"Una Metodología para Medir Calidad de Datos"

Tesis de Licenciatura

Tesista: Silvia N. Morillaz
Directora: Dra. Martina Marré

Junio de 1999

Abstract:

La toma de decisiones de las organizaciones se basa en la información que poseen. Esto marca la importancia de los datos informáticos. Tener buenos datos puede significar una ventaja competitiva importante para la organización. El reconocimiento de estas razones por el mundo de los negocios han provocado un aumento en la importancia brindada al tema Calidad de Datos.

Ante la importancia adquirida por el tema Calidad de Datos Informáticos surge la inquietud de que la Ingeniería del Software incorpore este tema dentro de sus áreas de interés. Para que esto sea posible es necesario poseer conceptos y metodologías estándares que permitan a los profesionales de esta materia hablar un lenguaje común que facilite el intercambio de experiencias y conocimientos que vayan adquiriendo.

Cada organización necesita conocer cuál es el grado de calidad que tienen sus datos. Tiene que saber si la toma de decisiones se alimenta de datos confiables o no. También tiene que poder decidir si necesita invertir dinero en mejorar sus datos o no. Para realizar este análisis se necesita Medir la calidad de los datos.

En particular, en esta tesis se trabaja sobre el tema Métricas para medir Calidad de Datos. El objetivo es lograr un conjunto estándar de métricas y una metodología de trabajo a aplicar en proyectos de medición de calidad de datos. Aplicar estas definiciones en un experimento concreto que permita evaluar y mejorar esta propuesta de acuerdo a los resultados que se van obteniendo.

CAPITULO I: Introducción

Hoy en día se reconoce que el valor de la información contenida en los sistemas pertenecientes a las empresas es muy significativo. Se sabe que, en muchos casos, la toma de decisiones importantes de la organización está basada en datos que pertenecen a sus sistemas de información; y también que se invierte tiempo y dinero adicional en obtener datos confiables por medio de terceras partes, que debieran poder ser recuperados a partir del uso de sus propios sistemas.

Es un hecho que los datos pueden ser usados como una ventaja competitiva. La información que poseen las organizaciones pueden ayudarlas a optimizar su trabajo, tomar buenas decisiones e incrementar sus ganancias [2]. Esta posible ventaja se vuelve en contra cuando la calidad de los datos no es buena. Esta situación se agrava por el hecho de que muchos usuarios de sistemas informáticos utilizan los datos confiadamente, sin saber si son datos de *buena calidad*. Como es bien sabido, las decisiones no son mejores que los datos en los que se basan [3].
Por lo tanto, es importante conocer el grado de calidad de los datos cuando la toma de decisiones está basada en ellos.

Actualmente, las empresas de negocios están reclamando datos de calidad. La proliferación de proyectos de Software Integrados, Data Warehousing y Data Mining, han dejado a la vista grandes problemas en las bases de datos que existen en las empresas.

El concepto de calidad no tiene un significado preciso, sino que depende del objeto al que se esté refiriendo y de la persona que lo está considerando. Cuando por ejemplo, hablamos de calidad de un restaurante, probablemente estamos considerando diferentes aspectos tales como atención, rapidez, calidad de la comida, ambiente, precio,.... Sin embargo, cuando hablamos de calidad de una prenda de vestir probablemente nos referimos a aspectos tales como calidad del género, confección, precio,.... Vemos con este ejemplo que la calidad dependerá del objeto al que nos estamos refiriendo. Por otro lado, una persona puede considerar que un restaurante es de calidad si su comida es buena y los precios bajos, y otra persona puede considerar que es de calidad si la comida y la atención son buenos, sin importarle los precios. Con este ejemplo vemos que el concepto de calidad depende de la persona que lo está considerando.

A pesar de falta de precisión en el concepto de calidad, en el caso de calidad de datos podemos basarnos en ciertos aspectos que han sido establecidos en investigaciones anteriores:

- Los datos deben ser relevantes para el entorno en donde se utilizan. Esto implica que se deben conocer bien los requerimientos de los usuarios que utilizan la información, para poder evaluar su calidad [4].
- La búsqueda de datos de calidad debe tener como objetivo que el conjunto de datos refleje la porción de la realidad que necesitan los sistemas que los utilizan. Las revisiones de los errores se deben realizar con esta consideración, de otra forma se obtendrían datos de una buena calidad teórica pero no los necesarios para las funciones que los utilizan.
- Existe un grupo de características de calidad relevantes (también llamadas *dimensiones)* en [2] que se deben considerar como un punto de partida para la definición del conjunto estándar de métricas de calidad de datos.

Según [1], desde el punto de vista de la Ingeniería del Software, la calidad de los datos puede evaluarse y mejorarse por al menos tres caminos:

- consideraciones en el desarrollo del software, generalmente se busca que el esquema de las bases de datos cumpla con las formas normales, pero además se debe considerar la funcionalidad de las transacciones de carga y modificación de datos para que aseguren, en el grado que sea posible, la calidad de los datos almacenados.
- métricas para determinar la calidad de los datos, es decir, tener una manera de poder evaluar el nivel de calidad de datos existentes en las bases.
- extensión del concepto de testing de software para que considere el testeo de los datos con los que trabajará el sistema. Normalmente, en la etapa de testing de los sistemas se utilizan conjuntos de datos ficticios que cubren la mayor cantidad posible de casos de prueba de cada función del sistema. Generalmente, estos nuevos sistemas heredarán datos de los sistemas que actualmente utiliza la organización, en estos casos consideramos que es muy importante que se realicen pruebas de las nuevas funciones con los datos existentes en las bases se datos actuales.

Nuestro trabajo estará focalizado en el segundo de los puntos anteriores: "métricas para determinar la calidad de los datos".

En ingeniería del software las métricas han sido utilizadas desde hace muchos años para controlar los procesos de desarrollo de software [9]. Existe un gran número de métricas estándares que permiten realizar el seguimiento del desarrollo, detectando los casos de necesidad de mejoras para garantizar calidad y cumplimiento de los plazos y presupuestos establecidos.
La importancia de realizar las mediciones con procedimientos y unidades estándares se debe a que esto permite a los profesionales de la Ingeniería del Software hablar un lenguaje común, mejorando el intercambio de información relevante y, consecuentemente, el proceso en sí mismo.

La definición de métricas para determinar la calidad de datos es un trabajo que requiere discutir y conocer en profundidad cuáles son las características que definen mejor a calidad de datos, para luego poder definir mediciones sobre estas características.
Para este proceso de definición de métricas se usará el *Paradigma GQM (Goal-Question-Metrics)* [8]. Este paradigma es un proceso de selección top-down que facilita la identificación de aquellas métricas que claramente responden a los objetivos de la organización.

En [2] se brinda un conjunto de métricas para determinar calidad de datos que establecen un punto de partida para el Análisis Sistemático de Calidad de Datos, y se propone refinarlas intentando re-utilizar las técnicas tradicionales de la Ingeniería de Software.

En este trabajo se toma a [2] como punto de partida. Consistirá en la definición de una metodología para llevar a cabo proyectos de medición de calidad de datos en una organización, en la cual a partir de un conjunto estándar de métricas y escalas de medición pueda obtenerse:

- el grado de calidad de los datos y
- los problemas de calidad de los datos más graves que se detectaron.

Esto debería permitir luego realizar un diagnóstico que permita derivar una propuesta de mejoras para calidad de la información de la empresa que se está analizando.
Además, este trabajo será validado por un experimento concreto, donde se aplicará la metodología a una base de datos conformada por la información de cinco agendas electrónicas personales.

De esta manera, estaremos obteniendo un *procedimiento reutilizable, estándar y probado empíricamente* al menos en forma preliminar, preparado para efectuar análisis de calidad de datos.

En el Capítulo II presentamos background necesario para la comprensión de esta tesis: Métricas, Mediciones , Paradigma GQM, DCF (Data Collection Form) .
Luego, en el Capítulo III se presenta una revisión de las dimensiones y métricas propuestas [2], se refinan estas métricas y se seleccionan las que pertenecerán al conjunto estándar.
En el Capítulo IV se presenta la metodología propuesta para llevar a cabo proyectos de medición de Calidad de Datos, en la cual se utiliza el conjunto estándar de métricas y escalas de medición definido previamente.
En el Capítulo V se desarrolla el experimento.
Por último, se presentan las conclusiones generales del trabajo desarrollado.

CAPITULO II: Background

En este Capítulo se brindarán algunas definiciones que se utilizarán luego en el presente trabajo.

a) *Métricas y Mediciones*

En nuestra vida cotidiana realizamos una gran cantidad de mediciones que nos permiten conocer y controlar nuestro entorno. Seguramente, la mayoría de nosotros las realizamos sin darnos cuenta que estamos aplicando procedimientos de medición.
Por ejemplo, nos enteramos de los datos meteorológicos y sabemos cuándo una temperatura dada indica si está frío o caluroso, porque podemos compararla con datos de experiencias anteriores. Otro ejemplo es nuestro peso, a partir del cual decidimos si estamos con sobrepeso o bajo peso porque lo comparamos contra nuestro peso ideal. Así podríamos continuar con una gran cantidad de acciones de medición que realizamos a diario.

Presentamos ahora las definiciones de *métricas* y *mediciones,* Fenton & Pfleeger [9]:

> Una *métrica* es un conjunto de *mediciones* relacionadas para obtener el grado de cierta característica de alguna entidad.
> Una *medición* es el proceso por el cual se le asignan números o símbolos a atributos de entidades del mundo real, de tal forma que, de acuerdo con reglas precisas, dan una descripción de la entidad.

Fenton & Pfleeger [9] indican que "uno de las ayudas de las ciencias es encontrar formas de medir atributos de las cosas en las cuales se está interesado", aludiendo a la frase de Galileo Galilei (1564-1642) que dice "lo que no es medible se hace medible". Las métricas han sido la base de las prácticas empíricas de los científicos para lograr comprender distintos hechos.

El hecho es que cuando medimos tomamos mayor conocimiento de la entidad sobre la cual realizamos la medición y por este motivo podemos tener una mayor comprensión sobre la misma. Sin embargo, no debemos olvidar que si bien las métricas indican cuál es la situación, no indican porqué se tiene esa situación. Siempre las métricas deben ir acompañadas de la reflexión sobre la causa que provoca la situación que están indicando.

b) Paradigma Goal Question Metrics (GQM):

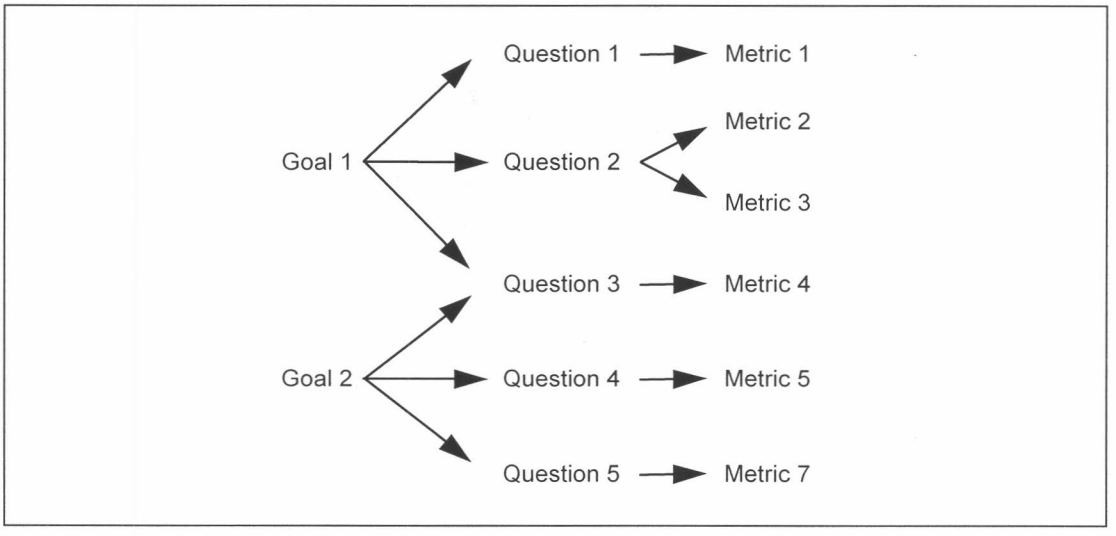
La definición de métricas para determinar la calidad de datos es un trabajo que requiere discutir y conocer en profundidad cuáles son las características que definen mejor a calidad de datos, para luego poder definir mediciones sobre estas características.

Para este proceso de definición de las métricas en este trabajo se usará el *Paradigma GQM (Goal-Question-Metrics)* [8]. Este paradigma es un proceso de selección top-down que facilita la identificación de aquellas métricas que claramente responden a los *objetivos(Goals)* de una organización.
En un primer paso identifica los objetivos y luego, desglosando estos objetivos en piezas manejables, el paradigma asegura que las métricas a recolectar responden a los intereses de la organización.

Este método brinda guías para la definición de métricas sin requerir conocimiento previo de mediciones específicas. Se lo ha elegido por su simplicidad, su reconocimiento y buenos resultados obtenidos en aplicaciones de ingeniería del software [10],[11],[12] .

El principio básico de GQM es que cada organización y cada equipo dentro una organización poseen objetivos (*Goals*) que deben lograr. Entonces, basándose en ellos se realizan Preguntas *(Questions)* acordes a cada uno, que permitan saber si se están cumpliendo o no estos objetivos. Estas Preguntas son respondidas por métricas específicas que proveen las respuestas apropiadas a dichas preguntas. Entonces, si las Métricas responden consistentemente a las Preguntas, y las Preguntas están bien relacionadas a los Objetivos, se puede saber, mediante las mediciones, si estos Objetivos se están cumpliendo o no.

La siguiente figura muestra la relación entre los componentes de este paradigma:



El *Goal* debe ser una breve descripción sobre qué es lo que se debe alcanzar. La redacción debe ser descriptiva, no prescriptiva para prever que esté desactualizada en un lapso corto de tiempo o que posea fechas o valores específicos.

Las Preguntas (*Questions*) para cada *Goal* deben focalizarse en los atributos a medir. Este proceso es útil para evitar el exceso de métricas que no están direccionadas a acciones ni a la toma de decisiones que busca la organización.
Las Preguntas a veces disparan dudas adicionales que realmente tienen importancia. El proceso de definir buenas Preguntas es un proceso iterativo. Una vez que se logra este paso, seleccionar las Métricas (*Metrics*) que soportan estas preguntas es, normalmente, más sencillo.
Todo este proceso de refinamiento ayuda a clarificar a la organización cuál es el objetivo de poseer la información.

