
Fuente: Diseño Propio

Entenderemos por Modelo de Gestión de la Información: *Al conjunto de acciones o procesos orientados a administrar la información contenida en los sistemas registrales a fin de garantizar la adecuada operación de las instituciones, la calidad de la información, la forma en que ésta se adquiere y enriquece y los procesos de generación y sistematización del conocimiento a fin de coordinar y armonizar estos recursos para el logro de los objetivos trazados por cada institución.*

Es importante resaltar que para los efectos de este trabajo, no se profundizará en los temas referentes al análisis de requerimientos de información, ni a las técnicas de modelado de datos, de análisis y de desarrollo de sistemas y modelos de información, esto por considerarse que son temas ampliamente desarrollados en la literatura de sistemas y que en general son llevados a cabo por las áreas de tecnología de la información de las instituciones.

Suele suceder que en las instituciones que llevan sistemas registrales, el modelo de gestión de la información, como aquí se propone, queda trunco y las áreas de tecnología se apropian del tema desarrollando sólo los procesos de análisis de requerimientos, adquisición de información e implementación de sistemas y modelos de información.

El planteamiento que se realiza pretende que la gestión de la información no sea exclusivamente un tema de las áreas de tecnología, sino que se transforme en una estrategia institucional. Por lo que en los capítulos siguientes se abordarán sólo los temas de Gestión de la Calidad, Gestión del Conocimiento y del Enfoque Estratégico de la Gestión de la Información en las Instituciones.

Los componentes del Modelo de Gestión de la Información

La adquisición de Información corresponde a todas aquellas acciones diseñadas para captar la información que se requiere en nuestro sistema de registro.

Es importante definir cuáles son los datos relevantes para que la institución opere. Los requerimientos de información principalmente se clasifican en:

1) Requerimientos Normativos: Son todos aquellos que son definidos por algún tipo de legislación o simplemente por las políticas de las instituciones.

2) Requerimientos Funcionales: Son todos aquellos indispensables para el funcionamiento de la organización en general están asociados a los procesos que en ella se desempeñan.

3) Requerimientos Temáticos: Son todos aquellos datos necesarios para atender necesidades muy especificas de alguna unidad de la organización y que no forman parte de los procesos de la misma.

4) Requerimientos Estratégicos: Son aquellos indispensables para conocer el desempeño de una institución, generalmente están asociados a la construcción de indicadores estratégicos.

Modelos de Información.

Las vertientes de servicios y control, nos llevan a la necesidad de crear los Modelos de Información que por un lado nos permita generar las condiciones tecnológicas necesarias para captar la información e integrar el sistema de registro y por el otro se diseñen los sistemas necesarios para la operación del mismo.

El diseño de los modelos de información deberá atender a los diferentes niveles que impactan a la operación de la institución, mismo que se muestran en la siguiente figura (Fig.10):

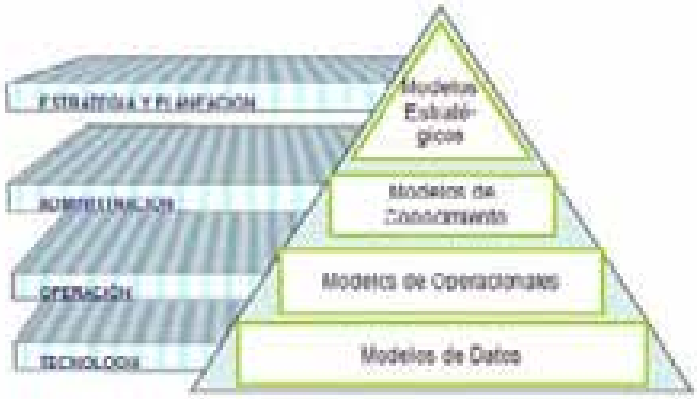


Fig. 10 Modelos de Información asociados a la organización

Los modelos de información estarán directamente asociados a los diferentes niveles de la pirámide informacional, siendo estos los habilitadores para que los datos se conviertan en información, estos a

su vez en conocimiento y a partir de éste se definan las estrategias a seguir en las organizaciones, la siguiente figura (Fig. 11) muestra este ciclo:

Fig. 11 Ciclo de la Información.

Entendamos como Modelos de Información al conjunto de conceptos, reglas, convenciones y algoritmos que son implementados a través de sistemas computacionales y nos permiten describir, manipular, analizar y explotar los datos que responden a las necesidades de información de una institución.

El nivel modelo de datos corresponde a la etapa de diseño de las estructuras de datos que responderán a alguno de los diversos tipos de requerimientos. Organizacionalmente es un tema cuya ejecución corresponde a las áreas de tecnología.

El nivel de los modelos operacionales, corresponde a todos aquellos sistemas desarrollados y operados para que la institución lleve a cabo sus procesos sustantivos.

Los modelos de conocimiento son aquellos que impactan en los niveles de dirección y toma de decisiones sobre los diversos procesos de la institución, aquí es donde se administra el conocimiento mediante la implementación de los sistemas informáticos que permiten a la institución establecer diferentes estrategias para actuar en función de las circunstancias que se presenten y que previamente fueron ya modeladas.

Los modelos estratégicos son aquellos enfocados a brindar información a la alta gerencia sobre el comportamiento de la institución, generalmente son modelos que permiten la alineación a los objetivos y estrategias institucionales de los procesos, proyectos y presupuestos.

La Gestión de la Calidad de la Información y del Conocimiento.

El potencial recaudatorio de la institución está directamente asociado a el número de contribuyentes que se logren identificar, a los esfuerzos para captar a los contribuyentes potenciales, al conocimiento que sobre los contribuyentes se tenga, al uso de modelos de incorporación y actualización de la información, así como a los modelos de enriquecimiento y consolidación de la calidad de la misma, sin dejar a un lado el propio andar de la economía y el marco legal vigente.

Es posible controlar un gran número de las variables asociadas, siempre y cuando se integren a las estrategias acciones encaminadas a mejorar la calidad y generar información que permita mejorar el conocimiento que sobre el contribuyente se tiene.

Gestión de la calidad de la información[9]: Conjunto de elementos mutuamente relacionados que interactúan para lograr el cumplimiento de los objetivos y políticas de calidad de la información. Comprende las actividades relacionadas con la obtención de la información, la consistencia de la misma, su disponibilidad y accesibilidad, su correspondencia con el mundo real y el mantenimiento que se le dé.

Gestión del conocimiento: Proceso mediante el cual se desarrolla, estructura y mantiene la información, con el objetivo de transformarla en un activo crítico y ponerla a disposición de una comunidad de usuarios definida con la seguridad necesaria. Incluye el aprendizaje, la información, las aptitudes y la experiencia desarrollada durante la historia de la organización y la generación de conocimiento. Para los fines de este trabajo, sólo abordaremos el tema de generación del conocimiento.

La Mejora Continua o ciclo Deming.

Finamente, se introduce el concepto de mejora continua, el cual se conoce como el proceso de Planificar, Hacer, Verificar, Actuar. Además de su importancia en el modelo de gestión, éste garantiza que el modelo

9 "Gestión de información, gestión del conocimiento y gestión de la calidad en las organizaciones", Lic. Lourdes Aja Quiroga.

madure conforme a las necesidades identificadas de la institución y a su propia evolución:

- **Planear**
    - Identificar el proceso a mejorar.
    - Recopilar datos para profundizar en el conocimiento del proceso.
    - Análisis e interpretación de los datos.
    - Establecer los objetivos de mejora.
    - Detallar las especificaciones de los resultados esperados.
    - Definir los procesos necesarios para conseguir estos objetivos, verificando las especificaciones.
- **Hacer**
    - Ejecutar los procesos definidos en el paso anterior.
    - Documentar las acciones realizadas.
- **Verificar**
    - Pasado un periodo de tiempo previsto de antemano, volver a recopilar datos de control y a analizarlos, comparándolos con los objetivos y especificaciones iniciales, para evaluar si se ha producido la mejora esperada.
    - Documentar las conclusiones.
- **Actuar**
    - Si es necesario, modificar los procesos según las conclusiones del paso anterior para alcanzar los objetivos con las especificaciones iniciales
    - Si se han detectado en el paso anterior, aplicar nuevas mejoras
    - Documentar el proceso.

Adaptando el "ciclo de Deming" en el contexto del modelo de gestión de la información y enfocándolo hacia la calidad de la información se tendrá correspondencia con los siguientes componentes:

Definir (planear). Determinar cuáles son los datos críticos para la organización

Medir (hacer). Auditar la calidad de esos datos

Analizar (verificar). Determinar los impactos

Mejorar (actuar). Implementar las acciones necesarias correctivas

IV. El Modelo de Gestión de la Calidad de la Información.

Una vez expuesto el proceso de generación del conocimiento a partir de la conversión de los datos en información y ésta a su vez en el conocimiento necesario para la toma de decisiones y la definición de las acciones a ejecutar, se podrá comprender la gran importancia que tiene el cuidar la calidad de los datos.

La idea de cuidar la calidad de los datos, parte de la necesidad de darle a estos una correcta interpretación, de ahí también nace el concepto de calidad de la información y más específicamente el de gestión de la calidad de la información.

Previo a la introducción del modelo, analizaremos conceptualmente los diferentes términos y definamos cuáles son las dimensiones de la calidad de la información.

IV.1 Calidad y gestión de la calidad

El concepto de Gestión de la Calidad de la Información se define a partir del concepto de Sistema de Gestión de la Calidad de la norma ISO 9000:2000. Esta norma establece que un Sistema de Gestión es "un conjunto de elementos relacionados o que interactúan para establecer la política y los objetivos y lograr dichos objetivos".

Por consiguiente definimos como Gestión de la Calidad de la Información al conjunto de elementos mutuamente relacionados o que interactúan para establecer los objetivos y políticas de la calidad de la información para una organización y dirigir y controlar ésta, garantizando la consecución de los objetivos formulados[10].

Implementando un sistema de gestión de la calidad adecuado, las instituciones, como es el caso del Servicio de Administración Tributaria, se logra por una parte, confianza en la capacidad de sus procesos y por otra, se establecen las bases para la mejora continua. Ambas conducen a la satisfacción del cliente y al éxito en su operación.

10 "Gestión del conocimiento y calidad total," Carlos A. Benavides Velasco, Cristina Quintana García

El Concepto de Calidad. Dentro de múltiples definiciones que de éste existen podemos adoptar por su sencillez y claridad la contenida en la norma ISO/CD2 9000:2000:

- de forma práctica: la satisfacción de necesidades y expectativas;

- de forma técnica: conjunto completo de características de la calidad especificadas e implícitas y sus correspondientes valores.

IV.2 Dimensiones de la calidad.

Definiremos principalmente 6 características de la calidad de la información, formalmente conocidas como dimensiones de la calidad de la información:

Exactitud (Accuracy). Mide el grado en que los datos son correctos y libres de error, refleja la correspondencia de la información del negocio comparándola contra la realidad.

Totalidad (Completeness). Denota el grado al cual los datos no faltan, refleja el grado en que las bases de datos cuentan con toda la información critica para el negocio.

Relevancia (Relevancy) denota el grado en que los datos son aplicables y útiles para la tarea actual, que la información le sirva a la persona que se le está proporcionando.

Oportunidad (Timeliness) denota el grado en el que los datos son actualizados oportunamente, medición de que la información esté disponible cuando se requiere para tomar una decisión.

Accesibilidad (Accessibility) denota el grado en el que la información está disponible y fácilmente accesible.

Consistencia (Consistency). Que la información sea la misma en todas las áreas o sistemas utilizados por la compañía.

Estas dimensiones son un subconjunto de todas aquellas que se consideran las importantes para los usuarios de la información[11],

11 Wang, R.Y. and Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12 (4). 1996

algunos autores llegan a manejar hasta dieciséis[12]. Se eligieron sólo éstas por considerarse que son las más representativas y las más críticas para los usuarios.

IV.3 Ciclo de vida de la información.

Un modelo de referencia que nos servirá en el manejo de los conceptos, es el ciclo de vida de la información en una organización[13].

Adquirir: Es el proceso por el cual se incorpora información a los sistemas de registro.

Crear: Es la fase en la que se produce información a partir de los datos existentes en los sistemas de registro, quien produce información es responsable de su calidad; sin embargo, la calidad puede ser afectada por la calidad de los datos existentes.

Catalogar: Posterior al proceso de adquisición, es necesario catalogar los datos en objetos de información, en esta fase existen dos elementos importantes: El primero es el relacionado con su representación y el nivel de detalle, el segundo corresponde a la necesidad de una representación de la información persistente que conlleva a elección de un identificador único para cada nuevo objeto de información.

Almacenar: Es el hecho de insertar nuevos objetos de información en archivos digitales. Un elemento importante es el proceso de selección de los medios de almacenamiento.

Preservar: Se refiere a las actividades de la gerencia que preservan el contenido, así como la apariencia y forma de los objetos de la información. En detalle, el formato en el cual los datos tienen que ser preservados se elige en esta etapa, que puede dar lugar a una necesidad previa de transformar los datos para guarda uniformidad.

Acceder: Asegurar el acceso continuo a los objetos de información almacenados en los archivos digitales.

12 Kahn, B.K., Strong, D.M. and Wang, R.Y. Information Quality Benchmarks: Product and Service Performance Communications of the ACM, 2002,

13 “Best practises for digital archiving”, Hodge G.M., D-Lib, January 2000.

IV.4 Implementación del modelo de gestión de la calidad.

La implementación del Modelo de Gestión de la Calidad de Información que se propone, con base en el "Ciclo de Deming", comprende las siguientes etapas:

1. Definir los objetivos y políticas de la calidad de la información.

2. Determinar cuáles son los datos críticos e identificar los procesos asociados.

3. Medir la calidad de los datos.

4. Identificar las acciones a realizar para prevenir la aparición de deficiencias.

5. Planificar las estrategias, procesos y recursos para llevar a cabo las mejoras identificadas.

6. Poner en práctica el plan.

7. Revisar las actividades de mejora para determinar la adecuación de las actividades de seguimiento.

Implementando un sistema de gestión de la calidad adecuado, las instituciones, como es el caso de la administración tributaria, generarán por una parte, confianza en la capacidad y fiabilidad de sus procesos y por otra las bases para la mejora continua. Ambas conducen a la satisfacción del cliente y al éxito en la gestión de las instituciones.

Así, tendremos dos grandes áreas de oportunidad dentro de las instituciones:

1) La gestión: que corresponde a la definición de objetivos y políticas, y el establecimiento de acciones permanentes, procesos dentro de la institución para cuidar la calidad de los datos y en consecuencia de la información que a partir de estos se genera.

2) El aseguramiento de la calidad de los datos. Corresponde a todas aquellas acciones que se llevan a cabo para garantizar que los datos y la información que se maneja en la institución

garanticen su calidad en cada una de las dimensiones que se definieron anteriormente.

Enseguida realizaremos un caso práctico el modelo de gestión definido, basados conceptualmente en su implementación para el SAT, pero, por fines de confidencialidad de la información, ejemplificaremos usando la base de datos del SIEM (Sistema de Información Empresarial Mexicano) de San Luis Potosí, obtenida del portal de Internet.

IV.4.1 Definir los objetivos y políticas de la calidad de la información.

El objetivo principal de la calidad de la información será: Producir los datos exactos, completos, consistentes, oportunos y con el nivel de detalle necesario, que sean accesibles para los usuarios y relevantes para la realización de sus tareas.

Las políticas de calidad deberán establecerse por cada institución y servirán de parámetro para determinar hasta qué punto los procesos orientados a que la información tenga la calidad requerida se lleven a cabo y determinarán qué acciones de mejora continua se deberán ejecutar.

Siguiendo el ejemplo del SAT, su política de calidad debe estar orientada a satisfacer cada una de las dimensiones de la calidad de la siguiente forma:

Exactitud. Facilitar los mecanismos de actualización de la información del RFC, para que esta sea actualizada en cada contacto con el contribuyente y diseñar campañas de actualización de la información; así como de difusión de las obligaciones de manifestar cualquier cambio en las circunstancias que originaron el registro del contribuyente en lo plazos que marca la ley, asegurando que los datos de los contribuyentes almacenados en el RFC sean correctos y que correspondan a la realidad.

Relevancia. Garantizar que la información que se le solicita al contribuyente sea sólo la relevante para la institución y que a partir de ella se puedan llevar a cabo los procesos internos de control del cumplimiento de obligaciones y fiscalización.

Oportunidad. Mejorar los procesos de adquisición y procesamiento de datos externos para ponerlos a disposición de los usuarios, en el menor tiempo posible.

Totalidad. Asegurar que se tiene al total de contribuyentes registrados en el RFC y que se cuenta con toda la información relevante de los mismos.

Accesibilidad y Consistencia. Rediseñar o mejorar los procesos internos, los sistemas de información Rediseñar y crear modelos de información basados en fuentes externas para garantizar que el RFC sea el eje de la operación de la Administración Tributaria.

IV.4.2 Identificar los procesos y determinar los datos críticos asociados.

Para determinar cuales son los datos relevantes de RFC que deben someterse a un modelo de aseguramiento de la calidad. Observemos cuales son los procesos en los que el RFC interviene como eje de la operación del SAT.

Partamos nuevamente del Modelo de Referencia de Ciclo Tributario (Fig. 12)

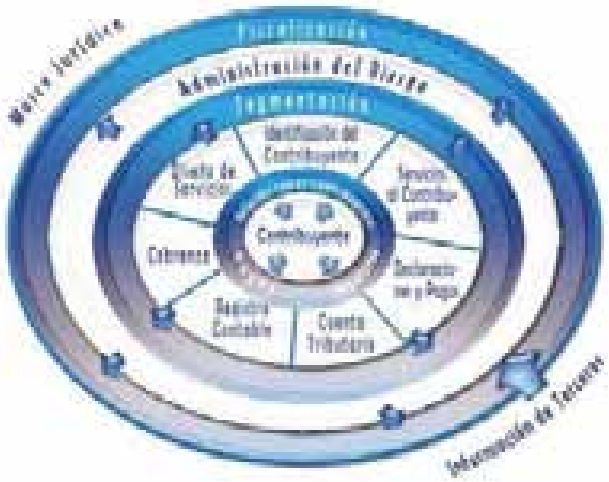


Fig. 12 Modelo de Referencia del ciclo tributario

Los procesos que podemos distinguir son los siguientes:

- Identificación del Contribuyente
- Servicios
- Declaraciones y Pagos
- Cuenta Tributaria
- Registro Contable
- Cobranza
- Fiscalización

Analizando la estructura orgánica (Fig. 13):

Fig. 13: Imagen de la estructura orgánica obtenida del portal del SAT

Se incorpora el proceso de Comercio Exterior, que no fue identificado en el modelo del ciclo Tributario.

Si la institución cuenta con un mapa de procesos o bien, con un mapa estratégico de los objetivos de la institución, es conveniente partir de esta información para determinar cuáles son los procesos sustantivos en los que está involucrado el sistema de registro que se va a analizar. En el caso del SAT seguramente existe esta información; sin embargo, no fue posible encontrarla publicada.

Se realiza un mapeo de los procesos y la información que estos utilizan. Primeramente se agrupan conjuntos de datos y se nombran bajo el objeto de información que estos representan, posteriormente se

determina si el grupo de información tiene o no relevancia para el proceso, como se muestra en la siguiente tabla (Tabla 4):

Mapeo de Proceos con Objetos de Información								
Procesos / Datos	Datos de Identificación	Datos de Ubicación	Actividad Económica	Obligaciones Fiscales	Cumplimiento de Obligaciones	Información de Pagos	Información de devoluciones	Créditos
1. Identificación del Contribuyente	x	x	x	x				
2. Servicios	x	x	x	x	x	x	x	x
3. Declaraciones y Pagos	x	x		x	x	x	x	x
4. Cuenta Tributaria	x	x		x	x	x	x	x
5. Registro Contable	x	x			x	x	x	x
6. Cobranza	x	x	x	x	x	x		x
7. Fiscalización	x	x	x	x	x	x	x	x
8. Comercio Exterior	x	x	x	x	x			
Cueta	8	8	5	7	7	6	5	6

Tabla 4: Mapeo de Procesos con Objetos de Información

Se identifican cuales son las variables que más repercuten en los procesos, en este caso son aquellas asociadas a:

- Datos de Identificación
- Datos de Ubicación
- Obligaciones fiscales
- Cumplimiento de Obligaciones

En esta etapa pueden incorporarse variables que por su importancia sea necesario considerar, se requiere determinar si éstas forman parte real del ámbito de estudio y, en su caso, si existen problemas de calidad en las mismas. Por ejemplo, las variables información de pagos y créditos son muy importantes para el SAT.

Posteriormente se realiza un análisis de dependencias entre las variables a fin de determinar si las elegidas no son generadas por otras; es decir, su producción o existencia depende de los valores de éstas. Un ejemplo se muestra en la siguiente tabla (Tabla 5):

Análisis de dependencias entre Objetos de Información									
	Datos de Identificación	Datos de Ubicación	Actividad Económica	Obligaciones Fiscales	Cumplimiento de Obligaciones	Información de Pagos	Información de devoluciones	Créditos	Número de Dependencias
Datos de Identificación	x								1
Datos de Ubicación		x							1
Actividad Económica			x						1
Obligaciones Fiscales	x		x	x					3
Cumplimiento de Obligaciones			x	x	x	x			4
Información de Pagos						x			1
Información de devoluciones					x	x	x		3
Créditos					x	x		x	1

Tabla 5: Análisis de Dependencias entre Objetos de Información

En este caso puede observarse que las variables que se eligieron de obligaciones fiscales y cumplimiento de obligaciones, dependen de otras. En el caso de las obligaciones fiscales que un contribuyente tiene, éstas son determinadas en función de la actividad económica que realiza y por su categoría de Persona Física o Moral. En este caso la variable de obligaciones fiscales se sustituye por la de actividad económica.

La variable cumplimiento de obligaciones es la que más dependencias tiene, es de gran utilidad para los procesos del SAT, y no es una variable que se obtenga del contribuyente, ya que esta es generada por diversos algoritmos durantes los procesos de control de obligaciones del SAT, por lo que su calidad dependerá de los algoritmos diseñados para generarla y de los datos con la que se procesa.

Para los fines de este trabajo, restringiremos el análisis a los siguientes objetos de información críticos:

- *Datos de Identificación*
- *Datos de Ubicación*
- *Actividad Económica*

El siguiente paso a realizar es el de descomponer cada uno de los objetos de información en los datos que lo constituyen:

Datos de Identificación:

Tipo de Persona (física, moral)
Nombre
Razón Social
Clave de RFC
Clave de CURP

Datos de Ubicación

Entidad
Municipio
Domicilio
CP

Datos de Actividad Económica.

Sector de la Economía
Actividad

Un análisis detallado de cómo la información es útil a lo largo de cada uno de los procesos, además de ayudarnos en la identificación de los datos críticos, servirá para medir qué tan relevante son los datos almacenados en los sistemas de registro para el desarrollo las actividades de la institución, con lo que se estaría en posibilidad dictaminar la dimensión de la calidad denominada relevancia, es decir en que la información está siendo útil a la organización.

IV.4.3. Medir la calidad de los datos.

El siguiente paso, corresponde a las acciones para medir la calidad de los datos, esta etapa es mejor conocida como el proceso de auditaría de datos.

Entre los métodos más empleados para medir la calidad de los datos y encontrar deficiencias en los mismos podemos señalar:

1. Comparar nuestros datos contra una serie de datos considerados verdaderos.

2. El uso de comandos SQL.

3. Usar una herramienta comercial de “Data Profiling”.

4. Muestrear los datos y aplicar revisiones de escritorio o de campo.

5. Contratar servicios de outsourcing para desarrollar esta actividad.

Con las 4 primeras opciones se atiende principalmente a las 2 primeras dimensiones de la calidad de la información: Exactitud y Totalidad.

¿Qué se busca en la medición de la calidad de los datos?

Exactitud:

Que no existan datos fuera de rango.
Que no existan errores ortográficos o una mala escritura.

