



UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y TECNOLOGÍAS DE LA
INFORMACIÓN

DEFINICIÓN DE UNA ONTOLOGÍA DE CALIDAD DE DATOS EN EL CONTEXTO DE LOS PROCESOS DE NEGOCIO

TESIS PRESENTADA POR LUIS FABIÁN ORTEGA RUIZ
PARA OBTENER EL GRADO DE MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN

Directora de Tesis Dra. María Angélica Caro Gutiérrez

Octubre 2016

Resumen

El modelado de procesos de negocio en las organizaciones permite representar el conjunto de actividades y las interacciones que se efectúan para lograr los objetivos del negocio. Por otra parte, uno de los factores que puede afectar el éxito durante la ejecución de un proceso de negocio es la calidad de los datos involucrados en este. Lo anterior originó la creación del método BPiDQ (Business Process including Data Quality) el cual permite la definición temprana de requisitos de calidad de datos, durante el modelado. Sin embargo, en esta tarea deben participar y colaborar los expertos del negocio y los expertos en tecnologías de información, quienes no suelen tener un lenguaje común para referirse a los mismos conceptos, por lo cual es posible que se generen confusiones y/o interpretaciones erróneas al momento de definir los requisitos de calidad de datos. Motivado por lo anterior, y como estrategia de apoyo al método BPiDQ, en esta tesis se crea una ontología de calidad de datos en el contexto de los procesos de negocio que permita establecer un lenguaje común entre los involucrados en la definición de requisitos de calidad de datos de forma temprana.

Abstract

Modeling business processes in organizations can represent all activities and interactions that take place to achieve business objectives. Moreover, one of the factors that can affect the success during the execution of a business process is the quality of the data involved in this. Above, it led to the creation of BPiDQ (Business Process Including Data Quality) method which allows early identification of data quality requirements during modeling. Whereas in this task must participate and cooperate business experts and experts in information technology, and they do not usually have a common language to refer to the same concepts, which may lead to confusion and / or misinterpretation when defining data quality requirements. Motivated by this, and as a strategy to support the BPiDQ method, this thesis an ontology data quality is created in the context of business processes in order to establish a common language among those involved in defining quality requirements data early.

*Dedico esta tesis a mi familia
en especial a mis padres,
mi hermano, mis primos
y a mis amigos.*

Agradecimientos

Comienzo agradeciendo a Dios quien me ha guiado y acompañado, dándome las habilidades y fortalezas necesarias para enfrentar los desafíos y dificultades que se me han presentado en mi vida.

A mis padres Eugenia Betzabet Ruiz García y José Luis Ortega Candía a los que amo, quienes han confiado y me apoyado han apoyado de forma incondicional en todo lo que me he propuesto haciendo grandes esfuerzos para ello.

A mi hermano Pablo Andrés Ortega Ruiz al que quiero demasiado, quien siempre ha estado para alegrarme, apoyarme y ayudarme en lo que necesite.

A mi prima Victoria Ester Ruiz Díaz por escucharme, aconsejarme y alegrarme la vida.

A Claudio Andrés Torres Fonseca por su amistad incondicional.

A Alex Danilo Díaz Quilodran un amigo quien ha sabido escucharme y aconsejarme siempre que lo necesito.

A la profesora Angélica Caro de quien he aprendido mucho en todo este tiempo y a quien agradezco sus consejos y paciencia.

Finalmente agradezco a Fernando Ramírez, Mauricio Sepúlveda Venegas, Nicole Escobar Albornoz y a Carlos Quijada Fuentes amigos con los que he compartido lindos momentos.

Índice general

Resumen	2
Abstract	3
Agradecimientos	5
1. Introducción	12
1.1. Motivación	13
1.2. Objetivo general	14
1.3. Objetivos específicos	14
1.4. Alcance de la investigación	14
1.5. Estructura del documento	15
2. Metodología de trabajo	16
2.1. Identificación de la terminología de calidad de datos	17
2.1.1. Revisión Sistemática de la Literatura	18
2.1.2. Revisión Sistemática de la Web	18
2.1.3. Validación de la terminología de calidad de datos	18
2.2. Desarrollo de la ontología	19
3. Estado del arte	24
3.1. Procesos de negocio	25
3.2. Calidad de datos	26
3.2.1. Norma ISO/IEC 25012	26
3.3. Ontologías	29
3.3.1. ¿Qué es una ontología?	29
3.3.2. Objetivos características y usos de las ontologías	30

3.3.3.	Clasificaciones de ontologías	31
3.3.4.	Metodologías para la creación de ontologías	32
3.3.5.	Lenguajes para crear ontologías	34
3.3.6.	Herramientas para el desarrollo de ontologías	35
3.4.	Método BPiDQ	36
3.5.	Ontologías, Calidad de Datos y Procesos de Negocio	37
3.5.1.	Ontologías y Calidad de Datos	38
3.5.2.	Ontologías y Procesos de Negocio	38
3.5.3.	Calidad de Datos y Procesos de Negocio	40
4.	Terminología de DQ	42
4.1.	Objetivo e Interrogante de Investigación	43
4.2.	Revisión Sistemática de la Literatura	43
4.3.	Revisión Sistemática de la Web	45
4.4.	Resultados de la revisión	46
4.5.	Validación de los resultados	49
4.5.1.	Objetivos del estudio	49
4.5.2.	Diseño del instrumento	49
4.5.3.	Desarrollo del instrumento	49
4.5.4.	Validación del instrumento	50
4.5.5.	Selección de los participantes	50
4.5.6.	Aplicación de la encuesta	50
4.6.	Análisis de los resultados del estudio	50
4.6.1.	Criterios de validación	51
4.6.2.	Análisis demográfico	51
4.6.3.	Análisis estadístico de los resultados referente a problemas de DQ	52
4.6.4.	Análisis estadístico referente a las características de DQ	52
4.7.	Conclusiones del estudio	53
4.8.	Análisis de la terminología	53
5.	Desarrollo de la Ontología	66
5.1.	Especificación de la Ontología	68
5.1.1.	¿Por qué se construye la ontología?	68
5.1.2.	¿Cuál será su uso?	68
5.1.3.	¿Quiénes serán sus usuarios finales?	69
5.2.	Conceptualización de la Ontología	69
5.2.1.	Tarea 1: Construcción de un glosario de términos	69
5.2.2.	Tarea 2: Taxonomía de términos	76
5.2.3.	Tarea 3: Diagramas de relaciones binarias	77
5.2.4.	Tarea 4: Diccionario de conceptos	77

5.2.5. Tarea 5: Describir las relaciones binarias ad hoc	79
5.2.6. Tarea 6: Describir los atributos de instancia	79
5.2.7. Tareas de la 7 a la 10	80
5.2.8. Tarea 11: Describir instancias	80
5.3. Implementación de la Ontología	87
5.3.1. Ambiente de desarrollo	87
5.3.2. Desarrollo de la ontología	88
5.3.3. Pruebas de las preguntas de competencias	91
5.4. Validación de la ontología	94
5.5. Vinculación de la ontología con el método PBiDQ	95
6. Conclusiones	96
6.1. Análisis del cumplimiento de objetivos	97
6.2. Principales aportes	98
6.3. Contrates de resultados	98
6.4. Líneas de trabajo futuro	98
A. Instrumento	99
Bibliografía	106

Índice de figuras

2.1. Tareas de la metodología METHONTOLOGY, extraído de (Corcho et al., 2005)	21
3.1. Las 4 etapas del método BPiDQ.	37
4.1. Método para la identificación de la terminología de DQ	43
5.1. Actividades de la metodología METHONTOLOGY (Corcho et al., 2005)	67
5.2. Tareas de la metodología METHONTOLOGY, extraído de (Corcho et al., 2005)	69
5.3. Taxonomía de Conceptos	76
5.4. Diagramas de relaciones binarias ad hoc	77
5.5. Interfaz gráfica de Protégé	87
5.6. Captura de la creación y configuración de la ontología	88
5.7. Captura de la jerarquía de clases	89
5.8. Captura de la definición de las propiedades de objeto	89
5.9. Captura de la definición de las propiedades de dato	90
5.10. Captura de las instancias agregadas a la ontología	90
5.11. Vista gráfica de la ontología desarrollada	91
5.12. Editor consultas SPARQL de Protégé	91
5.13. Ontología resultante de la primera y última interacción.	94