Existen métricas cuyas mediciones no se pueden realizar automáticamente, sino que dependen del conocimiento y la apreciación de los usuarios. Para estos casos existe una técnica que facilita la recolección y el procesamiento de la información que brindan los usuarios llamada *Data Collection Forms* (DCF). Para cada caso se define un formulario o DCF con preguntas adecuadas a la métrica que deberán ser respondidas por los usuarios.

CAPITULO III: Estudio, refinamiento y definición de las métricas que compondrán el conjunto estándar a aplicar en cada proyecto de evaluación.

Cualquier proyecto de medición de calidad de los datos de los sistemas de una organización debe, en primer lugar, definir el entorno en el que se encuentra la misma - la situación actual de la organización, sus puntos de interés, la composición de sus sistemas, ...- para luego poder planificar qué características de calidad se deberán medir para obtener el grado de calidad de los datos.

La finalidad de este capítulo es lograr un conjunto de características de calidad estándar a considerar en cualquier proyecto de medición de calidad de datos, teniendo en cuenta que en cada uno de ellos será necesario revisar este conjunto y decidir, dependiendo de la situación particular de la organización, cuáles de estas características se utilizarán y cuáles no. De esta forma nos estamos asegurando que siempre se evalúe la necesidad de medir a cada una de estas características contenidas en el conjunto estándar y cada una de ellas será considerada de la misma forma en los distintos proyectos debido a que también poseerán una definición estándar.

Para esta tarea se toma el conjunto de *dimensiones** y *métricas* definidos en *Measuring Data Quality* [2] para efectuar su estudio, comprensión, definición y posible inclusión/exclusión del conjunto resultante como estándar.

En [2] se estudia el tema de métricas para determinar calidad de datos. Este trabajo, se basa en la idea de que el valor de la información depende del uso que se le da a la misma. Tomando como base este concepto y un conjunto de dimensiones de calidad de datos definidos en [6] y [7], se utiliza GQM para obtener un conjunto de métricas lo suficientemente genérico como para poder ser considerado en distintos casos de medición de calidad de datos. Se definen técnicas a aplicar en las mediciones, que en algunos casos será necesario refinar, ya que las presentadas son sólo un punto de partida para futuras investigaciones. También se presenta un "algoritmo" a seguir en la ejecución de este plan de medición.

Estas definiciones de las dimensiones se harán en forma genérica, es decir sin considerar ninguna base de datos en particular. Se considerará el punto de vista del usuario de datos como criterio de definición de cada dimensión y se remarcará el concepto de que la calidad se mide en cuanto al valor que agrega la información para la organización a la cual pertenece.

* nomenclatura: dimensión de calidad = característica de calidad

1. Dimensiones de calidad tomadas como base:

El conjunto de dimensiones, junto con sus definiciones relacionadas, inicialmente propuesto en [2], [6] y [7] es el siguiente:

Completitud	Cada hecho del mundo real está representado. Es posible considerar dos aspectos diferentes de completitud: primero, ciertos valores pueden no estar representados en el momento; segundo, ciertos atributos no pueden ser representados.
Relevancia	Cada pieza de información almacenada es relevante para obtener una representación del mundo real.
Conciseness	El mundo real esta representado con la mínima cantidad de información requerida por el objetivo para el cual es utilizada.
Cantidad de Datos	Número de hechos almacenados.
Consistencia	No hay contradicciones entre los datos almacenados.
Correctitud	Cada conjunto de datos almacenados representa una situación del mundo real.
Precisión	Los datos son almacenados con la precisión que requiere su caracterización.
No Ambigüedad	Cada pieza de información tiene un único significado.
Exactitud	Cada datum(*) almacenado se relaciona con un hecho del mundo real en forma precisa.
Objetividad	Los datos son objetivos, es decir, no dependen de la opinión, interpretación o evaluación de las personas.
Vigencia	Los datos son actualizados en fecha; la frecuencia de modificación es adecuada.
Confiabilidad	Los datos almacenados son creíbles, es decir, pueden ser tomados como información verdadera.
Aplicabilidad	La información almacenada es aplicable a la organización.
Usabilidad	La información almacenada es usada por la organización.

2. Estudio y revisión de las definiciones. Discusión de la Relevancia de cada dimensión:

A continuación se revisarán y discutirán cada una de las definiciones de estas dimensiones, y a partir de este juicio se determinarán cuáles de ellas conformarán el conjunto estándar.

Completitud

En esta dimensión es sumamente importante considerar que tanto el esquema de datos como los datos en sí mismos, deben ser completos desde el punto de vista de las necesidades del negocio al cual está soportando el sistema. De otra manera podría intentarse representar hechos del mundo real que no estarían aportando valor a la organización, ya que no serían utilizados por los usuarios.
Con esta consideración, en el caso que se encuentren datos que son importantes de almacenar pero no pueden ser modelados, no se efectuarán mediciones sobre los mismos pero sí deben ser considerados al evaluar el grado final de *Completitud*.
La *Completitud* de los datos se mide por existencia o no de información dentro de los atributos indicados como necesarios de poseer en la base, no se mide por suficiencia o no de la información contenida dentro del campo (esto se mide en la dimensión *Precisión)*.
Esta dimensión será considerada en el conjunto estándar sobre los datos y sobre el esquema.

Relevancia

A la definición original le agregamos la consideración de que se debe medir *Relevancia* con respecto al uso final que se le de a la información. Es decir, puede haber datos relevantes en cuanto a la representación del mundo real, pero que no sirvan para la organización, en cuyo caso no debieran considerarse de tal forma.
El hecho de poseer atributos no relevantes en el esquema puede provocar pérdidas de tiempo en la carga de información por parte del usuario, con la posible pérdida de atención del mismo y las inconsistencias que esto puede significar.
Si la falta de Relevancia se da en los datos puede generar problemas de performance.
El hecho de no poseer datos relevantes puede ocasionar problemas, sobre todo si no se conoce esta falta de información, ya que se podrían tomar decisiones importantes considerando que los datos representan toda la realidad necesaria y en verdad está faltando parte de ella. Esta última condición la medimos en la dimensión *Completitud.*
Esta dimensión será considerada en el conjunto estándar sobre los datos y sobre el esquema.

Conciseness	Se toma la definición propuesta en los trabajos de investigación que basamos esta tarea, que indica que los datos son Concisos si son la mínima cantidad de información que se necesita para representar la realidad, con la aclaración de que debe ser para representar sólo la realidad que necesita la organización. Se debe aplicar a los datos y al esquema. En el esquema se debe revisar que no haya atributos dependientes de otros(es decir que se pueden inferir -en este punto se debe considerar si existen para resolver problemas de performance del sistema-), o atributos que no son necesarios para la organización. En los datos se debe revisar que no existan tuplas que no son necesarias para la organización, o datos demás como por ejemplo información duplicada. De la definición anterior se deduce que esta dimensión será medida considerando las mediciones de *Relevancia*, ya que los datos no importantes para la organización se detectan en *Relevancia* y también forman parte de esta dimensión. *Esta dimensión será considerada en el conjunto estándar sobre los datos y sobre el esquema.*
Cantidad de Datos	No parece relevante, ya que el volumen de datos no indica ningún aspecto de calidad por sí solo. En el caso en que haya una "cantidad que se suponga demasiado grande", si hay datos que están de más debieran detectarse al medir *Conciseness*; y en el caso de una "cantidad que se suponga demasiado chica", si hay datos que faltan se deben medir en *Completitud*. *Esta dimensión no será considerada en el conjunto estándar.*
Consistencia	Es una dimensión de los datos y del esquema, ya que en los datos se debe revisar que no existan contradicciones entre los mismos. En muchas ocasiones esta medición deberá contemplar más de un atributo, y tal vez atributos de más de una entidad, que relacionados de alguna manera determinada, sus valores puedan significar inconsistencias. Y en cuanto al esquema se debe revisar que las propiedades de los atributos de la base no impliquen inconsistencias (por ejemplo validaciones que no puedan coexistir). *Esta dimensión será considerada en el conjunto estándar sobre los datos y sobre el esquema.*

Correctitud	Adoptaremos la definición de Fenton & Pfleeger [9], para quienes la *Correctitud* de los datos indica si estos representan un hecho de la realidad ó no. Es una dimensión aplicable sólo a los datos. *Esta dimensión será considerada en el conjunto estándar sobre los datos.*
Precisión	Es aplicable al esquema y a los datos. Se debe revisar que el esquema permita ingresar la información con la *Precisión* necesaria para el negocio y además que los datos hayan sido cargados de la misma forma. No se considera solamente la precisión de los campos numéricos, sino cualquier tipo de dato que necesite una determinada forma de almacenarse para que no haya dudas sobre el significado de esta información. Vale la aclaración de que el caso de un campo que contenga información, pero que esta no sea suficiente para interpretar su significado en la realidad es considerado un caso de falta de *Precisión.* Puede ser una falla de las funciones de carga de datos el no asegurar que sean cargados con la *Precisión* necesaria, cuando el esquema lo permite. *Esta dimensión será considerada en el conjunto estándar sobre los datos y sobre el esquema.*
No Ambigüedad	Este concepto se considera incluido dentro de las dimensiones *Completitud y Consistencia*, ya que un dato es ambiguo si no contiene la cantidad suficiente de información para caracterizarlo unívocamente *(Completitud)* o cuando para un valor buscado se obtiene más de una tupla con diferentes opciones y no se puede determinar cuál se debe seleccionar *(Consistencia).* *Esta dimensión no será considerada en el conjunto estándar.*

Exactitud	Se toma la definición de Fenton y Pfleeger [9], quienes consideran que al comparar el contenido del datum (*) con el mundo real se obtiene un "grado" de *Exactitud*, que representa la semejanza encontrada entre ambos. No se considera esta dimensión dentro del conjunto estándar ya que es una variante de la *Correctitud* y preferimos esta última ya que por la definición de *Exactitud*, se observan uno a uno los valores si reflejan la realidad o no independientemente de los atributos de la base relacionados, en cambio en *Correctitud* se observan las tuplas de información para ver si el conjunto representa la realidad o no. *Esta dimensión no será considerada en el conjunto estándar.*
Objetividad	Esta dimensión no será tomada en cuenta ya que para obtener el grado de *Objetividad* se deberían realizar mediciones sobre la funcionalidad de los sistemas que soportan la carga de datos, para ver si éstos son cubiertos por usuarios que cumplen la función con objetividad o no. En cuanto al impacto que esto puede tener en los datos, se determinará si existen errores (en la dimensión Correctitud) por falta de *Objetividad* o por otra razón. *Esta dimensión no será considerada en el conjunto estándar..*
Vigencia	Es sólo para los datos, es decir no tiene sentido aplicada al esquema. Se debe considerar en aquellos que tienen una validez temporal, es decir representan hechos de la realidad que varían con el tiempo. En [3]??? Se marca la diferencia entre un dato fuera de vigencia y uno incorrecto, considerando al primero a aquel que en algún momento fue correcto y en este momento no lo es, del que nunca fue correcto. *Esta dimensión será considerada en el conjunto estándar sobre los datos.*

Confiabilidad

Tomamos la definición del trabajo inicial [2] que propone que los datos almacenados son confiables si son creíbles, es decir, pueden ser tomados como información verdadera.
Por lo tanto podemos decir que los datos son confiables si son *Completos* (si se realiza una consulta para obtener todos los casos posibles de una determinada selección y los datos no son completos, el resultado no va a reflejar la realidad debido a que le van a faltar casos), *Concisos* (si no lo fuesen, ante una consulta reflejaría mayor cantidad de posibilidades que la real), *Correctos* (por definición de Correctitud: representación de la realidad) , *Vigentes* (es un caso particular de *Correctitud*) , *Precisos* (para que no haya dudas sobre el significado de la información) *y Consistentes* (si son contradictorios no pueden ser confiables).
Por lo tanto, esta dimensión debe ser medida en base a los valores obtenidos en las dimensiones relacionadas.
Este cálculo nos da el porcentaje total (100%) menos: el porcentaje de datos necesarios que no existen en la base (Completitud), el de datos que sí existen en la base pero no son necesarios (Conciseness), y el máximo porcentaje de error obtenido al medir las demás dimensiones que se consideran sobre el set de datos relevantes dentro de la base.
Esta dimensión será considerada en el conjunto estándar sobre los datos.

Aplicabilidad

Esta característica está muy relacionada con la *Relevancia*, es decir si los datos son relevantes para la organización, podemos considerar que son aplicables a la misma. El mismo razonamiento podemos aplicar para el modelo, por lo cual, como consideramos *Relevancia*, *Aplicabilidad* no será considerada dentro del núcleo estándar.
Esta dimensión no será considerada en el conjunto estándar.

Usabilidad

Si los datos son *Relevantes* pero no son utilizados por la organización, es muy probable que el problema resida en la aplicación que los interpreta o en la predisposición de los usuarios para con el sistema, o en el entrenamiento que ellos hayan tenido sobre el uso del mismo. Por lo tanto, *Usabilidad* no es una dimensión a considerar sobre el esquema de datos y/o los datos mismos.
Esta dimensión no será considerada en el conjunto estándar.

(*) Datum es la tripla de base de datos: (entidad, campo, valor)

3. Selección y reclasificación del conjunto estándar:

Fenton y Fleeger en [3] establecen dos criterios de clasificación de métricas de calidad del software que son perfectamente aplicables a las métricas de calidad de datos:

- Directas vs. Indirectas:
  Esta clasificación diferencia aquellas métricas cuyos resultados se obtienen directamente desde la lectura de las mediciones (Directas), de aquellas cuyo resultado se obtiene con algún cálculo sobre resultados de otras métricas realizadas anteriormente (Indirectas).

- Objetivas vs. Subjetivas:
  Esta clasificación diferencia aquellas métricas cuyas mediciones se basan en puntos de vista o interpretaciones personales (Subjetivas), de aquellas cuyas mediciones son un resultado de una medición automática, sin ambigüedades (Objetivas)

De esta clasificación podemos inferir que las dimensiones que son consideradas en cada métrica, también se pueden clasificar en directas e indirectas según sean soportadas por uno u otro tipo de métricas.
Es decir, las dimensiones se clasifican en:

- Directas vs. Indirectas:
  esta clasificación diferencia aquellas dimensiones cuyos resultados se obtienen directamente desde la aplicación de métricas directas (Directas), de aquellas cuyo resultado se obtiene con referencia a resultados de otras métricas que soportan otras dimensiones (Indirectas).

