

Uso de herramientas para el diagnóstico de calidad de datos: Desplegando tres hojas de la navaja del Ejército Suizo

Viviana Cotik¹, Pablo Luján², Damián Scotton³ y Daniel Yankelevich⁴

^{1,4} Pragma Consultores

San Martín 575 2º Piso (C1004AAK), Buenos Aires – Argentina
{vcotik, dyankele}@pragmaconsultores.com
<http://www.pragmaconsultores.com>

^{2,3} Pragma Consultores

San Martín 575 2º Piso (C1004AAK), Buenos Aires – Argentina
{pablo.lujan.bettati, dpscotton}@gmail.com

Abstract. El uso de herramientas para mejorar la calidad de datos y para diagnosticar problemas en los mismos es sumamente importante para automatizar tareas y detectar problemas que a simple vista podrían escaparse. Habitualmente, las herramientas utilizadas para medir calidad de datos (DQ) o corregir problemas no emplean una metodología general para la mejora de la calidad de datos a largo plazo. Para poder aprovechar las herramientas, creemos que estas deberían estar insertas en el contexto de una metodología general. En este trabajo presentamos un kit de herramientas que hemos desarrollado y su uso asociado con la metodología NEAT, que provee una forma sistemática de diagnosticar calidad de datos [2]. La novedad de la propuesta es la inserción de la herramienta en una metodología que sigue las mejores prácticas aceptadas en la comunidad [15] y la composición de esta en muchas funcionalidades independientes (en contraposición a una herramienta más grande, integrada y de desarrollo más complicado), que se combinan en un kit poderoso. Este kit nos ha permitido obtener mejores y más veloces resultados en proyectos de diagnóstico de calidad de datos realizados en diferentes mercados verticales.

Palabras Clave: calidad de datos; herramientas; metodología NEAT; diagnóstico de calidad de datos; DQ; calidad de la información

1 Introducción

“Según el sistema, usted está muerto señor López”. Así podría empezar una novela de terror de los nuevos tiempos, y así empieza en muchos casos reales una pesadilla para el que escucha estas palabras. Los problemas de calidad de datos son una realidad y muchas veces son la causa del fracaso de sistemas y productos informáticos, mucho más allá de los problemas de calidad del sistema en sí. Uno

de los autores de este trabajo, sin ir muy lejos, sufrió el haber recibido un informe crediticio negativo de una empresa de informes local, por haber sido “confundido” con otra persona por un dato erróneo. Las consecuencias en el caso personal pueden ser un crédito denegado, un alquiler que no se concreta, o tener que demostrar que uno está vivo. En el caso de una organización, las consecuencias pueden ser incluso peores: demandas por miles o millones de pesos, pérdidas millonarias, multas. Un ejemplo sencillo: tener datos duplicados en una base de marketing (del estilo “Pablo Luján, San Martín 575 p1”, “P Luján, San Martín 575, primero” y “Sr Luján, San Martín 575”) puede significar un gasto de un 30% más en una campaña por correspondencia [3]. Otro: dos bancos locales fueron condenados a pagar a clientes indemnizaciones por varios cientos de miles de pesos por haber sido incluidos erróneamente en bases de datos de morosos (ver, por ejemplo, Diario Clarín 14/02/2003: Un banco debe pagar \$ 120.000 por incluir mal a un cliente en Veraz).

Tradicionalmente, siempre se señaló una clara separación entre los problemas de calidad de los sistemas de información y de procesamiento de datos y calidad de la información o los datos en sí. El primer tema es parte de la ingeniería del software, el segundo, es un problema del negocio, de aquel que usa el sistema, del usuario final. “*Garbage in/garbage out*” reza el dicho: si entra basura en el sistema, uno no puede esperar que funcione bien. Este paper no tiene el objetivo de discutir si los profesionales de sistemas debemos ocuparnos o no de establecer controles de calidad de información y de que los sistemas gestionen la calidad de los datos, sino que intenta proponer una solución que, desde las buenas prácticas de ingeniería de software y el uso de soluciones informáticas, brinde respuestas al problema.