Que no exista desorden en las claves asociadas a catálogos, o que estén fuera de dominio.
Que no existan datos duplicados.
Que no existan datos nulos, donde no deben aparecer.
Que la información esté catalogada, donde sea susceptible.
Que los datos estén atomizados, es decir que no sean datos compuestos.

Totalidad:

Que no existan datos faltantes.

Tratándose de sistema de registro, es importante contar con el total de los objetos –en el caso del SAT el total de contribuyentes que componen el universo susceptible de registrar.

IV.4.3.1 Comparación contra datos verdaderos.

Sin lugar a duda el mejor método para comparar la información del registro contra datos reales es preguntando directamente al dueño de los datos. Sin embargo, resulta difícil tener a nuestra disposición la totalidad de los miembros de un registro para cuestionarlos sobre la veracidad de sus datos. No obstante, es un mecanismo al que podría recurrirse cuando la calidad de nuestros sistemas de registro es realmente desfavorable, en cuyo caso el mejor método realizar un levantamiento de toda la información –borrón y cuenta nueva- para lo cual se sugiere la adquisición de la información mediante técnicas en campo de tipo censal.

Comencemos por realizar un ejercicio de comparación contra datos verdaderos o que se conocen que son de mejor calidad. El análisis comparativo que aquí se presenta se centra en un universo de datos ya conocidos. Para esto se debe tener a disposición una fuente de información externa que se considere más alineada a la realidad, o bien medir la información mediante el desarrollo de encuestas.

Para ejemplificar este método, utilizaremos la información de actividad económica, definido como dato crítico para el SAT.

El SAT durante muchos años utilizó un catálogo de actividades económicas basado en la CMAP (Clasificación Mexicana de Actividades y Productos) del INEGI. Si comparamos la información de este catálogo contra el SCIAN (Sistema de Clasificación Industrial de América del

Norte), que es el catálogo de clasificación de actividades económicas que utiliza el INEGI para el levantamiento de sus censos y encuestas, y con fines de comparación internacional.

La clasificación CMAP está constituida por 9 sectores con sus respectivos subsectores, ramas, subramas y clases, como se muestra en el siguiente cuadro (Tabla 6):

SECTOR	CONCEPTO
1	Agricultura, ganadería, caza, silvicultura y pesca.
2	Minería y extracción de petróleo.
3	Industrias manufactureras.
4	Electricidad y agua.
5	Construcción.
6	Comercio.
7	Transporte y comunicaciones.
8	Servicios financieros, de administración y alquiler de bienes muebles e inmuebles.
9	Servicios comunales y sociales hoteles y restaurantes; profesionales, técnicos y personales. Incluye los servicios relacionados con: agricultura, ganadería, construcción, transporte, financiamiento y comercio.

Tabla 6: Sectores en los que se divide la clasificación CMAP-INEGI

El diseño del SCIAN prestó especial atención en clasificaciones orientadas hacia la producción, consideradas como características particulares:

- Actividades económicas nuevas y emergentes.
- Actividades de servicios en general.
- Actividades enfocadas a la producción de tecnologías avanzadas.

El SCIAN es considerado como clasificador único de actividades económicas, elaborado en base al principio fundamental de agregación. Divide la economía en 20 sectores, las actividades que hay dentro de cada sector están agrupadas de acuerdo con el criterio de producción, aunque la distinción de bienes y servicios no se refleja en forma explícita en la estructura; 5 sectores son productores de bienes y 15 productores de servicios.

Una característica importante del SCIAN es la creación de los siguientes sectores:

- Sector de información en medios masivos.
- Servicios profesionales, científicos y técnicos.
- Servicios de esparcimiento cultural y deportivo, y otros servicios recreativos.
- Servicios de salud y de asistencia social.
- Industrias manufactureras.

Un primer ejercicio comparativo se realizó entre el entonces catálogo actividades del RFC publicado en el DOF (Diario oficial de la Federación) y la propia CMAP encontrando lo siguiente:

CRITERIO DE COMPARACIÓN	COMPARACIÓN DE CLAVES DE LOS 9 SECTORES DE ACTIVIDAD ECONÓMICA			
	RFC	CMAP	RFC	CMAP
CLAVES Y CONCEPTOS IGUALES	279	279	37.4%	37.0%
CLAVES IGUALES CON CONCEPTOS DIFERENTES	104	104	13.9%	13.8%
CLAVES QUE EXISTEN EN UN CATÁLOGO Y NO EN EL OTRO	195	203	26.1%	26.9%
CLAVES CON CONCEPTOS QUE VARIAN PERO SE ASUME QUE ES LA MISMA ACTIVIDAD	168	168	22.5%	22.3%
TOTAL DE CLAVES	**746**	**754**	**100%**	**100%**

Tabla 7: Concentrado diferencias entre claves del RFC y CMAP

“Claves y conceptos iguales” en ambos catálogos tenemos 279 claves y conceptos iguales representando el 37.4% del total de RFC y el 37% del total de la CMAP.

“Claves iguales con conceptos diferentes” en este recuadro la cifra será la misma 104 para ambas clasificaciones, pero si existe una mínima variación porcentual de 13.9% en RFC y 13.8% en la CMAP.

“Claves que existen en un catálogo y no en el otro” en este punto tenemos 195 en RFC y 203 en la CMAP, con una representación

porcentual de 26.1% y 26.9% respectivamente; es importante mencionar que la mayor parte de estas claves son diferentes, pero representan conceptos o actividades existentes en ambos catálogos.

"Claves con conceptos que varían pero se asume que es la misma actividad", este rublo se refiere a las actividades con la misma clave pero en su concepto o descripción omite o agrega una a más palabras y se observa un total de 168 en ambos catálogos representando un 22.5% y el 22.3% respectivamente.

Ejemplo:

> 356006 Fabricación de piezas industriales modeladas con resina (RFC).
>
> 356006 Fabricación de piezas industriales modeladas con diversas resinas y los empaques de polietileno expandibles

También se observa en el catálogo de RFC un total de 746 claves y en la CMAP 754 que comparativamente existe una diferencia de 8 unidades.

Para concluir el análisis de este comparativo, es necesario comentar que en ambos catálogos se manejan 9 sectores con conceptos similares.

El siguiente paso, fue comparar estos catálogos contra el SCIAN, cabe señalar que el INEGI establece las correspondencias entre el catálogo CMAP y el SCIAN, por lo que la comparación entre RFC y CMAP se realizó para poder tener un puente entre el RFC y el SCIAN.

Grandes números:

NIVEL	CATEGORÍAS CMAP	CATEGORÍAS SCIAN
Sector	9	20
Subsector	85	92
Rama	130	294
Subrama	201	622
Clase	754	1,021

Conceptos exclusivos del SAT:

CLAVE	CONCEPTO
100000	RETENEDOR PURO
200000	COPROPIETARIO
300000	ASALARIADO

El manejo de estos conceptos predispone, sin duda, a no poder conocer cuál es el sector de actividad en el que se desarrollan los retenedores, copropietarios y asalariados.

La comparación de CMAP y SIAN arrojó los siguientes resultados (Tabla 8):

CONCEPTO	SECTORES DEL 1 AL 9			
	CMAP/SCIAN	SCIAN/CMAP	CMAP/SCIAN	SCIAN/CMAP
TOTAL DE CLAVES	754	1021	100%	100%
CONCEPTOS QUE EXISTEN EN UNO Y NO EN EL OTRO	1	6	0.10%	0.60%
CLAVES DIFERENTES Y CONCEPTOS IGUALES	501	637	66.50%	62.40%
CLAVES DIFERENTES CON CONCEPTOS QUE VARIAN PERO SE ASUME QUE ES LA MISMA ACTIVIDAD	252	378	33.40%	37.00%
TOTAL	754	1021	100%	100%

Tabla 8: Resultados de la comparación de claves y conceptos entre CMAP y SCIAN

Dentro de algunas de las situaciones que se observaron entre los diferentes catálogos fue la mayor apertura de claves, por ejemplo (Tabla 9):

CMAP		SCIAN	
CLAVE	ACTIVIDAD	CLAVE	ACTIVIDAD
111105	Cultivo de árboles frutales	111310	Cultivo de naranja
		111321	Cultivo de limón
		111329	Cultivo de otros cítricos
		111332	Cultivo de plátano
		111333	Cultivo de mango
		111334	Cultivo de aguacate
		111336	Cultivo de manzana
		111339	Cultivo de otros frutales no cítricos y de nueces
		111991	Cultivo de coco

Tabla 9: Ejemplo de apertura de claves del CMAP al SCIAN

Es fácil imaginar las dificultades a enfrentar y los esfuerzos que el SAT tendría que realizar si el día de mañana existiera una legislación fiscal especial para los exportadores de aguacate.

El siguiente paso a realizar, y que por motivos de confidencialidad de la información no se desarrolló, sería realizar una serie de consultas sobre la base de datos para conocer el número de registros que son impactados por deficiencias en el catálogo de actividades económicas.

IV.4.3.2 Uso de comandos SQL y exploración de datos.

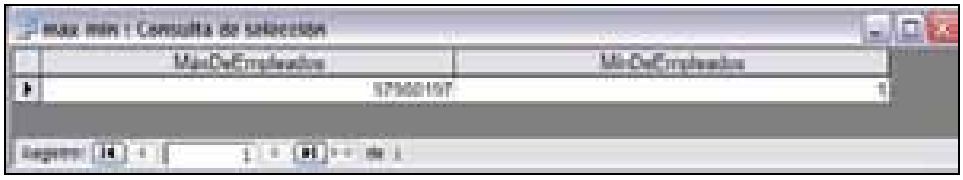
Uno de los métodos menos costosos y muy efectivo para tener un buen acercamiento a la calidad de nuestra información es el uso de los comandos SQL que acompañan al servidor de bases de datos en el que está almacenado el sistema de registro, así como una simple exploración visual puede ayudarnos a encontrar inconsistencias en la información.

Con la finalidad de demostrar el uso de estos comandos, se trabajó con un segmento de base de datos del SIEM que se descargó en Internet, en seguida se muestran una serie de ejemplos del uso de comandos SQL para obtener una medición de la calidad de la información.

a) Que no existan datos fuera de rango

Si se conocen los rangos en que los datos de una variable deben estar contenidos, las funciones de máximo y mínimo son muy útiles:

> SELECT Max(Empleados) AS MáxDeEmpleados, Min(Empleados) AS MínDeEmpleados FROM slp;
>
> Resultados:

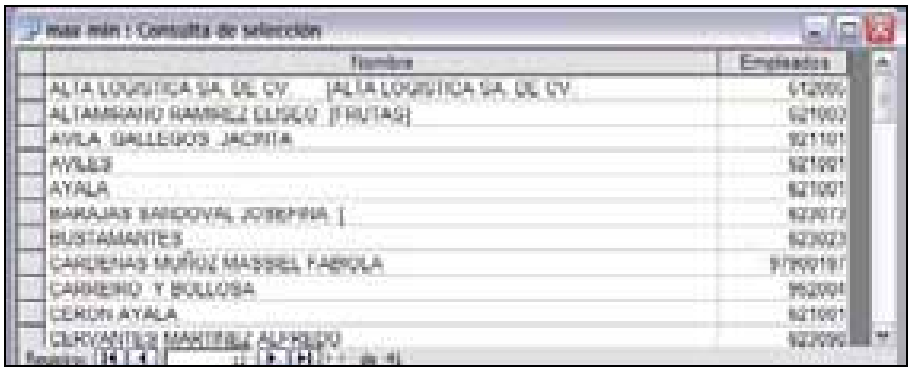


En este caso si se conoce que el número de empleados en ningún caso excede a un millón, se está ante un claro “Fuera de Rango”.

Ahora bien si el caso fuera que ninguno debe tener más de 500,000 empleados, el siguiente query es de utilidad:

SELECT Nombre, Empleados FROM slp GROUP BY Nombre, Empleados WHERE Empleados > 500000;

Resultados:



Es muy común que del resultado de la observación de las bases de datos, se vayan descubriendo otros errores, como el de escritura que se muestra en el nombre.

b) Que no existan errores ortográficos o de mala escritura

Con el siguiente query podríamos encontrar registros con caracteres no deseados:

SELECT Nombre FROM slp WHERE Nombre Like '%&%';

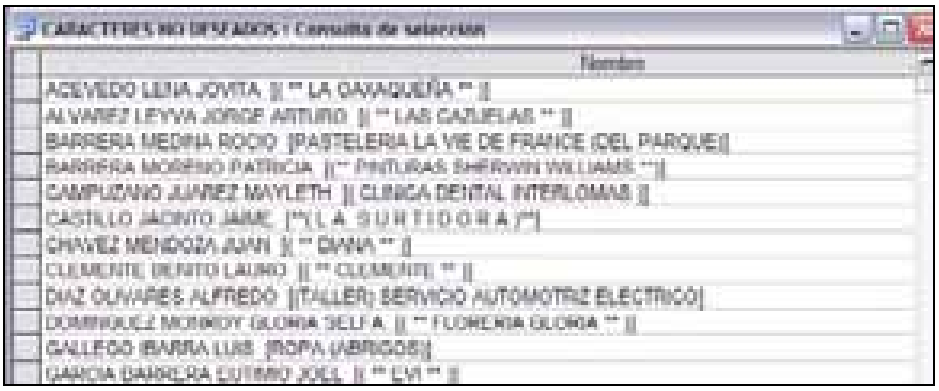
Resultados:

Observamos la presencia del caracter & en vez de la letra Ñ, situación que es muy común en los sistemas de registro antiguos o que tiene problemas en la definición de la tabla de caracteres a utilizar en el manejador de bases de datos.

Otro ejemplo,

SELECT Nombre FROM slp WHERE Nombre Like ' %(%';

Resultados:

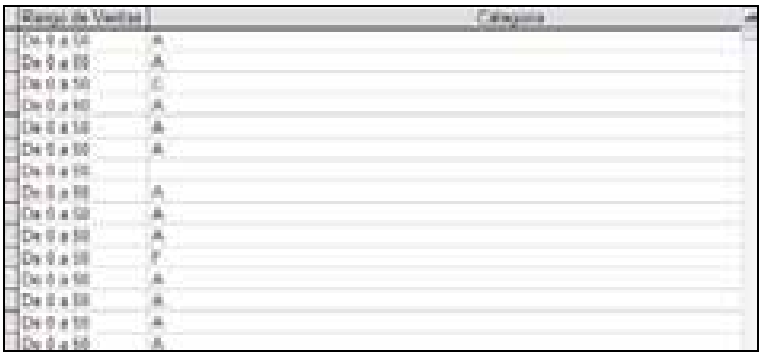


Aquí encontramos otro problema muy común y es que ante una mala definición de la información relevante para el negocio, no se incorporan en las estructuras de datos los campos necesarios. En este ejemplo observamos que se incorpora el nombre comercial entre paréntesis, lo que contamina el dato.

c) Que no exista desorden en las claves asociadas a catálogos, o que estén fuera de dominio.

SELECT Rango_de_Ventas, Categoria FROM slp WHERE Rango_de_Ventas="De 0 a 50"

Resultados:



Es evidentemente que la categoría “A” pertenece a aquellos negocios con un rango de ventas entre 0 y 50 mil pesos, pero como se observa existen errores de asignación y algunos datos nulos

d) Que no existan datos duplicados

Resolver esto mediante el uso de comados SQL es muy complicado, ya que en el mayor números de los casos los duplicados existentes se deben a que los datos fueron capturados con algún error. Para resolver esto, discutiremos más adelante el uso de herramientas para encontrar duplicados. Quizá una simple exploración visual sobre los datos ordenados por el campo de interés, nos permita encontrar uno que otro duplicado (Fig. 14).

Figura 14: Imagen de la exploración de la base de datos de San Luis Potosí ordenada por nombre

e) Que no existan datos nulos, donde no deben aparecer

SELECT Nombre, Rango_de_Ventas FROM slp WHERE Rango_de_Ventas Is Null

Resultados:

f) Que la información esté catalogada, donde sea susceptible

Una fuente frecuente de error en los datos, son los cometidos durante la adquisición de los mismos, específicamente durante la captura, si la información se cataloga y se asignan claves a conceptos, estos errores se disminuyen y se facilita la estandarización de la información.

Quizá de todos los ejemplos anteriores, éste es el que más aporta al trabajo de descubrir problemas de calidad de los datos, veamos el siguiente segmento de datos que se encontró al explorar la información (Fig. 15).

Fig. 15: Imagen de la exploración de la base de datos de San Luis Potosí que muestra los problemas en catálogos

Si observamos con detalle (Fig. 15), encontraremos que existen diferentes formas de escribir la información del nombre del estado, lo

cual se evitaría con un adecuado uso de catálogos, lo mismo sucede en el domicilio.
g) Que los datos estén atomizados, es decir que no sean datos compuestos.

El ejemplo anterior nos permite ver otro problema: El domicilio por si mismo representa una serie de problemas, el primero es que no está atomizado, es decir no está descompuesto en sus partes esenciales, ya que el tipo de vialidad, el nombre de la misma y en número interior o exterior están contenidos en un mismo campo, lo cual debería corregirse y además crear el catálogo para el tipo de vialidad, con lo que se evitarían los errores cometidos en la forma de escribir avenida o boulevard.

Al igual que el domicilio, la colonia no está atomizada, ya que en ella se incorpora la información del tipo de asentamiento (colonia, rancho, pueblo, etc.).

IV.4.3.3 Usar una herramienta comercial de "Data Profiling".

El ejercicio anterior es muy útil en los primeros acercamientos hacia la calidad de los datos, requiere de un amplio conocimiento sobre la información y habilidades para encontrar problemas, pero sobre todo que la persona que lo realice esté ajena a los procesos del ciclo de la información –sobre todo los relativos a la adquirir, crear, catalogar y almacenar.

Si bien, se puede crear todo un sistema para descubrir errores en la información utilizando las herramientas con las que cuenta el negocio, también existen comercialmente las herramientas que pueden ayudarnos a descubrir los problemas de calidad de nuestros datos.

Estas herramientas comercialmente conocidas como Data Profiling son las que apoyarán a la institución en los procesos de detección de errores en las bases de datos. Se basan en los ejemplos vistos en el apartado anterior del uso de queries y esencialmente lo que realizan es presentar una serie de estadísticas del estado que guarda la información.

En general los distribuidores de este tipo de herramientas, también venden las herramientas para corregir los errores, conocidas como software de "Data Quality", no se encontraron referencias sobre cual de éstas fue primero, probablemente nacieron en paralelo como un

conjunto de desarrollos que se fueron diseñando tanto para descubrir los problemas de datos, como para mejorar la calidad.

Data Profiling, o en español “perfilamiento de datos”, es el proceso de examinar y proporcionar información estadística acerca de la calidad de éstos, lo que nos permite tener un acercamiento inicial a la calidad de los datos y realizar un monitoreo frecuente de la misma.

La empresa TRILLIUM, cuyas herramientas se encuentran al alcance del SAT, cuenta con una herramienta llamada “TS Insight”, que se apoya en los resultados de TS Discovery para presentar una serie de métricas sobre la calidad de los datos.

TS Discovery automatiza el análisis de los datos implicados en su sistema para identificar cualquier error, permitiéndonos:

- Analizar e identificar los valores de los datos, su frecuencia, formatos y máscaras.

- Identificar estadísticas importantes de los campos como valores mínimos, máximos, nulos, etc.

- Encontrar redundancias y datos duplicados.

- Descubrir ligas ocultas, claves y dependencias de datos.

Mediante esta herramienta se pueden ver de forma gráfica los resultados sobre la calidad de los datos, cuyas reglas del negocio hayan sido definidas en TS Discovery. Podemos definir las reglas de negocio para la comprobación de irregularidades o inexactitudes en cualquier tipo de datos. Por ejemplo:

Claves de identificación o números en formato inválido.

Variables que están fuera de rango.

Valores de datos ilógicos, definidos por la combinación de múltiples campos.

Ejemplo de las Reglas de Negocio.

En este apartado se describe un caso práctico común en el SAT y que es solucionado definiendo las Reglas de Negocio.

- Escenario - Se cuentan con 20,000 registros de personas físicas, los cuales tienen un campo de teléfono y se necesita evaluar la calidad éstos tomando en consideración que los datos no contengan algún caracter alfabético.

- Regla de Negocio – Ningún datos deberá contener caracteres alfabéticos

- Proceso –Se codifica la regla del negocio al lenguaje TS Discovery (Fig. 16), se calendariza y se ejecuta. La herramienta procesa la regla y muestra los resultados en un lenguaje técnico. Posteriormente, se exportan los datos a TS Insight por medio de un script que se ejecuta desde el sistema operativo y finalmente se muestran los datos en un lenguaje de graficas e índices dentro de TS Insight, el cual nos facilita el análisis y evaluación de los resultados.

La siguiente es la sentencia que se codifica en TS Discovery para representar la regla de negocio:

PATTERN(SUBSTR(Attr2,1,10),"default") LIKE "*a*"

Donde:

- PATTERN – Regresa el patrón que cumpla con la sub-cadena especificada.
- SUBSTR – Regresa el la sub-cadena del atributo “Attr2” (Teléfono).
- 1, 10 – Especifica que inicia la sub-cadena en 1 y termina en la posición 10.
- Default – Específica el estilo que usará el patrón para la comparación.
- LIKE – Regresa los registros que son como el parámetro de la derecha.
- “*a*” – Es la representación del caracter alfabético.

La siguiente figura muestra como se incorpora la regla a la herramienta:

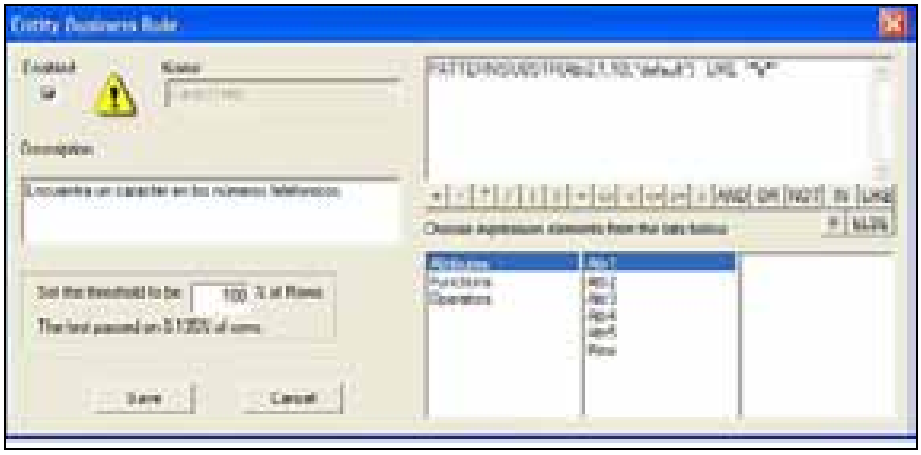


Fig. 16: Imagen del ingreso de la regla del negocio a TS-Dicovery

- Los resultados de este análisis muestran el 0.135% de los registros cumplieron con la regla del negocio (Fig. 17).

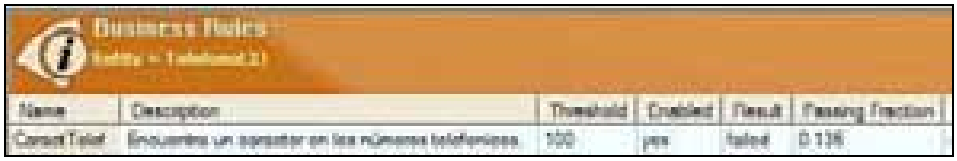


Fig. 17: resultados de ejecutar la regla del negocio

- Interpretación – El 0.135% del total de los registros tienen por lo menos 1 carácter del alfabeto.