Según este último concepto, el ***conjunto estándar de dimensiones*** a considerar en cada proceso de medición de calidad de datos está conformado por:

Dimensiones Directas:

Completitud
Relevancia
Consistencia
Correctitud
Vigencia
Precisión

Dimensiones Indirectas:

Conciseness (Depende de Relevancia)
Confiabilidad (Depende de Completitud, Conciseness, Correctitud, Vigencia, Precisión y Consistencia)

4. Estudio y definición de las métricas para las dimensiones seleccionadas:

Primero se presentan las métricas propuestas en [2], [6] y [7] para las dimensiones del conjunto estándar definido en el punto anterior. Se revisan y discuten las métricas existentes y se definen métricas para aquellas dimensiones que no tenían ninguna propuesta.
Para lograr una definición de métricas confiables y coherentes con las ya existentes se utiliza el paradigma GQM, que fue el utilizado para los trabajos de investigación anteriores. Este método basa su estudio en los *goals* detectados, en nuestro caso de calidad de datos, y mediante *preguntas (questions)* acerca del objetivo, llegar a descubrir qué *métricas* podrían responderlas y en qué entornos se debieran ejecutar. Se dividen en dos conjuntos, según el objeto a medir: el conjunto de datos y el esquema de datos.

A. Métricas propuestas en [2], [6] y [7]:

- Para el conjunto de datos

GOAL	QUESTION	METRIC	TECHNIQUE
Object: Conj. de datos **Purpose**: Evaluar **Quality**: *Reliability* **Perspectiva**: Usuario **Entorno**: -fixed data set -fixed query set	Do the obtained answers conform the expected answers?	Number of answers that conform the expected answers / Total number of answers	"Functional Data Test" DCF (Data Collection Forms)
Object: Conj. de datos **Purpose**: Evaluar **Quality**: *Relevance* **Perspectiva**: Usuario **Entorno**: -fixed data set -fixed query set	Is there data never queried?	% of tuples never returned as answers	Query set
Object: Conj. de datos **Purpose**: Evaluar **Quality**: *Usefulness* **Perspectiva**: Usuario **Entorno**: -fixed data set	How many times is the stored information queried every day?	Number of accesses to query the data (not including modifications)	LOG of database activities

	Is the data stored used at decision time?	% of decisions made using stored data	DCF (Data Collection Forms)
	Is there any difference in having or not having the data at decision time (i.e., does data help to make "profitable" decisions)?	$ earned in decisions made using the data stored (per time unit) - $ earned in decisions made without using data stored (per time unit)	DCF (Data Collection Forms)
Object: Conj. de datos **Purpose**: Evaluar **Quality**: *Timeliness* **Perspectiva**: Usuario **Entorno**: -fixed data set -fixed query set - *T*: Set of temporal attributes	How often is data updated?	Number of update operations per unit of time	LOG of database activities
	Which percentage of data is updated?	Number of records with attributes in *T* updated (per time unit) / Number of records in the database	LOG of database activities
	How much data has passed its deadline?	Number of records with at least one attribute in *T* not updated (per time unit) / Number of records with at least one attribute in *T*	"Temporal Testing"

- Para el modelo de datos

GOAL	QUESTION	METRIC	TECHNIQUE
Object: Data model **Purpose**: Evaluar **Dimensión**: *Conciseness* **Perspectiva**: Usuario **Entorno**: -fixed data model -fixed query set	Are there attributes (e.g., tuples, columns) that are never accessed?	Number of attributes never accessed	Query set
	Does dependency exist between certain attributes (i.e., may one attribute be computed in terms of others)?	Number of dependent attributes / total number of attributes	Query set

Object: Data model **Purpose**: Evaluar **Quality**: *Completeness* **Perspectiva**: Usuario **Entorno**: -fixed data model	May all the data be represented in the model?	Number of times data could not be stored in the database (does not fit in the model)	DCF (Data Collection Forms)
	Is every field format the right one to store the expected data?	Number of times a value for an attribute could not be stored in the correspondent field	DCF (Data Collection Forms)

B. Discusión de las métricas existentes y definición de las inexistentes:

- Para el **objeto SET de DATOS:**

GOAL	QUESTION	METRIC
Dimensión: *Completitud* **Perspectiva:** Usuario **Entorno:** set de datos	¿Están representados todos los casos necesarios de la realidad?	% de consulta que no obtienen respuestas. (Considerando consultas relevantes para la organización)
Dimensión: *Relevancia* **Perspectiva:** Usuario **Entorno:** - set de datos	¿Existe alguna diferencia en poseer o no los datos al momento de las decisiones? (en las decisiones que generan ganancias)	Llamemos : X:$ ganados en decisiones basadas en los datos almacenados (por unidad de tiempo) Y: $ ganados en decisiones tomadas sin utilizar datos almacenados (por unidad de tiempo). Entonces, el porcentaje de Relevancia será: 100 * X /(X+Y)
	¿Son relevantes los datos almacenados en la base?	# de datos que no tienen Relevancia / # datos revisados.
Dimensión: *Conciseness* **Perspectiva:** Usuario **Entorno:** set de datos	¿Existen datos que no se utilizan?	Es el % de Relevancia .
	¿Existen datos duplicados?	el % de información Duplicada que existe en la base de datos
Dimensión: *Consistencia* **Perspectiva:** Usuario **Entorno:** set de datos fijos	¿Existen contradicciones entre los datos?	Seleccionar set de datos y verificar que no existen inconsistencias entre ellos.
Dimensión: *Correctitud* **Perspectiva:** Usuario **Entorno:** atributos seleccionados como más importantes.	¿Los datos representan la realidad ?	Selección y verificación de datos con el usuario, o contra listas de valores reales.
Dimensión: *Precisión* **Perspectiva:** Usuario **Entorno:** atributos detectados de ingresar con una cierta precisión.	¿Los datos están cargados con la precisión necesaria?	Selección y verificación (por visualización o impresión) de datos

Dimensión: *Vigencia* **Perspectiva**: Usuario **Entorno**: - set de datos fijos - *T*: Set de atributos temporales.	¿Con qué frecuencia se actualizan los datos?	Número de operaciones de modificación por unidad de tiempo.
	¿Qué porcentaje de los datos se actualizan?	Número de registros en *T* modificados (por unidad de tiempo) / Número de atributos en la base de datos.
	¿Cuántos datos han sobrepasado su deadline?	Número de registros con al menos un atributo en T no modificado (por unidad de tiempo) / Número de registros
Dimensión: *Confiabilidad* **Perspectiva**: Usuario **Entorno**: -Set de datos	¿Puedo confiar en que las respuestas obtenidas a las consultas representan la realidad, que si existe la respuesta en el mundo real la obtendré, y si no existe me indicará que no existe?	Se calculará como el 100% menos (el porcentaje de falta de Completitud + el de falta de Conciseness + el mayor dentro de los porcentajes de Incorrectitud, falta de Vigencia, falta de Precisión, Inconsistencias) Se obtiene el mayor % dentro de los % de errores de los datos que contiene la base (se consideró Completitud) y que son necesarios para la representación de la realidad que requiere la organización (se consideró Conciseness). Se determinó utilizar el máximo porcentaje de error porque es la función, dentro de las posibles de realizar, que más se aproxima al cálculo real de este porcentaje de error que buscamos -que es el grado de error que tiene la base de datos sobre todas estas dimensiones juntas-. La forma de obtenerlo sería considerar juntos todos los casos de atributos con errores correspondientes a alguna de estas dimensiones para lograr eliminar los casos de un mismo atributo contado más de una vez por estar afectado a más de una dimensión, y es muy complicada de realizar.

- Para el objeto **MODELO de DATOS:**

GOAL	QUESTION	METRIC
Dimensión: *Completitud* **Perspectiva**: Usuario **Entorno**: -Modelo de datos	¿Todos los datos relevantes pueden ser representados en el modelo?	Porcentaje de datos que no pueden ser almacenados porque no están previstos en el modelo.
Dimensión: *Relevancia* **Perspectiva**: Usuario **Entorno**: -Modelo de datos	¿Existen atributos que no son relevantes para la organización?	Porcentaje de atributos no relevantes sobre el total de atributos..
Dimensión: *Conciseness* **Perspectiva**: Usuario **Entorno**: -Modelo de datos -Set de queries	¿Existen atributos que nunca son consultados?	Es igual al porcentaje de falta de Relevancia.
	¿Existen dependencias entre ciertos atributos? (por ej., ¿puede un atributo ser computado en términos de otros?	Número de atributos dependientes / número total de atributos
Dimensión: *Precisión* **Perspectiva**: Analista de Sistemas **Entorno**: -Modelo de datos	¿Tiene cada atributo el formato correcto para almacenar los datos esperados?	Porcentaje valores correctos que no pueden ser almacenados en los campos correspondientes.

CAPITULO IV: Definición de una metodología para proyectos de medición de calidad de datos.

En este capítulo se define una metodología para llevar a cabo proyectos de medición de calidad de datos. Tiene como objetivo mostrar los pasos a seguir en cualquier proyecto de esta naturaleza.

En esta Metodología se utiliza, como base de las mediciones a efectuar, el conjunto estándar de métricas definido en el Capítulo III de este trabajo. A este conjunto de métricas se lo deberá revisar y decidir la aplicación o no de cada medición propuesta, ajustándolo a cada proyecto en particular.

En cada implementación de proyectos bajo la metodología aquí propuesta, el resultado final será el *grado de calidad de los datos* y los *problemas de calidad de los datos* más graves que se detectaron.

1 Descripción del Plan:

El primer paso de estos proyectos es la tarea de *Relevamiento a Usuarios*, marcando la focalización del concepto de Calidad de Datos según el punto de vista del usuario.
El usuario de la información es quien se nutre de la misma y en ella basa su trabajo cotidiano, marcando la importancia de poseer buenos datos para obtener buenos resultados en su trabajo, con el que puede agregar o quitar valor a la organización a la cual pertenece.

Una vez detectados las funciones importantes que soportan a la organización para la cual se realiza el proyecto, se debe analizar, mediante un estudio de la situación de los sistemas informáticos, cuáles son los datos que soportan esta operatoria para identificar el objeto de nuestras mediciones.

El siguiente paso es analizar las Métricas propuestas en el Capítulo III para decidir si serán incluidas en la planificación de las mediciones del proyecto o no, de acuerdo a los puntos de interés y a los datos objetivos detectados antes.

Se realizará la planificación de las mediciones, se llevarán a cabo las mismas y como último paso de esta metodología se armarán las conclusiones y recomendaciones de mejoras para la base de datos.

El plan para el desarrollo de proyectos de medición se compone de las siguientes fases:

I. Evaluación del estado actual
   I.i. Relevamientos
   I.ii. Evaluación de resultados de los relevamientos
II. Ajuste del plan de métricas a utilizar.
III. Ejecución de las mediciones y Obtención de los resultados.
IV. Conclusiones del Proyecto.

2 Descripción de cada Fase:

I. Evaluación del estado actual:
   I.i. Relevamientos
   Es necesario conocer características tanto del negocio de la organización como de los sistemas y bases de datos que soportan su operatoria, para poder decidir cuáles de las métricas estándares propuestas en el Capítulo III se deberán ejecutar y en qué entorno debemos hacerlo.

   Con la finalidad de obtener estos datos y además detectar la necesidad o no de incluir métricas adicionales, se deben realizar relevamientos a los usuarios.

   La recolección de datos se organiza en tres grupos, según la fuente de la información:

   a nivel management:
   - detección de los puntos de interés del negocio al que pertenece la organización,
   - usuarios más importantes,
   - sistemas que brindan los Informes Gerenciales (son los que proveen información para la toma de decisiones)

   a nivel usuario (grupo de usuarios detectados como los más importantes en el relevamiento al management):
   - cuáles son los sistemas/funciones de los mismos que utilizan con más frecuencia,
   - que función de negocio soportan,
   - qué opinión tienen acerca de la calidad de los datos que utilizan,

- poseen alguna metodología de carga de información en los sistemas? (en el caso que realicen esta operación),

a nivel informática:

- cuáles son los sistemas que posee la organización,
- cuáles son las bases de datos que posee la organización,
- esquemas de bases de datos de cada sistema ,
- cómo interactúan entre ellas,
- cuáles no interactúan pero tienen datos conceptualmente relacionados,
- planes de depuración de datos,
- volúmenes,

Entre las reuniones de relevamiento a un grupo y el siguiente se debe realizar una revisión de la información obtenida en la primer reunión, ya que pueden haber surgido nuevas inquietudes que se deban incorporar a la planificación de la siguiente reunión.

I.ii. Evaluación de resultados de los relevamientos:

Una vez obtenida la información de los relevamientos, será revisada por especialistas funcionales y técnicos. La finalidad de esta tarea es detectar cuáles son los datos de la base que soportan las funciones más importantes de la organización, ya que se debe realizar la planificación de las mediciones y para ello se necesita conocer sobre qué entorno se deben efectuar.

Es importante que en esta tarea se considere la opinión de los usuarios sobre la calidad de los datos que ellos utilizan, ya que al tener el contacto día a día, normalmente detectan casos de errores que afectan su operatoria, por lo que puede ser un punto inicial importante de considerar en este proceso de detectar problemas de calidad en los datos.

Para cada función a revisar se realizará una estructura de árbol, tratando de detectar no sólo las entidades/atributos que la soportan sino también aquellas entidades/atributos que estén relacionados por interfaces o conceptualmente, en otros sistemas de la organización.

Nivel del árbol	Datos a recuperar	Responsable
Primero	Sistema/s que cubre/n cada función de negocio	Especialista Funcional
Segundo	Transacción/es involucrada/s dentro de cada sistema	Especialista Funcional y Especialista Técnico
Tercero	Identificación de los datos importantes en estas transacciones	Especialista Funcional
Cuarto	Entidad/es que 'alimentan' estas transacciones	Especialista Técnico
Quinto	Interfaces que 'alimentan' estas entidades	Especialista Técnico
Sexto	Atributos que 'alimentan' los datos identificados como importantes	Especialista Técnico

Una vez realizada esta tarea tendremos detectados cuáles son las entidades-atributos de la base de datos que soportan las funciones más importantes de la organización, para poder considerarlas el entorno en el cual se deben realizar las mediciones.

II. Ajuste del plan de métricas a utilizar:

Se toma como base de métricas a utilizar el *conjunto de métricas* definido como estándar en el Capítulo III de este trabajo.
De los resultados obtenidos en los relevamientos a usuarios, se pueden obtener cuáles son las funciones y los datos más importantes para la organización para la cual se efectuará el proyecto.

A partir de esta información se ajustará el conjunto de métricas a aplicar. Para cada una de ellas se definirán los *entornos* en los cuales se efectuarán las mediciones y las *técnicas de recolección de datos*.