Índice de cuadros

2.1. Etapas de una RSL (Kitchenham, 2004)	18
2.2. Etapas de la revisión sistemática de la web	19
2.3. Actividades para el desarrollo de la encuesta	19
3.1. Características de calidad ISO/IEC-25012	27
4.1. Términos de búsqueda para la RSL y RSW	44
4.2. Combinaciones de términos de búsqueda para la RSL y RSW	44
4.3. Resumen de los resultados de búsqueda para la RSL	45
4.4. Resumen de los resultados de búsqueda para la RSW	46
4.5. Lista de problemas de Calidad de Datos	47
4.6. Lista de características de Calidad de Datos (1/3)	55
4.7. Lista de características de Calidad de Datos (2/3)	56
4.8. Lista de características de Calidad de Datos (3/3)	57
4.9. Análisis respecto de la experiencia de los encuestados	57
4.10. Análisis del sector industrial al que pertenecen los encuestados	58
4.11. Análisis de los resultados del estudio de los tipos de problemas de calidad de datos	58
4.12. Análisis de los resultados del estudio de las características de calidad de datos	59
4.13. Relaciones entre tipos de problemas DQ y características DQ de negocio (1/3)	60
4.14. Relaciones entre tipos de problemas DQ y características DQ de negocio (2/3)	61
4.15. Relaciones entre tipos de problemas DQ y características DQ de negocio (3/3)	62
4.16. Relaciones entre características de DQ de negocio y IT (1/3)	63
4.17. Relaciones entre características de DQ de negocio y IT (2/3)	64
4.18. Relaciones entre características de DQ de negocio y IT (3/3)	65
5.1. Glosario de términos (1/6)	70
5.2. Glosario de términos (2/6)	71

5.3. Glosario de términos (3/6)	72
5.4. Glosario de términos (4/6)	73
5.5. Glosario de términos (5/6)	74
5.6. Glosario de términos (6/6)	75
5.7. Diccionario de conceptos de la Ontología	78
5.8. Descripción de las relaciones binarias ad hoc	79
5.9. Descripción de los atributos de instancia	80
5.10. Descripción de instancias(1/6)	81
5.11. Descripción de instancias(2/6)	82
5.12. Descripción de instancias(3/6)	83
5.13. Descripción de instancias(4/6)	84
5.14. Descripción de instancias(5/6)	85
5.15. Descripción de instancias(6/6)	86
5.16. Resumen de las etapas de BPiDQ (extraído de (Rodríguez y Caro, 2012))	95

Capítulo 1

Introducción

Este capítulo se estructura de la siguiente manera, en la Sección 1.1 se expone la motivación de la tesis, en la Sección 1.2 y Sección 1.3 se muestran los objetivos generales y específicos. Luego en la Sección 1.4 se dan a conocer los alcances de la investigación. Finalmente, en la Sección 1.5 se describe la estructura del documento.

1.1. Motivación

Los procesos de negocio (BP por sus siglas en inglés Business Process) se definen como un conjunto de actividades que se realizan de manera coordinada para cumplir un objetivo de negocio en un contexto, tecnológico y organizacional (Weske, 2012). En la actualidad, las organizaciones están inmersas en un entorno competitivo y cambiante, resulta primordial hacer visibles sus BP de forma que puedan ser estudiados, mejorados y optimizados continuamente. Los BP se hacen visibles mediante el modelado, el que se realiza utilizando un lenguaje o notación, típicamente gráficos, siendo los lenguajes más conocidos UML (Unified Modeling Language) (OMG, 2007) y BPMN (Business Process Model and Notation) (OMG, 2011). Actualmente, BPMN es el estándar de facto para el modelado de procesos de negocio, debido a que es la notación más utilizada por la industria (Harmon y Wolf, 2011).

Por otro lado, la calidad de datos (DQ por sus siglas en inglés Data Quality), se define como el grado en que las características de los datos satisfacen necesidades implícitas, establecidas cuando estos son utilizados en condiciones específicas (ISO/IEC-FDIS-25012, 2008). Teniendo presente que, actualmente, uno de los activos más importantes dentro de las organizaciones son los datos (Redman, 2008), la definición temprana de los requisitos de calidad de datos puede ser de gran relevancia para el éxito de sus procesos de negocio (el Abed, 2009).

Consecuentemente con lo anterior, se ha propuesto el método BPiDQ (Rodríguez y Caro, 2012) que aborda la especificación temprana de requisitos de calidad de datos en el modelado de BP. El método BPiDQ está compuesto por cuatro etapas: (1) Modelado de procesos de negocio consciente de la calidad de datos, (2) Especificación de requisitos de calidad de datos, (3) Análisis y mejora de Procesos de Negocio relacionado con la Calidad de Datos y (4) Generación de Diagramas de Casos de Uso. Como su nombre lo dice, la primera etapa de BPiDQ está dedicada a la especificación de los requisitos de DQ sobre un BP. Esta tarea es realizada por los expertos y/o analistas de negocio, quienes tienen la responsabilidad de definir las necesidades del negocio y, desde esa perspectiva, la importancia que tiene la DQ para el desempeño del mismo. Posteriormente, en las siguientes etapas, estas especificaciones son llevadas a un nivel más técnico, por los expertos en tecnologías de información (TI), y transformadas de tal modo que como resultado de la aplicación del método, será posible obtener casos de uso.

Sin embargo, en el interior de las organizaciones, las personas encargadas del negocio y las personas encargadas de las TI, tienen diferentes puntos de vista del dominio de un problema y

suelen usar lenguajes diferentes para expresarse sobre los mismos conceptos, lo cual puede generar malentendidos e interpretaciones erróneas (Jenz, 2003). Particularmente esto puede ocurrir, en el contexto de la primera etapa de BPiDQ, cuando se definen los requisitos de DQ. Lo anterior ha motivado el desarrollo de una estrategia para facilitar la comunicación y el intercambio de información entre los expertos de negocio y los expertos en TI.

Las ontologías favorecen la comunicación entre personas, organizaciones y aplicaciones, porque proporcionan una comprensión común de un dominio, de modo que se eliminan confusiones conceptuales y terminológicas (Guzmán et al., 2012a).

Considerando que la literatura no provee una ontología como la necesitada por el método BPiDQ, esta tesis se centra en la creación de una ontología de calidad de datos en el contexto de los procesos de negocio que permita establecer un lenguaje común entre los expertos de negocio y los expertos en TI.

1.2. Objetivo general

Crear una ontología que provea un modelo conceptual, que facilite la comunicación, respecto de la especificación de requisitos de calidad de datos en los procesos de negocio, entre los expertos de negocio y los expertos en TI.

1.3. Objetivos específicos

Los objetivos específicos definidos para alcanzar el objetivo global de la tesis son los siguientes:

1. Estudiar la literatura para establecer la terminología utilizada por los expertos y/o analistas de negocio para expresar problemas y/o requisitos de DQ en el contexto de los BP.
2. Identificar conceptos y el vocabulario utilizado para describir requisitos y/o problemas de DQ en los BP, por parte de los expertos de negocio, mediante entrevistas.
3. Definir una metodología y una herramienta para el desarrollo de la ontología a realizar.
4. Desarrollar una ontología de calidad de datos para el contexto de los procesos de negocio.

1.4. Alcance de la investigación

La investigación se centrará en la creación de una ontología que permita definir y relacionar conceptos, términos, dimensiones y atributos para representar requisitos y problemas de calidad de datos en procesos de negocio.

Se usará como referencia del lenguaje de los expertos de TI la norma ISO/IEC-25012 (ISO/IEC-FDIS-25012, 2008), que establece un modelo de calidad de datos, el cual, es aplicable a aquellos datos que se encuentran almacenados de manera estructurada y forman parte de un Sistema de Información.

Finalmente, esta tesis se enmarcará en la primera etapa del método BPiDQ (Rodríguez y Caro, 2012) que contempla el modelado de procesos de negocio consciente de la calidad de datos.

1.5. Estructura del documento

El resto del documento se organiza de la siguiente manera, en el capítulo 2 se describe la metodología de trabajo empleada en la tesis. Luego, en el capítulo 3 se presenta el estado del arte, el cual, incluye el marco conceptual. Después, en el capítulo 4 se expone la identificación de la terminología de calidad de datos y en el capítulo 5 se da a conocer el desarrollo de la ontología. Finalmente, en el capítulo 6 se muestran las conclusiones finales.

Capítulo 2

Metodología de trabajo

Al modelar procesos de negocio es importante asegurar la DQ desde un comienzo, en esta tarea participan actores con distintos puntos de vistas, algunos centrados en el negocio y otros en el aspecto más técnico. Estos suelen utilizar terminologías diferentes para expresar situaciones similares. Lo anterior puede generar discrepancias e interpretaciones erróneas, por esta razón, se ha propuesto desarrollar una ontología que ayude a mejorar la comunicación entre estos.

Para abordar el desarrollo de la ontología se establecen las siguientes actividades:

- Identificación de la terminología utilizada por los analistas y/o expertos del negocio para manifestar problemas y características de Calidad de Datos, con las cuales ellos podrían expresar requisitos de DQ.
- Validación de la terminología de DQ identificada con expertos de negocio mediante un instrumento de consulta.
- Desarrollo de la ontología de calidad de datos en el contexto de los procesos de negocio.

Para el desarrollo de las actividades mencionadas anteriormente se han usado diversos métodos y/o metodologías que se describen a continuación de acuerdo a la actividad en que fueron utilizadas.