De forma gráfica, también se presentan los patrones de distribución (Fig. 18):

Fig. 18: Gráfica de los resultados de ejecutar la regla del negocio, vistos desde TS-INSIGHT

- Interpretación:

  d8: caracteres numéricos (55281560)

  d3pd4: caracteres numéricos con otro dato (552-8156)

  d3pd3pd4: caracteres numéricos con otro dato (552-886-8025)

La arquitectura que se propone para utilizar estas herramientas es la siguiente (Fig. 19):

Fig. 19: Arquitectura propuesta para medir la calidad de los datos

Los datos provienen del canal de entrada de la información al sistema de registro, en este caso un CRM (Custommer Relationship Management), mediante un proceso se cargan los datos y se llevan al servidor en el que se revisarán las reglas del negocio con TS Discovery, la publicación de resultados quedará a cargo de la herramienta TS Insight. La utilidad de este último producto consiste en la facilidad de interpretación de los resultados obtenidos y que estos pueden ser monitoreados también por los directivos del SAT.

Reglas del Negocio.

El aspecto más importante para la utilización de este tipo de productos, sin duda lo constituye las reglas del negocio sobre la calidad de los datos. Las reglas del negocio serán un conjunto de convenciones que la

empresa definirá sobre la estructura, apariencia, relaciones, rangos, etc. que deben seguir sus datos.

Como se observó en el ejemplo anterior, las reglas del negocio se expresan de una manera coloquial, las cuales finalmente deberán programarse para que estas trabajen sobre los datos.

Ya sea empleando una herramienta de Data Profiling o bien, desarrollando alguna aplicación, lo importante es que las reglas se determinen.

Las reglas deberán orientarse a:

Tipo: Los datos no deberán contener caracteres que no corresponden a la naturaleza del dato (Ej: El número telefónico no debe contener caracteres alfabéticos).

Tamaño: Los datos almacenados deben tener siempre la misma longitud (Ej: Los códigos postales con menos de 5 caracteres).

Intervalo: Los datos deben estar entre cierto rango de valores (Por ejemplo la edad: 0<= edad <= 120).

Código: La información se capturará solamente con valores contenidos en un catálogo de datos (Ej: clave de la entidad).

Existencia: El dato deberá tener un valor, es decir que no existan valores nulos o en su caso espacios.

Fórmula: Que los datos calculados, puedan reproducirse.

Consistencia entre datos: Que los datos sean consistentes con la información contenida en otros campos (Ej: edad y fecha de nacimiento)

Conformación: Que responda a las reglas de conformación del dato (Ej: clave de RFC)

Patrones: Que el dato responda al patrón definido (Ejemplo clave del RFC, formada por 4 caracteres alfabéticos, 6 numéricos y 3 alfanuméricos)

Atomicidad: Que los datos estén lo más desagregados posibles (Ej: en vez del campo domicilio, tener calle, numero exterior, número interior, etc.)

IV.4.3.4 Muestrear los datos y aplicar revisiones de escritorio o de campo.

El análisis estadístico de la información mediante un muestreo de datos es muy útil para tratar de establecer una medición confiable de la calidad de los datos, sobre todo en las dimensiones exactitud y totalidad.

En el capítulo 2 en el que se planteo la problemática que enfrenta el SAT planteamiento del problema, mencionamos como esta Institución midió la calidad y la cobertura del RFC aplicando este tipo de técnicas.

El diseño muestral para levantar en campo encuestas sobre la calidad de la información es muy apropiado para conocer que tan exactos son los datos en un Registro, y permite contrastar los datos contra la información del mundo real.

Las técnicas de muestreo se utilizan cuando se quieren conocer las características de una población a partir de una muestra –subconjunto- de esa población – en nuestro caso el universo de datos. Las encuestas por muestreo son investigaciones que tienen como propósito conocer algo respecto a una determinada población humana, estudiando sólo una parte de esta[14].

En nuestro caso una revisión de escritorio será cuando se realice una selección muestral, del universo de datos a revisar, para conocer sus características mediante la exploración y análisis de los datos. A diferencia de un trabajo de campo, que se realizará sobre una determinada población humana, para comparar la información de la que disponemos contra la de la realidad.

Encuesta por muestreo

En una encuesta por muestreo se debe realizar un trabajo conceptual que determine entre otras cosas: ¿qué se quiere conocer? y ¿cuál es la población?

14 “Conceptos Básicos de Muestreo, Monografías del IIMAS”, Ignacio Méndez, Guillermina Eslava, Patricia Romero.

La población debe contar con un medio físico que identifique directa o indirectamente a todos los elemento de la población, a ese medio se le llama marco de muestreo. En nuestro caso:

> El marco de muestreo para conocer la calidad de los datos en la dimensión exactitud, será el total de registros -activos o no, en caso de que sea una variable a medir- de nuestro sistema de registro.
>
> El marco de muestreo para medir la calidad de la dimensión totalidad, no es fácil de integrar, ya que la fuente de información para su conformación sería el propio sistema de registro, por lo que se debe recurrir a otro tipo de marco, como lo es el cartográfico, en el que los elementos a seleccionar serán áreas geográficas que representen la distribución de los sectores económicos del país.

Existen varios métodos para tomar muestras, el más común es el probabilístico en el cual se seleccionan elementos con probabilidades conocidas y mayores a cero. De igual forma existen varios métodos de hacer la selección, entre estos están el muestreo aleatorio simple, el de probabilidad proporcional, el estratificado y combinaciones de estos.

Otro aspecto a considerar para la realización del diseño muestral, consiste en hacer que la muestra sea representativa del objeto de estudio, para lo cual se debe realizar el cálculo del tamaño de muestra.

Es importante contratar a un especialista para el diseño de la muestra, los objetivos:

1. Entender el objetivo de la encuesta y apoyar en la integración del marco muestral.

2. Diseño de la muestra y cálculo de la muestra.

3. Muestreo y selección de la muestra.

4. Cálculo de los ponderadores, ajuste y calibración de la muestra.

5. Análisis preliminar de los resultados.

Además se requiere la participación de expertos en el diseño de los instrumentos de captación y en la organización de los operativos de

campo, en muchas ocasiones puede ser el mismo muestrista, o bien, una empresa que brinde todos los servicios.

Muestreo para trabajo en escritorio

El diseño de una muestra para realizar una revisión de los datos en escritorio, es decir, sin un operativo de campo que implique captar la información, se sujeta a las mismas consideraciones que el diseño de una muestra para trabajar en campo; sin embargo, si lo que se quiere es determinar si una variable tiene o no una característica (p.e. presencia o ausencia de información), el tamaño de muestra y el diseño de la misma puede ser relativamente simple.

Este tipo de análisis es muy útil cuando la revisión a realizar para conocer la presencia o ausencia de una característica depende sólo de la observación del dato. Por ejemplo, si deseamos saber:

> ¿Cuántos contribuyentes tienen en el nombre de la calle incorporado el nombre de la vialidad?
>
> ¿Cuántos contribuyentes tienen abreviaturas en el nombre?
>
> ¿Cuántos contribuyentes en la referencia del domicilio especifican otra cosa diferente a las entrecalles?

Para este tipo de revisión reproducimos una tabla calculada por el Dr. Ignacio Méndez del IIMAS, publicado en la “Monografía, Conceptos Básicos de Muestreo”:

δ	n
.001	*1,000,000*
.01	*10,000*
.02	*2,500*
.025	*1,600*
.03	*1,111*
.035	*816*
.04	*625*

Tamaños de muestra dada una precisión

Donde δ es la precisión o error de estimación y n el tamaño de la muestra necesaria.

Esta tabla es de utilidad para muestra donde Y (U_i) es una medida o indicador de la presencia o ausencia de una característica en la unidad Ui con valor 1 si la característica está presente y 0 si no es así.

IV.4.4. Identificar las acciones a realizar para prevenir la aparición de deficiencias.

Una vez identificados los problemas de calidad en los datos se deben comunicar los resultados a la alta gerencia y se debe escribir un informe recomendando las estrategias a seguir para mejorar la calidad de la información.

Continuando con el modelo de arquitectura propuesto, en esta etapa es necesario incorporar acciones para garantizar que la nueva información que se integra al sistema de registro cumpla con las reglas del negocio definidas. Para esto es necesario que las reglas programadas en la herramienta de Data Profiling se integren a todos los sistemas en los que se opera el Registro.

En el caso del SAT, toda la relación con los contribuyentes y la base de datos se administra desde una aplicación CRM, por lo cual es necesaria que al captar la información, ésta se sujete al procesamiento de las reglas del negocio.

Así el modelo de arquitectura quedará de la siguiente forma (Fig. 20):

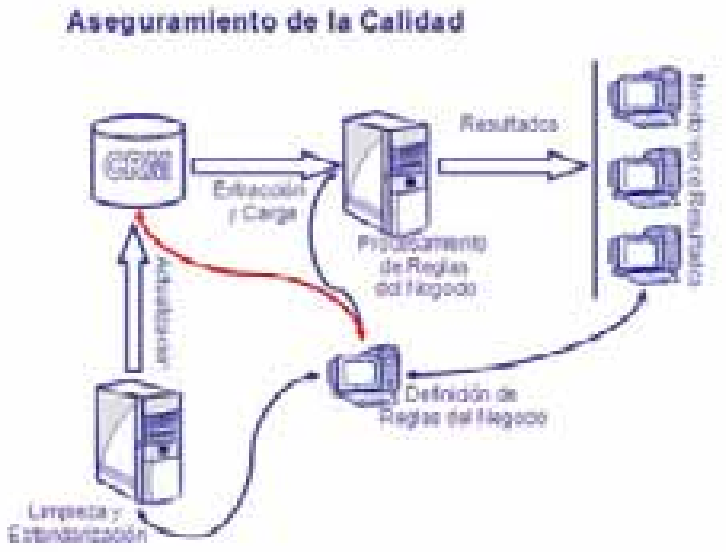


Fig. 20: Arquitectura propuesta para asegurar la calidad de los datos

Bajo este esquema y extendiéndolo a todo el modelo de operación del RFC, las principales acciones a realizar para prevenir la aparición de deficiencias se pueden identificar en el siguiente diagrama que resume la arquitectura propuesta para la Gestión de la Calidad (Fig. 21):

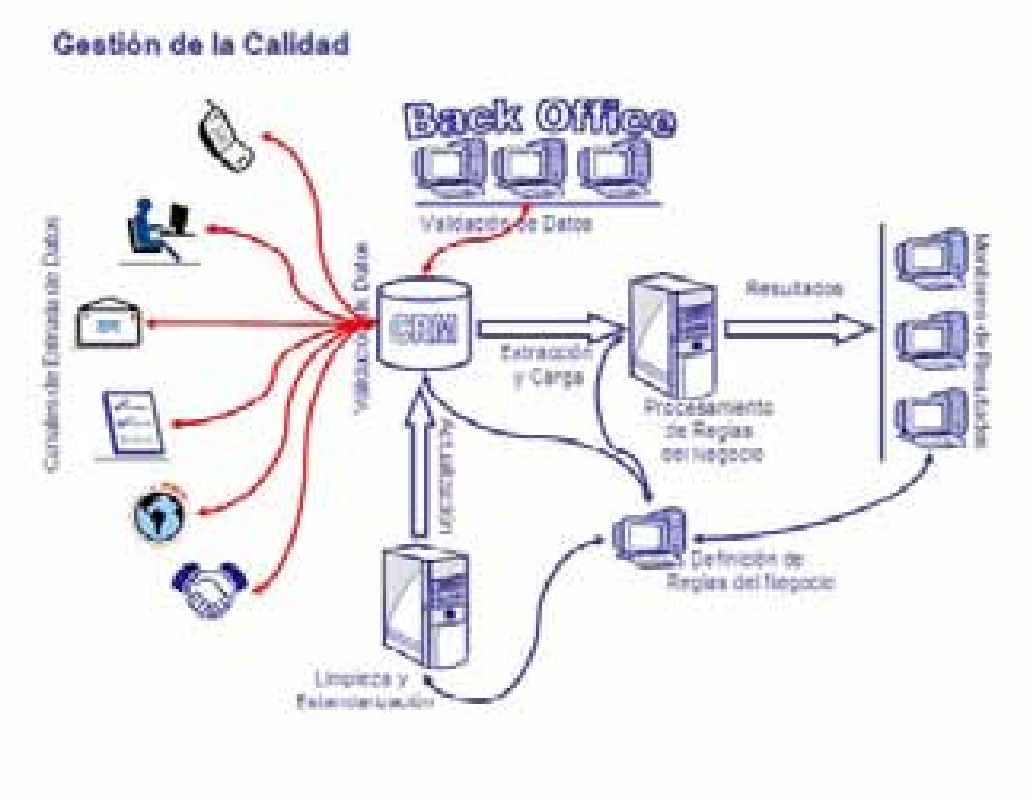


Fig. 21: Arquitectura propuesta del modelo de gestión de calidad

En resumen, deben existir acciones de validación de datos en cualquier proceso o medio por el cual se adquieran estos. Los procesos de validación deben responder a las reglas del negocio definidas.

De igual forma los datos generados o procesados, en los que se conoce como “backoffice” (operación interna de la institución), también deben responder a las reglas del negocio que se definan.

Debe existir un proceso permanente de monitoreo sobre la calidad de los datos, es decir, se debe continuamente llevar mediciones sobre la calidad y generar las acciones de mejora sobre la información, que básicamente se dividen en acciones preventivas y correctivas.

Las acciones preventivas tienen que ver con la definición de nuevas reglas del negocio y su implementación en los procesos para captar información, y las correctivas, que son aquellas encaminadas a mejorar la calidad de la información que ya se tiene en las bases de datos de los sistemas de registro.

Las acciones recomendadas para mejorar la calidad de los datos son las siguientes:

1) Preventivas:

Establecer reglas del negocio para definir los estándares de los datos.

Adecuar los sistemas y formatos mediante los cuales se captan los datos para que respondan a las reglas del negocio.

2) Correctivas:

Realizar los procesos de limpieza y estandarización de la información, acorde a las reglas del negocio definidas.

Hasta este momento, el modelo responde sólo a la calidad de los datos, para que éste responda a la calidad de la información, se deben llevar a cabo acciones para que la información que representan los datos corresponda a la realidad y que ésta sea lo más completa y actualizada posible.

Lo anterior, entre otras posibilidades, se puede lograr captando la información mediante dos métodos:

Indirectos: Procesos de actualización de la información mediante el uso de fuentes externas, es decir, captar la información proveniente de otras fuentes.

Directos: Implementar procesos permanentes de actualización de la información como pueden ser levantamientos censales, uso del correo para solicitar información, aprovechar todos los canales de entrada de datos para corroborar la información, etc.

IV.4.5. Planificar las estrategias, procesos y recursos para llevar a cabo las mejoras identificadas.

Volviendo al caso del SAT, la gran cantidad de datos internos y externos, a los que esta Institución puede tener acceso, son una fuente de información referente a los ciudadanos que debe ser aprovechada al máximo para mejorar la eficiencia de la Administración Tributaria. Así el limpiar, estandarizar, cruzar y concentrar la información en un repositorio único, permitirá reducir los costos inherentes en la operación por las deficiencias de la calidad de la información y se apoyará sustancialmente a los procesos de generación del conocimiento sobre el contribuyente, y en general de las personas que realizan algún tipo de actividad económica y que no forman parte del padrón de contribuyentes.

El componente tecnológico para poner en práctica el plan es muy importante, por lo que es necesario analizar diversos escenarios para mejorar la calidad de la información. Además, es importante que, en el caso del SAT, se consideren los aspectos de la legislación vigente en materia de acceso a la información.

Los posibles Alternativas de solución:

i. Que el SAT realice todos los procesos de higiene, estandarización, desduplicación y cruces de información, así como, el desarrollo o cambios a los sistemas para adoptar las reglas del negocio definidas.

ii. Contratar una empresa que preste todos los servicios tecnológicos, además que facilite información propia de la empresa para enriquecer los datos del RFC.

iii. Un esquema mixto en donde se dé parte del procesamiento, consultoría y desarrollo a terceros y otra parte la realice el SAT.

iv. Alternativa cero: Mantenerse con los esfuerzos aislados de cruces de información que actualmente se llevan en el SAT.

En el siguiente cuadro muestra un costeo considerando el valor de mercado de los servicios de limpieza, estandarización, desduplicación y cruces de bases de datos, con un costo de 1.70 peso por registro y suponiendo, que el SAT pueda contar con algunas fuentes de datos (Tabla 10).

Fuentes	Número de registros	Costo de $1.70 por registro
IMSS	15,658,480	26,619,416
SIEM	711,243	1,209,113
TELMEX	18,202,000	30,943,400
RFC	13,446,797	22,859,555
Total	**48,018,520**	**81,631,484**

Tabla 10: Cálculo de costos de los servicios de limpieza, estandarización y desduplicación de datos

Como se observa (Tabla 10), si se eligieran las alternativas ii o iii, se requeriría una inversión de poco más de 81 millones de pesos más IVA. Con la alternativa i se pueden abatir considerablemente los costos, al hacer uso de la plataforma tecnológica del SAT y se guarda la confidencialidad de la información, así la información adquirida de fuentes externas queda protegida bajo los mismos esquemas del RFC: Por los mecanismos de seguridad del SAT y legalmente por las disposiciones del Código Fiscal Federal y por la Ley Federal de Acceso a la Información.

El cuarto escenario no es adecuado debido a los costos internos asociados a la mala calidad de la información y al no desarrollar toda la potencialidad de la información que se obtiene de fuentes externas. Además, las áreas sustantivas no cuentan con la infraestructura tecnológica adecuada ni la experiencia requerida para llevar a cabo los procesos necesarios para generar el conocimiento con base en el uso de fuentes externas.

Estrategias:

- Es indispensable que la institución considere un eje de política y los altos mandos patrocinen la gestión de la calidad de la información.

- Debido al impacto de los beneficios que se podrían obtener en la recaudación así como el abatimiento de los costos de operación, es importante crear un área especializada en limpieza, estandarización, desduplicación y cruces de bases de datos, la cual deberá encargarse de monitorear la calidad de lo información y de ejecutar las acciones necesarias de mejorar continua y de aseguramiento de la calidad de la información.

- Contar con un repositorio en el que se almacene la totalidad de la información, así como los resultados de los procesos de calidad de datos y de los catálogos que apoyan a dichos procesos.

- Establecer las reglas del negocio que normen la calidad de los datos.

- Crear los catálogos de apoyo a los procesos de calidad de la información.

- Adquirir las herramientas de Software necesarias para llevar a cabo los procesos.

- Codificar las reglas y realizar los procesos de calidad de la información del padrón del RFC. Además llevar a cabo acciones de enriquecimiento de la información con los datos provenientes de fuentes externas como lo son: CFE, Catastros, Teléfonos, Registro Vehiculares, Bancos, Licencias de Funcionamiento, etc.

- Difundir las reglas entre los usuarios de los sistemas, principalmente entre los administradores de bases de datos y de los sistemas y propios los desarrolladores, a fin de que cualquier medio de adquisición de la información (sistemas, formatos, encuestas, etc.) cumplan en lo posible con las reglas del negocio.

- Crear la línea de producción para la higiene, estandarización y cruce de las bases de datos, a fin de crear un proceso permanente.

- Adquirir y Analizar la información de fuentes externas en conjunción con la información del SAT mediante técnicas de análisis estadístico y de minería de datos, a fin de modelar las funciones de riesgo y segmentación de contribuyentes. Así como, para la creación de modelos predictivos y en general para la explotación y uso de la información a fin para de mejorar el conocimiento del contribuyente

- Crear los modelos de información basados en fuentes externas y ponerlos a disposición de las áreas usuarias del SAT.

Procesos:

En los siguientes diagramas se esquematizan los procesos a los cuales debe ser sometida la información (Fig. 22 y 23).

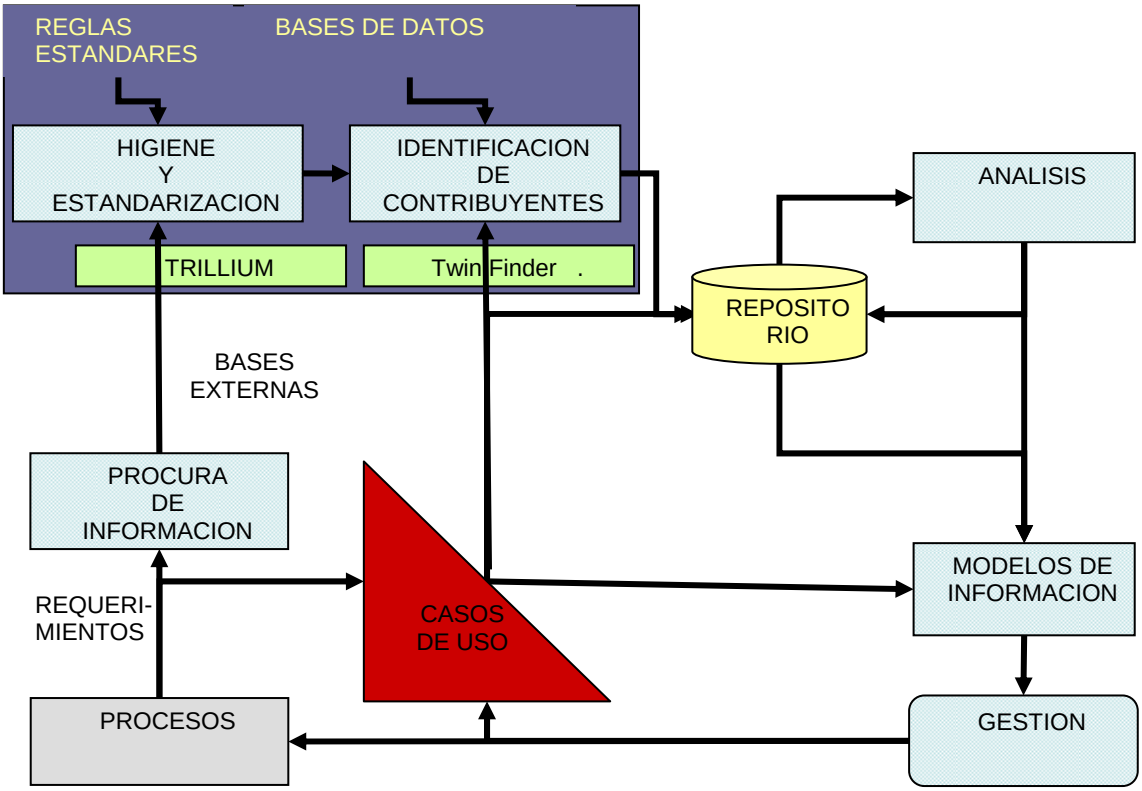


Fig. 22: Interacción entre los procesos de calidad de la información y la operación diaria de la operación

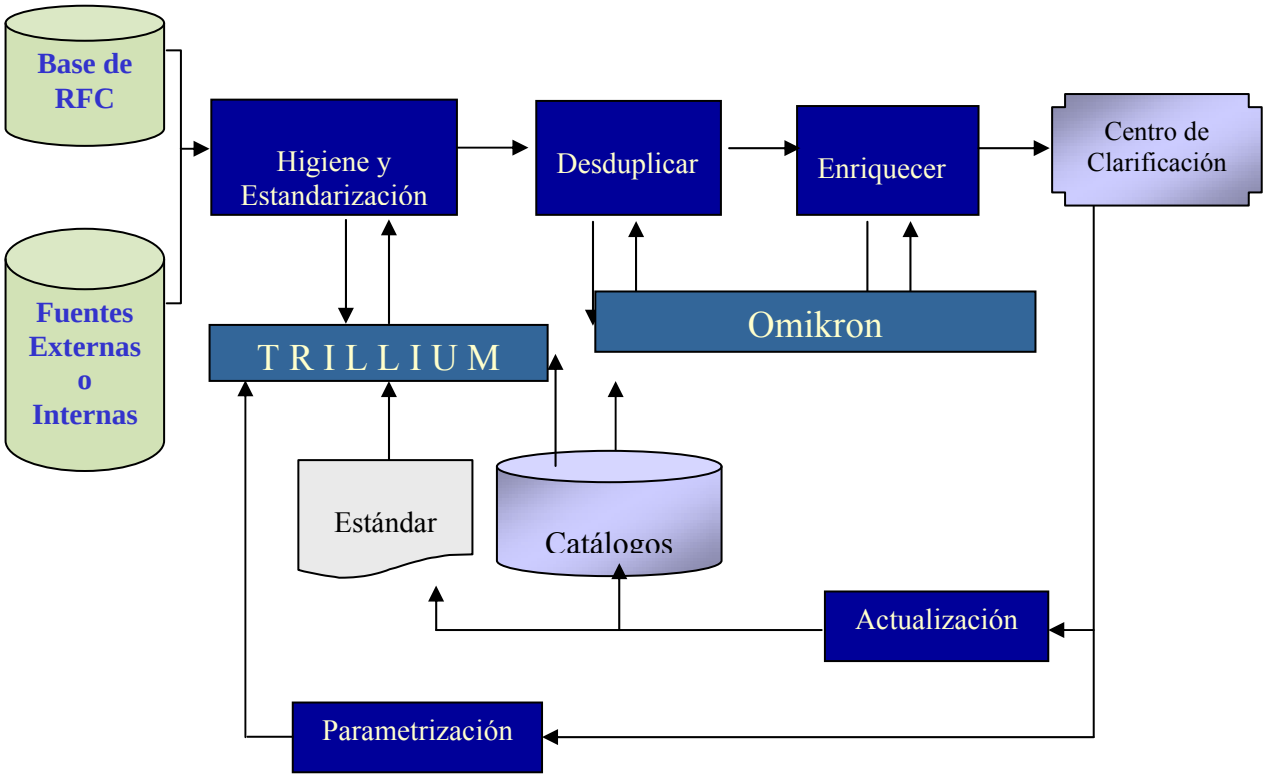


Fig. 23: Detalle de los procesos de calidad de la información

Se puede observar que los procesos de calidad de la información se inician únicamente con tres productos:

- La base de datos del sistema registro.
- La base de datos de las fuentes externas.
- La base de conocimientos (catálogos y estándares).