Nuestra organización cuenta desde hace varios años con una práctica de Calidad de Información, que brinda servicios en esta área. Si bien las distintas organizaciones tienen particularidades y por lo tanto requieren soluciones distintas, contamos con una metodología, llamada NEAT, que permite organizar y estructurar el trabajo en esta área, siguiendo un esquema que considera el ciclo de vida del dato [3].

A alto nivel, la metodología NEAT tiene tres fases: diagnóstico, tratamiento y mantenimiento. Adicionalmente, se analiza información funcional, modelos de datos y de procesos, que en muchos casos proveen información más objetiva aunque no más real. En el fondo, tal como identificaron varios autores, cada grupo de interés cuenta con información parcial y acotada sobre los procesos y problemas de gestión de los datos [6] [8] [11] [14] [12] [13] [14].

Resulta bastante claro que muchas de estas actividades pueden automatizarse o al menos contar con soporte de herramientas. De hecho, existen herramientas para el diagnóstico, mejora o corrección, en calidad de datos. Sin embargo, al momento de utilizar estas herramientas, nos hemos encontrado con varios problemas:

- pocas herramientas se insertan en una metodología general de solución de problemas de calidad de datos, por lo general atacan un problema puntual o una serie de problemas, pero en forma aislada. Uno de los principales problemas de esto es el “síndrome del eterno retorno”: los datos de una base se

mejoran mediante el uso de estas herramientas, pero dado que las causas de la falta de calidad no se eliminan, al poco tiempo la calidad vuelve a empeorar, generalmente en las mismas dimensiones, y es necesario un nuevo trabajo de corrección,

- las herramientas disponibles en el mercado son muy caras y el uso aún en casos específicos requiere una compra de un paquete que pocas veces se utiliza en alto porcentaje,
- las herramientas apuntan a mercados muy específicos, consideran formatos de dirección americanos o europeos y en general el idioma inglés.

Por estos motivos, encaramos el desarrollo de una herramienta que diera soporte al trabajo concreto realizado por consultores capacitados utilizando la metodología NEAT. Esta herramienta tiene las características de una verdadera “navaja del ejército suizo” para calidad de datos: cuenta con diferentes algoritmos y funciones que uno puede utilizar, dependiendo de su habilidad y de la situación en la que se encuentra, para enfrentar una gama importante de problemas que no se conocen a priori.

Este artículo no pretende mostrar resultados de investigación a nivel técnico, sino un caso de desarrollo y uso de un kit de herramientas por la industria local, del cual se pueden desprender varios resultados que consideramos de interés para la industria.

En la próxima sección describiremos las características de esta herramienta en el contexto de herramientas disponibles. Luego, ilustraremos el funcionamiento de la herramienta mostrando tres de las técnicas implementadas en casos concretos. Finalmente, presentamos las conclusiones y trabajos futuros.

2 Herramientas de Calidad de Datos

Las herramientas de calidad de datos pueden clasificarse en diversos grupos de acuerdo a sus objetivos y forma de trabajo [4] [7] [10].

En lugar de construir una herramienta más, con las características deseadas de integración, lo que necesariamente iba a requerir una gran inversión para lograr una herramienta pesada, con importante esfuerzo dedicado a la interfaz humana, definimos una estrategia alternativa: construir varias herramientas simples que implementen técnicas simples y prácticas que resulten útiles en distintos contextos.

Para explicar mejor la viabilidad de esta elección, es importante señalar un punto mencionado en la introducción: el principal objetivo de esta herramienta es dar apoyo a consultores que trabajan en el área. Para esto, proponemos contar con herramientas simples, cada una implementando un algoritmo o técnica que pueda ser utilizada en diferentes situaciones y a la vez ser combinadas para atacar problemas más complejos. La metáfora es la de una navaja de ejército suizo: cuenta con varias hojas, cada una útil para una situación distinta. Un usuario naïve puede hacer poco con esta navaja, pero un profesional entrenado va a poder elegir las herramientas adecuadas para cada tarea. La coherencia de este kit de

herramientas está dada por la metodología subyacente y el entrenamiento adecuado en su uso.