2.1. Identificación de la terminología de calidad de datos

Para identificar la terminología de DQ utilizada por personas del negocio, se decidió indagar la forma en que ellos expresan los problemas y las características de DQ. Para esto se definió usar el método de Revisiones Sistemáticas de la Literatura (RSL) (Kitchenham, 2004) para la revisión de literatura científica del área de negocios, el cual se describe en forma detallada en la Subsección 2.1.1 y una variante de este, para la revisión de foros, portales y blogs en la Web dedicados a temas de negocio, el cual se describe en forma detallada en la Subsección 2.1.2. En la primera etapa se definieron los objetivos y la interrogante de investigación. Luego en la segunda etapa correspondiente a la RSL, que abordó la búsqueda de literatura científica y en una siguiente etapa se realizó una Revisión Sistemática en la Web (RSW), que corresponde a la búsqueda en sitios Web. Luego de esto, se procesaron los resultados de ambas búsquedas, obteniendo una terminología preliminar. Finalmente, dicha terminología fue validada, obteniendo como resultado una lista de características de DQ y una lista de tipos de problemas de DQ expresados en un lenguaje en el ámbito del negocio. El método para validar la terminología se describe en forma detallada en la Subsección 2.1.3.

2.1.1. Revisión Sistemática de la Literatura

Una revisión sistemática de la literatura es una manera de indagar y analizar todos los trabajos disponibles relacionados a una interrogante de investigación concreta de un área de interés. Los trabajos individuales que forman parte de una RSL se denominan estudios primarios y una RSL se considera un estudio secundario (Kitchenham, 2004).

Esta metodología propone 3 etapas principales: comenzando con la planificación, luego el desarrollo de la revisión y finalizando con la publicación de resultados. Por cada una de las etapas se incluyen tareas específicas las cuales se pueden ver en la Tabla 2.1.

Etapa 1: Planificación de la revisión
Identificar la necesidad de revisión
Definir un protocolo de revisión
Etapa 2: Desarrollo de la revisión
Identificación de la Investigación
Selección de los estudios primarios
Extracción y monitoreo de datos
Síntesis de datos
Etapa 3: Publicación de los resultados

Tabla 2.1: Etapas de una RSL (Kitchenham, 2004)

En la Sección 4.2 del Capítulo 4 se muestra una descripción detallada de cómo se aplicó esta metodología.

2.1.2. Revisión Sistemática de la Web

La revisión sistemática de la Web es una adaptación de la RSL (Kitchenham, 2004) para desarrollar una búsqueda abierta y de esa manera llegar a sitios Web especializados en la temática de interés.

Esta metodología posee 3 etapas: la planificación de la búsqueda, el desarrollo de la búsqueda y la publicación de los resultados. Cada una de las etapas tiene tareas específicas que se muestran en la tabla 2.2. El detalle de su aplicación se puede apreciar en la Sección 4.3 del Capítulo 4.

2.1.3. Validación de la terminología de calidad de datos

Para validar la terminología de DQ se aplicó el método de encuestas. Una encuesta es un método empírico que se utiliza para recopilar información de o sobre personas para describir,

Etapa 1: Planificación de la BSW
Identificar la necesidad de búsqueda Definir un protocolo de búsqueda
Etapa 2: Desarrollo de la BSW
Identificación de la Investigación Selección de las páginas Web Extracción y gestión de la información Síntesis de datos
Etapa 3: Publicación de los resultados

Tabla 2.2: Etapas de la revisión sistemática de la web

comparar o explicar su conocimiento, sus actitudes o su comportamiento. También se puede utilizar encuestas para la descripción de las características de métodos y herramientas (Genero et al., 2014).

La encuesta para validar la terminología de calidad de datos del ámbito de los negocios tiene como fin consultar a los expertos y/o analistas de negocio sobre la utilización de la terminología encontrada en la literatura y en sitios especializados de negocio.

En base a lo establecido en (Genero et al., 2014) referente a encuestas se definió una serie de actividades para la preparación de esta que se pueden apreciar en la tabla 2.3.

Actividades
La definición de objetivos del estudio
El diseño del instrumento
La preparación del instrumento
La validación del instrumento
La elección de los participantes
La definición del método de aplicación
La aplicación del estudio
El análisis de los resultados del estudio
La difusión de los resultados

Tabla 2.3: Actividades para el desarrollo de la encuesta

El detalle del desarrollo de cada una de las actividades definidas de la encuesta para la validación de la terminología de DQ se puede apreciar en la Sección 4.5 y 4.6 del Capítulo 4.

2.2. Desarrollo de la ontología

La metodología elegida para el desarrollo de la ontología es METHONTOLOGY (Juristo et al., 1997), la cual proporciona una guía sobre cómo llevar a cabo su desarrollo estableciendo en forma

general las siguientes actividades:

- **La especificación**, que permite determinar por qué se construye la ontología, cuál será su uso, y quiénes serán sus usuarios finales.
- **La conceptualización**, que se encarga de organizar y convertir una percepción informal del dominio en una especificación semi-formal, para lo cual utiliza un conjunto de representaciones intermedias, basadas en notaciones tabulares y gráficas, que pueden ser fácilmente comprendidas por los expertos de dominio y los desarrolladores de ontologías. El resultado de esta actividad es el modelo conceptual de la ontología.
- **La formalización**, que se encarga de la transformación de dicho modelo conceptual en un modelo formal o semi-computable.
- **La implementación**, donde se construyen modelos computables en un lenguaje de ontologías (RDF Schema, OWL, etc.). La mayor parte de las herramientas de ontologías permiten llevar a cabo esta actividad de manera automática.
- **Finalmente el mantenimiento**, que se encarga de la actualización y/o corrección de la ontología, en caso necesario.

Como ayuda para el desarrollo del modelado de la ontología, la metodología METHONTOLOGY propone varios componentes: los conceptos, las relaciones, las constantes, los atributos, las axiomas y las reglas. Estos componentes se explican a continuación.

- **Los conceptos:** son objetos o entidades, considerados desde un punto de vista amplio. Los conceptos de una ontología están normalmente organizados en taxonomías en las cuales se pueden aplicar mecanismos de herencia.
- **Las relaciones:** permiten representar un tipo de asociación entre conceptos del dominio.
- **Las instancias:** se utilizan para representar individuos en la ontología.
- **Las constantes:** son valores numéricos que no cambian en un largo período de tiempo.
- **Los atributos:** permiten describir propiedades. Se pueden distinguir dos tipos de atributos: de instancia y de clase.
- **Los axiomas formales:** que son expresiones lógicas siempre verdaderas que suelen utilizarse para definir restricciones en la ontología.
- **Las reglas:** se utilizan normalmente para inferir conocimientos en la ontología, tales como valores de atributos, instancias de relaciones, etc.

La metodología propone varias tareas para desarrollar la conceptualización de la ontología, un resumen de estas se pueden visualizar en la Fig. 2.1 y se describen en detalle a continuación.



Fig. 2.1: Tareas de la metodología METHONTOLOGY, extraído de (Corcho et al., 2005)

Tarea 1: construir el glosario de términos. El desarrollador de la ontología en esta tarea debe construir un glosario de términos que incluye todos los términos relevantes del dominio (conceptos, instancias, atributos, relaciones entre conceptos, etc.), sus descripciones en lenguaje natural y sus sinónimos.

Tarea 2: construir taxonomías de conceptos. Una vez que el glosario de términos contiene suficientes términos, el desarrollador de la ontología debe construir la taxonomía de concepto. Para ello, se seleccionan términos del glosario del tipo conceptos y se analiza cómo estos se vinculan utilizando los siguientes tipos de relaciones:

- Subclase-de, un concepto C_1 es Subclase-de un C_2 si y solo si todas las instancias de C_1 también son instancias de C_2 .
- Descomposición-Disjunta de un concepto C es un conjunto de sub-conceptos de C que no tienen instancias comunes y que no cubren C , es decir, puede haber instancias del concepto C que no son instancias de ninguno de los conceptos que forman la descomposición.
- Descomposición-Exhaustiva de un concepto C es un conjunto de sub-conceptos de C que lo cubren, es decir, tal que no existe ninguna instancia de C que no sea instancia de al menos uno de los conceptos de la descomposición, los conceptos que pertenecen a este conjunto pueden tener instancias y sub-conceptos comunes.
- Partición de un concepto C es un conjunto de sub-conceptos de C que no tienen instancias ni sub-conceptos comunes y que cubren C .

Tarea 3: construir diagramas de relaciones binarias ad hoc. Una vez construida y evaluada la taxonomía, la actividad de conceptualización propone construir diagramas de relaciones binarias ad hoc. El objetivo de este diagrama es establecer las relaciones ad hoc existentes entre conceptos de la misma o de distintas taxonomías de conceptos.

Tarea 4: construir el diccionario de conceptos. Una vez que las taxonomías de conceptos y los diagramas de relaciones binarias ad hoc se han generado, se debe especificar cuáles son las propiedades que describen cada concepto de la taxonomía, así como las relaciones identificadas en el diagrama de la tarea anterior y las instancias de cada uno de los conceptos.