Los principales procesos de la metodología son:

- La higiene y estandarización de datos.
- Desduplicación.
- Enriquecimiento.

Resultados esperados del proceso de higiene y estandarización

La información en general debe presentar para cada uno de los campos de la base de datos, las siguientes características:

- Sin abreviaturas, excepto cuando la normatividad lo permita.
- Con mayúsculas.
- Sin palabras o caracteres no deseados.
- Sin errores de captura.

De forma general la información debe presentar las siguientes características en su estructura:

- Si no existe originalmente la colonia, se debe asignar en base al código postal.
- Tipo de asentamiento para la colonia.
- Tipo de vialidad para la calle.
- Dos campos para el número exterior, un campo con información numérica y otra que tenga la información alfanumérica.
- En caso de existir el campo referencia, se debe partir en dos, ya que por lo general se refieren las entrecalles.
-

Adicional a esto, la información de colonia y calle debe homologarse en base a los catálogos del SEPOMEX.

Resultados esperados de la desduplicación

Utilizando las herramientas tecnológicas adecuadas, en este proceso se tratarán de identificar los registros duplicados dentro de una misma base de datos, haciendo uso del nombre y del domicilio para tratar de identificar si se trata de una misma persona.

Los registros deberán ser identificados de la siguiente forma

C → Registros coincidentes en nombre.

F → Registros coincidentes en nombre con el mismo domicilio.

CA → Registros coincidentes en nombre con domicilio diferente.

M → Registros coincidentes sólo en el nombre.

F' → Registros coincidentes en nombre y otro campo, con distinto domicilio.

NC→ Registros no coincidentes.

P → Registros que sólo coinciden en domicilio.

La lógica con la que se obtienen los resultados anteriores es la siguiente.

$F = \{ A / (ANom1, Adom1) = (ANom2, Adom2) \}$

$F' = \{ A / (ANom1, Acampo1) = (ANom2, Acampo2) \} - F$

$M = \{ A / (ANom2) = (ANom2) \} - (F \cup F')$

$P = \{ A / (Adom1) = (Adom2) \} - (F \cup F')$

$CA = F' \cup M$

$C = F \cup CA$

$NC = A - (F \cup F' \cup M \cup P \cup CA \cup C)$

Donde:

- A es la base de datos.
- Anom son los nombres de las personas físicas o morales en A.
- Adom son los domicilios de las personas físicas o morales en A.

Resultados esperados del enriquecimiento de la información.

El proceso es similar al de desduplicación de la información; sin embargo, a diferencia de la desduplicación que se realiza sobre un mismo archivo, el enriquecimiento se obtiene al aplicar dichos procesos sobre dos o más archivos. Con esto, se logra realizar una identificación de los datos de una persona sobre las fuentes de información externas.

La lógica con la que se obtienen los resultados anteriores es la siguiente:

$F = \{ A / (ANom, Adom) = (BNom, Bdom) \}$

$F' = \{ A / (ANom, Acampo) = (BNom, Bcampo) \} - F$

$M = \{ A / (ANom) = (BNom) \} - (F \cup F')$

$P = \{ A / (Adom) = (Bdom) \} - (F \cup F')$

$CA = F' \cup M$

$C = F \cup CA$

$NC = A - (F \cup F' \cup M \cup P \cup CA \cup C)$

Donde:

- A y B son base de datos.
- Anom y Bnom son los nombres de las personas físicas o morales en A y B.
- Adom y Bdom son los domicilios de las personas físicas o morales en A y B.

Recursos:

El proyecto tiene un gran componente tecnológico, los recursos que se requieren son los siguientes:

1. Línea de Producción:
    a) Servidores (1 Datos, 1 procesamiento).
    b) Estaciones de trabajo o PC (2 procesamiento, 6 operación).
    c) Infraestructura de Redes y Comunicaciones.
    d) Personal capacitado.

2. Repositorio:
    e) Espacio en el Data Warehouse.
    f) Personal técnico.
    g) Herramientas de acceso (Consultas no planeadas).

3. Herramientas de análisis estadístico y minería de datos:
    h) Software.
    i) Capacitación para la administración de las herramientas.
    j) Capacitación para uso de las herramientas.

IV.4.6. Implementación: Poner en práctica el plan.

IV.4.6.1 Higiene y estandarización de datos.

El **proceso de higiene** es utilizado para dar calidad a los datos mediante: la expansión de abreviaturas, la eliminación de caracteres no deseados, la conversión de minúsculas a mayúsculas y la eliminación de los errores de captura.

Se comienza realizando un análisis de las variables involucradas, que son: nombre, apellido paterno, materno, entidad, localidad, municipio, colonia, calle, número exterior, número interior, referencia y CP. Lo que permite determinar cuáles son las variables que requieren de un proceso de higiene durante el cual se podrán determinar lo siguiente:

- Abreviaturas.
- Caracteres no deseados.
- Palabras más usadas.
- Análisis de errores de escritura.

El proceso se realiza con el CONVERTER de Trillium como herramienta principal, utilizando las siguientes funciones.

INP_TRAN01: Permite cambiar los caracteres no deseados dentro del proceso de cruce de información.

Ejemplo:

El & se sustituye por Ñ.

El Ð se sustituye por Ñ.

El ð se sustituye por Ñ.

OUT_CHANGE_RECODE: Se utiliza para expandir abreviaturas, cambiar o eliminar palabras no deseadas por otras que sean de mayor utilidad en el proceso de cruce de información.

Ejemplo:

MA en el campo nombre se sustituye por MARIA.

OTE en el campo calle se sustituye por ORIENTE.

OUT_SCAN_RECODE: Al usar esta función se puede eliminar información de un campo y moverla a otro campo. También se utiliza cuando existe dentro de un solo campo la información de calle y número exterior, en este caso se utilizan máscaras para poder detectar el punto de separación entre la calle y el número exterior.

Ejemplo:

Boulevard y sus abreviaturas dentro del campo calle, se elimina y se inserta en un campo que se utiliza para almacenar el tipo de vialidad.

OUT_MULTI_RECODE: Permite recodificar datos, es utilizada en conjunto con un catálogo que contiene dos campos: el campo1 que contiene el valor incorrecto que debe sustituirse por el valor que exista en el campo2.

Ejemplo:

M HIDALGO MIGUEL HIDALGO

OUT_BUILD_OR_LIST: Esta función es utilizada entre otras cosas, para convertir letras minúsculas a mayúsculas o viceversa.

Esto no significa que son las únicas funciones a utilizar, ya que la herramienta Trillium es muy amplia, por lo que se pueden usar otras funciones, mientras se cumpla con la finalidad de higiene.

El **proceso de estandarización** se lleva a cabo para poder asignar el valor correspondiente a cada campo, en función a la información contenida en los catálogos. Por ejemplo, se requiere que las colonias y municipios se escriban siempre conforme a los catálogos de SEPOMEX. También es utilizado para desatomizar los campos y separar los domicilios y nombres en estructuras de datos definidas.

Se utiliza principalmente para homologar la información de colonia y la calle a los catálogos de SEPOMEX, en base a la entidad y el municipio, también es utilizado para completar la información de:

- Colonia en base al código postal, en caso de no existir información de colonia.
- Tipo de vialidad, calle, no. exterior y no, interior
- Tipo de asentamiento.

El proceso se realiza utilizando el MATCHER de Trillium para comparar la información contra la base de datos que se está procesando y la que contiene los catálogos del SEPOMEX.

El tipo de vialidad y asentamiento se genera utilizando el CONVERTER de Trillium, basándose en las reglas del negocio que se establezcan.

En caso de que la información original no contenga colonia, se debe generar en base al código postal, para poder generar el tipo de asentamiento, además de aumentar la calidad de los resultados finales del cruce de información.

Al final de los procesos de higiene y estandarización es necesario revisar los resultados y determinar si es necesario ajustar parámetros y volver a realizar los procesos, esto se realiza mediante:

Selección de muestra aleatoria de registros

Se genera una muestra aleatoria para revisar los resultados de la higiene, utilizando el método aleatorio simple.

Revisión manual de registros seleccionados

Se analiza manualmente la muestra obtenida para poder determinar si continua existiendo palabras no deseadas, errores de captura, palabras abreviadas

Se puede utilizar como algoritmo un conteo de frecuencias, para auxiliarnos en la revisión manual.

Con base en los resultados se realiza lo siguiente:

- Propuestas de ajuste de parámetros: Se realiza para determinar que funciones o que parámetros de Trillium se deben modificar para corregir los errores detectados en la revisión manual.

- Reporte de incidencias: Se realiza para documentar las incidencias detectadas, con la finalidad de evitar repetir los errores encontrados y retroalimentar las reglas del negocio.

IV.4.6.2 Desduplicación y cruces de información.

Se utiliza para poder detectar registros duplicados en una base de datos y también para determinar asociaciones entre bases de datos que no tienen una clave en común.

Aunque éste puede ser realizado con el MATCHER de Trillium, existe una herramienta mucho más sencilla y eficiente para realizar este proceso denominada Omikron Adress Center (conocida también como Twin Finder).

En el software recomendado se pueden establecer porcentajes de coincidencia entre los campos, principalmente utiliza algoritmos de proximidad de cadenas de caracteres, con base en los porcentajes definidos, se podrán determinar tres grandes grupos de información (Fig. 24): duplicados, dudosos y no duplicados.

Con diferentes combinaciones de variables y sus respectivos porcentajes, se podrá determinar el valor de similitud. Los datos que se identificaron como dudosos requerirán un proceso asistido de revisión (centro de clarificación) cuyo fin es decidir si los datos identificados son o no realmente duplicados.

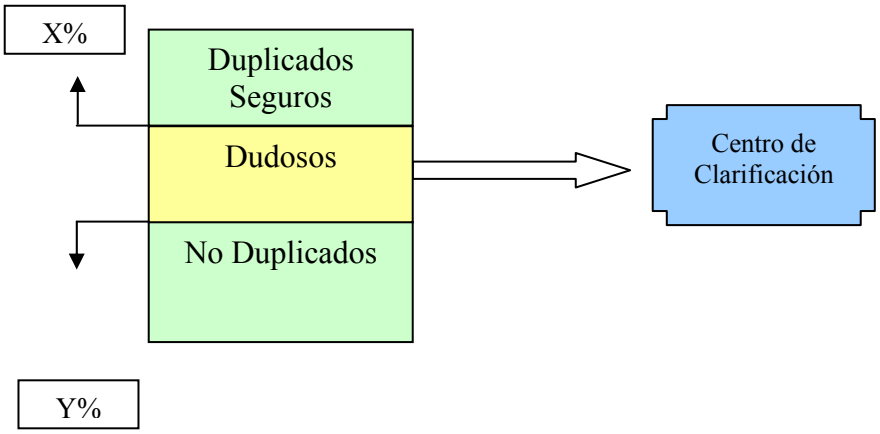


Fig. 24: Grupos de información generados por los procesos de desduplicación

Utilizando los resultados del cruce asistido, tenemos tres resultados adicionales, los registros que tienen el mismo nombre y el mismo domicilio, los que tienen el mismo nombre y el domicilio distinto y los que tienen el nombre distinto y el mismo domicilio.

Como se mencionó, el proceso para enriquecer la información es el mismo que se realiza para el de desduplicación. En el caso de Twin Finder, sólo requiere cambiar una especificación en la definición de realizar el proceso sobre la misma base de datos (Intra) o sobre una externa a la base de comparación (Extra)

Al terminar la desduplicación o cruce de base de datos, se debe contar con los siguientes resultados:

- Registros con el mismo nombre y el mismo domicilio.
- Registros con el mismo nombre y el domicilio distinto.
- Registros con el nombre distinto y el mismo domicilio.

IV.4.6.3 Definición de las Reglas del Negocio.

Las siguientes son un compendio de reglas negocio que pueden ser utilizadas para llevar a cabo los procesos de higiene y estandarización, para fines de este trabajo sólo se muestra un conjunto de reglas para ejemplificar a las mismas y no son exhaustivas.

Reglas del Negocio para domicilio

El domicilio debe atomizarse, es decir, dividirse en los siguientes elementos, por lo que todos sus componentes deberán ser asignados a campos separados cuyo tipo de dato sea definido de acuerdo al tipo de información que lo contenga:

Tipo de vialidad

Nombre de vialidad

No. exterior

No. interior

Tipo de asentamiento

Nombre de asentamiento

Código postal

Referencias de entre calles

Localidad

Municipio ó delegación

Entidad

Los tipos de caracteres deberán ser numéricos o alfabéticos de acuerdo a su información.

Los únicos caracteres especiales que se usarán son el apóstrofe (´) y el ampersan (&) toda vez que el nombre de la vialidad o asentamiento lo contenga.

Todos los caracteres extraños referidos a la letra ”Ñ” se deberán cambiar por dicha letra.

No deberán existir abreviaturas, a menos que se disponga de información oficial y así se determine.

Toda información deberá escribirse en mayúsculas y sin acentos.

Tipo de vialidad

Se refiere a la clasificación de vialidades y deberán de considerarse los siguientes (Tabla 11):

Tipo de Dato: alfanumérico

TIPO DE VIALIDAD
AUTOPISTA
ANDADOR
AVENIDA
BOULEVARD
CIRCUNVALACION
LIBRAMIENTO
PASAJE
CALLE
CALLEJON
CALZADA
CERRADA
CORREDOR
EJE VIAL
PRIVADA
PROLONGACION
CAMINO
CARRETERA
CIRCUITO

Tabla 11: Catálogo de tipo de vialidad

NOMBRE DE VIALIDAD o CALLE

Nombre propio asignado a la vialidad para su identificación. Este catálogo es difícil de conseguir y generalmente, se construye de la misma información contenida en las bases de datos, al cual se le debe dar un tratamiento de higiene, estandarización y desduplicación. No existe una normatividad de cómo el nombre de las calles es escrito y éste es asignado por las autoridades municipales, lo cual dificulta el proceso de obtener los nombres formales de las calles.

Tipo de Dato: alfanumérico

Todos aquellos nombres de calles que se refieren a personajes históricos o nombres propios, deberán complementarse.

Los nombres de vialidad que contengan la palabra “calle”, continuarán escribiéndose de la misma forma.

No se modificarán diminutivos ni superlativos.

El uso de los artículos se respeta tal cual se escribe el nombre de la calle.

En caso de que el nombre de vialidad contenga la orientación, ésta deberá escribirse sin abreviaciones y al final del nombre. (NORTE, SUR, ORIENTE, PONIENTE).

Existe un catálogo de abreviaciones (Tabla 12), las cuales por su importancia y frecuencia son válidas de utilizar para conformar el nombre de la calle.

ABREVIATURAS PERMITIDAS	
1a.	PRIMERA
2a.	SEGUNDA
3a.	TERCERA
1o.	PRIMERO
4a.	CUARTA
5a.	QUINTA
...	...
PTO.	PUERTO
ING.	INGENIERO
GRAL.	GENERAL
LIC.	LICENCIADO
ADMON.	ADMINISTRACION
DIAG.	DIAGONAL
FF CC	FERROCARRILES
F.C.	FERROCARRIL
DR.	DOCTOR
PROFRA.	PROFESORA
C.P.	CONTADOR PUBLICO
HNOS.	HNOS
NUM.	NUMERO
MTRO.	MAESTRO
EDO.	ESTADO
TNTE.	TENIENTE
DRA.	DOCTORA

Tabla 12: Fragmento del catálogo de abreviaturas

NÚMERO EXTERIOR

El número exterior es el número que identifica el predio en una vialidad.

Tipo de Dato: Alfanumérico

- El campo de número exterior sólo deberá contener información numérica.

- En caso de que la numeración sea una combinación de caracteres alfanuméricos, las letras se deberán enviar al campo número exterior complementario.

  Ejemplo:

Número exterior	Número exterior complementario
23	BIS
15	A
(NULL)	Lote 4 / Manzana 23
22	y 23
2	Lote 8 / Manzana 7

- Cuando se cuente con la combinación de un número exterior con BIS o alguna letra del alfabeto, éste se capturará junto con el número exterior, separado por un espacio en blanco.

- Cuando el domicilio reporte la numeración de manzana y lote, éstas deberán registrarse en el campo número exterior complementario, con la siguiente estructura:

  Número Exterior Complementario: Lote 4 / Manzana 23

- Cuando no se reporte el dato, no se registrará la información, es decir, en la base de datos existirá un NULL.
-

Número Interior

Es el número que identifica la vivienda al interior de un predio.

Tipo de caracteres: alfanumérico.

El contenido no se modificará.

TIPO DE ASENTAMIENTO

Se refiere a la clasificación del lugar en el que se establece el domicilio, destacando su cualidad espacial (Tabla 13).

Tipo de caracteres: alfanumérico

Deberán de considerarse como tipo de asentamientos los siguientes:

TIPO ASENTAMIENTO
COLONIA
FRACCIONAMIENTO
UNIDAD HABITACIONAL
CONJUNTO HABITACIONAL
PUEBLO
RANCHO
BARRIO
EJIDO
HACIENDA
AMPLIACIÓN
CONDOMINIO
CORREDOR INDUSTRIAL
CUARTEL
EX-HACIENDA
RESIDENCIAL
RINCONADA
SECCION
SECTOR
SUPERMANZANA
UNIDAD
ZONA INDUSTRIAL
ZONA MILITAR
RANCHERIA
AEROPUERTO
CAMPO MILITAR
ZONA FEDERAL
ZONA RURAL
BASE NAVAL
INGENIO
FRACCIONAMIENTO
UNIDAD HABITACIONAL

Tabla 13: Catálogo de tipo de asentamiento

NOMBRE DE ASENTAMIENTO

El nombre asentamiento se refiere a la identificación oficial que recibe un tipo de población definida espacialmente por una misma característica.

Tipo de caracteres: alfanumérico

Todos los caracteres extraños referidos a la letra ”Ñ” se deberán cambiar por dicha letra.

Todos aquellos nombres de asentamientos que se refieren a personajes históricos o nombres propios, deberán complementarse.

El uso de los artículos se respeta tal cual se escribe el nombre de la calle.

Todos aquellos nombres de asentamientos que hagan referencia a una fecha, deberán escribirse con número y letra.

Si existe un catálogo de abreviaciones, las cuales por su importancia y frecuencia son válidas, se deberá utilizar para conformar el nombre del asentamiento.

CÓDIGO POSTAL

Clave compuesta por cinco dígitos asignada a claves de repartos determinadas por el Servicio Postal Mexicano.

Tipo de caracteres: Alfanumérico

Se normalizará de acuerdo a la base de datos de códigos Postales que se obtenga de SEPOMEX, basados en su identificación por

Entidad Federativa
Ciudad
Municipio
Tipo de asentamiento
Nombre de asentamiento

REFERENCIAS

Nombre de las calles entre las que se encuentra un domicilio fiscal.
Tipo de caracteres: alfanumérico
Mismas reglas que para el campo calle.

LOCALIDAD, MUNICIPIO Y ENTIDAD

Se refiere a los límites administrativos establecidos por decreto.

Tipo de caracteres: alfanumérico.

La normalización de estos tres campos se realizará contra catálogo de SEPOMEX.

Reglas del Negocio para Nombres

El tipo de dato será alfanumérico

La información de nombres deberá atomizarse en:

Nombre (o Razón Social): Contiene la información del nombre de una persona moral y los nombres “de pila” de una persona física

Apellido Paterno

Apellido Materno

Tipo de Persona: Deberá identificarse si se trata de una persona Física o Moral.

Tipo de Sociedad: Para el caso de personas morales, se almacenará en este campo la información de tipo de sociedad y de régimen de capital

Los tipos de caracteres deberán ser numéricos o alfabéticos de acuerdo a su información.

Los únicos caracteres especiales que se usarán son el apóstrofe (´) y el ampersan (&) toda vez que el nombre cotejado contra el documento probatorio así lo contenga.

Todos los caracteres extraños referidos a la letra ”Ñ” se deberán cambiar por dicha letra.

No deberán existir abreviaturas, a menos de que se disponga de información oficial y así se así se determine.

Toda información deberá escribirse en mayúsculas y sin acentos.

Es necesario separar los registros de personas físicas del de personas morales.

Se requiere separar a otro campo el Régimen de capital y tipo de sociedad cuando se trata de Personas Morales.

Programando las reglas del negocio en Trilium.

Trillium permite definir las reglas del negocio desde le punto de vista de una gramática de un lenguaje, por lo que previo a mostrar esta proceso se introduce brevemente el concepto de gramática.

Gramática es el estudio de las reglas y principios que regulan el uso del lenguaje. También se denomina así al conjunto de reglas y principios que gobiernan el uso de un lenguaje determinado, por lo que puede decirse que cada lenguaje tiene su propia gramática. Mediante el uso de autómatas, esta gramática puede definirse formalmente:

Un autómata (AF) puede ser descrito como una 5-tupla (S, Σ, T, S0,A)

- *S* un conjunto de estados
- Σ es un alfabeto
- *T* es la función de transición
- S_0 es el estado inicial
- *A* conjunto de estados de aceptación o finales

En un autómata cada estado tiene una transición a otro estado por cada símbolo del alfabeto, la transición entre un estado y otro se norma por la función de transición del estado.

Un símbolo del alfabeto puede ser una letra número u otro tipo de caracteres o bien, un conjunto de ellos que al interpretarlo pueden constituir palabras, frases e incluso oraciones.

Los símbolos que conforman al alfabeto pueden ser nombrados para representar ciertos objetos o conceptos, a estos se les denomina Tokens. Cuando una sucesión de tokens, a través de una función de transición llevan a un estado de aceptación, la gramática se considera como válida.