Al analizar cómo aplicar cada métrica podemos encontrar casos que requieran cambios en los procedimientos de trabajo de los usuarios de la información. Se debe especificar el cambio detectado y plantearlo al management de la

organización para discutirlo y lograr su aprobación. "A veces es necesario modificar nuestro entorno o nuestras practicas, con el fin de medir algo diferente" [9].

Si se detectase una probabilidad alta de que el proyecto no logre resultados de valor para la organización, ya sea debido a características del negocio o informáticas o de presupuesto o de tiempos disponibles, se deben generar reuniones con los directivos de la organización para plantear esta situación y, en caso de ser posible, modificar las factores restrictivos para lograr un proyecto con un resultado que agregue valor a la organización.

Al planificar, se debe tener en cuenta que :

i. existen dos clases de mediciones, según la agrupación definida en el Capítulo III: *directas e indirectas*, y que las *directas* deben realizarse en primer lugar, ya que las *indirectas* se calculan en base a las primeras.
ii. Es importante que la planificación de las *mediciones* no interfiera en la operación normal de los sistemas, ni perjudique el desempeño del personal de la organización involucrado en este proyecto.
iii. Se revisará si es necesario el *entrenamiento* del personal que participará en el proyecto y en ese caso también se incluirá en la planificación.
iv. Si se utiliza una herramienta específica para efectuar las mediciones y recolectar la información, se debe planificar y ejecutar la implementación de la misma y el entrenamiento de los usuarios que la deban utilizar.

Es necesario obtener el compromiso del management de la organización y realizar un plan de comunicación sobre esta iniciativa a todo el personal involucrado para lograr la cooperación que se necesita para llevarlo a cabo.

III. Ejecución de las mediciones y Obtención de los resultados:

En primer lugar se deberán definir los DCFs (Data Collection Forms) necesarios para cada Métrica.
Una vez definidos, se debe contactar a los usuarios involucrados e informarlos sobre los DCFs que deberán completar. Se pueden generar reuniones para guiarlos en el llenado de la información o enviárselos para que los completen ellos mismos. Esto dependerá del tiempo disponible, la distancia a la que se encuentren, y posiblemente más factores que determinarán la forma de llevar a cabo esta tarea.

Si es necesario se deberán definir los queries que cada métrica indique realizar.

Una vez que se posee la información necesaria por parte de los usuarios y las herramientas para llevar a cabo las mediciones, se irán ejecutando según el orden de la planificación y recolectando la información obtenida.

Al realizar las mediciones es posible que encontremos que la técnica de recolección aplicada no nos devuelve un resultado representativo de la realidad que podemos observar, con lo cual se debe revisar nuevamente desde la definición de la métrica que se quiere aplicar para elegir una técnica que la represente mejor. Una vez definida se volverá a aplicar y revisar qué tan bien resultó.

IV. Conclusiones del Proyecto:

Una vez finalizadas las mediciones se deben realizar las conclusiones de la ejecución del proyecto. En ellas se debe resumir el nivel de calidad encontrado en la base de datos según los resultados de las métricas aplicadas.

Como se indicó en un principio, las métricas nos muestran el estado actual pero no nos muestran las causas de dicho estado, por lo tanto es necesario que se realice un análisis cuidadoso de cuáles son las causas de los problemas encontrados en los datos y si es posible mejorar esa situación cuál sería la forma de hacerlo.
Cuando se realizan los planteos de mejoras para los problemas detectados, se debe mantener siempre la orientación al usuario de la información.

Al revisar las causas de lo problemas encontrados se debe analizar cuidadosamente si el problema tiene alguna relación con temas de performance de la aplicación que alimentan (por ejemplo casos de falta de conciseness), y en ese caso evaluar costo beneficio entre la mejora de la calidad de la información y mantener la performance deseada de la aplicación.

Se deben revisar estas conclusiones con el management de la organización, y una vez consensuadas se deben publicar al personal. De esta forma se informará y además se demostrará la aplicación del proyecto realizado.

También se revisará como continuar con este proceso de medición de la calidad de los datos en el futuro. Se puede armar un plan de métricas a aplicar cada cierto período de tiempo con el fin de que muestren la evolución

de la calidad de los datos de la organización. Se debe prever una revisión previa a cada ejecución del plan para ajustarlo a los posibles cambios que realice el cliente.

3 Comentarios finales sobre la metodología propuesta:

Esta Metodología intenta ser general y lo más abarcativa posible, de tal forma que cada proyecto aplique de ella lo que la estructura y complejidad de la organización en la cual se esta trabajando lo permita.

En esta Tesis, se *experimentará* sobre un base de datos conformada por la información contenida en cinco agendas electrónicas personales, brindando un buen conjunto de datos para medir la aplicación de las dimensiones y métricas propuestas en el Capítulo III de este trabajo, pero muy sencillo en cuanto a estructura de la organización y complejidad informática como para probar y refinar en un alto grado esta *Metodología*. Por esta razón recomendamos que se realicen experimentaciones en trabajos posteriores en los cuales sea posible probar más intensamente su aplicación y lograr el mejoramiento de la misma.

CAPITULO V: Armado y ejecución de un experimento concreto para validar las métricas definidas en el Capítulo III y la metodología del Capítulo IV.

El experimento que se presenta en este capítulo, tiene por objetivo evaluar la calidad de los datos de una base conformada por la información contenida en cinco (5) agendas electrónicas personales, usando la metodología propuesta en este trabajo.

La información considerada de las agendas son el directorio telefónico, el scheduler y las notas. La elección de esta base de datos, y de las partes a estudiar se debe a:

I. Consideramos que es altamente probable que la información contenida sea de mala calidad, debido a que cada usuario puede cargar los datos con una metodología diferente o es más, sin ninguna.
II. Las funciones de ingreso de datos de las agendas no poseen verificaciones de formato ni de contenido, por lo tanto la estructura o esquema de la base ****.

Esto probablemente brindará muchos puntos de 'baja calidad', con la posibilidad de verificar la eficiencia de la metodología propuesta y, a medida que avance el experimento, adaptar la misma a las necesidades que vayan surgiendo.

Al solicitar las agendas a los dueños, se los consulta sobre la confidencialidad de sus datos, es decir si requieren que algunos ítems sean quitados de la base a conformar para el proyecto, ya que es posible su publicación en la documentación del experimento.

La información de las agendas se baja a archivos en PC, luego se revisan los formatos en que se pueden unificar los datos de las distintas agendas y se arman los archivos con la descripción necesaria. El paso siguiente es consolidar la información de cada una de las agendas en estos archivos finales. En este momento ya tenemos la base de datos sobre la cual aplicar la metodología de medición.

En cuanto a los roles a desempeñar dentro del proyecto, son los siguientes:

i. Los dueños de las agendas cumplirán con el papel de personal del management de la organización para la cual se está realizando el proyecto, y también con el de usuario de los datos.
ii. El rol de personal de informática lo asumiremos nosotros.

1 Planificación del experimento

La planificación del experimento se lleva a cabo siguiendo los lineamientos de la Metodología propuesta en el Capítulo III de este trabajo. Es decir, cumple con las siguientes etapas:

I. Evaluación del estado actual
I.i. Relevamientos
I.ii. Evaluación de resultados de los relevamientos
II. Ajuste del plan de métricas a utilizar.
III. Ejecución de las mediciones y Obtención de los resultados
IV. Conclusiones del Experimento.

2 *Ejecución del experimento.*

En esta sección se documenta el experimento tarea a tarea, tratando de explicar todas los hechos que se van sucediendo, el significado para el proyecto y su impacto en la metodología utilizada. Además, el feedback de la ejecución del experimento se utiliza para ajustar la metodología propuesta.

I. Evaluación del estado actual

I.i. Relevamiento a usuarios:

En esta etapa se trata de conocer cuáles son las necesidades de los usuarios en cuanto a las funciones más utilizadas de sus agendas, cuáles no utilizan porque consideran que los datos contenidos no son confiables, cuáles sí utilizan confiando en los datos, cuáles considera de mayor impacto negativo en caso de que se alimente de datos erróneos.

Algunos puntos propuestos en la metodología para el relevamiento al personal de informática no se están considerando debido a que no existe una estructura de sistemas relacionados ni una complejidad en las funciones que soporta esta base de datos.

Debido a la falta de reglas de ingreso de la información y de formatos específicos para el tipo de datos que debe almacenar se pide a cada usuario que explique cuál es la forma en que normalmente carga los datos. Es decir, en el caso de directorio telefónico se quiere saber si utiliza mayúsculas y minúsculas, si tiene alguna regla para los nombres como por ejemplo primero el apellido y luego el nombre, cómo utiliza los campos 'libres' , en el caso que el ítem posee más de un número telefónico cómo los almacena, cómo carga los código de área, etc. . Conocer estas características será de utilidad al

momento de realizar las mediciones.

Para llevar adelante esta tarea se obtuvo, en primer lugar, una lista de los usuarios con el número adonde se los pueda ubicar y se planificaron reuniones con cada uno de ellos para llevar a cabo el relevamiento, informándolos sobre el objetivo de la misma para que al momento del encuentro tengan pensadas sus opiniones.

La documentación de los relevamientos se presentará sobre un formato de tabla por usuario conteniendo las respuestas observadas en cada reunión. (Ver ANEXO 1)

II. Evaluación del estado actual – evaluación de los resultados de los relevamientos:

En base a los resultados de los relevamientos, se deben poder detectar cuáles son los datos más importantes para el usuario, y sobre ellos planificar las mediciones de calidad utilizando las métricas propuestas como estándares.

También se deben detectar cuáles funciones actualmente no se utilizan pero serían de mucha utilidad para los usuarios. Se debe encontrar la causa que provoca esta falta, que en principio puede ser entrenamiento incompleto del usuario o falta de amigabilidad de la función que realiza la carga de datos, o simplemente no está considerada dentro de la operatoria de la agenda.

De los relevamientos se observa que de todas las funciones que brindan estas agendas, la más utilizada es sin dudas el Directorio Telefónico, por lo tanto focalizaremos nuestras mediciones sobre los datos que la soportan.

Sobre las funciones que ofrecen estas agendas y no se utilizan se pueden ver las siguientes causas que generan esta falta de uso:

- Falta de espacio en memoria
- Dificultad para ingresar la cantidad de datos que cubre la función

También se ve que cada usuario tiene una norma distinta para cargar sus datos en el Directorio Telefónico. Este hecho debe ser considerado al efectuar las mediciones, ya que nos afectará al realizar comparaciones de valores de atributos.

No poseemos logs de actividades sobre la base de datos, que es una de las

herramientas pensadas para ejecutar las mediciones; esto se debe a que la base está recién conformada y en los sistemas orígenes de la información tampoco existe este tipo de logs, con lo cual no hay manera de obtenerlos. Esta condición nos obliga a interactuar más con los usuarios para obtener la información que en otro caso podríamos obtener directamente a partir de estos logs.

El próximo paso es revisar cada una de las dimensiones estándares para ver su inclusión o no dentro del proyecto y definir las técnicas de recolección de datos. Se armará, también, un plan de ejecución de estas métricas.

III. Ajuste del Plan de Métricas:

En esta etapa se genera la definición de cuáles métricas del conjunto estándar se deberán utilizar en el proyecto en base a los datos y funciones detectados en el paso anterior, como importantes para los usuarios y la organización. También se definen las técnicas de recolección de datos para cada una de las métricas a efectuar.

Al momento de evaluar cómo adaptar el conjunto estándar de métricas para armar el plan de medición, notamos que en los relevamientos a los usuarios deben incluirse algunas preguntas importantes relacionadas directamente con algunas métricas del plan, como por ejemplo de las funciones más utilizadas, cuáles son los datos más importantes o para qué utiliza dichas funciones. Esto lo debemos reflejar en la metodología propuesta en el Capítulo IV.

Existen mediciones que requieren muestreos de datos para que el usuario revise la información. Estos muestreos se efectuarán sobre las bases de datos de cada agenda personal (no sobre la base unificada) para que el resultado de la revisión sea más representativo.

A partir de la información compilada en **I.ii.**, en esta sección, presentamos una revisión del conjunto estándar de métricas propuesto en el Capítulo III:

- Para el objeto **SET de DATOS:**

GOAL	QUESTION	METRIC	TECHNIQUE
Dimensión: *Completitud* **Perspectiva:** Usuario **Entorno:** set de datos personales	¿Están representados todos los casos necesarios de la realidad?	% de consultas que no obtienen respuestas. (Considerando consultas relevantes para la organización).	N° de consultas que no obtienen respuestas * 100 /N° de consultas propuestas.
Se aplica Se debe consultar a cada usuario cuál es el universo de información que considera necesario poseer en esta base de datos. Para esto, se distribuye entre los usuarios un pedido de armado de una lista conteniendo los nombres más importantes que considera debieran estar presentes en la agenda. Se solicita que esta lista se compile sin consultar a la agenda. Se pide que luego indique todos los datos que cree necesarios (*TE, DIRECCIÓN,...*) de cada nombre. Una vez obtenida esta información, se realizarán los queries necesarios y se verificará, por un lado la existencia de cada Nombre y por otro que los datos asociados a cada nombre y marcados como necesarios estén cargados en la base.			
Dimensión: *Relevancia* **Perspectiva:** Usuario **Entorno:** -set de datos	¿Existe alguna diferencia en poseer o no los datos al momento de las decisiones? (en las decisiones que generan ganancias)	Llamemos : X:$ ganados en decisiones basadas en los datos almacenados (por unidad de tiempo) Y: $ ganados en decisiones tomadas sin utilizar datos almacenados (por unidad de tiempo). Entonces, el porcentaje de Relevancia será: 100 * X /(X+Y)	No se aplica.
	¿Los datos de la base son relevantes para la organización?	% de atributos que el usuario indica que no utiliza.	# atributos marcados como no utilizados * 100 / # atributos propuestos para revisión.