Tarea 5: describir las relaciones binarias ad hoc. El objetivo de esta tarea es describir en detalle todas las relaciones binarias ad hoc identificadas en el diagrama de relaciones binarias e incluidas en el diccionario de conceptos. Para cada relación binaria ad hoc, se debe especificar su nombre, los nombres de sus conceptos de origen y destino, su cardinalidad y su relación inversa, si existe.

Tarea 6: describir los atributos de instancia. El objetivo de esta tarea es describir en detalle todos los atributos de instancia incluidos en el diccionario de conceptos, donde cada fila de la tabla contiene la descripción detallada de estos. Por cada atributo de instancia, se debe especificar su nombre, el concepto al que pertenece el atributo (los atributos son locales a los conceptos), su tipo de valor, su rango de valores (en el caso de atributos numéricos), y sus cardinalidades mínima y máxima.

Tarea 7: describir los atributos de clase. El objetivo de esta tarea es describir en detalle todos los atributos de clase incluidos en el diccionario de conceptos. Para cada atributo, se debe rellenar la siguiente información: nombre del atributo, nombre del concepto, donde el atributo se define, tipo de valor, cardinalidad y valor.

Tarea 8: describir las constantes. El objetivo de esta tarea es describir en detalle cada una de las constantes identificadas en el glosario de términos. Para cada constante, se debe especificar su nombre, tipo de valor, valor y unidad de medida en el caso de constantes numéricas.

Tarea 9: definir las axiomas formales. Para realizar esta tarea, el desarrollador de ontologías debe identificar los axiomas formales que son necesarios en la ontología y describirlos de manera precisa. Para cada definición de axioma formal, se propone especificar la siguiente información: nombre, descripción en lenguaje natural, expresión lógica que define de manera formal el axioma usando lógica de primer orden, y los conceptos, atributos y relaciones ad hoc utilizadas en el axioma, así como las variables utilizadas.

Tarea 10: definir las reglas. De manera similar a la tarea previa, en esta tarea el desarrollador de la ontología debe identificar en primer lugar qué reglas se necesitan en la ontología, y describirlas en una tabla. Para cada regla, se propone incluir la siguiente información: nombre, descripción en lenguaje natural, expresión que describe formalmente la regla, y

conceptos, atributos y relaciones ad hoc utilizados en la regla, así como las variables usadas. METHONTOLOGY propone especificar las expresiones de las reglas utilizando el formato si <condiciones> entonces <consecuente>. La parte izquierda de la regla es una conjunción de condiciones simples, mientras que la parte derecha es una simple expresión de un valor de la ontología.

Tarea 11: describir las instancias. Una vez que el modelo conceptual de la ontología ha sido creado, se pueden definir las instancias que aparecen en el diccionario de conceptos. Para cada instancia se define: su nombre, el nombre del concepto al que pertenece y los valores de sus atributos de instancia, si se conocen.

Las tareas para el desarrollo de la conceptualización de la ontología mencionadas anteriormente se desarrollaron en el Capítulo 5 donde se muestra una descripción detallada de su ejecución.

Capítulo 3

Estado del arte

El capítulo se estructura de la siguiente manera, comienza con un marco conceptual en el cual se describen las temáticas de interés. En la Sección 3.1 se presentan los procesos de negocio, en la Sección 3.2 la calidad de datos, en la Sección 3.3 las ontologías y en la Sección 3.4 el método BPiDQ. Luego en la Sección 3.5 se desarrolla el estado del arte, donde se expone lo existente en la literatura que relaciona las ontologías, la calidad de datos y los procesos de negocio.

3.1. Procesos de negocio

En esta sección se presentan los aspectos generales de los procesos de negocio, sus definiciones y notaciones.

En forma genérica se puede definir un proceso como: *“una concatenación lógica de actividades, a través del tiempo y lugar, impulsadas por eventos y que a través de su proceso de transformación cumplen un determinado fin”* (Freund et al., 2013).

Los procesos de negocio se pueden definir como un conjunto de actividades que toman uno o más tipos de entradas y crean una salida que es de valor para un cliente. Según la definición, un proceso esta obligado a la creación de valor para un cliente (Hitpass, 2012). Los BP pertenecen a una organización particular, pero pueden interactuar con los BP realizados por otras organizaciones (Weske, 2012).

El modelado de BP es ampliamente utilizado para documentar el funcionamiento de los BP dentro de una organización, mostrando una vista simplificada de estos, para difundirlo a los distintos actores que participan dentro de esta (Recker, 2010). Además, suele ser utilizado como referencia para el desarrollo de software, como parte de la preparación para la implementación de un ERP (Enterprise Resource Planning), y para satisfacer los requisitos de documentación de procesos, entre otros (Harmon y Wolf, 2011).

En la industria, los lenguajes de modelado de BP más utilizados que permiten describir BP de forma gráfica son Diagramas de Actividad UML (Unified Modeling Language) (OMG, 2007) y BPMN (Business Process Modelling Notation) (OMG, 2011), siendo BPMN la notación más utilizada en la industria lo que la ha convertido en el estándar *de facto* para el modelado de BP (Harmon y Wolf, 2011).

Los Diagramas de Actividad son uno de los tres diagramas de UML (Unified Modeling Language), junto con los Diagramas de Estado y los Diagramas de Secuencia, utilizados para la descripción del comportamiento dinámico de un sistema. Estos diagramas utilizan clases del metamodelo de UML que se encuentran en los paquetes de la especificación dedicados a la descripción de comportamientos (Pérez, 2007).

BPMN es una notación gráfica para representar los procesos de negocio. La notación ha sido diseñada específicamente para coordinar la secuencia de los procesos y los mensajes que fluyen

entre los diferentes participantes que realizan un conjunto relacionado de actividades (OMG, 2011).

3.2. Calidad de datos

En esta sección se define calidad de datos mostrando su importancia, además se hace referencia a la norma ISO/IEC 25012 que establece un modelo de calidad de datos.

La DQ se define como el grado en que las características de los datos satisfacen las necesidades implícitas establecidas cuando son utilizados en condiciones específicas (ISO/IEC-FDIS-25012, 2008). Además se ha determinado que la calidad de datos es un concepto multidimensional y subjetivo (Klein, 2001), ya que se ha establecido que los datos deben ser “apropiados para el uso” (Strong et al., 1997), lo que implica un contexto específico y la perspectiva de un usuario (o un grupo de estos) en particular. Algunas de las dimensiones o características que se han propuesto son las siguientes: exactitud, integridad, consistencia, credibilidad, inherente, actualidad, accesibilidad, entre otras, permitiendo una mayor precisión a la hora de fijar los requisitos de DQ (ISO/IEC-FDIS-25012, 2008). La importancia de la DQ radica en que se dice que sólo las organizaciones que logran incorporar una estrategia de gestión de DQ efectiva, en su estrategia de negocio, serán capaces de convertir sus datos en una ventaja competitiva real, entregando valor a corto y largo plazo para asegurar el éxito y la sustentabilidad de su negocio (el Abed, 2009). También se dice que los datos son el activo más importante de las organizaciones (Piattini y Del Peso, 2001), teniendo en cuenta eso, el cuidado de la DQ se convierte en una tarea importante dentro de estas.

3.2.1. Norma ISO/IEC 25012

La norma ISO/IEC 25012 (ISO/IEC-FDIS-25012, 2008) define un modelo de calidad de datos general para datos estructurados pertenecientes a un sistema de información. Esta norma se centra en la calidad de los datos como parte de un sistema de información y define características que deben poseer los datos que serán utilizados por sistemas y por seres humanos.

El modelo de calidad de datos definido en la norma describe 15 características de DQ de acuerdo con los puntos de vista inherente y dependiente del sistema, como se puede apreciar en la Tabla 3.1.

Las características de calidad de datos están clasificadas en dos grandes categorías.

Calidad de datos Inherente: Se refiere al grado con el que las características de calidad de los datos tienen el potencial intrínseco para satisfacer las necesidades establecidas y necesarias

Características	Calidad de datos	
	Inherente	Dependiente del sistema
Precisión	X	
Compleitud	X	
Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Conformidad	X	X
Confidencialidad	X	X
Eficiencia	X	X
Exactitud	X	X
Trazabilidad	X	X
Comprensibilidad	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

Tabla 3.1: Características de calidad ISO/IEC-25012

cuando los datos son utilizados bajo condiciones específicas. Desde el punto de vista inherente, la Calidad de Datos se refiere a los datos en sí, en particular a:

- Valores de dominios de datos y posibles restricciones (por ejemplo, reglas de negocio gobernando la calidad requerida por las características en una aplicación dada).
- Relaciones entre valores de datos (por ejemplo, consistencia).
- Metadatos.