Los autómatas son normalmente representados de forma gráficas con los siguientes símbolos:

Representa un estado del autómata

Representa un estado de aceptación del autómata

Representa la transición de un estado a otro, sobre la flecha se asocia el token que lleva a otro estado del autómata

Volviendo al caso de TRILLIUM, una forma de definir las reglas del negocio de forma gramaticalmente válida, es mediante la posibilidad que éste tiene para definir los patrones de datos a los que la información puede responder. Así por ejemplo, el nombre de una persona puede estar configurado con las siguientes sucesiones de datos:

Nombre, Apellido, Apellido

Nombre, Apellido

Apellido, Apellido, Nombre

Nombre, Apellido, “de”, Apellido

Nombre, Apellido, “Viuda de”, Apellido

Estos patrones no son más que una forma de representación de los autómatas

Así, si definimos los siguientes Tokens:

Nombre

Apellido

El siguiente alfabeto:

“Luis”, “Pedro”, “Maria”, “Martha”, “Juárez”, “Domínguez”, “Rosales”, “de”, ”viuda de”

Los valores válidos para cada Token:

Nombre {“Luis”, “Pedro”, “Maria”, “Martha”}

Apellido {“Juárez”, “Domínguez”, “Rosales”}

Conector{“de”, ”viuda de”}

Y definimos los estados y las transiciones, logramos tener la siguiente gramática, que se ilustra mediante un autómata (Fig. 25):

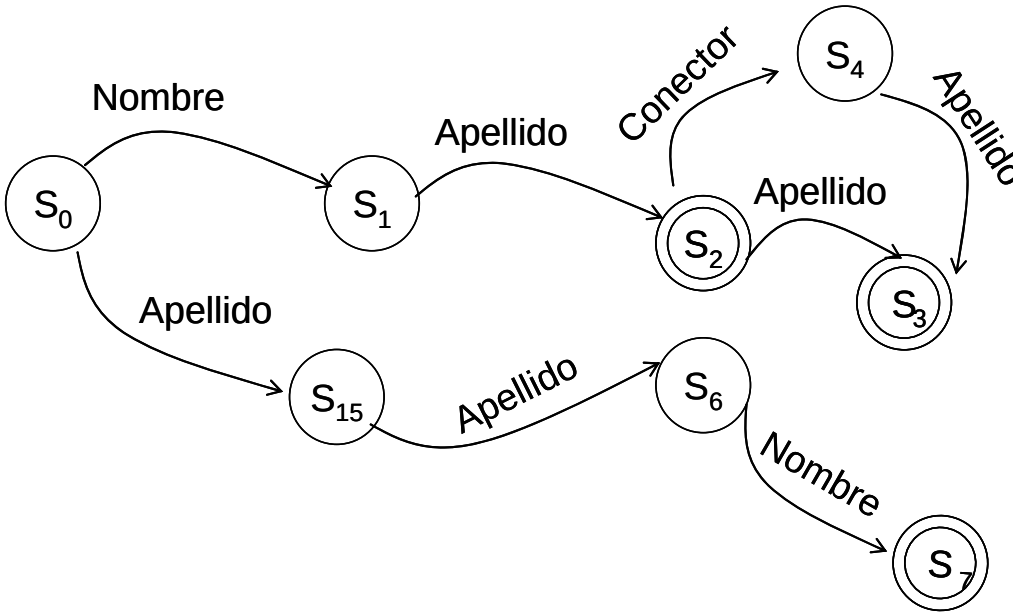


Fig. 25: Ejemplo de un autómata que representa la gramática del nombre

Así, siguiendo el autómata, oraciones como:

María Domínguez viuda de Rosales

Nos llevarán a la siguiente etiquetación de la oración:

Nombre{“María”}

Apellido{“Domínguez”}

Conector{“viuda de”}

Apellido {“Rosales”}

Siguiendo el autómata pasaríamos por los siguientes estados:

Estado Inicial	S0
Nombre{“María”}	S1
Apellido{“Domínguez”}	S2
Conector{“viuda de”}	S4
Apellido {“Rosales”}	S3

Como S3 es estado de aceptación la oración es gramaticalmente válida.

Sin embargo una oración de este tipo:

Rosa Maria Domínguez viuda de Rosales

Nos llevarán a la siguiente etiquetación:

ALFA{“Rosa”}

Nombre{“Maria”}

Apellido{“Domínguez”}

Conector{“viuda de”}

Apellido {“Rosales”}

Al no encontrar “Rosa” en el alfabeto, y en consecuencia no estar éste asociado a un TOKEN, se identifica como un tipo de datos no reconocido, en cuyo caso podría identificarse por el tipo de dato (ALFA). El autómata llega hasta el estado S0 (inicial) que no es estado de aceptación, por lo tanto se rechaza la oración como válida.

De igual forma funciona TRILLIUM, se requiere definir una serie de tokens y el conjunto de valores que éste puede tomar, esto se ingresa en forma de catálogos.

Por lo que respecta a las reglas de transición, al ser un producto en inglés, éstas se ingresan como patrones de la siguiente forma:

First, Last, Last
First, Last
Last, Last, First
First, Last, Userdefine, Last

La ejecución de las reglas del negocio siempre estará ligada al tema de catálogos. Estos son muy importantes para el desarrollo de los procesos de higiene y estandarización, ya que, además de permitirnos escribir la información de manera estandarizada, nos facilita la identificación de los diferentes componentes de un dato (tokens). Desde el punto de vista de una gramática, los catálogos representarán los datos válidos que un

Token puede adquirir, y, en consecuencia, se pueden determinar diferentes patrones (reglas de transición) a los que se puede responder.

Así, si los catálogos de nombres, calles, colonias, etc. no son lo suficientemente exhaustivos, no será posible el proceso de etiquetación y en consecuencia no se llegará a un estado de aceptación.

Del mismo modo, sucede con los patrones (representación del autómata en TRILLIUM), a un patrón no definido, existirá un dato no factible de procesar.

IV.4.6.4 Ejemplos de los resultados del procesamiento con TRILLIUM.

Para ejemplificar los procesos de higiene y estandarización de datos, se definieron las reglas del negocio y se generaron los catálogos necesarios sobre TRILLIUM, a fin de procesar el segmento de la base de datos del SIEM correspondiente a San Luis Potosí.

La siguiente imagen (Fig. 26) muestra un ejemplo del estado inicial de la información:

id	nombre
1718	SANDRA ARTEAGA LOPEZ
1719	ANTONIO ARELLANO LOPEZ
1720	JOSE MARLENE AVENDAÑO LEDEZMA
1721	ARTURO ALMENDARIZ AGUILA
1722	IRMA ARELLANO LUCERO
1723	IRMA ARELLANOS LUCERO
1724	NADIA IRAIS ALEJANDRO LOPEZ
1725	MARIA DEL ROSARIO ALEGO LUCANA
1726	CARMEN DEL ANGEL MAR
1727	CESAR ANGELES MARTINEZ
1728	EZEQUIEL ALEJOS MONTELONGO
1729	MA. ESTEFANIA DEL ANGEL MORATO
1730	GLORIA ACEVEDO MARTINEZ
1731	EUSEBIO ARMIDA MARTINEZ
1732	MA. GUADALUPE ANGELES MORAN
1733	GLORIA ACEVEDO MARTINEZ
1734	GABRIELA ARREDONDO MEDINA
1735	MARIA GUADALUPE ARTEAGA MENDOZA
1736	GLORICELA DEL ANGEL MARTINEZ
1737	JOSE ISAIAS ANGELES MATA

Fig. 26: Imagen del estado inicial de la tabla de nombres antes de procesar.

Iniciamos el procesamiento de la información incorporando un solo patrón y considerando que ya fue realizada la integración los catálogos correspondientes a nombres y apellidos para reconocer estos Tokens.

El patrón se define de forma específica para la categoría de datos nombre (name) (Fig. 27).

```
***
'Patrones de Nombre
***
'FIRST LAST LAST'
    INSERT PATTERN NAME DEF
    RECODE='USER1 USER2 USER3'
    EXPORT='USER1(1) USER2(1) USER3(1)'
```

Fig. 27: Ejemplo del patrón ingresado a TRILLIUM

Obteniéndose la siguiente salida de información (Fig. 28:)

Fig. 28: Imagen de los resultados de procesar los nombres con una sola regla.

A primera vista vemos una serie de patrones que no fueron identificados y que al no existir la regla, no fueron procesados, observamos que en el

43.78% de los casos, la información responde a un patrón en el que el dato del nombre corresponde a dos nombres propios y dos apellidos, situación muy común en México.

Al seleccionar el patrón de referencia, el producto nos muestra una serie de registros asociados a dicho patrón. Si al explorar los datos, descubrimos que se trata de un patrón válido, la herramienta tiene la posibilidad de generar automáticamente el código correspondiente, como se muestra en la siguiente figura (Fig. 29).

Fig. 29: Imagen que muestra como se inserta una nueva regla de un patrón no reconocido a TRILLIUM

También se puede presentar que un tipo de datos no haya sido identificado para alguno de los tokens, como se muestra a continuación (Fig. 30); en este caso, también se ingresa la información de forma automática a los catálogos, seleccionando el tipo de Token al que corresponde la información.

De esta forma se van alimentando las reglas del negocio en lo que respecta a la gramática.

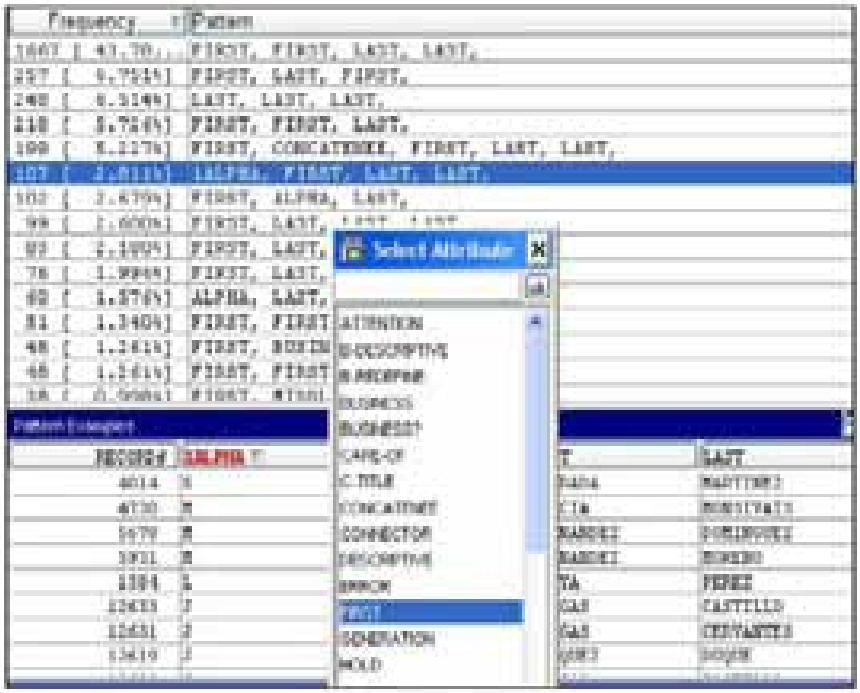


Fig. 30: Ejemplo de categorización de un dato no identificado a su respectivo Token

Las siguientes imágenes (Fig. 31 y 32) representan un fragmento del código que se genera para alimentar los catálogos y definir los patrones:

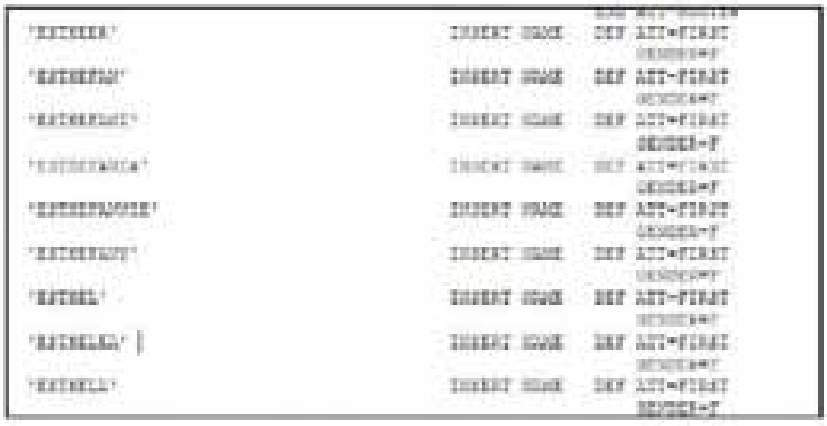


Fig. 31: Imagen de un fragmento del catálogo de nombres asociados al token first.

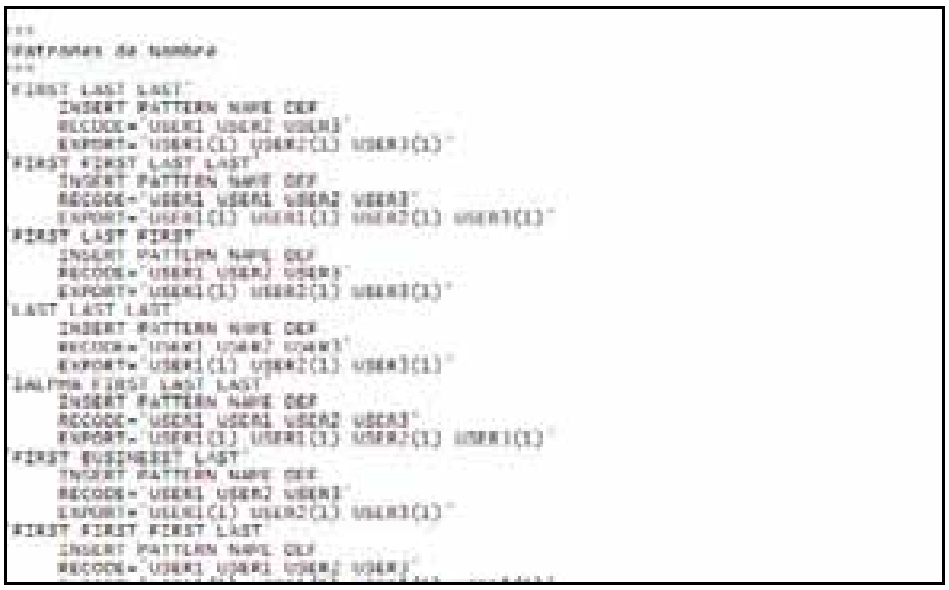


Fig. 32: Imagen de un fragmento de los patrones definidos para cumplir las reglas del negocio del dato nombre.

La identificación de las personas morales se logra al procesar todos aquellos registros para los cuales las reglas del nombre de personas físicas no fueron válidas y que en ellas se identifica el token “tipo de sociedad (Fig. 33)”. Enseguida se muestra un fragmento del código necesario para definir el catálogo correspondiente.

```
1.  "nombre","L"," SOCIEDAD DE PRODUCCION RURAL S DE PR DE R L","EC","tipo_soc",""
2.  "nombre","L"," S A DE R L DE C V","EC","tipo_soc",""
3.  "nombre","L"," SOCIEDAD ANONIMA","EC","tipo_soc",""
4.  "nombre","L"," S DE R L DE C V","EC","tipo_soc",""
5.  "nombre","L"," S DE P P DE R L","EC","tipo_soc",""
6.  "nombre","L"," S DE RL DE C V","EC","tipo_soc",""
7.  "nombre","L"," S DE PR DE R L","EC","tipo_soc",""
8.  "nombre","L"," S DE RL DE CV","EC","tipo_soc",""
9.  "nombre","L"," S DE RC DE CV","EC","tipo_soc",""
10. "nombre","L"," DE R L DE C V","EC","tipo_soc",""
11. "nombre","L"," S P R DE R L","EC","tipo_soc",""
12. "nombre","L"," S A D E C V","EC","tipo_soc",""
13. "nombre","L"," S C DE R L","EC","tipo_soc",""
14. "nombre","L"," S A DE C V","EC","tipo_soc",""
15. "nombre","L"," SA DE C V","EC","tipo_soc",""
16. "nombre","L"," S L P S C","EC","tipo_soc",""
17. "nombre","L"," S A DE CV","EC","tipo_soc",""
18. "nombre","L"," SC DE RL","EC","tipo_soc",""
19. "nombre","L"," SA DE VC","EC","tipo_soc",""
20. "nombre","L"," SA DE CV","EC","tipo_soc",""
21. "nombre","L"," SA DE CB","EC","tipo_soc",""
22. "nombre","L"," S DE R L","EC","tipo_soc",""
23. "nombre","L"," S A DE C","EC","tipo_soc",""
24. "nombre","L"," DE SA CV","EC","tipo_soc",""
25. "nombre","L"," SA DE V","EC","tipo_soc",""
26. "nombre","L"," S A C V","EC","tipo_soc",""
27. "nombre","L"," S R L","EC","tipo_soc",""
28. "nombre","L"," S L P","EC","tipo_soc",""
29. "nombre","L"," S C","EC","tipo_soc",""
30. "nombre","L"," S A","EC","tipo_soc",""
```

Fig. 33: Imagen de un fragmento del catálogo del tipo de sociedad asociado a su Token

Al concluir el procesamiento tendremos perfectamente estandarizada y atomizada la información en los campos correspondientes (Fig. 34).

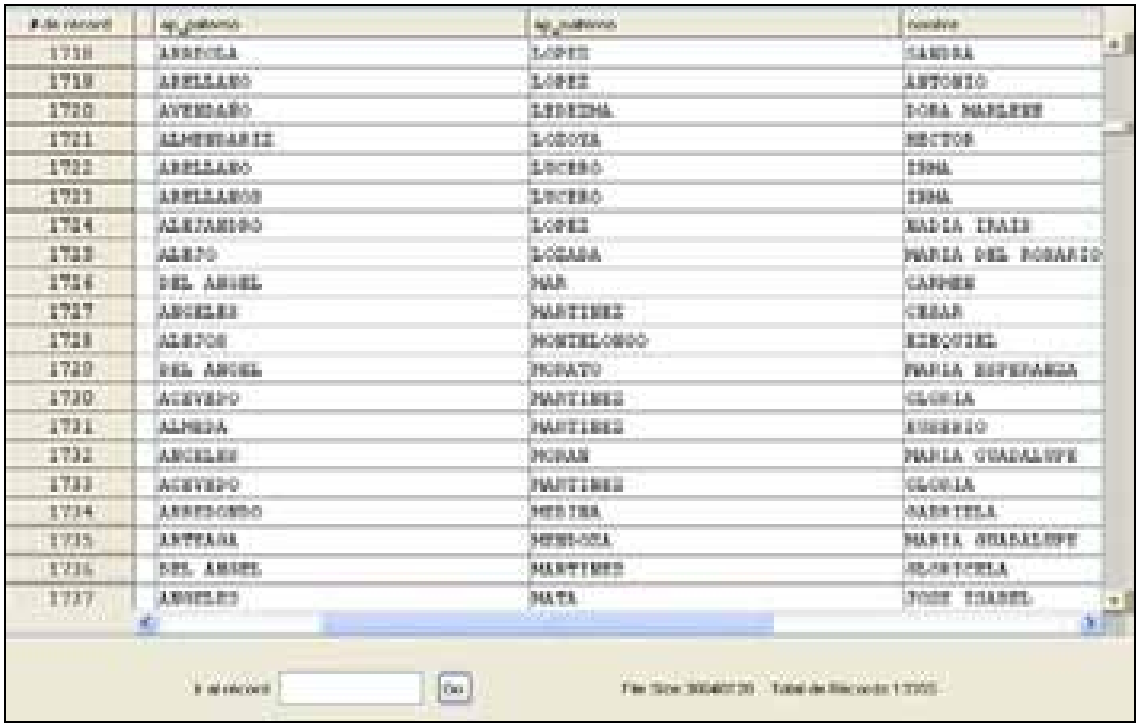


Fig. 34: Imagen del estado final de la tabla, una vez concluido el procesamiento sobre los nombres

Todo el procesamiento realizado se conoce como “PARSER” que desde el punto de vista de una gramática corresponde a un analizador sintáctico mediante el cual se analizan secuencias de tokens para determinar su estructura gramatical respecto a una gramática formal dada.

Este proceso nos permite mejorar la calidad de los datos ya que ésta queda estructurada y es reconocida como una oración válida de conformidad al autómata que se definió para el tipo de dato, de lo contrario la información que no fue aceptada como válida tendrá que ser tratada mediante procesos asistidos.

Dependiendo del grado de definición que se alcance a nivel de reglas y catálogos, el proceso puede resolver hasta un 95% de los datos contenidos en una base de datos. En el ejemplo mostrado se logró resolver el 89% de los casos.

Es claro que previo a todo este proceso y para hacer una adecuada identificación de los tokens que componen a una oración, se debió realizar un proceso de higiene de la información. Para ello el software TRILLIUM también cuenta con herramientas que nos permiten lograr esto.

El proceso de higiene se orienta a lograr una adecuada clasificación del alfabeto que conforma los tokens, es decir, se orienta a tener las palabras adecuadas para que puedan ser etiquetadas correctamente, básicamente se puede realizar de dos formas:

1. Con catálogos de reemplazo.

2. Definiendo sinónimos.

Los catálogos de reemplazo son de la forma “dice”, “debe decir”, y se utilizan en rutinas de sustitución de datos.

Lo sinónimos son definiciones que se realizan dentro de los tokens para decir que una palabra significa lo mismo que otra.

Los efectos son los mismos; sin embargo, mientras que el primero modifica los datos de forma previa al Parser, el segundo, opera durante la ejecución del mismo.

En el caso de nuestro objetivo de mejorar la calidad de los datos, utilizamos el método de catálogos de reemplazo, ya que la definición de las reglas del negocio nos orienta en el sentido de no tener información abreviada.

En las siguientes figuras (Fig. 35 y 36) se muestran fragmentos del código que genera los catálogos de reemplazo en TRILLIUM y que a su vez clasifica las nuevas palabras como miembros de un Token.

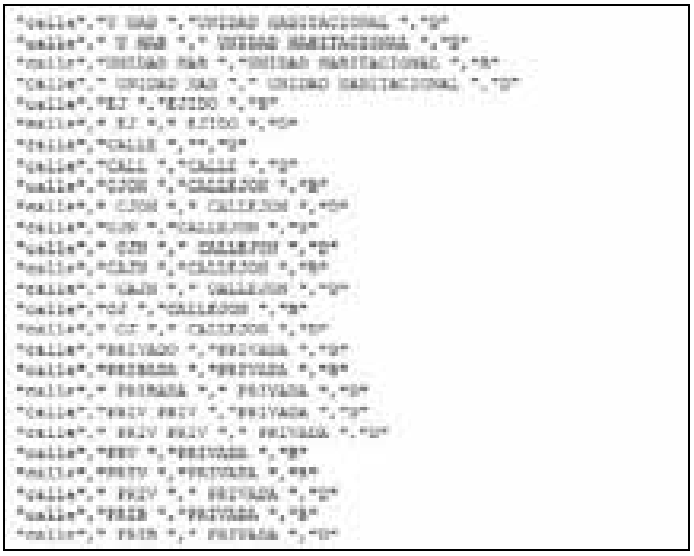


Fig. 35: Imagen que muestra un fragmento del catálogo de reemplazo de calles.

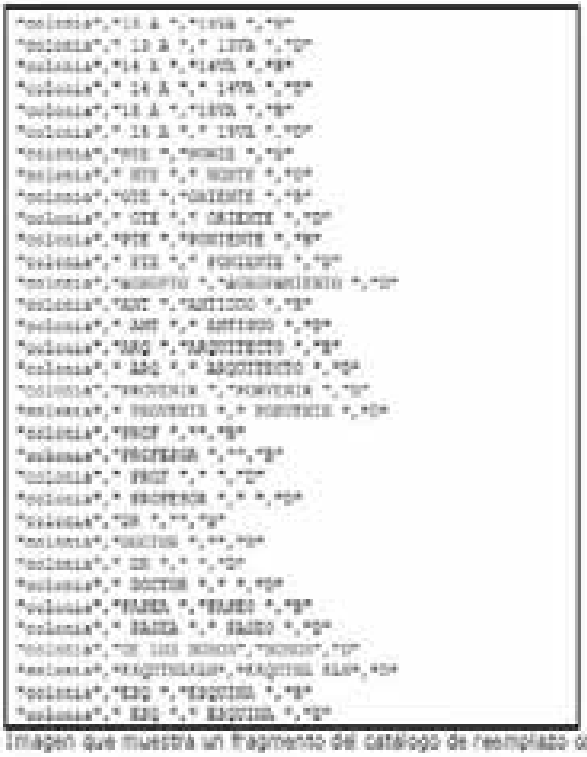


Fig. 36: Imagen que muestra un fragmento del catálogo de reemplazo colonias.