Entre las ventajas de esta estrategia, además de su costo y facilidad de implementación, está la facilidad con la que se pueden agregar nuevas funcionalidades. Por otro lado, las herramientas más grandes e integradas requieren un cierto esfuerzo de instalación y configuración, que en este caso es casi inexistente.

Como desventaja, es claro que el uso de un kit requiere más trabajo manual y profesionales más capacitados que, adicionalmente a usar la herramienta, elijan una estrategia de uso y combinen las herramientas. Dado que por lo general no se pueden predecir los problemas puntuales de la organización, contar con diferentes herramientas que puedan combinarse en una estrategia específica de resolución de problemas ha resultado muy práctico.

El kit está conformado por herramientas que no modifican datos en forma directa, sino que generan *scripts* SQL que modifican los datos. Estos *scripts* pueden usarse directamente, o en un ambiente controlado, o modificarse antes de ser usados.

Técnicamente, la herramienta corre en Windows® y accede a distintas bases de datos a través de una conexión ODBC. El desarrollo del kit fue parte de un proyecto de investigación y desarrollo, financiado en parte por un subsidio del Gobierno de la Ciudad de Buenos Aires.

La siguiente figura muestra los pasos en que la herramienta puede ser aplicada dentro de la metodología NEAT. Como vemos, aún no existen hojas para todos los pasos, pero la evolución natural de la herramienta nos permite cubrir cada vez más necesidades.



Fig. 1. El uso de la herramienta en el contexto de NEAT

3 Desplegando Tres Hojas de la Navaja del Ejército Suizo

En el resto del trabajo, presentaremos el kit (la navaja suiza) a través de ejemplos que ilustran la aplicación práctica de tres de sus hojas. Cada hoja será aplicada a una situación práctica en la cual presenta una solución concreta sea en etapa de diagnóstico, de tratamiento o de mantenimiento. Los ejemplos son autocontenidos y permiten entender la aplicación de las herramientas en situaciones reales de uso.

2.1 Uso de los Controles de Estandarización y Matching

El kit tiene una librería de funciones de matching, que permiten detectar en un conjunto de datos algunos que son parecidos entre sí. Algunas de estas son soundex, un algoritmo fonético que toma un término como entrada y produce una cadena que identifica a un conjunto de palabras que son fonéticamente similares [9], distancia entre texto y n-gramas. El soundex está implementado en el kit para los idiomas inglés y español. También tiene funciones de parsing y de estandarización [5]. Estas últimas convierten datos en un valor estándar.

Mostraremos las principales funcionalidades de matching y estandarización del kit mediante un ejemplo, en que utilizaremos el subconjunto de datos de clientes de una compañía que viven en la Ciudad de Buenos Aires, previamente denominada Capital Federal.

Utilizando la funcionalidad *frecuencia*, que agrupa conceptos similares y cuenta la cantidad de ocurrencias de los mismos, agrupamos el campo *ciudad* de la tabla de Clientes (ver Tabla 1).

Tabla 1. Ciudades en donde viven los clientes, agrupadas por nombre de ciudad.

Ciudad	Cant.	Ciudad	Cant.
BS.AS	1	CAP.FEDERAL-PTO MADE	1
BUENO AIRES	3	CAPFED	1
BUENOS AIRES	1	CAPI	1
BUENOS AIRES	1	CAPITAL	2
BUENOS AIRES	152	CAPITAL FEDERAL	1
BUENOS AIRES – CAPIT	1	CAPITAL FED	1
BUENOS AIRES - CAPITAL FEDERAL	1	CAPITAL FEDERAL	33
BUENOS ARIES	3	CAPITAL FEDERAL – BUENO AIRES	1
CAP FED	35	CAPITAL FEDERAL- BUENOS AIRES	1
CAP FED REP	1	CAPITAL FEDRAL	1
CAP FED.	1	CAPITOL FEDERAL	1
CAP FEDERAL	1	CIUDAD DE BUENOS AIR	2
CAP. FED.	14	CIUDAD DE BUENOS AI- RES	1
CAP.FED.	5	CPAITAL FEDERAL	1
		Total	268

Nuestro objetivo es tener la posibilidad de corregir los datos de forma automática o semi automática con el uso de las funcionalidades de estandarización y matching.