Calidad de Datos Dependiente del Sistema: Se refiere al grado con el que la calidad de datos es alcanzada y preservada a través de un sistema informático cuando los datos se utilizan bajo condiciones específicas. Desde el punto de vista dependiente del sistema, la Calidad de Datos depende del dominio tecnológico en el que los datos se utilizan, y se alcanza mediante las capacidades de los componentes del sistema informático, tales como: dispositivos hardware (por ejemplo, respaldo software para alcanzar la recuperabilidad) y otro software (por ejemplo, herramientas de migración para alcanzar la portabilidad).

Consecuente con la clasificación definida anteriormente, se comenzará definiendo las características de calidad de datos inherentes:

- **Precisión:** Grado en el que los datos representan correctamente el verdadero valor del atributo deseado de un concepto o evento en un contexto de uso específico. Tiene dos principales aspectos:

- *Precisión Sintáctica*: Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado sintácticamente correcto.
- *Precisión Semántica*: Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado semánticamente correcto.
- **Completitud**: Grado en el que los datos asociados con una entidad, tienen valores para todos los atributos esperados e instancias de entidades relacionadas en un contexto de uso específico.
- **Consistencia**: Grado en el que los datos están libres de contradicción y son coherentes con otros datos en un contexto de uso específico. Puede ser analizada en datos que se refieran tanto a una, como a varias entidades comparables.
- **Credibilidad**: Grado en el que los datos tienen atributos que se consideran ciertos y creíbles en un contexto de uso específico. La credibilidad incluye el concepto de autenticidad (la veracidad de los orígenes de datos, atribuciones, compromisos).
- **Actualidad**: Grado en el que los datos tienen atributos que tienen la edad correcta en un contexto de uso específico.

Continuando con las definiciones ahora se presentarán de las características de calidad de datos que están clasificadas como inherentes y dependientes del sistema:

- **Accesibilidad**: Grado en el que los datos pueden ser accedidos en un contexto específico, particularmente por personas que necesiten tecnologías de apoyo o una configuración especial por algún tipo de discapacidad.
- **Conformidad**: Grado en el que los datos tienen atributos que se adhieren a estándares, convenciones o normativas vigentes y reglas similares referentes a la calidad de datos en un contexto de uso específico.
- **Confidencialidad**: Grado en el que los datos tienen atributos que aseguran que son sólo accedidos e interpretados por usuarios autorizados en un contexto de uso específico.
- **Eficiencia**: Grado en el que los datos tienen atributos que pueden ser procesados y proporcionados con los niveles de rendimiento esperados mediante el uso de cantidades y tipos adecuados de recursos en un contexto de uso específico.
- **Exactitud**: Grado en el que los datos tienen atributos que son exactos o proporcionan discernimiento en un contexto de uso específico.
- **Trazabilidad**: Grado en el que los datos tienen atributos que proporcionan un camino de acceso auditado a los datos o cualquier otro cambio realizado sobre los datos en un contexto de uso específico.

- **Comprensibilidad:** Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante meta-datos.

Finalizando con las definiciones, ahora corresponde presentar las características de calidad de datos clasificadas como dependientes del sistema:

- **Disponibilidad:** Grado en el que los datos tienen atributos que permiten ser obtenidos por usuarios y/o aplicaciones autorizados en un contexto de uso específico.
- **Portabilidad:** Grado en el que los datos tienen atributos que les permiten ser instalados, reemplazados o eliminados de un sistema a otro, preservando el nivel de calidad en un contexto de uso específico.
- **Recuperabilidad:** Grado en el que los datos tienen atributos que permiten mantener y preservar un nivel específico de operaciones y calidad, incluso en caso de fallas, en un contexto de uso específico.

3.3. Ontologías

En esta sección se presentará una descripción general de las ontologías, exponiendo sus definiciones y analizando sus funcionalidades. Además, se estudian las clasificaciones y las metodologías que han surgido para el desarrollo de estas.

3.3.1. ¿Qué es una ontología?

Lo primero que conviene establecer es que las ontologías son un campo de investigación de la inteligencia artificial y, más específicamente, de la rama relacionada con la representación del conocimiento, que se ocupa de la construcción de sistemas expertos (Nickles et al., 2007).

Las ontologías son un procedimiento basado en la lógica de primer orden desarrollado para codificar adecuadamente el sistema de términos utilizados en dichas declaraciones, de forma que se exprese adecuadamente la relación entre los términos. También se pueden definir como las herramientas para construir sistemas conceptuales, por utilizar una terminología común y vocabularios estructurados, en los que se explicitan todas las relaciones entre los términos que se utilizan y otras restricciones de significado (García-Marco, 2007).

El concepto de ontología ha estado presente desde hace mucho tiempo en la filosofía, recientemente se utiliza en informática para definir vocabularios que las máquinas puedan

entender y que son especificados con la suficiente precisión, como para permitir diferenciar términos y referenciarlos de manera precisa.

La definición declarativa más consolidada es la propuesta por Gruber que la describe como: “Una especificación explícita y formal sobre una conceptualización compartida” (Gruber, 1993).

El término ontología en informática hace referencia a la formulación de un exhaustivo y riguroso esquema conceptual dentro de uno o varios dominios dados; con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades. Aunque toma su nombre por analogía, esta es la diferencia con el punto de vista filosófico de la palabra ontología.

Las ontologías se pueden ver como modelos conceptuales que tratan de soportar de una manera no ambigua y consistente el conocimiento y su intercambio de un modo que permita la integración del mismo, mediante la unión de etiquetas conceptuales a sus interpretaciones (Spasic et al., 2005).

El significado de esta palabra ha cambiado durante el paso del tiempo, actualmente en el contexto de la Web Semántica puede intentar explicarse como “la especificación de una conceptualización”, de una forma muy corta y compleja. Para entenderlo un poco más, hay que pensar que una conceptualización es una abstracción, una visión simplificada del mundo que queremos representar para algún propósito, cada base de conocimiento, cada sistema basado en conocimiento, o cada agente de conocimiento está comprometido con alguna conceptualización, implícita o explícita (Gruber, 1993).

3.3.2. Objetivos características y usos de las ontologías

Las ontologías tienen como principales objetivos los siguientes (Noy y McGuinness, 2001):

- Compartir la comprensión común de las estructuras de información entre personas y/o agentes de software.
- Permitir la reutilización del conocimiento perteneciente a un dominio.
- Permitir hacer explícitos los supuestos de un dominio.
- Separar el conocimiento de un dominio del conocimiento que se puede denominar operacional.
- Hacer posible analizar el conocimiento de un dominio determinado.

Dentro de las principales características de las ontologías están el suministrar un vocabulario compartido para expresar información, además incluir axiomas para especificar relaciones entre conceptos y trazar una estructura conceptual tan detallada y completa como sea necesario.

Los principales usos de las ontologías son los que se enumeran a continuación:

- Para favorecer la comunicación entre personas, organizaciones y aplicaciones. Porque proporcionan una comprensión común de un dominio, de modo que se eliminan confusiones conceptuales y terminológicas.
- Para lograr la interoperabilidad entre sistemas informáticos. Porque sirven para traducir los términos usados por una aplicación a otra.
- Para facilitar el razonamiento automático, es decir, sin intervención humana.
- Para la ingeniería de software. Donde ayudan a la especificación de los sistemas de software al establecer un entendimiento común entre los desarrolladores y los usuarios.

3.3.3. Clasificaciones de ontologías

Existe más de una clasificación para las ontologías, propuesta en la literatura. Estas clasificaciones pueden ser por: la cantidad y el tipo de estructura de lo que se quiere conceptualizar, el nivel de generalidad, el mundo real que se quiere modelar, la riqueza del tipo de estructura a modelar, entre otros. Siendo las principales las dos que se describen a continuación.

Clasificación de ontologías de acuerdo a la riqueza de su estructura interna (Aranda y Ruiz, 2005):

- **Vocabularios controlados:** Lista finita de términos, por ejemplo, un catálogo.
- **Glosarios:** Son listas de términos con sus significados expresados en lenguaje natural.
- **Tesauros:** Proveen semánticas adicionales entre términos, como por ejemplo información referida a sinónimos.
- **Jerarquías Informales “Es-Un”:** Son jerarquías de términos que no corresponden a una subclase estricta, por ejemplo los términos “auto de alquiler” y “hotel” podrían ser modelados informalmente bajo la jerarquía “viaje” ya que se considerarían partes clave de un viaje.
- **Jerarquías Formales “Es-Un”:** En este caso existe una relación estricta entre instancias de una clase y de las superclases correspondientes. Su objetivo es explotar el concepto de herencia.
- **Marcos:** Son ontologías que incluyen tanto clases como sus propiedades, las cuales pueden ser heredadas por otras clases en los niveles más bajos de una taxonomía formal “es-un”.
- **Ontologías que expresan restricciones de valor:** Por ejemplo expresan restricciones de acuerdo al tipo de dato de una propiedad (por ejemplo, tipo fecha).

- **Ontologías que expresan restricciones lógicas generales:** Son las ontologías más expresivas. Especifican, por medio de lógica de primer orden, restricciones entre los términos de la ontología.