<table>
<tr><td colspan="4">Se aplica sólo una de las dos métricas.
La primer métrica propuesta no se aplica porque en esta "organización" no existen decisiones que generen pérdidas o ganancias.
La segunda métrica sí se aplica. Para medirla se realizan muestreos de datos para que cada usuario realice una revisión de la relevancia de los mismos. Estos muestreos se obtienen de los archivos de datos de cada agenda personal (en vez de la base unificada) para aumentar la efectividad de la muestra.
Una vez que los usuarios devuelven esta información, se revisa la cantidad de atributos marcados como innecesarios y se obtiene el porcentaje de esta cantidad sobre el total de atributos revisados.</td></tr>
<tr><td rowspan="2">Dimensión:
Conciseness
Perspectiva:
Usuario
Entorno: set de datos personales</td><td>¿Existe información almacenada que no se utiliza?</td><td>% de atributos que el usuario indica que no utiliza.</td><td>Es el % de Relevancia de la información.</td></tr>
<tr><td>¿Existe información de más?</td><td>% de información duplicada</td><td># atributos con información duplicada *100 / # atributos revisados.</td></tr>
<tr><td colspan="4">Se aplica.
Para la primer métrica se adopta el % de Relevancia de datos (ver definición de la dimensión Conciseness en el Capítulo III).
Para la segunda métrica se obtiene el % de información duplicada detectada en los muestreos de datos realizados para medir Precisión.
Como porcentaje final de Conciseness se suman los resultados de estas dos mediciones.</td></tr>
<tr><td>Dimensión:
Consistencia
Perspectiva:
Usuario
Entorno: set de datos fijos</td><td>¿Existen contradicciones entre los datos?</td><td>Seleccionar set de datos y verificar que no existen inconsistencias entre ellos.</td><td>Cantidad de inconsistencias encontradas * 100 /Cantidad de atributos en el set de datos a revisar.</td></tr>
</table>

Por las características de la base que poseemos –no poseemos propiedades, reglas de validación de input, ...- , aplicamos las siguientes heurísticas:

1. se revisan casos de más de una tupla con igual número de *TELEFONO*, considerando que no todos estos casos necesariamente significan inconsistencias ya que por ejemplo, puede ocurrir que dos personas tengan el mismo número telefónico de trabajo, por lo tanto será necesario asegurarse de que no sea un caso correcto con la participación del conocimiento del usuario. También pueden encontrarse casos de duplicidad de información, los cuales pasarán a afectar a la dimensión *Conciseness*.
2. se revisan tuplas con igual *NOMBRE* (considerar variación de mayúsculas minúsculas) y verifican los datos de cada una (si fueran iguales pasaría a ser un caso de redundancia de información afectando a la dimensión *Conciseness*).

En las dos revisiones se consideran *todos* los atributos de la base.

Dimensión: *Correctitud* **Perspectiva**: Usuario **Entorno**: atributos seleccionados como más importantes.	¿Los datos representan la realidad ?	Selección y verificación de datos con el usuario, o contra listas de valores reales.	Cantidad de casos incorrectos / cantidad de atributos revisadas.

Se aplica.
En la revisión se consideran todos los atributos disponibles en el esquema. Se realizan muestreos de datos para que cada usuario efectúe su revisión.

Dimensión: *Precisión* **Perspectiva**: Usuario **Entorno**: atributos detectados de ingresar con una cierta precisión.	Los datos están cargados con la precisión necesaria?	Selección y verificación (por visualización o impresión) de datos	# atributos con falta de precisión/# de atributos del set seleccionado para revisar.

Se aplica.
Se revisa que el campo *TELEFONO* posea datos numéricos de al menos cuatro (4) bytes de longitud. No importa se posee o no datos alfa, ya que en algunos casos se indica si es un interno o a qué lugar pertenece ese número (part., oficina, celular,...) dentro de este mismo atributo. En los casos en que TELEFONO esté en blancos se revisa con el usuario para detectar si es un caso correcto o falta la información (en este caso afecta a Completitud), o está mal cargada (afecta a Precisión). **(**)**

Dimensión: *Vigencia* **Perspectiva**: Usuario **Entorno**: - set de datos fijos - *T*: Set de atributos temporales.	Con qué frecuencia se actualizan los datos?	Número de operaciones de modificación por unidad de tiempo.	No se aplica.
	¿Qué porcentaje de los datos se actualizan?	Número de registros con al menos un atributo en *T* modificado (por unidad de tiempo) / Número de atributos en la base de datos.	No se aplica
	¿Cuántos datos han sobrepasado su deadline?	Número de registros con al menos un atributo en T no modificado (por unidad de tiempo) / Número de registros	% de atributos desactualizados sobre el total de atributos revisados
Las dos primeras métricas propuestas no se aplican. Esto se debe a que no poseemos logs de actividad de la base de los cuales obtener la información que requiere la medición y el usuario no sabe responderlas. Por lo tanto, los problemas de *Vigencia* en esta base los detectamos con la tercer métrica aquí propuesta. Para esta medición se utilizan los muestreos de datos armados para medir *Correctitud.* Se pide a los usuarios que al revisar si los datos son incorrectos, observe si no son casos de falta de vigencia, es decir el dato en algún momento fue correcto y ahora no lo es porque está desactualizado. Una vez obtenidas las respuestas de los usuarios se calcula el promedio de casos de falta de Vigencia detectados sobre la cantidad de atributos revisados.			
Dimensión: *Confiabilidad* **Perspectiva**: Usuario **Entorno**: -set de datos	Puedo confiar en que las respuestas obtenidas a las consultas representan la realidad, que si existe la respuesta en el mundo real la obtendré. y sino existe me indicará que no existe?	Será en relación a las métricas: *Completitud, Conciseness, Correctitud, Vigencia, Precisión y Consistencia*	100 % -(% de *Incompletitud* + % de falta de *Conciseness* + MAX(% *Inconsistencia*, % falta de *Vigencia*, % falta de *Precisión*, % *Incorrectitud*)
Se obtendrá el porcentaje de *Confiabilidad* como el total (100 %) menos la suma de: porcentaje de falta de *Completitud* más % de falta de *Conciseness* más el máximo porcentaje entre los de *Correctitud, Vigencia, Precisión y Consistencia.*			

- Para el objeto **MODELO de DATOS:**

GOAL	QUESTION	METRIC	TECHNIQUE
Dimensión: *Completitud* **Perspectiva:** Usuario **Entorno:** -Modelo de datos	¿Todos los datos relevantes pueden ser representados en el modelo?	Número de veces que un dato no puede ser almacenado porque no está previsto en el modelo.	DCF (Data Collection Forms)
Se aplica Se consulta al usuario sobre casos en los cuales no haya podido almacenar los datos necesarios en la base de datos (DCF). También se consulta sobre qué tipo de información se almacena en los campos *LIBRES1..6* . El porcentaje de Completitud del esquema estará dado por la cantidad de casos en que no se haya podido almacenar la información. Si el tipo de información almacenada en los campos LIBRES1..6 es repetitiva, podemos inferir que el esquema puede mejorarse incluyendo este tipo de atributo. Esto se revisará al plantear las mejoras a la base al final del experimento, *no se considerará en el cálculo del % de Completitud.*			
Dimensión: *Relevancia* **Perspectiva:** Usuario **Entorno:** -Modelo de datos	¿Existen atributos que no son relevantes para la organización?	Porcentaje de atributos no relevantes sobre el total de atributos.	DCF
Se aplica Se consulta a los usuarios aquellos atributos disponibles en el Directorio Telefónico que nunca utiliza. Se obtienen el porcentaje de atributos no utilizados sobre el total de atributos del esquema.			
Dimensión: *Conciseness* **Perspectiva:** Usuario **Entorno:** -Modelo de datos -Set de queries	¿Existen atributos que nunca son consultados?	Es igual al porcentaje de falta de Relevancia.	% de falta de Relevancia
	¿Existen dependencias entre ciertos atributos? (por ej., ¿puede un atributo ser computado en términos de otros?	Número de atributos dependientes / número total de atributos	No se aplica.
Se aplica sólo la primera métrica. En nuestro caso, el esquema de datos es muy sencillo y no tenemos posibilidades de que existan dependencias entre los datos, por lo tanto el porcentaje de Conciseness del Esquema será igual al de Relevancia del Esquema.			

Dimensión: *Precisión* **Perspectiva**: Analista de Sistemas **Entorno**: -Modelo de datos	¿Tiene cada atributo el formato correcto para almacenar los datos esperados?	Número de veces que no puede ser almacenado un valor correcto en el campo correspondiente.	No se aplica.
No se aplica. La base no tiene formatos con una precisión específica, son campos tipo Character con longitud suficiente para cargar la información.			

Consideraciones generales de la revisión:

- Al definir estas métricas se mantiene el concepto de respetar el punto de vista del usuario, en consecuencia las mediciones deben ser significativas para los usuarios de los datos. Por lo tanto, al realizar las mediciones de *Correctitud, Consistencia, Vigencia* y *Precisión* de los datos no se deben considerar aquellos datos marcados como "no utilizables" al medir falta de *Relevancia.*. La razón es que a los usuarios no les interesan los datos *irrelevantes*, por lo tanto tampoco les va a interesar si los mismos contienen errores o no.

- Para almacenar los datos obtenidos de las mediciones se opta por la herramienta Excel de MS-Windows, por su facilidad de uso y adaptabilidad a las funciones de almacenamiento y cálculo que necesitamos.

- **(**)** En la dimensión *Precisión* de datos se podría medir la precisión en la información contenida en los campos *NOMBRE,* ya que pueden existir casos en los cuales no se pueda identificar a quién corresponde una tupla de información con los datos que contiene este campo.
  Como se mencionó anteriormente, el uso de cada una de estas agendas es monousuario y el volumen de información que contiene cada una es bajo, por lo cual es muy probable que cada usuario pueda identificar a quién pertenece cada tupla de información.
  Con lo cual, si consideramos que en esta medición debemos detectar aquellos casos en los cuales el campo NOMBRE no contiene la información suficiente para que *el usuario pueda identificar* a quién pertenecen los datos de la tupla, no se prevé que se detecte un grado de error significativo. Por lo tanto esta medición no se realizará.

Ya detectadas cuáles métricas aplicaremos en el experimento, se comienza con la definición de los DCF (Data Collection Forms) para cada métrica que los necesite (Completitud de los datos, Completitud del esquema y Relevancia del esquema) .
Se deben generar los reportes necesarios para que los usuarios puedan revisar los muestreos de datos para las dimensiones Relevancia, Consistencia de los datos y Correctitud. Una vez diseñados y generados se planifican reuniones con los usuarios para lograr completar la información de los mismos.

IV. Ejecución de las mediciones y obtención de los resultados:

En esta etapa se detalla la forma en que se llevan a cabo las mediciones para cada métrica planificada en el paso anterior:
En el ANEXO 2 se pueden consultar los resultados de cada medición realizada.

Mediciones aplicadas a los Datos:

Completitud de los datos:

Para medir *Completitud de los datos* se distribuyó entre los usuarios un DCF (ver DCF - Completitud de los datos en ANEXO 3) con una lista a ser completada por ellos con los Nombres de mayor interés para cada uno y la lista de los datos más importantes en cada caso; después se unificaron, se convirtieron a ACCESS y mediante Queries se detectaron los casos que no matcheaban con la base de datos.
Observando estos casos se notaba que algunos registros sí existían en la base pero con alguna diferencia ortográfica o sintáctica, por lo cual, considerando la libertad del ingreso de la información a esta base de datos, se decidió solicitar la participación de los usuarios para confirmar la existencia o no de estos registros en la base y de esta forma lograr un grado de Completitud más cercano a la realidad.

En el DCF sobre *Completitud de datos* se debería haber aclarado al usuario que ingrese los datos de Nombres con las mismas reglas con que lo ingresa en su agenda, o si recordaba cómo lo tiene registrado que lo escribiese de la misma manera. De esta forma hubiese resultado más automático el matcheo de la información.

Las dos mediciones que se efectuaron para esta dimensión fueron *Incompletitud de tuplas* de información e *Incompletitud de atributos* dentro de las tuplas existentes. En el primer caso (tuplas inexistentes) se contó la

cantidad de atributos que el usuario había marcado como necesarios para cada una esta tuplas marcadas como no utilizables (considerando TE, Dirección y Dato adicional). En el segundo caso, para cada tupla existente se contó la cantidad de atributos marcados como necesarios e inexistentes en la base.
El resultado final de *Completitud de datos* se obtuvo sumando todos los casos de datos faltantes detectados en las dos mediciones y calculando el porcentaje de esta cantidad sobre el número de atributos revisados (cantidad de tuplas revisadas multiplicada por tres, que es la cantidad de atributos que se revisaron de cada tupla).

Relevancia y Correctitud de los datos:

Para medir Relevancia y Correctitud de los datos se generaron muestreos de las bases de datos personales para cada usuario, con el pedido de que señale cuáles registros de la lista no utiliza y de los que sí utiliza revise el contenido de los datos para ver si existían algunos que no utilizaba, o eran incorrectos, o estaban desactualizados.
Las muestras se tomaron de las bases personales para evitar el caso en que un usuario marque un registro como no usado porque en realidad era de interés para otro usuario y no para él.

En *Relevancia* se consideraron tanto los casos de tuplas completas que no se utilizaban como los de tuplas que sí se utilizaban pero poseían algunos datos no utilizables. En ambos casos se contaron, por cada usuario, la cantidad total de atributos no utilizados (*)[SM1] y se obtuvo el promedio entre esta cantidad y el total de atributos revisados por usuario (cantidad de tuplas multiplicado por cantidad de atributos revisados por tupla).
En el caso de tuplas no utilizables se consideró el campo NOMBRE como un atributo a contar, ya que al medir Relevancia estamos revisando toda la información que el usuario no necesita y este dato cumple esta condición, por lo tanto cuenta dentro del porcentaje de esta característica de calidad.
El porcentaje final de *Relevancia* está dado por el promedio de los porcentajes obtenidos por cada usuario en la medición recientemente explicada.

La medición de *Correctitud* de los datos se efectuó de la misma forma que la de *Relevancia* en cuanto a la contabilización por cantidad de atributos incorrectos en relación a la cantidad de atributos revisados. En este conteo no se consideraron los casos marcados en falta de *Relevancia* ya que en estos datos no estaría afectando a los usuarios el hecho de tener *Incorrectitud.*
Se obtuvo el porcentaje por cada usuario de estas mediciones sobre el muestreo de datos.

El porcentaje final de *Correctitud* se obtuvo con el promedio de los porcentajes por usuario más el porcentaje de datos en tuplas totalmente incorrectas que se detectó al medir *Precisión* de los datos.

Conciseness de los datos:

Esta dimensión es indirecta. Según la definición a la que se llegó en el Capítulo I, la mediremos en base al porcentaje obtenido en la medición de *Relevancia* más el porcentaje de información duplicada (obtenida al medir Consistencia de los datos).

Consistencia de datos:

Para medir *Consistencia* de datos, se realizaron queries para obtener los casos de tuplas con igual valor en el campo NOMBRE, con el objetivo de revisar la relación entre los demás campos, ya que podía ser un caso de duplicidad de datos (cuando la información fuese la misma), o de inconsistencia (cuando la información fuese distinta).
En este último caso también encontramos registros que tenían información que se complementaba, es decir no se duplicaba y tampoco generaba inconsistencia, un ejemplo puede ser el caso de tener dos tuplas con el mismo NOMBRE y distintos valores en el campo TELEFONO y que uno corresponda al particular y otro a la oficina. Si bien no se puede afirmar que este caso sea correcto, ya que tenemos más de una tupla para la misma clave y con distinta información, no se considerará en el conteo de ninguna dimensión, debido a que por el tipo de uso que se le da a estas agendas (cada usuario utiliza sólo la información que él mismo cargó, que no es un gran volumen por lo tanto la conoce, y además puede acceder a más de un registro en cada búsqueda) no existe posibilidad de que genere errores de interpretación de la información.