Primero, utilizamos la funcionalidad estandarización. Esta estandariza datos basándose en el conocimiento almacenado en la base de conocimiento de la herramienta.

Una vez utilizada la función estandarización, se obtienen los resultados mostrados en la Tabla 2. Podemos ver que, si bien los resultados mejoraron, seguimos teniendo muchos valores distintos que representan la misma ciudad.

Tabla 2. Entradas de la base de datos de conocimiento de ciudades que estandarizan los datos al valor “BUENOS AIRES”.

Ciudad Estandarizada	Cant.
BUENO AIRES	3
BUENOES AIRES	1
BUENOS AIRES	247
BUENOS AIRES – BUENO AIRES	1
BUENOS AIRES – BUENOS AIRES	1
BUENOS AIRES – CAPIT	1
BUENOS AIRES REP	1
BUENOS ARIES	3
CAP.FEDERAL-PTO MADE	1
CAPÍ	1
CAPITAL	2
CAPITAL FEDERAL- BUENOS AIRES	1
CAPITAL FEDRAL	1
CAPITOL FEDERAL	1
CIUDAD DE BUENOS AIR	2
CPAITAL FEDERAL	1
Total	268

Ahora, a los valores estandarizados se le aplica la funcionalidad soundex en inglés¹. Los códigos asignados a cada ciudad de acuerdo a su pronunciación pueden verse en la Tabla 3.

Tabla 3. Ciudad estandarizada con su código soundex y cantidad de apariciones. Se resaltan los registros con el mismo código soundex.

Ciudad estandarizada	Código Soundex	Cant.
BUENO AIRES	B562	3
BUENOES AIRES	B526	1
BUENOS AIRES	B526	247
BUENOS AIRES - BUENO AIRES	B526	1
BUENOS AIRES - BUENOS AIRES	B526	1
BUENOS AIRES – CAPIT	B526	1

¹ Antes de aplicar la funcionalidad soundex en inglés se aplicó la versión en español. Esta utiliza la fonética de este idioma para determinar el parecido de dos frases, por ejemplo, sugiere que Buenos Aires y Vuenos Aires son lo mismo.

BUENOS AIRES REP	B526	1
BUENOS ARIES	B526	3
CAP.FEDERAL-PTO MADE	C136	1
CAPI	C1	1
CAPITAL	C134	2
CAPITAL FEDERAL- BUENOS AIRES	C134	1
CAPITAL FEDERAL	C134	1
CAPITOL FEDERAL	C134	1
CIUDAD DE BUENOS AIR	C331	2
CPAITAL FEDERAL	C134	1
Total		268

De esta forma se detectaron 255 registros que corresponden a la misma ciudad (aquellos con código soundex B526). Esto arroja un 95,1% verdaderos positivos y 4,9% falsos negativos; no hay falsos positivos. Estos valores dependen de cuán completa es la base de conocimiento.

El proceso puede iterarse más de una vez. Nótese que el código soundex C134 está repetido en muchos registros. Si normalizamos el conjunto, cambiando el código soundex C134 a Capital Federal y reejecutamos la función de estandarización, obtenemos un nuevo conjunto de datos con sólo un 2,6% de falsos negativos (si se vuelve a repetir el procedimiento se encuentran 6 registros más). Se podría generar un script SQL para modificar los valores del campo ciudad con mismo soundex, reemplazándolos con el valor estandarizado más frecuente o aquél que el usuario seleccione por defecto (ver Tabla 4).

Tabla 4. Valores corregidos.