Clasificación de ontologías de acuerdo al sujeto de conceptualización (Aranda y Ruiz, 2005).

- **Ontologías de representación de conocimiento:** Capturan primitivas de representación utilizadas para formalizar conocimiento bajo un paradigma de representación de conocimiento dado.
- **Ontologías comunes o generales:** Representan conocimiento de sentido común y reutilizable en distintos dominios, por ejemplo sobre vocabulario relacionado a cosas, eventos, tiempo, espacio, etc.
- **Ontologías de alto nivel:** Son ontologías que describen conceptos y nociones generales bajo las cuales pueden enlazarse los términos raíces de todas las ontologías. Un problema existente es que hay varias de estas ontologías de alto nivel que difieren en los criterios para clasificar la mayoría de los conceptos generales de la taxonomía.
- **Ontologías de dominio:** Son aquellas ontologías reutilizables en un dominio particular (medicina, ingeniería, etc.). Proveen vocabularios sobre conceptos dentro del dominio y sus relaciones.
- **Ontologías de tareas:** Describen vocabulario relacionado a actividades genéricas. Proveen un vocabulario sistemático de términos utilizados para resolver problemas que pueden o no pertenecer a un mismo dominio.
- **Ontologías de tareas de dominios:** A diferencia de las ontologías de tareas, estas ontologías son re utilizables en un dominio dado, y no entre dominios diferentes.
- **Ontologías de métodos:** Proveen definiciones de conceptos relevantes y relaciones aplicables a un proceso de razonamiento específico a fin de cumplir una tarea particular.
- **Ontologías de aplicaciones:** Son dependientes de las aplicaciones. A menudo extienden y especializan vocabulario de una ontología de dominio o de tareas para una aplicación particular.

3.3.4. Metodologías para la creación de ontologías

Para el diseño y desarrollo de cualquier ontología es necesario contar con una metodología específica que se adecue con las necesidades de desarrollo. Existen en la literatura una gran variedad de métodos y metodologías para el desarrollo de ontologías (Guzmán et al., 2012b), a continuación se muestran algunas de las más importantes.

La metodología CYC: Fue publicada por Lenat y Guha desde 1990 (Lenat y Guha, 1993), en la que divulgaron algunos pasos generales para la construcción de ontologías; el primero consiste en extraer manualmente el conocimiento común que está implícito en diferentes fuentes para después, cuando se tenga suficiente conocimiento en la ontología, adquirir nuevo conocimiento común usando herramientas de procesamiento de lenguaje natural o aprendizaje computacional.

La Metodología de USCHOLD Y KING: En 1995, con base en la experiencia recaudada en el desarrollo de una ontología de la Empresa, surge la Metodología de USCHOLD Y KING (Uschold y King, 1995) que recrean una serie de pasos que permiten plasmar y especificar los conocimientos que se tienen sobre un dominio específico, centrandos sus esfuerzos en la forma en la cual representar los conocimientos. Entre sus pasos para desarrollar ontologías propone: (1) identificar el propósito; (2) capturar los conceptos y relaciones entre estos conceptos y los términos utilizados para referirse a estos conceptos y relaciones; (3) codificar la ontología. La ontología debe ser documentada y se pueden usar otras ontologías para crear la nueva.

La metodología de GRÜNINGER Y FOX: La metodología de GRÜNINGER Y FOX (Grüninger y Fox, 1995) tiene como primer paso el identificar intuitivamente las aplicaciones posibles en las que se usará la ontología. Luego, se utilizan un conjunto de preguntas en lenguaje natural, llamadas preguntas de competencia, para determinar el ámbito de la ontología. Estas se usan para extraer los conceptos principales, sus propiedades, relaciones y axiomas, los cuales se definen formalmente en Prolog.

La metodología METHONTOLOGY: La metodología METHONTOLOGY (Gómez-Pérez et al., 1996) es una de las propuestas más completas, ya que toma la creación de ontologías como un proyecto informático. Además de las actividades propias de la generación de la ontología, esta abarca actividades para la planificación del proyecto, la calidad del resultado, la documentación, entre otras. Permitiendo construir ontologías totalmente nuevas o reutilizar otras ontologías. El entorno incluye la identificación del proceso de desarrollo de la ontología donde se incluyen las principales actividades (evaluación, conceptualización, configuración, integración, implementación, etc.), un ciclo de vida basado en prototipos evolucionados y la metodología propiamente dicha, que especifica los pasos a ejecutar en cada actividad, las técnicas usadas, los productos a obtener y su forma de evaluación. Esta metodología está parcialmente soportada por el entorno de desarrollo ontológico WebODE y propone las siguientes etapas: (1) especificación, (2) conceptualización, (3) formalización, (4) implementación y (5) mantenimiento.

La metodología SENSUS: En 1997, se propuso un nuevo método para construir ontologías, este estaba basado en la ontología de SENSUS (Swartout et al., 1996). La cual constituye un enfoque top-down para derivar ontologías específicas del dominio a partir de grandes ontologías. En esta metodología se identifican un conjunto de términos semilla que son relevantes en un dominio particular. Tales términos se enlazan manualmente a una ontología de amplia cobertura. Los usuarios seleccionan automáticamente los términos relevantes para describir el dominio y

acotar la ontología Sensus. Consecuentemente, el algoritmo devuelve el conjunto de términos estructurados jerárquicamente para describir un dominio, que puede ser usado como esqueleto para la base de conocimiento.

La metodología ON-TOKNOWLEDGE: Algunos años más tarde, aparece la metodología de ON-TOKNOWLEDGE como resultado del proyecto con el mismo nombre (Staab et al., 2001). Esta aplica ontologías a la información disponible electrónicamente para mejorar la calidad de la gestión de conocimiento en organizaciones grandes y distribuidas. Además, incluye la identificación de metas que se deberían conseguir por medio de herramientas de gestión de conocimiento y está basada en el análisis de escenarios de uso y en los diferentes papeles desempeñados por trabajadores de conocimiento y accionistas en las organizaciones.

La metodología TERMINAE: En el 2002, aparece la metodología TERMINAE (Corcho et al., 2003) la cual aporta tanto una metodología como una herramienta para la construcción de ontologías a partir de textos. Se basa en un análisis lingüístico de los textos, el cual se realiza mediante la aplicación de diferentes herramientas para el procesamiento del lenguaje natural. En particular, se usan dos herramientas: (1) Syntex para identificar términos y relaciones; y (2) Caméléon para identificar roles o relaciones.

3.3.5. Lenguajes para crear ontologías

Según (Garcia, 2004; Guzmán et al., 2012b; Pereira y Aular, 2007) los lenguajes más usados para la definición de ontologías son, DAML (DARPA Agent Mark-Up Language) + OIL(Ontology Inference Layer), OWL (Ontology Web Language), XML(Extensible Markup Language) y RDF(Resource Description Framework), los cuales, se describen a continuación.

DAML + OIL: En inglés el significado de su sigla es (DARPA Agent Markup Language) + (Ontology Inference Layer). Su objetivo es capturar las primitivas del modelo que proveen un adecuado poder expresivo a la ontología, especificando su semántica y haciendo viable la inferencia, además de describir e intercambiar datos en la red distribuida.

Respecto de su aplicabilidad se utiliza en el área de Inteligencia Artificial (IA) para facilitar compartir y rehusar el conocimiento.

OWL: En inglés el significado de sus siglas es Web Ontology Language y este deriva de DAMAL + OIL. Su objetivo es proveer un mayor conjunto de primitivas para representar el significado de los elementos y sus relaciones con otros elementos en una ontología, permitiendo a los usuarios dar una mayor expresividad a sus sentencias y tener la libertad de usar la sintaxis en RDF (Resource Description Framework).

Con respecto a su aplicabilidad y uso, se utiliza en la Web Semántica y ontologías, integración empresarial y brinda apoyo a las decisiones médicas, en la creación de contenidos, herramientas para la administración de sitios Web, reutilización y compartición de datos.

XML: El significado de sus siglas en inglés es *eXtensible Markup Language* y consiste en un formato estándar para la estructuración de datos. Dentro de los objetivos del lenguaje están el permitir la interoperabilidad, puesto que todo el mundo puede crear sus propios vocabularios basados en una sintaxis para documentos potentes y flexibles, no imponiendo restricciones semánticas al significado de esos documentos, además de ser un metalenguaje.

RDF: El significado de sus siglas en inglés es *Resource Description Framework* y corresponde a un modelo para definir relaciones semánticas entre distintas URLs, basados en la sintaxis de XML. Dentro de los objetivos de RDF esta proporcionar interoperabilidad semántica entre aplicaciones que intercambian información entendible y legible por máquinas, en diferentes comunidades para la descripción de recursos como las páginas Web.

3.3.6. Herramientas para el desarrollo de ontologías

Según (Capote, 2013; Contreras y Martínez-Comeche, 2007) las herramientas más utilizadas para desarrollar ontologías son Protégé, KAON, NeOn Toolkit, WebOnto y OntoEdit, estas se describirán a continuación.