Otra medición que se realizó para medir *Consistencia* de datos fue obtener las tuplas con TELEFONOS duplicados para efectuar el mismo análisis que en el caso anterior, es decir revisar si era un caso de duplicidad de información, o de inconsistencia o de datos complementarios.
En este caso se notó que debido al reciente cambio de la numeración telefónica nacional, muchos casos de TELEFONOS duplicados no se detectaron ya que uno estaba actualizado y otro no, con lo cual el porcentaje de *Inconsistencia* o Duplicidad de datos puede ser aún mayor al obtenido.

Para cada una de las mediciones de *Consistencia* explicadas arriba, se contó la cantidad de atributos inconsistentes (no se consideró a la clave de búsqueda como un atributo) y se obtuvo el porcentaje de esta cantidad sobre la cantidad de atributos totales de la base.

Para la contabilización de casos Duplicados se contó la cantidad total de atributos duplicados (incluyendo la clave de búsqueda) sobre la cantidad total de atributos de la base.
En ambos casos no se consideraron aquellos datos afectados a la medición de *Relevancia*.

El porcentaje final de *Consistencia* se obtuvo sumando los porcentajes de realizar las dos mediciones que describimos antes para esta dimensión.

Precisión de los datos :

Al realizar la selección de tuplas con TELEFONO en blanco para la medición de *Precisión,* descontamos los casos que habían sido marcados como no utilizados por falta de *Relevancia.*
Al realizar la revisión encontramos casos que eran correctos, es decir el TELEFONO no existía en la realidad o no le interesaba al usuario, por lo tanto no afectaban a esta dimensión; otros casos con falta de precisión en la carga - por ejemplo el número de teléfono estaba cargado en otro campo de la tupla- que sí se contabilizaron como casos de falta de *Precisión*, y también un tercer grupo de registros con toda su información incorrecta los cuales fueron contabilizados a la dimensión *Correctitud.*

Otra selección que se realizó para esta dimensión fue obtener los registros que tenían datos en el campo TELEFONO pero con menos de cuatro caracteres numéricos, ya que esto indicaba que el número telefónico estaba mal cargado o se habían cargado datos que no correspondían al campo. Todos estos casos resultaron tener falta de *Precisión* en la carga de al menos uno de sus datos.
En las dos selecciones se revisó la *Precisión* de los campos NOMBRE, TELEFONO y DIRECCION, por lo tanto se contaron los casos de falta de *Precisión* y se obtuvo el promedio sobre la cantidad total de atributos bajo revisión.

El porcentaje total de *Precisión* se obtuvo realizando la suma de los dos porcentajes obtenidos según la descripción de las mediciones anteriores.

Vigencia de los datos:

Como a los usuarios se les envió el pedido de que diferencien los casos de *Incorrectitud* de los casos de falta de *Vigencia,* se pudo contabilizar esta dimensión por separado.
Al revisar estos casos de falta de *Vigencia*, no se consideraron los casos ya marcados en la medición de *Relevancia.*

El conteo se realizó por cantidad de atributos desactualizados sobre la cantidad total de atributos revisados, primero se obtuvo el porcentaje por usuario y luego el porcentaje final se obtuvo como promedio de estos porcentajes por usuario.

Confiabilidad de los datos:

Esta dimensión es indirecta. Se mide en base a las dimensiones: *Completitud, Conciseness, Correctitud, Consistencia, Vigencia y Precisión.* Se calcula en base a la fórmula dada en la definición de la dimensión:

100 – (% Incompletitud + % de falta de Conciseness + MAX(% Incorrectitud, % Inconsistencia, % falta de Vigencia, % falta de Precisión))

Mediciones aplicadas al Esquema:

Completitud del Esquema:

Para realizar esta medición se distribuyen DCF – Completitud del Esquema de datos (ver ANEXO 3) entre los usuarios, que constan de dos partes:

1. consultando sobre la ocurrencia de no haber podido cargar algún tipo de información necesaria en el Directorio Telefónico.
2. solicitando información sobre los datos que carga en los campos LIBRES1..6 y si es posible, el porcentajes de veces que esto le ocurre (sobre el total de veces que carga información).

La primer consulta está orientada a determinar el porcentaje de *Incompletitud* del esquema, en cambio la segunda intenta detectar mejoras al esquema, el resultado de esta segunda parte no está involucrado en el cálculo de *Completitud*.

El resultados de todos los formularios indican que nunca habían tenido un caso de querer almacenar algún dato en el directorio y no haber podido realizarlo. Un sólo usuario indicó un caso por error de interpretación ya que después lo colocó también en la segunda parte, con lo cual demostró que sí pudo cargar esa información en la agenda. Esto nos está indicando que el Esquema del Directorio Telefónico es *Completo*.

La segunda parte de los DCFs, en las que se consulta por la frecuencia en que se almacena el mismo tipo de información en los campos LIBRES1..6, demuestra que sería una mejora para el esquema del Directorio Telefónico agregar algunos de estos atributos, ya que mejoraría el orden de la información dentro de la agenda y facilitaría las consultas de los usuarios.

Para obtener el grado de frecuencia de cada ítem descripto, se realizó por cada uno de ellos el promedio entre todos los usuarios. El resultado, que se considerará al presentar las propuestas de mejoras, se muestra en la siguiente tabla:

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO	PROMEDIO
Teléfono de los padres **(*)**	5%	
Celulares **(*)**	15%	
Teléfonos de los trabajos **(*)**	10%	
Casas de Veraneo	5%	**1%**
Otros lugares donde ubicarlos	2%	**0.4%**
Números de TE adicionales **(*)**	20 % 30% 40% 25%	**28%**
Direcciones adicionales	10 % 20% 10%	**8%**
Personas de contacto (al ingresar un dato no personal por ejemplo una empresa, institución, ...)	10%	**2%**
N° de documentos o de credenciales	2%	**0.4%**

(*) El porcentaje de estos cuatro ítems se unifican ya que todos se refieren a la necesidad de tener campos para el ingreso de números telefónicos adicionales . Sería más preciso distribuir el porcentaje de Teléfonos Adicionales entre los otros tres ítems, pero para eso debería haber solicitado al usuario mayor precisión al cargar el DCF o solicitarles más información.

Relevancia del Esquema:

Esta medición se realizó en base a la información devuelta por los usuarios en los DCF – Relevancia del Esquema de datos- (ver Anexo 3), en el cual se solicitaba a los usuarios que identificaran cuáles atributos del Directorio Telefónico no utilizaban nunca al cargar sus datos.
Se calculó el promedio de los porcentajes por usuario como resultado final de *Relevancia del Esquema.*

V. Conclusiones de la ejecución del experimento.

Las conclusiones se organizarán en dos secciones:

a) **Conclusiones de la medición de calidad**, que estarán orientadas a los usuarios y contendrán la información que se obtuvo en el experimento sobre la calidad de los datos de la base.
b) **Conclusiones sobre la ejecución del experimento**, que estarán orientadas a los lectores de esta tesis y se focalizarán en la

experimentación de los estándares propuestos en los capítulos anteriores de este trabajo.

a) Conclusiones de la medición de calidad:

En estas conclusiones remarcamos el concepto presentado inicialmente sobre la importancia de aplicar las métricas considerando el punto de vista del usuario y las necesidades de la organización que utilizará los datos.

Las mejoras propuestas estarán orientadas al uso del Directorio Telefónico de las agendas personales y no de la base de datos que conformamos. Se tendrá en consideración que las agendas son monousuario, que cada uno de ellos reconoce cómo está cargada la información por más que no esté en un formato preciso, y que sea completa, y que no es necesario que un usuario utilice información de otro.

Los problemas más graves detectados en las mediciones de calidad sobre esta base de datos son los niveles de *Completitud* y *Conciseness* de los datos, ya que se obtuvieron porcentajes del 10,91% en falta de *Completitud* y del 8,46% en *Conciseness*.

El porcentaje de *Incompletitud* nos está indicando que falta un 10,91% de los datos marcados por los usuarios como importantes de poseer en la agenda. En la mayoría de los casos, faltan datos del tipo Teléfono Celular, Teléfono de la Oficina o algún tipo de información adicional al NOMBRE y TELEFONO. Como tenemos los casos identificados, se podría distribuir entre los usuarios la lista de la información faltante para que ellos mismos se ocupen de averiguarla y cargarla en el Directorio.

En cuanto a Conciseness, podría pensarse que esta característica medida sobre los datos sería de esperar que indique un porcentaje alto de error, ya que la forma en que se conformó la base de datos (unificando las bases de datos de cinco agendas personales) podría estar generando mucha información duplicada. Pero el porcentaje de falta de *Conciseness* en los datos está conformado por:

Detección de registros/atributos sin relevancia:	7,15%
Detección de información duplicada:	1,31%

Con lo cual si bien la anterior suposición es válida, no es la causa más importante de la falta de *Conciseness* de los datos, sino que la causa de este problema es que los usuarios no realizan procedimientos de depuración de la información que ya no utilizan y por lo tanto quedan en la base sin agregar valor y ocupando espacio que, por lo relevado a los usuarios, es sumamente requerido por la baja capacidad de almacenamiento que poseen estas agendas.

De las mediciones sobre los datos existentes y relevantes de la base (*Consistencia, Correctitud, Vigencia y Precisión*), se encuentra que el de mayor índice de error es la falta de *Precisión* en la carga de los datos (0,91%), esto se debe a que la funcionalidad del Directorio Telefónico de las agendas no posee ningún tipo de validación de la información que se ingresa, con lo que queda a disposición del usuario el respetar o no la forma de cargar la información que propone, en teoría, el esquema de datos.

De las mediciones de *Correctitud* y falta de *Vigencia* se observa que el problema mayor es la desactualización de la información (0,67%), que en realidad se debe a falta de constancia del usuario en mantener sus datos actualizados, ya que en la gran mayoría de los casos tenía conocimiento de que la información no era vigente pero no la actualizaba en la agenda.

Se detectaron muy pocas *Inconsistencias* en la información de la base (0,17%).

Hablaremos ahora sobre las mediciones aplicadas sobre el esquema de datos. La información acerca de la falta de *Conciseness* del Esquema nos indica que los dos últimos campos LIBRES generalmente no se utilizan en la carga de datos. Realmente esta información no es muy significativa al momento de proponer mejoras, debido a que el esquema es sumamente sencillo y el hecho de tener estos dos atributos de más no implica necesidad de mantenimiento ni uso de memoria de más. Por lo tanto se propone que en caso de existir una revisión del esquema de datos, se considere la eliminación de estos atributos detectados como irrelevantes.

El porcentaje obtenido de *Completitud* del Esquema es 0%, que nos indica que es totalmente completo. Pero de la información obtenida en los DCF(Data Collection Forms) asociados a esta medición, se extrae que se carga un volumen importante de información en los campos LIBRES1..4 y en muchos casos se repite el mismo tipo de información. Por esta razón

se recomienda agregar en el esquema de datos, al menos los atributos que corresponden a los casos de mayor frecuencia. Esto facilitaría la carga de datos del usuario (por ejemplo, le evitaría el ingreso de leyendas aclarando qué es el dato que está ingresando) y mejoraría la organización de la información almacenada en la agenda.

En conclusión se puede indicar que los problemas más importantes detectados en esta base de datos se deben a que no existen procedimientos de depuración ni de actualización de la información que contienen y que por el tipo de funcionalidad que ofrecen estas agendas, las mejoras dependen de la voluntad de los usuarios para realizarlos.

Otro hecho importante para mejorar el ingreso de la información y también para facilitar las ejecuciones de estas mediciones en el futuro, sería modificar el esquema de datos para que contenga campos específicos para el ingreso de distintos números telefónicos, por ej. Particular, Oficina, Celular... De esta forma se ordenaría el ingreso de los datos en los campos correspondientes y tendrían más eficacia las comparaciones de valores de los atributos en las mediciones futuras.

b) Conclusiones sobre la ejecución del experimento:

En la realización de este experimento se aplicaron los estándares de dimensiones a considerar definidos en el Capítulo I de este trabajo. Se encontró que la mayoría de ellos eran aplicables sobre la base de datos de estudio y por lo tanto se planificó su medición y recolección de información.

Por poseer una base de datos con poco volumen se pudo tener una apreciación sobre la calidad de sus datos revisándola visualmente y al comparar con los resultados en las mediciones, se encontró que son representativos de lo que se vio en la base.

Al inicio del experimento, teníamos definiciones sobre las métricas de las dimensiones *Conciseness* y *Relevancia* diferentes de las que tenemos ahora, que están finalizadas las mediciones:
En un principio se definieron las métricas de *Conciseness* como

- % de información que el usuario no necesita mantener en la base.
- % de información duplicada.

Las métricas de *Relevancia* poseía una definición en relación a las mediciones de *Completitud* y *Conciseness,* dada por la fórmula:

- 100% - % de falta de *Completitud* - % de falta de *Conciseness*

Se realizaron las mediciones para estas métricas y al revisarlas se encontró que no eran muy representativas de la realidad, entonces se efectuó una revisión desde la definición de las dimensiones y se encontró que las métricas propuestas para cada dimensión no se ajustaban a las definiciones de las mismas.

Por la definición de *Relevancia*, que indica que es *el porcentaje de información relevante para la representación de la realidad que poseemos en la base*, se decidió que la métrica que más se ajustaba a la misma era *el porcentaje de datos relevantes para representar la realidad necesaria para la organización*. Por lo tanto, esta dimensión quedó dentro del subconjunto de dimensiones Directas y su medición se realizó sobre el muestreo de datos enviados a los usuarios para indicar cuáles de los datos eran importantes de poseer en la agenda.

La definición de *Conciseness* indica que es *la representación de la realidad necesaria para la organización con la mínima cantidad posible de información*, con lo cual se mantuvieron las métricas pero se cambió la medición del *porcentaje de información que el usuario no necesita mantener en la base*, para que tome el porcentaje obtenido en la dimensión *Relevancia*. De esta manera la dimensión *Conciseness* para a formar parte del subconjunto de dimensiones Indirectas dentro del conjunto estándar.