Ciudad	Cant.
BUENO AIRES	3
BUENOS AIRES	255
CAP.FEDERAL-PTO MADE	1
CAPI	1
CAPITAL FEDERAL	6
CIUDAD DE BUENOS AIR	2
Total	268

2.2 Uso de los Controles de Chequeo de Integridad, Enriquecimiento y Validación

La funcionalidad de control de integridad posee tres funciones principales: control de integridad referencial, control de atributos y control de clave primaria.

El control de integridad referencial detecta, dadas dos tablas, si existen problemas de integridad referencial (i.e. si las foreign keys de una tabla hija no son clave primaria de la tabla padre). Este problema nos llevaría a tener registros huérfanos en la tabla hija.

El control de clave primaria determina si un determinado conjunto de atributos de una tabla puede llegar a ser clave primaria de la misma (un conjunto de atributos es candidato a ser clave primaria, si la misma es única y ninguno de sus valores es *null*). Con esta funcionalidad podemos detectar fácilmente información errónea, por ejemplo números de DNI duplicados, en casos donde este atributo debiera identificar unívocamente a una persona.

La funcionalidad de enriquecimiento nos permite aumentar y mejorar la información de la base de datos. Este proceso se efectúa tomando información desde fuentes externas o internas. Una de estas funcionalidades de la herramienta nos permite obtener, en campos separados, el nombre, apellido y género de una persona, a partir de un sólo campo que contenga el nombre completo.

2.3 Uso del Análisis de Datos de Referencias Cartográficas

Con el crecimiento de Sistemas de Información Geográficos (GIS, por sus siglas en inglés), muchas industrias informatizaron datos cartográficos importantes para el negocio. Particularmente, las petroleras están en la actualidad refinando su capacidad de análisis de los datos y la capacidad de detectar problemas de integridad de los mismos cobra una gran relevancia.

Como se mencionó anteriormente, uno de los recursos de la herramienta es su base de conocimiento. En ella, es posible almacenar puntos representando localizaciones cartográficas mediante latitud y longitud, y, con al menos tres puntos, construir polígonos, los cuales representarán los límites de un país, una provincia, o cualquier distrito identificable. Mediante consultas a la base de conocimiento, se puede saber si un dato, con coordenadas de latitud y longitud, se encuentra dentro o fuera de un polígono.

Así diseñado, es posible representar lo siguiente. Dado el polígono representando una provincia argentina, y dada una lista de datos (ver Tabla 5), se pueden deducir los siguientes errores: Pozos off-shore que se encuentran sobre tierra firme (pozo 5). Pozos on-shore que se encuentran en el mar (pozo 3). Pozos ubicados en una provincia diferente a la que tienen en el campo "Well_Province" (pozo 6)

Tabla 5. Nombre de Pozos, tipo, provincia en que se encuentran, latitud y longitud.

Nombre de pozo	Tipo de pozo	Provincia	Latitud	Longitud
1	Offshore	Chubut	-45.96890833	-67.35483611
2	Offshore	Chubut	-45.98957222	-67.41477222
3	Onshore	Chubut	-45.89564444	-67.37914167
4	Onshore	Chubut	-45.95831667	-67.65584167

5	Offshore	Chubut	-45.88393889	-67.60993889
6	Onshore	Chubut	-46.03311389	-67.67387778



Fig. 2. Pozos de la Tabla 6 en *Microsoft Live Local*.

La funcionalidad provista para dibujar los mapas está basada en llamadas a las aplicaciones cartográficas de Microsoft (*Microsoft Live Local*, <http://local.live.com>) y Google Maps (<http://maps.google.com>). Existe una opción en el menú para seleccionar en cuál de las dos aplicaciones dibujar.

La visualización de datos resulta ser una herramienta práctica, sencilla y eficaz para detectar errores en los datos que de otra forma requerirían análisis mucho más complejos.