Como resultado final tendremos a los datos estandarizados y de mejor calidad.

IV.4.6.5 Ejemplos de los resultados del procesamiento con Omikron Adress Center (TWIN FINDER).

Enseguida se muestra un ejemplo del uso de la herramienta Omikron Adress Center para realizar los procesos de cruce de bases de datos, que son en los que se basa la actividad de desduplicación de datos y de enriquecimiento de la información.

Cabe mencionar que el software TRILLIUM tiene su propia herramienta para realizar la desduplicación y cruces de bases de datos; sin embargo, se prefiere utilizar una herramienta especializada en este tipo de procesos.

El éxito de un buen proceso de desduplicación o de cruce de bases de datos, está dado en gran medida por los resultados alcanzados durante la estandarización e higiene de los datos.

Omikron basa su funcionamiento en la identificación de grupos de información cuyos datos tienen un cierto porcentaje de similitud, que es definido por el usuario. El producto es muy eficiente; sin embargo, no existe forma de decirle que tipo de algoritmo de proximidad de cadenas utilizar y tampoco se localizó documentación al respecto.

El ejercicio que continuación se muestra, toma como salida los resultados obtenidos de TRILLIUM y tiene como objetivo identificar los posibles duplicados en la base de datos.

El proceso comienza con la carga de los datos, para lo cual se puede hacer conectándose a prácticamente cualquier base de datos siempre que se tenga el ODBC correspondiente. También, de manera predefinida identifica diversos tipos de formatos.

En el proceso de carga de los archivos, se puede incorporar un solo archivo, si se va a realizar el proceso de desduplicación y más de uno si se va a realizar un cruce entre bases de datos. La siguiente figura (Fig. 37) muestra esta acción:

Fig. 37. Imagen que muestra la carga de un archivo para su procesamiento en TIGN FINDER

El parámetro INTRA permite diferenciar si se trata de una desduplicación (INTRA=SI) o un cruce de tablas (INTRA=NO), en cuyo caso se especifica cual es la tabla principal o de referencia y cuáles las secundarias.

Debido a que el proceso para desduplicar o enriquecer la información, mediante el cruce de base de datos, sólo difiere de cómo se especifica el parámetro INTRA, en los siguientes ejemplos sólo se muestra un proceso de desduplicación.

El siguiente paso a realizar en la herramienta, consiste en definir la matriz de porcentajes de similitud que guiará al proceso como se muestra a continuación (Fig. 38):

Fig. 38: Imagen de la pantalla de definición de la matriz de coincidencias TWIN FINDER.

Se definieron 4 columnas de porcentajes, que en orden de izquierda a derecha, nos permitirán identificar:

1. Registros con el mismo nombre y domicilio.
2. Registros con gran similitud en el nombre y domicilio.
3. Registros con gran similitud en el nombre.
4. Registros con gran similitud en el domicilio.

Una vez que se define la matriz y se identifican los campos correspondientes en la tabla, se especifican las ventanas de búsqueda. Éstas se crean al definir un orden lógico dentro de la tabla (índices) y se determina el entorno de búsqueda, que corresponde al número de registros a revisar antes y después del registro que se está procesando. La siguiente figura (Fig. 39) muestra, los entornos definidos.

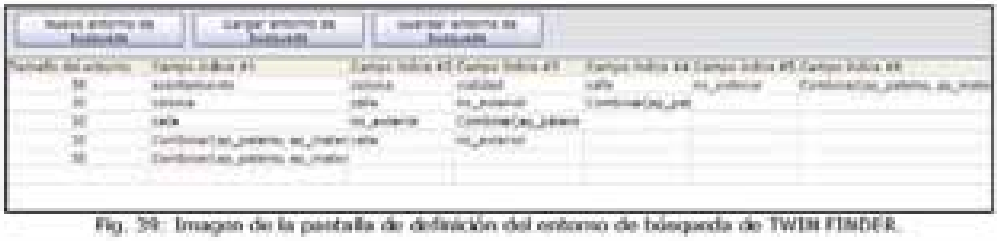


Fig. 39: Imagen de la pantalla de definición del entorno de búsqueda de TWIN FINDER.

El resto lo realiza el producto, aunque será necesario probar diversas configuraciones hasta obtener un buen resultado. En la siguiente imagen se muestra la pantalla resultante del proceso (Fig. 40).

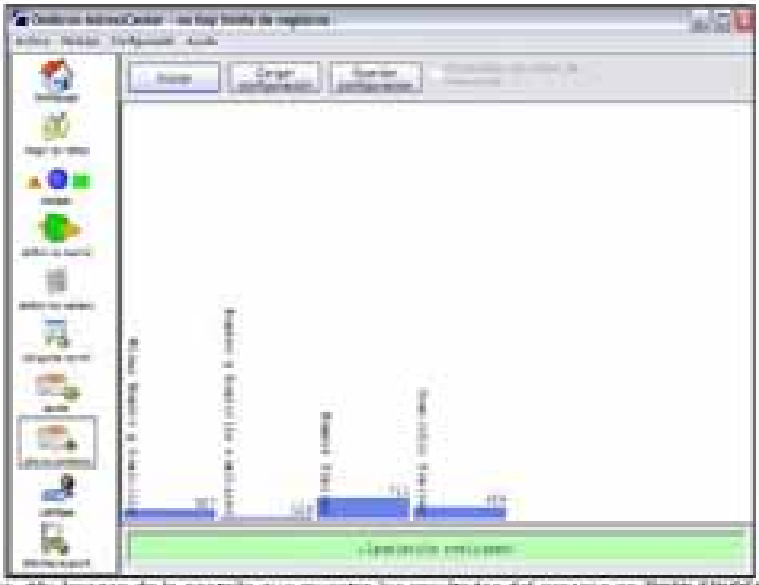
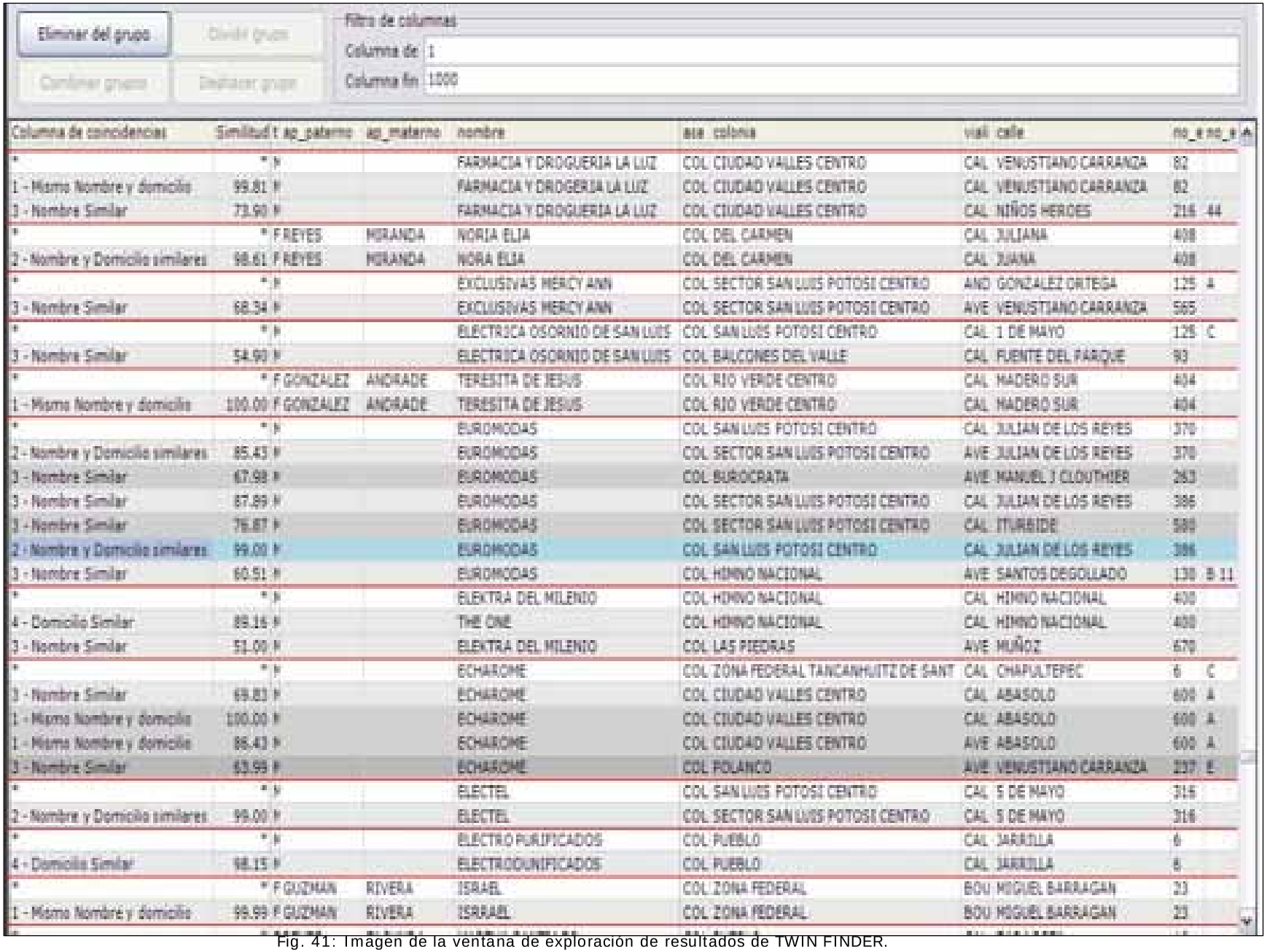


Fig. 40: Imagen de la pantalla que muestra los resultados del proceso en Data Finder.

En este caso de 13,355 registros se encontraron 327 duplicados seguros (mismo nombre y domicilio), 113 posibles duplicados por similitud en nombre y domicilio, 711 con el mismo nombre y 436 con el mismo domicilio.

En la siguiente página se muestra un fragmento de la pantalla de exploración de datos (Fig. 41) en la que se puede observar los grupos de asociación de los registros identificados como duplicados:



Columna de coincidencias	Similitud	t	ap_paterno	ap_materno	nombre	asa	colonia	vial	calle	no_e	no_i
*	*	M			FARMACIA Y DROGUERIA LA LUZ	COL	CIUDAD VALLES CENTRO	CAL	VENUSTIANO CARRANZA	82	
1 - Mismo Nombre y domicilio	99.81	M			FARMACIA Y DROGERIA LA LUZ	COL	CIUDAD VALLES CENTRO	CAL	VENUSTIANO CARRANZA	82	
3 - Nombre Similar	73.90	M			FARMACIA Y DROGUERIA LA LUZ	COL	CIUDAD VALLES CENTRO	CAL	NIÑOS HEROES	216	44
*	*	F	REYES	MIRANDA	NORIA ELIA	COL	DEL CARMEN	CAL	JULIANA	408	
2 - Nombre y Domicilio similares	98.61	F	REYES	MIRANDA	NORA ELIA	COL	DEL CARMEN	CAL	JUANA	408	
*	*	M			EXCLUSIVAS MERCY ANN	COL	SECTOR SAN LUIS POTOSI CENTRO	AND	GONZALEZ ORTEGA	125	A
3 - Nombre Similar	68.34	M			EXCLUSIVAS MERCY ANN	COL	SECTOR SAN LUIS POTOSI CENTRO	AVE	VENUSTIANO CARRANZA	565	
*	*	M			ELECTRICA OSORNIO DE SAN LUIS	COL	SAN LUIS POTOSI CENTRO	CAL	1 DE MAYO	125	C
3 - Nombre Similar	54.90	M			ELECTRICA OSORNIO DE SAN LUIS	COL	BALCONES DEL VALLE	CAL	FUENTE DEL PARQUE	93	
*	*	F	GONZALEZ	ANDRADE	TERESITA DE JESUS	COL	RIO VERDE CENTRO	CAL	MADERO SUR	404	
1 - Mismo Nombre y domicilio	100.00	F	GONZALEZ	ANDRADE	TERESITA DE JESUS	COL	RIO VERDE CENTRO	CAL	MADERO SUR	404	
*	*	M			EUROMODAS	COL	SAN LUIS POTOSI CENTRO	CAL	JULIAN DE LOS REYES	370	
2 - Nombre y Domicilio similares	85.43	M			EUROMODAS	COL	SECTOR SAN LUIS POTOSI CENTRO	AVE	JULIAN DE LOS REYES	370	
3 - Nombre Similar	67.98	M			EUROMODAS	COL	BUROCRATA	AVE	MANUEL J CLOUTHIER	263	
3 - Nombre Similar	87.89	M			EUROMODAS	COL	SECTOR SAN LUIS POTOSI CENTRO	CAL	JULIAN DE LOS REYES	386	
3 - Nombre Similar	76.87	M			EUROMODAS	COL	SECTOR SAN LUIS POTOSI CENTRO	CAL	ITURBIDE	580	
2 - Nombre y Domicilio similares	99.00	M			EUROMODAS	COL	SAN LUIS POTOSI CENTRO	CAL	JULIAN DE LOS REYES	386	
3 - Nombre Similar	60.51	M			EUROMODAS	COL	HIMNO NACIONAL	AVE	SANTOS DEGOLLADO	130	B 11
*	*	M			ELEKTRA DEL MILENIO	COL	HIMNO NACIONAL	CAL	HIMNO NACIONAL	400	
4 - Domicilio Similar	89.16	M			THE ONE	COL	HIMNO NACIONAL	CAL	HIMNO NACIONAL	400	
3 - Nombre Similar	51.00	M			ELEKTRA DEL MILENIO	COL	LAS PIEDRAS	AVE	MUÑOZ	670	
*	*	M			ECHAROME	COL	ZONA FEDERAL TANCANHUITZ DE SANT	CAL	CHAPULTEPEC	6	C
3 - Nombre Similar	69.83	M			ECHAROME	COL	CIUDAD VALLES CENTRO	CAL	ABASOLO	600	A
1 - Mismo Nombre y domicilio	100.00	M			ECHAROME	COL	CIUDAD VALLES CENTRO	CAL	ABASOLO	600	A
1 - Mismo Nombre y domicilio	86.43	M			ECHAROME	COL	CIUDAD VALLES CENTRO	AVE	ABASOLO	600	A
3 - Nombre Similar	63.99	M			ECHAROME	COL	POLANCO	AVE	VENUSTIANO CARRANZA	237	E
*	*	M			ELECTEL	COL	SAN LUIS POTOSI CENTRO	CAL	5 DE MAYO	316	
2 - Nombre y Domicilio similares	99.00	M			ELECTEL	COL	SECTOR SAN LUIS POTOSI CENTRO	CAL	5 DE MAYO	316	
*	*	M			ELECTRO PURIFICADOS	COL	PUEBLO	CAL	JARRILLA	6	
4 - Domicilio Similar	98.15	M			ELECTRODUNIFICADOS	COL	PUEBLO	CAL	JARRILLA	6	
*	*	F	GUZMAN	RIVERA	ISRAEL	COL	ZONA FEDERAL	BOU	MIGUEL BARRAGAN	23	
1 - Mismo Nombre y domicilio	99.99	F	GUZMAN	RIVERA	ISRAEL	COL	ZONA FEDERAL	BOU	MIGUEL BARRAGAN	23	

Fig. 41: Imagen de la ventana de exploración de resultados de TWIN FINDER.

No obstante a que este producto genera muy buenos resultados, es importante que los registros calificados como similares sean revisados por un grupo de personas, en lo que se ha denominado el centro de clarificación. Este equipo de personas tendrá la responsabilidad de determinar si un grupo de registros son o no duplicados reales.

Una vez concluido el trabajo de calidad de datos, estos deben ser devueltos a la fuente de datos original para que sean sustituidos y la organización pueda disfrutar de los efectos de mejora en la calidad de los datos.

IV.4.7. Revisar las actividades de mejora para determinar la adecuación de las actividades de seguimiento.

Finalmente, deberá realizarse una supervisión a las mejoras realizadas, esto permitirá determinar si las actividades que se realizaron para mejorar la calidad son las adecuadas y se está llegando a los resultados esperados.

El no alcanzar los resultados esperados significará que debe mejorarse y planificarse mejor las estrategias que se han definido o bien, que se deben retroalimentar los resultados a fin de corregir las deficiencias.

El siguiente gráfico (Fig. 42) muestra los primeros resultados alcanzados en el procesamiento de la información previo a la definición de los patrones y la mejora en la parametrización del Software TRILLIUM.

Fig. 42: Resultados de TRILLIUM cuando se proceso la información con un solo patrón.

En este caso, el proceso de identificación y estandarización de nombres sólo se alcanzó a procesar el 60% de la base de datos.

Después de un proceso de revisión de los resultados y de mejorar la parametrización del software, se alcanzó un 89% de porcentaje de registros estandarizados. (Fig. 43)

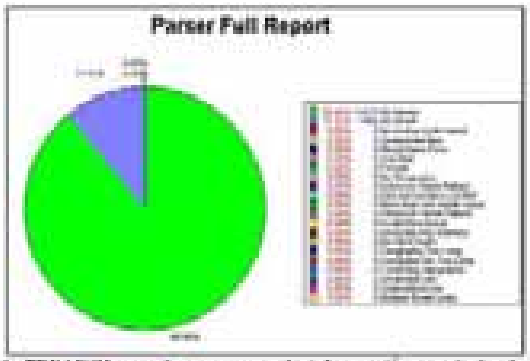


Fig. 43: Resultados de TRILLIUM cuando se procesa la información con todos los patrones definidos.

En el caso de OMIKRON, el primer proceso de desduplicación arrojó sólo la identificación de 247 nombres y domicilios duplicados, 157 registros identificados con el mismo nombre y 535 con el mismo domicilio, como se vé en la siguiente figura (Fig. 44):

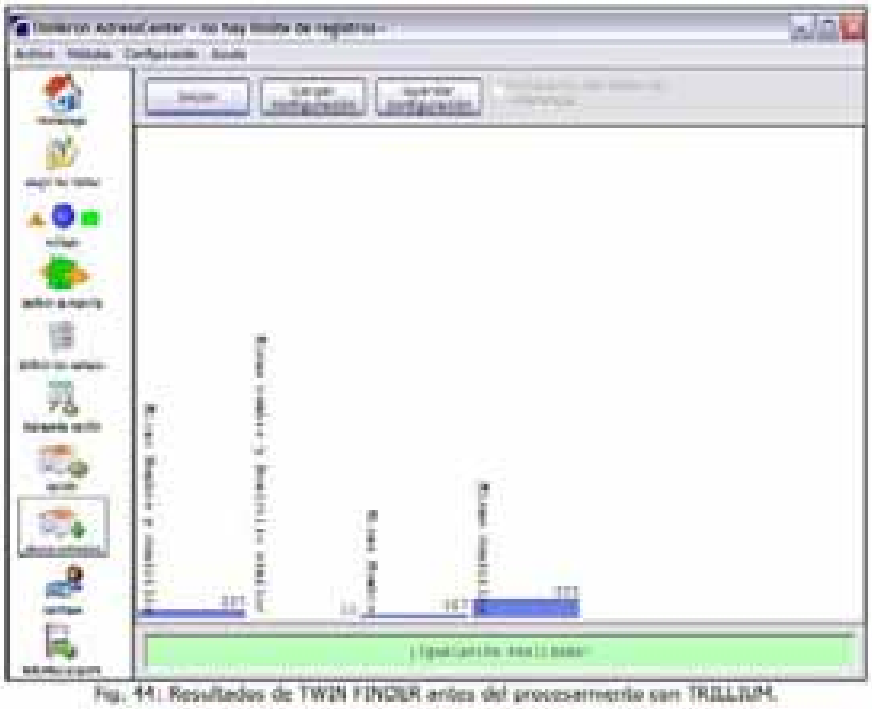


Fig. 44: Resultados de TWIN FINDER antes del procesamiento con TRILLIUM.

Posterior a los procesos de revisión y de las acciones emprendidas para mejorar el proceso se logró, como ya se había mencionado, encontrar a 440 posibles duplicados por nombre y domicilio, 711 con el mismo nombre y 436 con el mismo domicilio (Fig. 40).

V. Hacia la Gestión del Conocimiento.

Una organización basada en la información tiene su eje central en el manejo y difusión de la misma y en el empleo de las comunicaciones electrónicas, más que en la generación de conocimiento. Se organiza sobre la base "del uso generalizado de información a bajo costo, del almacenamiento de datos y de las tecnologías de la transmisión"; en cambio, las organizaciones basadas en el conocimiento y el aprendizaje se centran en la "capacidad de innovar y crear valor más rápido en base al conocimiento y a su rápida actualización en diversos ámbitos por medio del aprendizaje"[15].

Las estrategias de desarrollo a largo plazo visualizan al conocimiento como factor estratégico, por ello la resolución de problemas y la toma de decisiones se realizan básicamente por medio de operaciones cuyo soporte son[16]:

- La disponibilidad de la información y conocimiento.
- La capacidad de analizar, clasificar, modelar y relacionar sistémicamente datos e información.

Una serie de transformaciones claves en las organizaciones deben acompañar el cambio hacia una organización basadas en la gestión del conocimiento. Estos cambios y transformaciones se focalizan en[16]:

- La forma en cómo se hacen las cosas (se tiende a administrar por competencias más que por puesto de trabajo).
- Las formas de encarar la combinación del uso de la tecnología con los saberes individuales y organizacionales acumulados.
- Los nuevos modelos de formación profesional (por ejemplo: la certificación por competencias, la formación modular basada en competencias).

15 Conceptos expresados sobre la economía por MONTUSCHI, L. "La economía basada en el Conocimiento: importancia del conocimiento tácito y del Conocimiento Codificado", CEMA, Buenos Aires, 2000.

16 Martha Beatriz Peluffo A. y Edith Catalán Contreras, "Introducción a la gestión del conocimiento y su aplicación al sector público", CEPAL, 2002.

- Las nuevas formas de comunicar el conocimiento y de construirlo (conocimiento tácito almacenado, técnicas para el análisis de la información, los bancos de ideas, de conocimiento, las mejores prácticas y lecciones aprendidas entre otros).

- El cambio cultural experimentado por la aceptación de los beneficios del nuevo modelo sobre el tradicional entre otros (nuevas formas de valorización del trabajo, el papel del factor humano, la mayor autonomía para el desarrollo tareas, el alineamiento entre los intereses individuales y los organizacionales).

- En definitiva el cambio se generó por la necesidad de búsqueda de mejores estrategias para aumentar la calidad y la eficiencia en el manejo de los recursos almacenados en las propias organizaciones: el conocimiento y en la capacidad para absorber nueva información.

En los siguientes apartados analizaremos las opciones para resolver los temas de disponibilidad de la información y del conocimiento, por una parte y por otra el uso de la minería de datos para analizar, clasificar, modelar y relacionar los datos y la información a fin de crear conocimiento.

V.1 Enriquecimiento de la información.

Entendemos por enriquecer la información a la acción de asociar datos que no existan en nuestros sistemas de registros. Ya sea porque no se dispone de estos, ya que no fueron proporcionados por las fuentes de información, o bien porque estos no fueron captados en su oportunidad o quizá porque son datos que se derivan de dar cierto tratamiento a la información existente.

Como ejemplo de datos que no se solicitan y que se infieren de otras variables podemos citar el caso del nivel socioeconómico de las personas, que puede inferirse de una serie de preguntas como son: grado de estudios, número de habitaciones del hogar, número de baños completos, total de autos en la familia, etc. Como se mencionó anteriormente, la Asociación Mexicana de Agencias de Investigación de Mercados y Opinión Pública, A.C. tiene una batería de preguntas para definir niveles socioeconómicos.

Este tipo de información se puede captar mediante encuestas o mediante la asociación de variables socioeconómicas que proporciona el INEGI, como resultado de los Censos y Conteos de Población y Vivienda.