Con este hecho se pone en práctica la condición de iteración entre los pasos de *Questions* y *Metrics* que posee el *paradigma GQM* [8] utilizado en las definiciones de las métricas, donde plantea que existe un refinamiento de las métricas a medida que se las aplica en el cual se revisa desde las preguntas (*Questions*) planteadas para el Objetivo (*Goal)* para ver si son respondidas por las métricas (*Metrics*) planteadas.

Otro cambio que se realizó en las mediciones fue considerar, en todos los casos, los porcentajes por cantidad de atributos medidos y no por tuplas, ya que en algunas mediciones se consideraba el porcentaje de tuplas sobre tuplas revisadas y se sumaba esta cantidad con otra obtenida por porcentaje de atributos sobre atributos revisados. La diferencia se daba porque en cada tupla no estaban todos los campos cargados con información, por lo cual se estaban contando casos de más,

ya que el porcentaje por tuplas considera a todos los campos de la tupla en la misma condición de error que se consideró a la tupla.

Por las características de la base de datos sobre la que se desarrolló el experimento, es posible, mediante una revisión, obtener una apreciación del grado de calidad de datos que contiene. Esto nos permite comparar los datos obtenidos en la realización del experimento contra los que se aprecian efectuando esta revisión. Dada la similitud encontrada en los resultados de ambos procesos, podemos concluir que el conjunto de Métricas y la Metodología aplicados en el experimento son apropiados para realizar mediciones de calidad de datos.

En cuanto a la aplicación de la Metodología propuesta en el Capítulo II, se pudo implementar en forma reducida debido a la falta de estructura de la organización dueña de la base de datos y a la baja complejidad informática que soporta a los datos bajo experimentación.
Sería interesante aplicar esta propuesta en trabajos de experimentación posteriores que cumplan con estas complejidades, para poder aplicarla y refinarla según los resultados.

CAPITULO VI: Conclusiones generales del trabajo

El objetivo de esta Tesis fue lograr, mediante una profundización de la definición del concepto de Calidad de Datos, un conjunto de Dimensiones de Calidad y una Metodología a aplicar en distintos proyectos de medición de Calidad de Datos.
Una vez realizadas estas definiciones, experimentar la propuesta en un caso concreto de medición de Calidad de Datos.

El tercer Capítulo, en el cual se discuten y se seleccionan las Dimensiones de Calidad, nos brindó la posibilidad de comprender más profundamente el concepto de Calidad en general, de entender su falta de precisión y por lo tanto la necesidad de elaborar definiciones estándares de las características (dimensiones) de Calidad de Datos para lograr una base común de conceptos sobre los cuales se pueda continuar trabajando.

La definición de Métricas sobre estas dimensiones se realizó utilizando el paradigma GQM. En nuestro caso consideramos a cada Dimensión sobre el Conjunto de Datos o sobre el Esquema (cada caso especifica lo que corresponde) como un Goal a cumplir. La tarea siguiente fue definir las preguntas cuyas respuestas contestarían a cada dimensión, esto nos obligó a imaginar lo más generalmente posible cuáles serían estas preguntas sin limitarnos a casos particulares y tratando de imaginar diversas situaciones. La tarea siguiente consistió en definir cuáles métricas responderían estas preguntas. Esta última tarea fue más sencilla ya que las métricas se infieren directamente de la preguntas que responden.

En el cuarto Capítulo, donde se define la Metodología para proyectos de Medición de Calidad de Datos, se recalcó y comprendió la necesidad de orientar el concepto de Calidad de Datos hacia el punto de vista del usuario de la información. Por esta razón se basó la planificación del proyecto en el resultado de evaluar el conjunto estándar de Métricas propuesto en el Capítulo III sobre los datos de los relevamientos de los usuarios, para decidir su ejecución o no, y en caso de que se realizara, el entorno en el cuál se efectuaría.

En el quinto Capítulo se redactó la realización del experimento que consistió en aplicar el conjunto de Métricas y la Metodología definidas antes sobre una base de datos conformada por la información contenida en cinco agendas electrónicas personales.
Se optó por unificar estos datos en un archivo en MS-ACCESS, para luego tener la posibilidad de realizar queries y ordenamientos de los datos con una herramienta que brinde facilidades.
Se optó por realizar el registro de las mediciones en MS-EXCEL, por la facilidad para realizar cálculos y para representar los datos en forma visiblemente ordenada.

El experimento se planificó siguiendo la Metodología del Capítulo III, se ajustaron las métricas, se ejecutaron las Mediciones y se armaron las Conclusiones del proyecto.
Al planificar el experimento no se aplicó parte de la Metodología debido a la baja complejidad de estructura organizativa y de sistemas que poseíamos. Pero sí se aplicó la secuencia de pasos a seguir y la importancia asignada al punto de vista de los usuarios.

De acuerdo a las características de datos detectadas como importantes en los relevamientos a los usuarios, se realizó el ajuste de las Métricas revisando cada una de las métricas propuestas en el tercer Capítulo para decidir si era aplicable a nuestro proyecto o no.
En los casos en los que resultó aplicable se debió definir la técnica para realizar la medición. En este paso nos encontramos con que una técnica comúnmente pensada para realizar mediciones es la consulta a los logs de actividad de las bases de datos, que en nuestro caso son inexistentes. Por lo tanto tuvimos que darle una participación más activa a los usuarios en el proceso de medir las dimensiones de calidad utilizando DCFs y muestreos de datos.

La ejecución de las mediciones se vio en algunos casos complicada por la necesidad de realizar matcheo de datos en una base que no tiene formatos establecidos para sus datos. Esto se debe a que las agendas no poseen formatos de ingreso de información con validaciones de forma, tipo, etc. , por lo tanto cada usuario tiene su forma de cargar los datos y en algunos casos ni siquiera posee una regla estándar para efectuar sus propias cargas.
Con lo cual, viendo que se perdían casos de matching al realizarlo automáticamente, se decidió nuevamente la participación de los usuarios para realizar esta tarea manualmente y lograr un resultado más cercano a la realidad.

Por último, podemos concluir que se logró una primera experimentación de las Dimensiones, Métricas y Metodología propuestos en este trabajo, logrando, de esta forma, una primera revisión y refinamiento de las definiciones y obteniendo una muestra de la aplicabilidad que poseen. Sentando un base para futuras experimentaciones y refinamientos.

ANEXO 1:

Documentación de los Relevamientos a los usuarios.

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

Usuario 1:

Funciones de datos que posee la agenda:	
	Directorio Telefónico Memos Recordatorio Planificación de Tiempo
Funciones más utilizadas:	
	Directorio Telefónico
Funciones no utilizadas:	
	El resto. Al usuario no le interesa utilizar las demás funciones.
Funciones que utiliza y opina que los datos que la alimentan están bien:	
	Directorio Telefónico, no cree que los datos sean de buena calidad en un 100%, pero sí en un porcentaje alto.
Funciones más importantes:	
	Directorio Telefónico, para obtener números telefónicos y direcciones.
Forma en que carga sus datos:	
	En el Directorio, el usuario carga Nombre y apellido con mayúsculas en el campo NOMBRE, el número telefónico incluyendo el código de área en los casos de números fuera de Capital Federal en el campo TELEFONO. Si la persona a registrar tuviese más de un número telefónico, carga a estos en los campos LIBRE1..6, de la misma forma que el primer número y agregándole la leyenda de lugar del número (por ejemplo: particular, oficina, vecino, ...). Utiliza el guión para separar el número, y espacios en el caso de varios números consecutivos. Si conoce la dirección la carga en el campo DIRECCION. Este dato no es muy común que lo tenga.

Usuario 2:

Funciones de datos que posee la agenda:	
	Directorio Telefónico Memos Recordatorio Planificación de Tiempo
Funciones más utilizadas:	
	Directorio Telefónico Recordatorios Planificación de tiempos
Funciones no utilizadas:	
	Ninguna.
Funciones que utiliza y opina que los datos que la alimentan están bien:	
	Directorio Telefónico, cree que los datos son, en su gran mayoría de buena calidad.
Funciones más importantes:	
	Directorio Telefónico (para números telefónicos y direcciones) y Planificación de tiempos (sobre todo para recordar horarios de compromisos).
Forma en que carga sus datos:	
	En el Directorio, el usuario carga Nombre y apellido con mayúsculas la letra inicial y minúsculas el resto, en el campo NOMBRE. El número telefónico incluyendo el código de área en los casos de números fuera de Capital Federal son cargados en el campo TELEFONO. Utiliza guión como separador y barra para indicar más números consecutivos. Si la persona a registrar tuviese más de un número telefónico no tiene una forma normal de cargarlos. En la mayoría de los casos, si son un par de números más, los carga en el mismo campo de TELEFONO sino, en los campos LIBRES1..6. A los números adicionales, generalmente les agrega una leyenda de lugar al que pertenece el número (por ejemplo: particular, oficina, vecino, ...). Si conoce la dirección la carga en el campo DIRECCION. Si tiene información adicional la carga en los campos LIBRES1..6, sin ningún formato específico.

Usuario 3:

Funciones de datos que posee la agenda:	
	Directorio Telefónico Memos Gastos Recordatorio Planificación de Tiempo
Funciones más utilizadas:	
	Directorio Telefónico Gastos (no actualmente, tiene datos anteriores cargados) Memo
Funciones no utilizadas:	
	Scheduler Expensas (actualmente) Planificación de Tiempos: la razón es la dificultad para ingresar todos los datos que necesitaría para la planificación de tiempos.
Funciones que utiliza y opina que los datos que la alimentan están bien:	
	Directorio Telefónico Recordatorio
Funciones más importantes:	
	Directorio Telefónico
Forma en que carga sus datos:	
	En el Directorio, el usuario carga Nombre y apellido con mayúsculas la letra inicial y minúsculas el resto, en el campo NOMBRE. Si el nombre fuese una sigla, lo ingresa todo en mayúsculas. El número telefónico incluyendo el código de área en los casos de números fuera de Capital Federal son cargados en el campo TELEFONO. Utiliza guión como separador del número, barra para indicar varias números consecutivos. Si la persona a registrar tuviese más de un número telefónico pertenecientes al mismo lugar, los cargar en el campo TELEFONO. Si tiene más de un número pero pertenecientes a más de un lugar utiliza los campos LIBRES1..6. A los números adicionales, generalmente les agrega una leyenda del lugar al que pertenece el número (por ejemplo: particular, oficina, vecino, ...). Si conoce la dirección la carga en el campo DIRECCION. Si tiene información adicional la carga en los campos LIBRES1..6, sin ningún formato específico.

Usuario 4:

Funciones de datos que posee la agenda:	
	Directorio Telefónico Memos Gastos Tarjetas de Negocios Recordatorio Planificación de Tiempo
Funciones más utilizadas:	
	Directorio Telefónico
Funciones no utilizadas:	
	El resto. La razón es la capacidad de la máquina.
Funciones que utiliza y opina que los datos que la alimentan están bien:	
	Directorio Telefónico
Funciones más importantes:	
	Directorio Telefónico, para obtener tanto números telefónicos, direcciones o contactos.
Forma en que carga sus datos:	
	En el Directorio, el usuario carga Apellido y Nombre todo en mayúsculas, en ese orden y en el campo NOMBRE. El número telefónico incluyendo el código de área en los casos de números fuera de Capital Federal son cargados en el campo TELEFONO. Utiliza guión como separador del número, barra o la palabra "al" para indicar varios números consecutivos. Si la persona a registrar tuviese más de un número telefónico pertenecientes al mismo lugar, los carga en el campo TELEFONO. Si tiene más de un número pero pertenecientes a más de un lugar utiliza los campos libres. A los números adicionales, generalmente les agrega una leyenda de lugar al que pertenece el número (por ejemplo: particular, oficina, vecino, ...). Si conoce la dirección la carga en el campo DIRECCION. Si tiene información adicional la carga en los campos LIBRES 1..6, sin ningún formato específico.

Usuario 5:

Funciones de datos que posee la agenda:	
	Directorio Telefónico Memo Recordatorio Planificación de Tiempo
Funciones más utilizadas:	
	Directorio Telefónico Planificación Memo
Funciones no utilizadas:	
	Utiliza todas las funciones restringiendo la información debido a la poca capacidad de almacenamiento de la máquina.
Funciones que utiliza y opina que los datos que la alimentan están bien:	
	Directorio Telefónico Memo.
Funciones más importantes:	
	Directorio Telefónico, para obtener tanto números telefónicos, direcciones o contactos.
Forma en que carga sus datos:	
	En el Directorio, el usuario carga los datos en mayúsculas en el campo NOMBRE, no tiene un regla sobre el orden de los datos Apellido y Nombre. El número telefónico incluyendo el código de área en los casos de números fuera de Capital Federal son cargados en el campo TELEFONO. Utiliza guión como separador del número, barra para indicar varios números consecutivos. Si la persona a registrar tuviese más de un número telefónico puede cargarlos en el campo TELEFONO o en los LIBRES. A los números adicionales, generalmente les agrega una leyenda de lugar al que pertenece el número (por ejemplo: particular, oficina, vecino, ...). Si conoce la dirección la carga en el campo DIRECCION. Si tiene información adicional la carga en los campos LIBRES1..6, sin ningún formato específico.

ANEXO 2 :

Documentación de las mediciones realizadas en el experimento.

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

Datos		Usuario1	Usuario2	Usuario3	Usuario4	Usuario5	Total
	Completitud	Inexistencia de información					10,91%
		Total Completitud de los datos					**89,09%**
	Relevancia	5,10%	4,63%	8,07%	15,12%	2,84%	7,15%
		Total Relevancia de los datos					**92,85%**
	Conciseness	Falta de Relevancia de los datos					7,15%
		Información duplicada					1,31%
		Total Conciseness de los datos					**91,54%**
	Consistencia	Inconsistencia detectadas por duplicidad de NOMBRE					0,16%
		Inconsistencia detectadas por duplicidad de TELEFONO					0,01%
		Total Consistencia de los datos					**99,83%**
	Correctitud	0,73%	0,21%	0,00%	0,15%	0,00%	0,22%
		Registros con basura					0,06%
		Total Correctitud de los datos					**99,72%**
	Vigencia	0,44%	1,21%	0,79%	0,93%	0,00%	0,67%
		Total Falta de Vigencia en los datos					**99,33%**
	Precisión	Falta de Precisión al cargar datos					0,91%
		Total Precisión en los datos					**99,09%**
	Confiabilidad	Falta de Confiabilidad de los datos					18,97%
		Total Confiabilidad de los datos					**81,03%**

		Usuario1	Usuario2	Usuario3	Usuario4	Usuario5	Total
Modelo de datos							
	Completitud	0	0	0	0	0	0,00%
		Total Completitud del Esquema					100,00%
	Relevancia	33,33%	33,33%	22,22%	0,00%	33,33%	24,44%
		Total Relevancia del Esquema					75,56%

ANEXO 3 :

Data Collection Forms

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF – Completitud de los Datos

USUARIO 1:

Completar la siguiente lista con los nombres que Ud. considera más importantes de poseer en su agenda. Indique la necesidad de tener su número telefónico y/o dirección y cualquier otra información que considere importante.
Por favor, no realice esta tarea consultando su agenda.