Protégé: Es una herramienta de código abierto desarrollada por la Universidad de Stanford (Research Stanford Center for Biomedical Informatics, 2015), la cual tiene una comunidad de miles de usuarios. Aunque el desarrollo de Protégé ha sido impulsado principalmente por aplicaciones biomédicas, el sistema es independiente del dominio y se ha utilizado con éxito en muchas otras áreas de aplicación (Gennari et al., 2003). Esta herramienta se encuentra separada en dos partes, una de modelo y otra de vista. El modelo es el mecanismo de representación interna de ontologías y bases de conocimiento. Los componentes de la vista proveen al usuario una interfase para visualizar y manipular el modelo. El modelo de Protégé está basado en un simple metamodelo comparable con sistemas orientados a objetos y sistemas basados en marcos, estos básicamente pueden representar ontologías que consisten en clases, propiedades, características de propiedades e instancias. La ventaja de esta herramienta es que su metamodelo es una ontología y la estructura del metamodelo de Protégé habilita fácilmente la extensión y adaptación a otras representaciones. Usando las vistas de la interfase de usuario de Protégé, los diseñadores de ontologías básicamente crean clases y les asignan propiedades. Usando la función `resulting ontologies`, Protégé es capaz de generar automáticamente interfaces de usuario que soporten la creación de instancias. En cada clase de la ontología el sistema crea un formulario con componentes de edición para cada propiedad de la clase. Protégé actualmente puede ser usado para cargar, editar y guardar ontologías en varios formatos como RDF, XML y OWL (Fernandez, 2013).

KAON: Es una infraestructura para la gestión de ontologías específica para aplicaciones de negocios, esta incluye una suite completa de herramientas que permite la creación y gestión de

ontologías. Además, los mecanismos de persistencia de KAON se inspiran en bases de datos relacionales (Volz, 2015).

NeOn Toolkit: Es un kit de herramientas de código abierto, multiplataforma, que proporciona soporte completo para el ciclo de vida del desarrollo de una ontología. Este kit se basa en la plataforma Eclipse, un entorno de desarrollo que conduce, y proporciona un amplio conjunto de plug-ins (actualmente 45 plug-ins están disponibles) (Harth, 2014).

WebOnto: Es una herramienta que proporciona la visualización, navegación y edición, para el desarrollo y el mantenimiento de ontologías y modelos de conocimiento, basado en la Web. En WebOnto, una ontología puede ser vista, como un modelo de la estructura conceptual de un dominio, proporcionando los medios para su representación gráfica (Domingue, 1998).

OntoEdit: OntoEdit es una herramienta de edición de ontologías que apoya el desarrollo y mantenimiento de las mismas, utilizando medios gráficos en un entorno Web. Permitiendo la representación semántica de lenguajes conceptuales y estructuras mediante conceptos, jerarquías de conceptos, relaciones y axiomas (Carlos Carrascosa y Julián, 2005).

3.4. Método BPiDQ

En esta sección se presenta en forma general el método, sus componentes y sus etapas. El BPiDQ es un método para la obtención de requisitos de software de DQ desde las especificaciones de procesos de negocio.

Tiene como objetivo soportar la especificación temprana de requisitos de calidad de datos en BP y en base a dicha especificación obtener requisitos de DQ en forma de casos de uso.

El método posee tres componentes que apoyan el desarrollo de sus etapas, los cuales se describen a continuación.

1. La extensión dqBP (Rodríguez et al., 2012) que tiene por objetivo agregar capacidad expresiva a la notación BPMN 2.0, permitiendo la representación de requisitos de DQ en un modelo de BP.
2. El repositorio de actividades de DQ que contiene actividades en el nivel de BP orientadas a satisfacer requisitos de DQ.
3. El repositorio de casos de uso de DQ que contiene los casos de uso estándar para cada dimensión de DQ que puede ser especificada como requisito de DQ en un BP.

El método BPiDQ consta de 4 etapas, tal como muestra la Fig. 3.1, éstas inician con el modelado de un BP y termina con la generación de diagramas de casos de uso.



Fig. 3.1: Las 4 etapas del método BPiDQ.

A continuación se describe brevemente cada una de las etapas del método.

La etapa 1, modelado de procesos de negocio consciente de la calidad de datos, que está dedicada a la captura temprana de requisitos de DQ, los que son representados en un modelo de BP a nivel descriptivo de BPMN, incorporando marcas donde se estime que la calidad de los datos involucrados en el BP es relevante para el éxito del negocio.

La etapa 2, especificación de requisitos de calidad de datos, tiene como objetivo obtener una especificación detallada de los requisitos de DQ definidos en el proceso de negocio, siendo su entrada el modelo de BP con requisitos de DQ generado en la etapa anterior.

La etapa 3, análisis y mejora de BP relacionado con la calidad de datos, es donde se analiza y deciden las mejoras que se pueden hacer al modelo del BP teniendo en cuenta los requisitos de DQ especificados.

En la etapa 4, generación de diagramas de casos de uso, por parte de los analistas de sistemas y los expertos en DQ, usando las actividades relacionadas con la DQ, para seleccionar el conjunto de casos de uso de DQ estándar, teniendo como resultado un conjunto de casos de uso que pueden ser usados en un proceso de desarrollo de software.

3.5. Ontologías, Calidad de Datos y Procesos de Negocio

En esta sección se desarrollará el estado del arte de la tesis, en concreto los trabajos de interés para esta investigación son los que relacionan las ontologías con la calidad de datos y los procesos de negocio. Tras el desarrollo de una RSL se constató que no existen trabajos que relacionen estas tres temáticas. Sin embargo, se encontraron trabajos que abordan las temáticas mencionadas por separado. En la Subsección 3.5.1 se muestran los trabajos relacionados con ontologías y calidad de datos. Luego en la Subsección 3.5.2 se presentan trabajos que se relacionan con ontologías y procesos de negocio, por último en la Subsección 3.5.3 se abordan trabajos referentes a calidad de datos y procesos de negocio.

3.5.1. Ontologías y Calidad de Datos

En (Oliveira et al., 2005a) se presenta una taxonomía de los problemas de calidad de datos producidos en las bases de datos utilizadas en el mundo real. En esta taxonomía se organizan los problemas en diferentes niveles de abstracción. Además define métodos para detectar los problemas de calidad de datos. Estos métodos son representados como árboles de clasificación binaria, con la idea subyacente de que se puedan utilizar para detectar de forma semiautomática los problemas DQ que puedan existir.

Por otra parte, en (Oliveira et al., 2005b) se propone una taxonomía de problemas de calidad de datos, organizados por niveles de granularidad de ocurrencia. Por cada problema incluido se presenta una definición formal. Además la taxonomía proporciona definiciones rigurosas que son útiles para el desarrollo de una herramienta de calidad de datos que detecte automáticamente problemas identificados.

Dentro de estos trabajos están la taxonomía de calidad de datos y la taxonomía de problemas de calidad de datos, pero estos abordan la DQ de una manera muy técnica en comparación con las intenciones de esta investigación.

3.5.2. Ontologías y Procesos de Negocio

En esta subsección se presentarán trabajos relacionados con las ontologías y los procesos de negocios.

Respecto de ontologías relacionadas con los BP, en la literatura se encontraron trabajos como (Jenz, 2003) cuyo propósito es establecer un lenguaje común entre las personas encargadas de negocio y las personas encargadas de las TI, eliminando las brechas semánticas a la hora de diseñar los procesos de negocio de una organización. El medio con el cual se desarrolló esto, es una ontología para el intercambio de información en el mismo “lenguaje”. Esta ontología de procesos de negocio fue creada con dos finalidades distintas. La primera es, hacer que el conocimiento sea explícito permitiendo el intercambio de conocimientos entre los expertos del dominio y la gente de TI que participa en el diseño y desarrollo de software. La segunda es como una herramienta para la especificación de requisitos a partir del cual se puede generar una serie de artefactos de software.

Relacionado con la misma temática está (Hepp y Roman, 2007), un trabajo que esboza los requisitos para la representación semántica de Business Process Management, proponiendo un conjunto de ontologías. Entre ellas, una ontología de proceso superior que captura las nociones básicas de los procesos y de flujo de trabajo para un proceso. También una ontología de organización superior que proporciona el vocabulario y limitaciones para describir el entorno en el que los procesos se llevan a cabo. Además, una ontología de áreas funcionales superior que

incluye grupos funcionales típicos (contabilidad, venta) y una ontología de datos superior que describe las fuentes y los receptores de datos disponibles en una empresa, los metadatos y los eventos que generan o consumen datos.

Por otra parte, está (Rospocher et al., 2014) un trabajo que hace una descripción formal de una ontología de la notación de modelado de procesos de negocio (BPMN), esta ontología ofrece una clasificación de todos los elementos de BPMN, junto con la descripción formal de los atributos y cómo los elementos de BPMN se pueden combinar. La ontología se desarrolló con el lenguaje OWL-DL y se realizó basándose en la especificación de BPMN de la OMG.