Como ejemplos de datos no proporcionados o captados podemos citar: teléfono, CP, colonia, e-mail, datos vehiculares, otros domicilios, etc.

Para el caso del SAT, el nivel socioeconómico de sus contribuyentes es fácilmente determinable ya que es información que se capta en las declaraciones de los contribuyentes. Sin embargo, existen datos que no son captados o que bien se desactualizan rápidamente. Es por esto que la propuesta que se realiza para enriquecer la información se basa en:

1. Captar información de fuentes externas.
2. Realizar los procesos de higiene, estandarización y cruce de bases de datos para poder asociar la información de fuentes externas a la de los contribuyentes, recordemos que en general los sistemas de registro no utilizan llave de identificación comunes, como lo es la CURP.
3. Asociar la información a los contribuyentes y ponerla a disposición de las áreas usuarias.

Para lograr esto es indispensable llevar a cabo una serie de acciones que garanticen la disponibilidad de la información y, bajo el esquema del SAT, se garantice la confidencialidad de la misma, por lo que se proponen las siguientes acciones:

- Identificar la información de fuentes externas que es de utilidad para el SAT y que puede ser gestionada de forma oficial.

- Establecer mecanismos permanentes y sistemáticos de intercambio y actualización de la información, tanto tecnológicos, como administrativos.

- Asegurar la confidencialidad de la información implantado procedimientos administrativos e instrumentando las acciones tecnológicas necesarias.

- Realizar las acciones de identificación de contribuyentes en fuentes externas, basados en los procesos de higiene, estandarización y cruces de bases de datos.

- Impulsar el uso de estándares y catálogos compartidos en el SAT y entre las entidades externas.

- Crear modelos de información que respondan a las necesidades de las áreas sustantivas del SAT.

Las posibles bases de datos que pueden apoyar al SAT para enriquecer su información y generar conocimiento sobre el contribuyente son las siguientes:

Seguridad Social: Patrones y Asalariados.- Permitirá identificar a contribuyentes no registrados, además de contar con información sobre la actividad económica y datos sobre las nóminas, lo que será de utilidad para evitar posibles acciones de evasión fiscal. Es el mismo caso para los registros de **Impuestos Locales**

Catastros: Las bases de datos de los catastros, generalmente son poco actualizadas; sin embargo, los domicilios son por lo general permanentes, por lo que un cruce de bases de datos para asociar los catastros a los domicilios de los contribuyentes, permitirá detectar posible actividad económica, ya que en los catastros se almacena el uso del suelo.

Licencias de funcionamiento: Son aquellas expedidas por las autoridades locales para la apertura de negocios, a partir de esta información el SAT puede determinar contribuyentes que no estén inscritos o detectar actividades comerciales no registradas.

Padrón Vehicular: Con la información de estos registros, y una vez determinando el tipo de auto y cantidad de autos propiedad de una persona, se pueden inferir signos de riqueza y detectar posibles contribuyentes; además el tipo de vehículo permite detectar si se realiza alguna actividad comercial por la cual se deba estar inscrito.

Licencias de Conducir: El valor agregado de este tipo de licencias, está en que para emitirla los gobiernos estatales solicitan comprobantes de domicilios, lo que resulta en una fuente de información para conocer el domicilio particular de una persona, recordemos que en el caso del SAT, el domicilio registrado es el domicilio fiscal.

Consumos: El caso de los consumos de luz y agua, permiten detectar mediante métodos estadísticos a aquellos domicilios en

los que se esté realizando un consumo muy por encima de la media de una cierta área geográfica, lo que puede representar que existan comercios o fábricas que pudieran operar sin el adecuado registro en el SAT.

Registro Público de la Propiedad. Con esta información, además de contar con domicilios alternos de los contribuyentes para su localización, el SAT tendría a su disposición información sobre los bienes de los contribuyentes a fin de ejercer acciones de cobro.

Como se ve el cruce de datos, además de permitirnos obtener información faltante como son datos de domicilios o teléfonos, pueden aportarle al SAT información sobre:

Posibles contribuyentes no inscritos.
Signos de actividad económica.
Domicilios alternos para la localización de contribuyentes.
Indicios de posible evasión fiscal.
Información de activos.

El conocimiento de los funcionarios del SAT, unido a la información de fuentes externas, permitirá al SAT mejorar su funcionamiento y alcanzar sus metas.

La información una vez procesada deberá categorizarse y ponerse a disposición de las diversas áreas sustantivas. Esto puede hacerse utilizando un Repositorio Central de Información y generando los modelos de acceso al mismo.

Con al acceso a esta información los funcionarios del SAT podrán ejercer acciones como son:

Inscribir a contribuyentes no inscritos.
Localizar a contribuyentes no localizados.
Apoyarse en la planeación de auditorías.
Ejercer acciones de cobro.

V.2 Creando conocimiento mediante minería de datos.

El SAT es uno de los mejores ejemplos como organización para explorar todo el modelo de gestión propuesto. En este sentido, dada la complejidad de sus procesos y la gran disponibilidad de información que

puede tener, se dan las posibilidades para que esta Institución pueda crear conocimiento y ponerlo en práctica de forma institucional.

Son muy diversos los factores que se reconocen como potenciadores de los riesgos que debe encarar el SAT durante la administración del Registro Federal de Contribuyentes: La complejidad y frecuentes innovaciones en las estructuras de los negocios, nuevos productos financieros, el gran número de contribuyentes que son obligados a pagar diversos tipos de impuestos, el desarrollo del comercio electrónico, etc.

El uso de la tecnología en combinación con la información de la cual puede disponer el SAT, el conocimiento sobre el negocio y el capital humano de la organización le permitirá responder a los retos que se le presenten, siempre y cuando se trabaje hacia la generación de conocimiento y éste sea institucionalizado. Para esto, es necesario eliminar las dependencias personales y funcionales y crear modelos de información mediante los cuales se difunda el conocimiento y se aproveche el mismo mediante su incorporación a los sistemas operacionales.

Así, mediante el empleo de minería de datos el SAT contará con los mecanismos para responder a la dinámica del comportamiento de los contribuyentes y mejorar la calidad y la cobertura del padrón, al contar con modelos de información que le permitan detectar contribuyentes potenciales y el momento en que estos pasan a ser activos; al tener una segmentación de contribuyentes basados en diversas variables, a fin de llevar a cabo estrategias de control y actualización de forma dirigida; al definir funciones de riesgo que le permitan tratar a los contribuyentes de forma diferenciada, facilitando o incorporando controles en los trámites que se realizan.

V.2.1 Marco conceptual.

“Minería de datos” es el proceso de descubrir nuevas y significativas correlaciones, patrones y tendencias analizando gran cantidad de datos –como es el caso del RFC y de las fuentes de información externas – almacenados en repositorios, utilizando tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas (Gartner Group).

Al ser técnicas exploratorias, de entrada no se puede asegurar con cuál de ellas se trabajará y qué modelos descriptivos puedan convertirse a

modelos predictivos. Los productos para realizar Minería de Datos, entre otras emplean las siguientes técnicas:

- Descriptivas
- De clasificación
- De estimación
- Predictivas
- Análisis por grupos de afinidad
- Segmentación

A diferencia del análisis estadístico tradicional, las técnicas de “Minería de Datos” realizan la exploración y el análisis de la información de forma automática y semiautomática a fin de descubrir patrones y reglas significantes sobre la información sin la necesidad de partir de una hipótesis específica, por el contrario, permite descubrir las hipótesis.

Usando las técnicas de “Minería de Datos” es posible analizar cientos de miles de registros de los contribuyentes para identificar atributos comunes y crear perfiles de contribuyentes que representen diferentes tipos de fenómenos.

Estas técnicas, facilitan a las organizaciones liberar el conocimiento sobre sus datos al resto de la organización y que se lleven a la operación los resultados del análisis, así se podrá estar en posibilidad de realizar predicciones sobre el comportamiento de los contribuyentes.

En los aspectos correspondientes al manejo de grandes volúmenes de información y descubrimiento automático o semiautomático de hipótesis, ha radicado el éxito del empleo de éstas técnicas, ya que nos permite modelar diversos aspectos y comportamientos de la información que se está analizando.

Sin embargo, y he aquí la necesidad de que la minería de datos la realice el área sustantiva que es el la que radica el conocimiento de la organización, aunque se abre una gran gama de posibilidades de análisis sobre la información y aparentemente se pueden descubrir hipótesis y modelos de comportamiento de forma semiautomática, se requiere que el “Minero de Datos” tenga un conocimiento importante sobre el negocio, la información que se procesa y el perfil para interpretar los resultados que son en su mayoría de carácter estadístico, y así lograr

realizar aproximaciones sistemáticas a fin de buscar obtener mejores resultados.

No obstante que se pueden obtener una gran cantidad de descriptores del comportamiento de la información, es necesario someterlos a un análisis riguroso de discriminación, ensayos y aproximaciones sucesivas para poder determinar uno o varios modelos que al final describan cabalmente los comportamientos encontrados (modelos descriptivos) y que estos se sometan finalmente a pruebas para determinar si dichos modelos pueden o no predecir las condiciones que se están describiendo; es decir, que exista la factibilidad de implementarse como modelos predictivos.

Para el desarrollo de proyectos de "Data Mining" existe un estándar denominado: "The Cross-Industry Standard Process for Data Mining" (CRISP-DM). Éste es un modelo desarrollado en 1996 por un consorcio de consultores en Data Mining y por especialistas en tecnología de SPSS, Daimler-Benz (DaimlerChrysler) y NCR. Los desarrolladores de CRISP-DM definieron en base a su experiencia un proceso de seis fases que incorporan los objetivos y el conocimiento de la organización (Fig. 45)

Fig. 45: Fases del modelo "The Cross-Industry Standard Process for Data Mining" (CRISP-DM).

Las fases del modelo CRISP-DM

CRISP-DM es considerado *de facto* el estándar de la industria de Minería de Datos.

Entendimiento del negocio:

Esta primer fase, permite asegurarse que todos los participantes entiendan los objetivos del proyecto desde la perspectiva de la organización. Así los objetivos del negocio son incorporados a la definición del problema y en el plan detallado del proyecto. Esto comprenderá el entendimiento de los procesos involucrados, los roles y funciones desarrolladas, la información que es colectada y administrada, y los retos específicos para mejorar le eficiencia de la organización.

Entendimiento de los Datos

La segunda fase está diseñada para acceder a las fuentes de información, a la calidad y a las características de los datos. Esta exploración inicial puede proveer definiciones que ayuden a enfocar el proyecto hacia los resultados esperados. El resultado es un entendimiento detallado de los datos clave que serán utilizados en la construcción de modelos.

Esta fase puede requerir buena parte del tiempo del proyecto, sobre todo en el caso del SAT que tiene una gran diversidad de fuentes de información tanto internas como externas. Sin embargo, es un proceso crítico para el proyecto.

Preparación de los Datos

La siguiente fase de la metodología involucra el poner a disposición los datos en un formato adecuado para la construcción de modelos. El analista usa los objetivos del negocio determinados en la etapa de entendimiento del negocio para determinar cuál tipo de datos y de algoritmos se emplearán. En esta etapa también se resuelven los problemas no resueltos en la fase de entendimientos de los datos, tales como datos faltantes.

Modelado

La fase de modelado involucra la creación de los algoritmos que descubren el conocimiento en los datos. Hay una gran variedad de técnicas, ya descritas con anterioridad. Cada técnica requiere un tipo específico de datos. La cual puede requerir el regreso a la fase de preparación de datos.

Esta fase producirá un modelo o un conjunto de modelos que contendrán en si mismos el conocimiento descubierto en un formato apropiado para su análisis.

Evaluación

Esta fase se enfoca en evaluar la calidad del modelo o modelos obtenidos. Los algoritmos empleados pueden llevar a descubrir un número ilimitado de patrones; sin embargo, muchos de estos no tendrán sentido.

Esta fase ayuda a determinar cuáles de los modelos son útiles en términos de alcanzar los objetivos del proyecto y llevarlos a los procesos de toma de decisiones del día a día.

Implementación

Dependiendo de la significancia de los resultados, se puede requerir sólo modificaciones menores o incluso podría necesitarse un cambio mayor de reingeniería de procesos de los sistemas de soporte a la toma de decisiones. Esta fase también involucra la creación de procesos repetibles para enriquecer o recalibrar el modelo. Por ejemplo, las leyes fiscales cambian constantemente, por lo que se requiere de un proceso estandarizado para actualizar los modelos de acuerdo al desarrollo de nuevos resultados.

La presentación adecuada de los resultados asegura que los tomadores de decisiones le den uso a la información. Esto puede ser tan simple como crear un reporte o tan complejo que un solo proceso involucre a toda la organización. Por lo que es necesario tener en cuenta que podrían requerirse cambios o nuevas funcionalidades en los procesos actuales de identificación del contribuyente, o bien podrían tomarse acciones sólo con los reportes de los resultados.

V.2.2 Resultados esperados.

Con “minería de datos” se pretende encontrar asociaciones, patrones o tendencias entre las variables propias del RFC, el histórico de movimientos de los contribuyentes, la información de control de obligaciones, declaraciones, pagos y créditos, así como la información de

fuentes externas para tratar de determinar modelos que describan y, en su caso, prevean el comportamiento de los contribuyentes.

El alcance de la minería de datos aplicada al SAT, específicamente en el tema del Registro Federal de Contribuyentes, permitirá descubrir si existen correlaciones entre variables sobre temas que pueden abarcar los siguientes aspectos:

I. Modelos para descubrir tendencias de cambios de domicilios y de no localización de contribuyentes.

II. Segmentación de contribuyentes con base en signos de riqueza.

III. Modelo para detectar contribuyentes potenciales no inscritos en el RFC.

IV. Segmentación de los contribuyentes cumplidos y de los incumplidos.

V. Probabilidad de cobro de créditos con base en signos de riqueza.

VI. Probabilidad de cobro con base en el comportamiento del contribuyente en el padrón de RFC, control de obligaciones y situación en fuentes de información externa.

VII. Modelos de riesgo para determinar contribuyentes a verificar su domicilio durante el proceso de inscripción.

VIII. Relación entre el comportamiento de los socios y accionistas y el comportamiento de la persona moral en los temas de localización y cumplimiento de obligaciones del RFC y pago de créditos.

IX. Asociaciones entre los procesos de notificación, las características de los contribuyentes y la no localización de los mismos (riesgo del notificador).

X. Asociaciones entre el tiempo transcurrido de la omisión de una obligación, el acto de notificar el requerimiento y la respuesta del contribuyente a los mismos.

Es importante mencionar que los modelos descriptivos aportarán información importante para el conocimiento del contribuyente, que en sí mismo es de gran valía para la Institución, pero más importante aún es el identificar o crear modelos que prevean el comportamiento, mismo

que de existir deberán se implementados tecnológicamente para llevarlos a la operación diaria del SAT.

Se realizaran pruebas desde dos puntos de vista: Partiendo de hipótesis que requieran ser comprobadas y descubriendo el conocimiento mediante las técnicas de minería de datos.

Mediante las pruebas de hipótesis se partirá de ideas preconcebidas sobre la relación existente entre datos, por citar algunos ejemplos:

1. El monto de pago de impuestos está asociado al ingreso de los contribuyentes, y éste a su vez al número y valor de sus bienes.

2. Los contribuyentes no se localizan debido a que llevan a cabo prácticas de cambio de domicilios. Se piensa que a mayor número de domicilios identificados del contribuyente, es mayor la probabilidad de cambiarse de domicilio.

Las técnicas de búsqueda de conocimiento podrán apoyarnos en descubrir posibles asociaciones o patrones de comportamiento en:

1. Segmentos de contribuyentes que cumplen sus obligaciones.
2. Patrones de comportamiento de cambios de domicilios.
3. Patrones de pagos para evitar ser sancionados o bien auditados.
4. Patrones o tendencias de cambios de domicilios y de no localización de contribuyentes.
5. Categorización de contribuyentes con base en signos de riqueza.
6. Detectar contribuyentes potenciales no inscritos en el RFC.

Para esto, se prevé utilizar la información de fuentes internas y externas para descubrir patrones de comportamiento de interés; sin embargo, como se mencionó, esto no es suficiente. Una vez que se encontraron patrones de comportamiento, debemos caminar hacia la construcción del modelo de información para el conocimiento del contribuyente. Los siguientes pasos a dar, serán habilitar la información obtenida y crear conocimiento a fin de caminar hacia un modelo de conocimiento del contribuyente.

Es necesario que los modelos descriptivos que se determinen se lleven al nivel de la operación y que los patrones, asociaciones y en general modelos predictivos se lleven al nivel de la administración del RFC a fin

de crear conocimiento y que se tomen las acciones necesarias para actuar en consecuencia.

Existirán dos niveles de acción, los que vayan en concordancia con la estrategia del SAT, por ejemplo disminuir la no localización que deberán implementarse al nivel de la administración del RFC, y aquellos producto de las acciones de minería de datos y que tienen que ver con la acción de descubrimiento de conocimiento y que los resultados deberán llevarse al nivel de los modelos estratégicos del SAT.

Como se comentó anteriormente, se trata de descubrir conocimiento para que posteriormente éste sea utilizado, ya sea de forma directa o mediante su implementación tecnológica para definir estrategias y acciones a llevar a cabo a fin de eficientar la Administración Tributaria.
La minería de datos, permitirá al SAT mejorar su operación en los siguientes procesos:

Segmentación:

Además de la segmentación natural del padrón por ejemplo, régimen, obligaciones, etc. se podrá contar con una segmentación por tamaño del contribuyente, signos de riqueza, tipificarlos por su comportamiento en el cumplimiento de obligaciones, de patrones de localización o cambios de domicilio y probabilidad de localización, entre otros.

Riesgo:

Se contará con información para determinar cuando verificar o no el domicilio de un contribuyente. Se intentará generar un modelo para descubrir tendencias de cambio de domicilios, así como asociaciones entre el comportamiento de los contribuyentes y la de los notificadores. Así como las posibles relaciones de cumplimiento o incumplimiento entre socios y accionistas.

Incorporación y Enriquecimiento:

Se contará con un modelo para la identificación de contribuyentes potenciales y su incorporación al Padrón. Así mismo, se podrá identificar la presencia de nuevas obligaciones de los contribuyentes, así como la identificación de posibles cambios de domicilios.

Control:

Se podrán definir los criterios disparadores que hagan de un contribuyente potencial un contribuyente activo.

Además para el área de cobranza se podrán definir parámetros que les permita categorizar a los contribuyentes en función de la probabilidad de recuperación de un crédito, en función de la información contenida en fuentes externas.

Dentro de los beneficios esperados se encuentran: Mejorar y mantener la calidad del padrón de contribuyentes, detectar prácticas de cambios de domicilio, apoyar al abatimiento de la no localización, contar con información para determinar cuando un contribuyente deja de ser potencial y se convierte en activo, incorporar contribuyentes potenciales y no inscritos, y mejorar la efectividad de la cobranza al contar con información de probabilidad de cobro de créditos, factibilidad de localizar un domicilio y determinación de signos de riqueza.

VI. Requerimientos de software.

Los procesos descritos anteriormente fueron realizados principalmente con dos productos especializados: TRILLIUM y Omikron Address Center, este último también conocido como.

Estos dos productos están licenciados al SAT y fueron utilizados durante mi estancia como trabajador de esa institución para la elaboración del presente trabajo de Tesis. Si bien en el mercado existen otras soluciones, estas herramientas estaban a disposición.

Por lo anterior, sólo se realiza un análisis comparativo entre ambas a fin de explorar un poco sobre las características de cada una de ellas.

TWINFINDER

Software para encontrar duplicidades en Bases de Datos, verificando errores tipográficos, similitudes fonéticas abreviaciones, etc.

La búsqueda que efectúa es mediante SIMILITUDES, Twinfinder no busca registros idénticos sino similares, emplea para ello, la técnica de comparación de similitud fragmentaria que es un algoritmo lingüístico-matemático que imita la sensibilidad humana para captar similitudes.

Existen 4 Variaciones: R, L, Duo y Multi. Las funciones de cada uno se describen a continuación:

Twinfinder R: Analiza sus datos para encontrar duplicados y después suministra un listado en forma de texto. No es posible borrar automáticamente.

Twinfinder L: Permite obtener un listado de duplicados, puede borrar automáticamente los duplicados que encuentre.

Twinfinder Duo: Permite combinar dos bases de datos sin tener que reunirlas previamente. También permite trabajar con una sola base de datos, al igual que la versión L.

Twinfinder Multi: Realiza combinación de datos de diferentes fuentes, realizando la búsqueda con diferentes grados de prioridad.

CARACTERÍSTICAS (TWINFINDER MULTI)

Separación de la calle y el número	Tiene la posibilidad de separar el nombre de la calle y el número, para elevar la exactitud de la separación y con ello la calidad en el MATCH.
Ajuste del entorno del índice	El entorno del índice sirve para acelerar el MATCH, ya que permite comparar datos dentro de n entorno, sin necesidad de recorrer toda la base de datos.
Fusión de varios campos	Se utiliza para acelerar el MATCH, mediante una búsqueda abreviada mediante la unión de campos.
Conectividad con distintas bases de datos	Conexión con las principales bases de datos como Oracle, Informix, SQL Server, MS-Access, etc. Además permite conectividad mediante ODBC para los controladores no integrados en el software. Soporta hasta 999 Bases de Datos entrantes.
Asistente	Cuenta con un asistente que facilita tareas similares, básicamente es la creación de macros o “templates” que facilitan actividades futuras.

TRILLIUM 7

Software de higiene y estandarización de bases de datos, básicamente se utiliza para la geo-referenciación de datos, también permite la localización de registros similares.

CARACTERÍSTICAS

Sistemas Operativos	Debido a que esta aplicación está desarrollada en Java, soporta distintas plataformas como: Windows 9x, Windows NT, 2000 y XP, distintas variaciones de UNIX como Solaris, Linux, BSD, etc.
Archivos Planos	El manejo de la información en Trillium debe ser obligadamente en archivos planos, por esta razón, cada proyecto consume demasiado espacio de almacenamiento.
Soporta	La información de entrada para Trillium puede ser de

Distintas Fuentes	fuentes distintas, y con los procesos que posee se puede estandarizar para dar el mismo formato a todas las fuentes de datos.
Procesos Batch	La individualidad que posee cada módulo permite la generación de procesos por lotes que facilita la labor de generar "plantillas prediseñadas" que facilitan tareas futuras similares con modificaciones mínimas o nulas en las propiedades de los módulos.
API's para distintos lenguajes de programación	Ofrece extensión de API's para lenguajes como: C, JAVA, Visual Basic.
Completamente parametrizado	Es una herramienta con gran cantidad de parámetros en cada módulo que pueden ser variados para mejorar las concordancias (similitudes) o variar el resultado de acuerdo con los requerimientos solicitados por el usuario.
Adecuaciones para México	La versión 7 de Trillium que actualmente se maneja tiene modificaciones para mejorar las búsquedas y las concordancias en México, tiene adecuaciones en rutinas para calles o códigos postales, y para algunos módulos, pero no en su totalidad.
Módulos Independientes	A continuación, se mencionan los módulos de uso más frecuente para un proyecto estándar de limpieza en Trillium: • **Converter:** Búsqueda de datos, recodificación de información, formato de datos. • **Data Parser:** Identifica nombres para individuos, grupos de individuos, empresas y los agrupa para tener más y mejores concordancias al momento de ejecutar el Matcher. • **Geocoder:** Identifica nombres para calles, colonias, códigos postales y los agrupa para tener más y mejores concordancias al momento de ejecutar el Matcher. • **Window Key Generator:** Generador de claves definidas por el usuario para mejorar el proceso de identificación y concordancias. • **Matcher:** Con base a las claves generadas por el Window Key Generator, se procede al proceso de concordancias.