Nombre	Importa poseer TE?	Importa poseer Dirección?	Otros datos necesarios de poseer?
Carlitos pijuan	SI	NO	
Mauricio	SI	SI	
Graciela y Moroco	SI	NO	
Chelo	SI	NO	
Checha	SI	NO	
Frank	SI	NO	
Claudia	SI	SI	
Sebastian	SI	NO	
Alberto Jionin	SI	NO	
Walter Muños	SI	NO	
Dario Berraud	SI	SI	

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF - Completitud de los Datos

USUARIO 2:

Completar la siguiente lista con los nombres que Ud. considera más importantes de poseer en su agenda. Indique la necesidad de tener su número telefónico y/o dirección y cualquier otra información que considere importante.
Por favor, no realice esta tarea consultando su agenda.

Nombre	Importa poseer TE?	Importa poseer Dirección?	Otros datos necesarios de poseer?
Fabián Lamas	Si	no	N° de documento
Andersen Consulting	Si	Si	N° de Fax N° de Solution Center
Docthos	Si	Si	Centros de atención de urgencia N° TE para autorizaciones N° TE atencón domiciliaria
Dr. Vainstein	Si	Si	N° TE 2° consultorio N° TE celular
Dr Algarra	Si	Si	N° celular
Jardín Pequeño Mundo	Si	Si	
Josefina			N° de DNI
Graciela y Gastón	Si	Si	
Mauricio Y Flavia	Si	Si	
Analía Lamas	No tiene	Si	
Gabriel Morillaz	Si	Si	
Hospital Italiano	SI		N° de credenciales de papá y mamá N°TE médico de cabecera y nombre del mismo
Cecilia Rodriguez	Si	Si	
Andrea Castillo	SI	Si	TE de la oficna
Gabriela Guersyanick	Si	Si	TE celular
María Candina	Si	No	
Banco Francés	SI	Si	N° de Cuentas
Banelco	SI		N° de tarjeta Banelco
VISA	SI		N° de tarjeta VISA
Mastercard	SI		N° de Tarjeta Mastercard
Gabriela y Oscar	Si	Si	
Darío Berrau	Si		
Esteban	SI	SI	
Marcela (enfermera)	SI	No	
Martina Marre	SI	SI	N° de celular
Alberto Carneiro	Si	No	
Sonia de los Santos	Si	No	

Chelo Alvarez	Si	Si	N° de celular
Luisa	Si	No	N° Te de su tía
Italia	Si	No	
Claudia Gómez	Si	Si	N° de su oficina
Ana y Carlitos	Si	Si	

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF - Completitud de los Datos

USUARIO 3:

Completar la siguiente lista con los nombres que Ud. considera más importantes de poseer en su agenda. Indi
que la necesidad de tener su número telefónico y/o dirección y cualquier otra información que considere importante.
Por favor, no realice esta tarea consultando su agenda.

Nombre	Importa poseer TE?	Importa poseer Dirección?	Otros datos necesarios de poseer?
Alto Palermo	Si	Si	
ESSO	Si	No	
ASSA	Si	Si	
Mariana Abboud	Si	No	
Gustavo Scarafia	Si	No	
Fabian Lamas	Si	No	
Karina Falbo	Si	No	
Chelo Alvarez	Si	No	
Silvina Loccisano	Si	Si	
Sebastian	Si	Si	
Pablo Peirano	Si	Si	
Taller	Si	Si	
Silvia Morillaz	Si	No	
Mama y Papa	Si	No	
Cecilia Conte	Si	Si	
Tia Lydia	Si	Si	
Alicia Santice	Si	No	
Tio Beto	Si	No	
Tia Mirta	Si	No	
Suller	Si	Si	
Rolando Rojas	Si	No	
Viviana Altieri	Si	No	
Cecilia Suarez	Si	No	
EG3	Si	Si	
SER Empresario	Si	Si	
Y.P.F.	Si	No	
Graciela Trabucco	Si	No	
Graciela Criscuolo	Si	Si	
Viviana	Si	Si	

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF – Completitud de los Datos

USUARIO 4:

Completar la siguiente lista con los nombres que Ud. considera más importantes de poseer en su agenda. Indique la necesidad de tener su número telefónico y/o dirección y cualquier otra información que considere importante.
Por favor, no realice esta tarea consultando su agenda.

Nombre	Importa poseer TE?	Importa poseer Dirección?	Otros datos necesarios de poseer?
Cippitelli Raúl	No	No	Teléfono y dirección de la oficina y celulares
Cippitelli Marino	No	Si	Teléfono de la oficina
Corrales Shuska y Juanjo	Si	Si	Teléfono del trabajo y celulares
Tortosa Marcelo	Si	Si	Celulares y teléfono de los padres y suegros
Dentista	Si	Si	Nombre y horarios
Dermatólogo	Si	Si	Nombre y horarios
Doctho´s	Si	Si	Teléfonos para pedir turnos y de Urgencias. Nombre de mi médica de cabecera
Excelsitas	Si	Si	Nombre de mi ginecóloga y teléfonos para pedir turnos
Moni Claudio	No	Si	Celulares y teléfonos de oficina
Banco de Gallicia	Si	Si	Nros. De Cuentas, personas con las que me comunico en cada sucursal con las que opero
Waisberg Cora Lis	Si	Si	Celular y familia
Baran Laura	Si	No	Distintos teléfonos donde ubicarla en el resto del mundo y las direcciones
Lacanna Sandra	No	Si	Distintos teléfonos donde ubicarla en el resto del mundo y las direcciones
Zárate Carlos José	Si	Si	Otros teléfonos donde los puedo ubicar (Montevideo, padres, negocios)
Visa	Si	No	Claves y números de tarjetas
Mastercard	Si	No	Claves y números de tarjetas
American Express	Si	No	Claves y números de tarjetas, otros teléfonos de beneficios
Lefty	Si	Si	
Tourné Marcelo	Si	Si	Teléfono de USA y dirección y

			de la oficina
Rozados Maria Amalia	No	Si	Hospital y celular
Corrales Alejandro	Si	Si	Trabajo y celular
Galizia Gabriel	No	No	Celular y flia
Saccheri Victoria	Si	Si	Oficina y celulares
Mazzocchi Aldo	Si	Si	Oficina y celular
Cippitelli Graciela	No	Si	Celular
Goldstein Toto	Si	Si	Celular y negocio
Cippitelli Norma	Si	No	
Darolles Cecilia	Si	No	Celular, Estudio
Morillaz Silvia	Si	No	Celular
Gentile Paula	Si	Si	Trabajo

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF – Completitud de los Datos

USUARIO 5:

Completar la siguiente lista con los nombres que Ud. considera más importantes de poseer en su agenda. Indique la necesidad de tener su número telefónico y/o dirección y cualquier otra información que considere importante.
Por favor, no realice esta tarea consultando su agenda.

Nombre	Importa poseer TE?	Importa poseer Dirección?	Otros datos necesarios de poseer?
Celulares	X		
Fadda	X		Celular
Contador	X		Nro Particular
ANSES	X		
Lidia	X		
Mirta	X		
Nutricionista	X	X	
Azize	X	X	

USUARIO N°1

- ¿Le ha sucedido querer o necesitar almacenar datos en la agenda y no encontrar el lugar adecuado para hacerlo?

SI	NO
	X

(marque lo que corresponda)

Si su respuesta es afirmativa y lo recuerda, por favor explique el/los casos.

- Qué tipo de información almacena en los campos libres? ¿Se repiten frecuentemente algunos de estos casos?

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO
Otro n° de Télefono	30%
Otra dirección	20%

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF - Completitud del Esquema de datos

USUARIO N°2

- ¿Le ha sucedido querer o necesitar almacenar datos en la agenda y no encontrar el lugar adecuado para hacerlo?

SI	NO
	X

(marque lo que corresponda)

Si su respuesta es afirmativa y lo recuerda, por favor explique el/los casos.

- Qué tipo de información almacena en los campos libres? ¿Se repiten frecuentemente algunos de estos casos?

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO
N° de teléfono adicionales	25%
Personas de contacto (al ingresar un dato no personal por ejemplo una empresa, institución, ...)	10%
N° de documentos o de credenciales	2%

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF – Completitud del Esquema de datos

USUARIO N°3

- ¿Le ha sucedido querer o necesitar almacenar datos en la agenda y no encontrar el lugar adecuado para hacerlo?

SI	NO
X	

(marque lo que corresponda)

Si su respuesta es afirmativa y lo recuerda, por favor explique el/los casos.

Registracion de tareas y el tiempo insumido que realice en el dia.
Distintos campos para numeros de telefono. Por ejemplo: TE Casa, TE Oficina, Celular.

- Qué tipo de información almacena en los campos libres? ¿Se repiten frecuentemente algunos de estos casos?

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO
Numeros de TE adicionales	20 %
Direcciones adicionales	10 %

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF – Completitud del Esquema de datos

USUARIO N°4

- ¿Le ha sucedido querer o necesitar almacenar datos en la agenda y no encontrar el lugar adecuado para hacerlo?

SI	NO
	X

(marque lo que corresponda)

Si su respuesta es afirmativa y lo recuerda, por favor explique el/los casos.

- Qué tipo de información almacena en los campos libres? ¿Se repiten frecuentemente algunos de estos casos?

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO
Teléfono de los padres	5%
Celulares	15%
Teléfonos de los trabajos	10%
Casas de Veraneo	5%
Otros lugares donde ubicarlos	2%

USUARIO N° 5

- ¿Le ha sucedido querer o necesitar almacenar datos en la agenda y no encontrar el lugar adecuado para hacerlo?

SI	NO
	X

(marque lo que corresponda)

Si su respuesta es afirmativa y lo recuerda, por favor explique el/los casos.

- Qué tipo de información almacena en los campos libres? ¿Se repiten frecuentemente algunos de estos casos?

TIPO DE INFORMACIÓN	# DE VECES QUE LE HA SUCEDIDO
Otros teléfonos	40%
Direcciones relacionadas	10%

"Una Metodología para Medir Calidad de Datos"

Experimentación de lo propuesto

DCF - Relevancia del Esquema de Datos

USUARIO N° 1:

En la siguiente planilla deberá indicar si existen campos de la agenda que no utiliza nunca en la carga de información:

ATRIBUTOS DE LA BASE:	Marcar (X) aquellos que nunca utiliza
Nombre	
Número de TE	
Dirección	
Libre1	
Libre2	
Libre3	
Libre4	X
Libre5	X
Libre6	X

USUARIO N° 2:

En la siguiente planilla deberá indicar si existen campos de la agenda que no utiliza nunca en la carga de información:

ATRIBUTOS DE LA BASE:	Marcar (X) aquellos que nunca utiliza
Nombre	
Número de TE	
Dirección	
Libre1	
Libre2	
Libre3	
Libre4	X
Libre5	X
Libre6	X

USUARIO N° 3:

En la siguiente planilla deberá indicar si existen campos de la agenda que no utiliza nunca en la carga de información:

ATRIBUTOS DE LA BASE:	Marcar (X) aquellos que nunca utiliza
Nombre	
Número de TE	
Dirección	
Libre1	
Libre2	
Libre3	
Libre4	
Libre5	X
Libre6	X

USUARIO N° 4:

En la siguiente planilla deberá indicar si existen campos de la agenda que no utiliza nunca en la carga de información:

ATRIBUTOS DE LA BASE:	Marcar (X) aquellos que nunca utiliza
Nombre	
Número de TE	
Dirección	
Libre1	
Libre2	
Libre3	
Libre4	
Libre5	
Libre6	

Los uso todos si los necesito. A veces no me da la cantidad de información para usar alguno de los libres, es más, a veces tengo más de un teléfono en los campos libres por que no me alcanzan y trato de tener la menos cantidad de registros por personas

USUARIO N° 5:

En la siguiente planilla deberá indicar si existen campos de la agenda que no utiliza nunca en la carga de información:

ATRIBUTOS DE LA BASE:	Marcar (X) aquellos que nunca utiliza
Nombre	
Número de TE	
Dirección	
Libre1	
Libre2	
Libre3	
Libre4	X
Libre5	X
Libre6	X

Bibliografía:

[1] M. Bobrowski, M. Marré, D. Yankelevich: "*A Software Engineering View of Data Quality*", Proceedings Second International Software Quality Week in Europe, Bélgica, Noviembre, 1998.
Elegido ``best paper'' de la conferencia. Invitado para ser presentado en International Software Quality Week in San Franciso, May 1999.

[2] Bobrowski, M., Marré, M., Yankelevich, D.: "*Measuring Data Quality*", 1998.

[3] Redman, T.: "*The Impact of Poor Data Quality on the Typical Enterprise*", Communications of the ACM, Vol. 41, No. 2, pp. 79-82, February 1998.

[4] R.Wang, D.Strong, L.Guarascio "*Data consumer´s perspective of Data Quality*" , TDQM Research Program, Massachusetts Institute of Technology (MIT), Sloan School of Management.

[5] A. McKeating *"Quality can stop dirty data"*, Computerworld,Vol.26, no.49, (1992).

[6] Strong, D.; Lee, Y.;Wang,R.; *"Data Quality in context"*, Communications of the ACM, Vol. 40, No. 5, May 1997.

[7] Wand, Y.; Wang, R.; *"Anchoring Data Quality Dimensions in Ontological Foundations"*, Communications of the ACM, Vol. 39, No. 11, November 1996.

[8] Basili, V.R.; Rombach, H.D.; *"The TAME Project: Towards Improvement-Oriented Software Enviroments"*, IEEE Transaccions on Software Engineering, Vol. 14, No. 6, June 1988.

[9] Fenton, N. y Pfleeger, S.; "*Software Metrics, A Rigorous & Practical Approach*" , PWS, 2° Edition

[10]Fuggetta, Lavazza, Morasca, Cinti, Oldano, Orazi: "*Applying GQM in an Industrial Software Factory*", ACM TOSEM, Vol 7, No 4, October 1998

[11]Basili, Daskalantonakis, Yacobellis: "*Technology transfer at Motorola*", IEEE Software 11, pp. 70-76.

[12]Stark, Durst, Vowell: "*Using Metrics in Management Decision Making*", Computer, Sept. 1994.