4 Conclusiones

Como una alternativa al uso de herramientas de calidad de datos integradas proponemos el uso de un kit que combina muchas herramientas simples, como si fuera una navaja suiza para profesionales de DQ. El uso de la herramienta se propone en el marco de una metodología que sigue las mejores prácticas aceptadas en la comunidad [15].

La herramienta se utilizó en varios proyectos de diagnóstico de calidad de datos en compañías de distintos mercados verticales y organizaciones gubernamentales. A partir de los resultados de los mismos pudimos verificar que el uso de una herramienta adecuada en el momento apropiado puede ser de gran utilidad para detectar inconvenientes de DQ más rápidamente y de manera automática o semi automática.

La estructura flexible de la herramienta nos permite incorporar nuevos algoritmos o técnicas como hojas independientes de la navaja, que pueden o no interactuar con otras de las existentes.

Resumiendo, el kit tiene muchas de las funcionalidades que de acuerdo a Friedman and Bitterer [7] habría que evaluar al elegir una herramienta de DQ [16]: calidad de datos independiente del dominio, soporte a datos en formatos internacionales, facilidad de implementación y usabilidad y posibilidad de manejar varias actividades de calidad de datos

Agradecimientos

Queremos agradecer al Gobierno de la Ciudad de Buenos Aires, que mediante su apoyo nos permitió realizar parte de este proyecto de investigación y desarrollo.

Referencias

1. Basili V. and Rombach H.: 'The TAME Project: Towards Improvement-Oriented Software Environments', IEEE Transactions on Software Engineering, vol. 16, no. 6 (1988).
2. Bobrowski M., Marré M. and Yankelevich D.: 'A Homogeneous Framework to Measure Data Quality', Proceedings of IQ' 99, Boston (1999).
3. Bobrowski M., Marré M. and Yankelevich D.: 'A NEAT Approach for Data Quality Assessment', Information and Database Quality, Piattini M., Calero C., Género M. (Eds.), Chapter 7, Kluwer (2002).
4. Cotik V.: 'Survey de Herramientas y de Datos Disponibles en la Región', Technical Report, Pragma Consultores (2003).
5. Cotik V., Luján P., Scotton D., Yankelevich D.: 'A Swiss Army knife for Data Quality Assessments', International Journal on Information Quality, Inderscience Publishers, 2nd. Issue (2007).
6. Dravis F.: 'Why Categorize Data Quality Problems?', Business Objects Data Quality Weblog, <http://eimblog.businessobjects.com/dravis/2005/8/31/why-categorize-data-quality-problems.html> (2005).
7. Friedman T. and Bitterer A.: 'Magic Quadrant for Data Quality Tools', Gartner Group (2006).
8. Huang K., Lee Y. and Wang R.: 'Quality Information and Knowledge', Prentice Hall (1999).
9. Knuth D.: 'The Art Of Computer Programming', vol. 3, Addison-Wesley (1998).
10. Melgratti H., Yankelevich D.: 'Tools for Data Quality, Technical Report 99-005', Universidad de Buenos Aires, <http://www-2.dc.uba.ar/proyinv/arte/papers/ToolsDQ.zip> (1999).
11. Redman T.: 'Data Quality for the Information Age', Artech House (1996).
12. Strong D., Lee Y. and Wang R.: '10 Potholes in the Road of Information Quality', IEEE Computer (August 1997).
13. Strong D., Lee Y. and Wang R.: 'Data Quality in Context', Communications of the ACM, Vol. 40, No. 5 (May 1997).
14. Wang R., Lee Y., Pipino L. and Strong D.: 'Manage your information as a Product', Sloan Management Review, pp. 95-105 (1998).
15. Wang R., Strong D. and Guarascio L.: 'Beyond Accuracy: What data quality means to data consumers', Total Data Quality Management Program (1996).
16. Smalltree H.: 'Gartner names top data quality management software tools in new Magic Quadrant', SearchDataManagement.com, http://searchdatamanagement.techtarget.com/originalContent/0,289142,sid91_gci1186165_00.html?asrc=SS_CLA_302160&psrc=CLT_91 (2006).