Tabla comparativa de características principales de TRILLIUM 7 y TWINFINDER

CARACTERÍS-TICA	TRILLIUM	VENTAJAS	DESVENTAJAS	TWIN FINDER	VENTAJAS	DESVENTAJAS
Soporta distintos OS's	SI	- Funciona en prácticamente cualquier plataforma	- Instalaciones complejas - Interfaz de Usuario distinta a la tradicional	NO	- Funciona en la mayoría de las versiones de Windows -Instalación sencilla - Tiene la típica Interfaz de Usuario	- No funciona en ambientes UNIX
Archivos Planos	SI	- Gracias a esta característica, soporta cualquier Base de Datos	- Consume demasiado espacio de almacenamiento	SI	- No se basa únicamente en archivos planos	- Ninguna
Soporta distintas fuentes	SI	- La entrada en fuentes puede ser de 10 archivos planos, y la salida resultante puede ser una entrante para una nueva fuente - El archivo de salida toma un formato estándar sin importar de qué fuente provino	- Debido a la cantidad de parámetros de entrada, el proceso de inicialización se vuelve tardado y complicado para usuarios nuevos	SI	- Soporta hasta 999 fuentes distintas	- La estructura de las distintas fuentes debe ser muy similar o incluso idéntica
API's para programación	SI	- Se pueden desarrollar aplicaciones externas haciendo uso de las potentes herramientas de Trillium	- Inversión de mucho tiempo para parametrizar y definir la estructura en los módulos de Trillium	NO	- Es de fácil manejo, y su uso particular lo hace muy atractivo para usuarios finales	- Limitado en su totalidad al fabricante

Adecuaciones para México	SI	- En algunos de los módulos y rutinas hay adecuaciones para México que mejoran las coincidencias	- No todos los módulos ni rutinas se encuentran adaptadas	NO	- Pese a no tener adecuaciones el algoritmo de búsqueda que tiene genera porcentajes de coincidencia muy buenos	- Ninguna
Módulos Independientes	SI	- El resultado generado por cada módulo puede ser utilizado por el siguiente, y no hay necesidad de ejecutar todo el proceso si se desean hacer modificaciones en algún paso en particular	- El número de parámetros a manipular aumenta y se vuelve más complejo todo el proceso	NO	- El proceso de iniciación de parámetros se vuelve más rápido	- El proceso es único y los parámetros son generales por lo que no se puede particularizar en procesos distintos
Separación de campos	SI	- El resultado de separación es muy limpio - Separa cualquier tipo de campo	- Es un proceso muy tardado que requiere de procesos previos	SI	- Es un proceso sencillo	- Sólo separa calle y número - No es un proceso muy limpio, puede contener coincidencias falsas
Conectividad con Bases de Datos	NO	- Sólo maneja archivos planos.	- Procesos de preparación de archivos planos previos debido a que no hay conectividad con ninguna base de datos	SI	- Facilita el proceso de entrada de las fuentes, y pueden ser distintas	- Ninguna
Asistente	NO	- Es para usuarios experimentados.	- No cuenta con asistentes para guía o ayuda para el usuario	SI	- Ayuda al usuario abreviando algunas tareas repetitivas	- Ninguna

Fusión de Campos	SI	- Es un proceso rápido	- Manejo de muchos parámetros	SI	- Es un proceso sencillo	- Ninguna
Manejo de Índices	NO	- Tiene un equivalente a los índices que se llaman ventanas, y agiliza el proceso de búsqueda y coincidencias	- Debido a que no maneja índices algunos procesos son muy lentos, sobre todo en tareas repetitivas	SI	- Agiliza los procesos de búsquedas y coincidencias en tareas repetitivas	- El primer índice requiere mucho tiempo para su construcción
Tiempos de Ejecución	-	Ninguna	Requiere mucho tiempo en sus procesos Depende de la plataforma (OS, procesador, memoria y espacio en disco) sistema	-	Son proporcionales al número de registros y fuentes, en general se percibe my rápido	Ninguna
Espacio de Almacenamiento en Disco	-	Agiliza los procesos	El resultado final aumenta en tamaño exponencialmente	-	Los tamaños finales no aumentan tanto con respecto a las fuentes	Ninguna

RESUMEN DE LA COMPARACIÓN DE TRILLIUM 7 Y TWINFINDER

Debido al manejo masivo de información y las diferencias existentes entre las distintas fuentes de datos, los procesos para *higienizar*, *estandarizar* y *cruzar* información, no pueden ser genéricos, pero el desarrollo de un proceso o de varios para que se adecuen a la mayor cantidad posible de fuentes si puede ser elaborado.

En el mercado existen pocas herramientas para actividades tan especializadas como éstas, es por ello que sólo se evaluaron dos de las herramientas con más renombre, y después de haberlas evaluado, lo ideal para el desarrollo de un proceso general de *higiene, estandarización* y *cruces* de información es:

1. El uso de **Trillium 7** para higienizar y estandarizar.

2. El uso de **TwinFinder** para la búsqueda de duplicados y enriquecimiento de la información otras fuentes de datos.

Como un solo software no cubre con las necesidades del desarrollo de un proceso genérico para búsqueda de información en fuentes alternas se requiere de ambas herramientas. Las razones son las siguientes:

- Los procesos de *cruce* con **Trillium 7** son excesivamente lentos.

- **Trillium 7** es demasiado complejo incluso para usuarios experimentados.

- Las adecuaciones para México en **Trillium 7** no se encuentran actualizadas, por lo que no son confiables.

- **Trillium 7** tiene procesos de *higiene* y *estandarización* que se pueden adaptar a las necesidades de la institución.

- **TwinFinder** realiza procesos de *cruces* de manera muy precisa y con tiempos de ejecución razonables.

VII. Hacia un enfoque estratégico de la gestión de información en las instituciones.

Revisemos nuevamente el modelo de gestión de la información planteado (Fig. 44).

Fig. 44 Modelo de Gestión de la Información

¿Cuáles de estos componentes podríamos encontrar en cualquier institución que lleve un sistema de registro? Seguro se tiene un mecanismo de adquisición de la información y tendrán también diseñados algunos sistemas para hacer uso de la misma (modelos de información). Quizá alguna institución tenga procesos de mejora continua, aunque generalmente están asociados a procesos productivos y no a la calidad de la información.

Los mecanismos de adquisición de la información por lo general tienen una serie de validaciones de los datos, pero muy orientados hacia el tipo de dato, el rango y valores que estos pueden tener y se programan algunas reglas de integridad de la información en la base de datos y otras de consistencia entre datos.

En el mejor de los casos encontraremos que la información que se capta es perfectamente validad contra las reglas del negocio que se definen;

no obstante, no existe una estrategia para captar la información y para mejorar **la calidad y la cobertura** de la información, acciones estrechamente relacionadas con las dimensiones de la calidad ***Totalidad*** y ***Exactitud***.

Es importante ir a conseguir el principal activo de la institución: la información. Si queremos tener mejores datos y completos es necesario el diseño de una estrategia para que la institución se haga de ellos. Es importante dejar de ser espectadores y de permanecer pacientemente esperando a que la información se actualice.

Existen varias estrategias para obtener la información, tradicionalmente la adquisición de la información se realiza a través de:

1) Diseño de sistemas para captar la información de los clientes conforme se presenten ante la institución.
2) Adquisición de la información de otros sistemas registrales.
3) Levantamiento de la información en campo, como son las estrategias censales.
4) Contratación o convenio con terceros para que estos capten información relevante para la institución, e incluso la totalidad de ésta.

Estos mecanismos deben estar acompañados de las estrategias (Marketing) para captar a los clientes y promover su constante actualización, por ejemplo:

1) Campañas en medios de comunicación como radio, TV, publicaciones escritas, Internet, por ejemplo las campañas del IFE, para su registro de electores.
2) Establecer en Ley la obligación de incorporarse a algún sistema de Registro, como el caso del SAT en su registro de contribuyentes.
3) Otorgar servicios, como el caso del IMSS, en su sistema de registro de derechohabientes.

No hay que dejar de lado que existen otros mecanismos para obtener información que están asociados, más que a captar la información, a la producción y transformación de datos.

El problema de ***Tranportabilidad*** de la información obtenida de otras fuentes se resuelve fácilmente al aplicarle a la información los procesos de higiene y estandarización de datos, ya que esto permitirá que la información se sujete a las mismas reglas del negocio y en

consecuencia, se facilite la comparación mediante herramientas de cruce de base de datos.

Resuelto este problema, es importante que las instituciones establezcan convenios de colaboración con otros proveedores de datos para llevar a cabo un intercambio de información. El crear catálogos “puente” de claves de acceso a los datos ya identificados permitirá facilitar el intercambio subsecuente de datos.

Como estrategia, se recomienda se homologuen los catálogos que utilizan los sistemas de los principales socios de la información, quizá en este sentido los más importantes y que pueden presentar mayores diferencias son aquellos asociados a los domicilios: municipio, localidad, colonia, calles y CP´s. Esto facilitará la ***Sistematización del intercambio de la información*** y podrán definirse los medios tecnológicos para llevar a cabo el intercambio, que puede ir desde una simple transferencia de archivos hasta la conexión de las redes para tener acceso en tiempo real a la información, creando repositorios de información o creando servicios para acceso a los sistemas.

En los capítulos anteriores se planteó ampliamente la forma de llegar hacia un modelo de gestión del conocimiento basado en la colección de datos de otras fuentes para enriquecer la información y mediante el uso de técnicas de minería de datos, que por un lado permiten descubrir hipótesis sobre el comportamiento de los individuos en los sistemas de registro, y por otra parte es una herramienta muy útil para que mediante el uso de las diferentes técnicas de minería de datos en conjunción con la experiencia y conocimiento del personal de una institución, se puedan generar modelos que al programarse en un sistema garantice que el conocimiento generado en la institución sobre los sistemas de registro se logre conservar y así eliminar las dependencias funcionales y personales que en muchas ocasiones se tiene para el uso de la información.

Por otra parte, también mostramos la existencia de herramientas para llevar a cabo las actividades de gestión de la calidad de la información y aquellas que nos permiten explorar modelos de información para crear conocimiento.

Así el desarrollo de estos modelos, que denominamos modelos descriptivos y sus correspondientes modelos predictivos fortalecerán ampliamente el desarrollo de modelos de información.

Estrategia de Gestión de la Información

La Gestión de la información en los sistemas de registro es una actividad que no solo corresponde a las áreas de tecnología; si bien tienen una gran participación, la gestión de la información es un tema que debe preocupar a la alta gerencia y debe ser implementada desde la cabeza de la institución. Si no se asume así, el modelo puede operar, pero no se alcanzarán a plenitud los objetivos.

Sólo desde la Dirección de la Institución podrán definirse las estrategias para captar datos, podrán identificarse los impactos económicos y de operación en los procesos, podrán definir cuándo es necesario llevar a cabo campañas de actualización de los datos y establecer acuerdos con otras instituciones para compartir información. Así también es de suma importancia la creación de área que lleva a cabo los procesos tecnológicos asociados a la calida de la información.

Enseguida se propone una estrategia general para la gestión de la información en las instituciones, misma que debe acompañar la operación de los sistemas de registro:

1. Definición clara de los objetivos y alcances del sistema de registro.

   Deberán describirse claramente los objetivos y alcances del sistema de registro, esto permitirá identificar cuál es la información más importante para la Institución.

   Seguramente, como en el caso del SAT, se encontrará que los datos de identificación de las personas y los de localización de la misma son de suma importancia, y a estos se asocia la información propia de cada modelo de registro.

2. Identificación de actores participantes y principales usuarios y crear mecanismos de comunicación.

   Se requiere identificar perfectamente quienes son los usuarios de los sistemas de registro e identificar qué información es la más valiosa para ellos y cuál es el rigen de la misma.

   Así, deberá crearse una comunicación constante con los todos los usuarios de la información para poder identificar áreas de oportunidad para mejorar la calidad y la cobertura de la misma.

Es necesario tener en cuenta que los usuarios acostumbran resolver los problemas que se encuentran en los sistemas y en la propia información, e incluso desarrollan sus propias redes de comunicación para lograr este objetivo, por lo que es muy frecuente que las áreas de sistemas no se enteren de los problemas que los usuarios enfrentan y acaban siendo normales; sin embargo, esto sólo acarrea costos a las instituciones.

Una estrategia interesante y poco practicada es el enviar al personal de sistemas y los responsables de la calidad de los datos a operar los sistemas junto con los principales usuarios.

3. Crear una estrategia para captar la información.

   Este tema ha sido ampliamente tratado a lo largo de este trabajo, es claro que la información debe actualizarse permanentemente, para lo cual es necesario implementar estrategias de actualización permanente de la información.

   No basta con esperar a que la información se actualice por “acercamiento” de las personas que forman parte del sistema de registro, es necesario captar la información por todos los medios posibles, ya sea utilizando todos los puntos de contacto o acercamiento con las personas, creando convenios para intercambio de información, diseñando campañas de marketing o de actualización de la información en campo.

4. Crear el área dedicada a la calidad de la información.

   Será la responsable de la implementación del modelo de gestión de calidad de la información. Se sugiere que sea parte de las áreas de Data Warehouse, ya que estás son las que conocen los datos de toda la institución y por lo general, las herramientas que utilizan en los procesos de extracción, transformación y carga de datos, incorporan utilerías para la limpieza y estandarización de datos.

   Exista o no el área de Data Warehouse, deberá considerarse la realización de procesos de calidad sobre los datos con el empleo de software especializado o bien contratando servicios de outsourcing.

5. Constituir a la CURP como una clave de referencia para la Identificación de personas.

   Es necesario contar con una clave de uso general para la identificación de las personas, en el caso del registro de personas la CURP es una de ellas y se puede someter a validación en el RENAPO para la identificación de inconsistencias en la conformación de la clave. Un mecanismo simple, será el proporcionarles una copia de la base de datos integrada con la información que RENAPO considere indispensable para su validación.

   No obstante, que RENAPO participe en la validación de la base de datos, es recomendable considerar su participación durante la operación del Sistema de Registro. En este sentido, deberán definirse los mecanismos para tener una validación constante de la información que se incorpore o actualice durante la operación del Registro. Como ejemplo se puede citar el caso del SAT, quien con el fin de tener una adecuada identificación de las personas, tiene comunicación con los sistemas de RENAPO y se valida y comparte la información en línea.

6. Sistema de Información Geográfica.

   Aunque no fue objetivo de esta tesis el desarrollo de los modelos de registro que son explotados a través de sistemas de información geográfica, es posible que para enriquecer la información o mejorar el conocimiento que sobre los miembros de un sistema de registro se tiene, sea necesario hacer uso de los productos estadísticos y demográficos que genera tanto el INEGI, como otras instituciones con programas sociales. Por ejemplo, la calificación de una localidad como urbana o rural, la asignación de los índices de marginación de CONAPO o la densidad de población de una zona geográfica, es información de utilidad para analizar la información y generar conocimiento.

7. Incorporar en las áreas estratégicas mecanismos para “institucionalizar” el conocimiento y generar los modelos de información.

Muy diversos pueden ser los mecanismos de generación del conocimiento, pero lo más importante es que éste se lleve como algoritmos a los sistemas de información. Ya sea que se parta de la experiencia del personal, que se determinen reglas de asociación o correlaciones entre datos mediante análisis estadístico, o bien, que se haga uso de la minería de datos, como se describió en el capítulo correspondiente, lo importante es que el conocimiento generado se incorpore como reglas de operación en los sistemas de la institución, con lo cual se logrará "institucionalizar el conocimiento".

8. Indicadores de calidad y mejora continua

Es importante que la institución integre a sus sistemas de seguimiento de objetivos y metas los indicadores correspondientes a la calidad de la información.

En el caso del SAT, por ejemplo, se definió al inicio de la tesis un indicador compuesto de la calidad del RFC:

Índice de calidad del RFC = (Exactitud)(Totalidad)

Exactitud = Calidad = 60.5%

Totalidad = Cobertura =56.5%

Índice de calidad del RFC = (60.5%)(56.5%)

Índice de calidad del RFC = 34.18%

Indicadores de este tipo permitirán que la alta gerencia tenga plena referencia de la situación que guarda el sistema de registro y en consecuencia emprender acciones de mejora continua.

La estrategia aquí sugerida es un buen comienzo para desarrollar el modelo de Gestión de la Información en los sistemas de registro de las instituciones; no obstante, se sugiere la contratación de consultores expertos en el tema para que ésta sea implementada de la mejor forma y complementada con las mejores prácticas de la industria de la calidad y gestión de la información.

Conclusiones.

La información y el conocimiento son los principales activos de una institución, lo que justifica la inversión de tiempo, dinero y recursos humanos y tecnológicos para cuidar de estos activos.

El no cuidar de éstos activos puede ocasionar que la institución incurra en costos fijos de operación que no son controlables y que en muchas ocasiones los Directivos no se percatan que estos existen.

La estrategia definida para el modelo de gestión de la información partió del análisis de la situación que guarda el Registro Federal de Contribuyentes; no obstante, el trabajo aquí presentado sirve de base para que otro tipo de instituciones reflexionen sobre la calidad de su información y traten de medir los costos inherentes a la misma.

A lo largo del trabajo mostramos la existencia de muchos sistemas de registro, me pregunto si los responsables de estos tienen en mente cuál es la calidad de su información y cuánto les está costando operativamente. Me parece que difícilmente encontraremos una institución que tenga en su “War Room” un indicador de la calidad de su información.

Desde mi opinión, el gran reto que enfrentan los sistemas de registro tiene que ver con identificar esos costos asociados a la calidad de la información y preocuparse por mediar la misma y en consecuencia ejercer las acciones necesarias para su mejora.

El modelo de gestión que se propone muestra que se resuelve gran parte de los problemas inherentes a la calidad de los datos, Los resultados obtenidos son muy satisfactorios, ya que cerca del 80% de los datos se pueden resolver sin gran esfuerzo con las herramientas utilizadas para ejemplificar los argumentos que se expresaron (SQL, Trillium y TwinFinder).

El resto de los problemas requiere la intervención de un grupo de personas que apoyen en la definición de reglas del negocio y la conformación de los catálogos necesarios e incluso realizar “a mano” los procesos de calidad. Esto; sin embargo, es parte del modelo de gestión propuesto ya que debe existir un área que específicamente se dedique a cuidar la calidad de la información.

Los esquemas mostrados para medir la calidad de los datos nos dan una gran idea de la calidad que guarda la información y pueden ser fácilmente identificadas las deficiencias mediante el simple uso de comandos SQL, mediante estos se demostró que no sólo es necesario contar con la mejor herramienta tecnológica o la más costosa; lo importante es medir la calidad de los datos. También mostramos cómo puede medirse utilizando herramientas de “Data Profiling”, muestreo de datos y hasta validación en campo.

Con el modelo propuesto quedó demostrado que la problemática planteada se resuelve ampliamente. Se definieron esquemas para cuidar la calidad de los datos y se propuso un mecanismo para lograr tener la totalidad de la información (Cobertura). En el esquema propuesto, el acceso a fuentes de información externas a la institución es clave en el desarrollo de los procesos.

Con los procesos propuestos se mostró como resolver el problema de la transportabilidad de los datos al no existir una clave en común: los procesos de higiene, estandarización y desduplicación o cruce de bases de datos, aunados a las herramientas necesarias para tal fin, permiten resolver ampliamente el problema.

La implementación del modelo se inició en el SAT con éxito y se logró que esta Institución diera la importancia necesaria a los indicadores de Calidad y Cobertura del Registro Federal de Contribuyentes, se consideró tan importante que el año del 2001 se crea la Administración Central del RFC, encargada de llevar a cabo la primer medición de a calidad de la información y en consecuencia la gestión de la información del sistema de registro.

Un tema importante para el SAT, lo constituye el “Conocimiento del Contribuyente”, el tratar de predecir comportamientos, crear una segmentación de los mismos para poderos comparar y poder utilizar diversas fuentes de información como indicadores socioeconómicos y establecer diversa relaciones, permiten al SAT ser más eficiente en sus acciones de fiscalización y cobranza, e incluso en los servicios que presta al contribuyentes. Esta actividad también es muy exitosa si se utilizan los mecanismos aquí presentados para estandarizar y cruzar información, ya que ésta es la base para poder conjuntar datos y poderlos someter a los mecanismos de análisis utilizando software de minería de datos.

Con los mecanismos presentados para mejorar la calidad de los datos, en consecuencia se mejora la información y el conocimiento que se

genera a partir de la misma. Así la información se puede utilizar para descubrir asociaciones, tendencias y patrones de comportamiento de los contribuyentes, con lo cual se puede generar conocimiento que posiblemente no sabíamos que podría obtenerse.

Es importante mencionar que para los fines de este trabajo de tesis no se consideró el tema de desarrollo de sistemas, estructuración de bases de datos, comunicaciones y la tecnología asociada a la operación de los sistemas de registro; es claro que esto y los temas que tienen que ver con la construcción de un marco normativo, el diseño de los procesos, la articulación de los recursos y el propio diseño del modelo de negocio son parte del modelo de gestión de un sistema de registro; sin embargo, se parte de la hipótesis de que el sistema de registro ya existe y opera utilizando estos componentes, y el planteamiento que se realiza es que en general faltan los componentes para cuidar las calidad de la información y para la creación de conocimiento.

Por lo anterior, considero que incorporando a las modelos de registro actualmente existentes los planteamientos realizados en este trabajo, principalmente en los registros de la administración pública, se tendría completo el modelo de gestión de la información de los sistemas registrales y en consecuencia se mejoraría la eficiencia y eficacia en la operación de los mismos.

Bibliografía.

"La Gestión de la Calidad en la Administración Tributaria", Asociación Española de Normalización y Certificación.

"Gestión del conocimiento y calidad total," Carlos A. Benavides Velasco, Cristina Quintana García.

"8th International Conference on Information Quality: A Framework for Assessing Data Quality – from a Business Perspective" Pia Gustafsson, Åsa Lindström, Cecilia Jägerlind, Jevgenij Tsoi; Dep. of Industrial information and Control Systems Royal Institute of Technology / KTH, SE-100 44 Stockholm, Sweden.

"8th International Conference on Information Quality: Dealing With data quality", John rome & Susane Moore, Arizona State University.

"8th International Conference on Information Quality: HOW'S YOUR DATA QUALITY?" A CASE STUDY IN CORPORATE DATA QUALITY STRATEGY", Traci Campbell, Zack Wilhoit , Acxiom Corporation.

"8th International Conference on Information Quality: PANEL 1: FEDERAL DATA QUALITY ISSUES INFORMATIONMANAGEMENT AND DATA QUALITY", Position Statement, Context, Issues, Concepts, Strategy and Lessons Learned", Bryan Aucoin

"8th International Conference on Information Quality: PRESERVINGWEB SITES: A DATAQUALITYAPPROACH", Cinzia Cappiello, Chiara Francalanci, Barbara Pernici, Politecnico di Milano, Milano, Italy.

Wang, R.Y. and Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12.

Kahn, B.K., Strong, D.M. and Wang, R.Y. Information Quality Benchmarks: Productand Service Performance Communications of the ACM, 2002.

Hodge, G.M. Best practises for digital archiving, D-Lib, January 2000.

Sistema de Clasificación Industrial de America del Norte, INEGI, 2002.

Clasificación Mexicana de Actividades y Productos, INEGI.

MONTUSCHI, L. “La economía basada en el Conocimiento: importancia del conocimiento tácito y del Conocimiento Codificado”, CEMA, Buenos Aires, 2000.

Martha Beatriz Peluffo A. y Edith Catalán Contreras, “Introducción a la gestión del conocimiento y su aplicación al sector público”, CEPAL, 2002.

“Sistema de Información y sistema de calidad: relación y dependencia en las organizaciones empresariales”, Alicia Arias Coello, Isabel, Portela Fielgueras, EUBD, Universidad Complutense.

http://sepiensa.org.mx/contenidos/historia_mundo/antigua/fenicia/alfabeto/alfabeto.htm.

http://sepiensa.org.mx/contenidos/h_mexicanas/colonia/jesuitas_edu/jesuitasedu_1.html.

Manual en Línea Omikron Adress Center. http://www.omikron.com