Luego en (vom Brocke et al., 2014) se propone una ontología de dominio para modelar la relación entre los recursos de TI y procesos de negocio con el fin de medir el valor de negocio de TI para la empresa Hilti Corporation. Esta ontología permite establecer relaciones entre los recursos de TI y las actividades de los procesos de negocio para modelarlas y con ello confeccionar análisis de costos, basado en un conjunto coherente de vinculaciones que existen entre los elementos de infraestructura de TI y procesos de negocio.

El análisis que se puede hacer de los trabajos encontrados que relacionan las ontologías y los procesos de negocio, es que existen experiencias en la literatura que demuestran que las ontologías pueden ayudar a mejorar la comunicación entre los expertos de negocio y los expertos en TI.

3.5.3. Calidad de Datos y Procesos de Negocio

En esta subsección se presentan los principales trabajos que relacionan procesos de negocio y calidad de datos.

En el trabajo (Sánchez-Serrano et al., 2009) se analiza la capacidad de BPMN para representar temas de calidad de datos e información en procesos de negocio. Además en él se expone el desarrollo de una extensión para BPMN donde se incluye un símbolo en forma de estrella llamada “DQDim” para representar alguna de las dimensiones de calidad de datos definidas en la norma ISO /IEC-25012.

Igualmente relacionado con herramientas y extensiones en (Heinrich et al., 2011a) se discute los resultados de una revisión para presentar el estado del arte del modelado de la DQ dentro de los modelos de BP, centrándose en las notaciones de modelado gráfico. Este estudio encontró 16 herramientas que permiten la descripción de algún tipo de aspecto DQ en los procesos de negocio, sin embargo, según el estudio ninguna de ellas admite satisfactoriamente modelar DQ dentro de los modelos de BP de forma gráfica.

El autor del artículo anterior en (Heinrich et al., 2011b) continúa con su trabajo, donde da a conocer el desarrollo de un modelador. Este modelador permite agregar características de calidad de datos a un elemento de la notación para representar procesos de negocio, de una manera visual, arrastrando y soltando los iconos relacionados con características o atributos para la definición de requisitos de calidad de datos.

Por otro lado, está (Ofner et al., 2012) que postula una conceptualización de la calidad de datos, en el contexto de la gestión de BP. Esto lo realizan con la intención de incluir un enfoque de calidad de datos en el re-diseño de BP. En sí, lo desarrollado es un método de evaluación de la DQ por medio de la definición de reglas y políticas sobre entidades de información (objetos de datos y atributos de datos) que son dependientes del contexto de la organización. Esto con la idea de guiar el proceso de toma de decisiones durante el re-diseño de procesos de negocio con la ayuda de una herramienta de modelado contemplada en el método.

Continuando con la idea de conceptualizar la DQ, en el estudio (Falge et al., 2012) se trabajó en la identificación de los requisitos y problemas de calidad de datos en los procesos de negocios colaborativos (Business Networking), con la idea de definir clases de datos y dimensiones de la calidad de datos adecuadas por rubro del negocio. Esto se realizó con la revisión de doce casos de estudios de empresas de diferentes sectores industriales (cadenas de suministro, comercio, mantenimiento y reparación , finanzas), a los que se le realizó un análisis cualitativo donde se codificaban las clases de datos y las dimensiones de calidad de datos para determinar los requisitos de calidad de datos, por cada caso de estudio que luego se clasifica por tipo de sector industrial, llegando a las siguientes conclusiones, (1) las normas de datos necesitan ser acordados en una fase inicial de un proyecto,(2) hay una gran tendencia en las organizaciones por la automatización

de procesos y (3) las tres principales dimensiones DQ para las organizaciones analizadas son “oportunidad”, “precisión” y “exhaustividad”.

En el trabajo (Nasution y Albarda, 2013) se proporciona una guía para la realización de mejora de procesos de negocio que se centran en la gestión de la calidad de la información. Como parte del trabajo se aborda el desarrollo de una investigación conceptual con el objetivo de definir los requisitos y parámetros de calidad de información y un caso de estudio. Además propone separar en dos partes el proceso para establecer calidad de datos. La primera la producción de información, y la segunda la entrega de información. El trabajo en sí centra su preocupación en la producción de información.

En (Glowalla y Sunyaev, 2013) se hizo una revisión respecto a los lenguajes de modelado de procesos utilizados dentro de las organizaciones. Esto con el objetivo de proponer dos enfoques de integración de DQ en modelos de procesos existentes en las organizaciones, que son la integración dentro del modelo y a través de un modelo. También, se evaluó el impacto en la complejidad los modelos de BP existentes al integrar calidad de datos y además se estudió la aplicabilidad de los patrones de reducción de complejidad de procesos disponibles.

Al analizar los trabajos, se pudo constatar que existe interés en los investigadores por abordar temáticas que relacionen los procesos de negocio y la DQ, sin embargo, las investigaciones actuales no se preocupan del léxico o la terminología que se utiliza para especificar DQ en los procesos de negocio, teniendo en consideración los diferentes actores que participan en el modelado o documentación de éstos (analistas de negocio y de TI).

Capítulo 4

Terminología de DQ

En este capítulo se describe de forma detallada la manera en que se llevó a cabo la identificación de la terminología de DQ utilizada por las personas de negocio. Para la realización de esta tarea, se utilizó el método descrito en la Sección 2.1 del Capítulo 2, el cual, se puede apreciar en la Fig. 4.1.

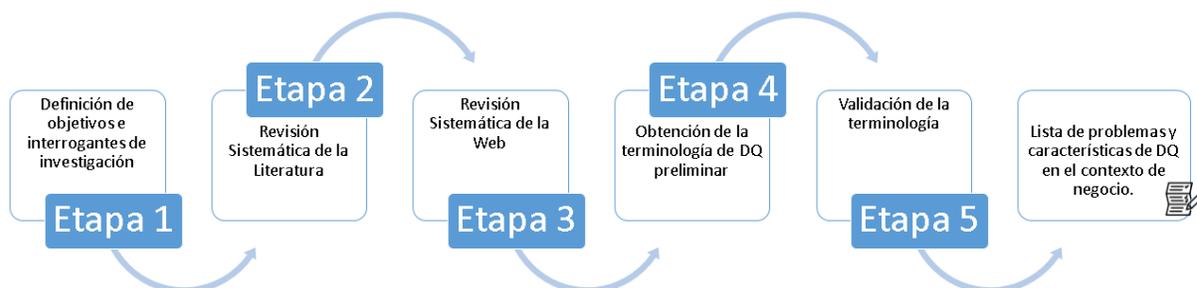


Fig. 4.1: Método para la identificación de la terminología de DQ

4.1. Objetivo e Interrogante de Investigación

El objetivo de investigación se definió como:

“Analizar estudios y trabajos existentes, en el área de negocios, que contengan la terminología usada para la definición de requisitos y problemas de DQ con lenguaje cercano al ámbito del negocio”.

En base a dicho objetivo se definió como interrogante de investigación:

“¿Qué términos se utilizan en el ámbito empresarial para expresar características y problemas de calidad de datos?”.

Una vez establecida la interrogante de investigación se comenzó con el desarrollo de la RSL y la RSW, tal como se describe a continuación.

4.2. Revisión Sistemática de la Literatura

Siguiendo la metodología RSL (Kitchenham, 2004), se desarrollaron cada una de sus etapas. Durante la **Planificación de la Revisión**, y en base al objetivo e interrogante de investigación planteados, se definió el protocolo de revisión. En este se consideraron como fuentes de búsqueda: Google académico, Scopus, Springer Link, SciELO y Science Direct. Además, se establecieron los términos de búsqueda y sus combinaciones para utilizarlos en los buscadores, que se muestran en las Tablas 4.1 y 4.2.

Se estableció un protocolo de revisión donde se determinaron los criterios de inclusión que se describen a continuación: (i) el título contiene uno de los términos definido para la búsqueda, (ii) dentro de las palabras claves del artículo se encuentra al menos uno de los términos definidos

Palabras claves (términos de búsqueda)	
Inglés	Español
“Data quality”	“Calidad de datos”
Requirements	Requisitos
“Information Quality”	“Calidad de información”
Problem	Problemas
Business	Negocios

Tabla 4.1: Términos de búsqueda para la RSL y RSW

Combinaciones	
Inglés	Español
Problem + “Data quality” + Business	Problemas + “Calidad de datos” + Negocios
Requirements + “Data quality” + Business	Requisitos + “Calidad de datos” + Negocios
Problem + “Information Quality” + Business	Problemas + “Calidad de información” + Negocios
Requirements + “Information Quality” + Business	Requisitos + “Calidad de información” + Negocios

Tabla 4.2: Combinaciones de términos de búsqueda para la RSL y RSW

para la búsqueda, en el caso de tratarse de un artículo. Los criterios de exclusión son los que se muestran a continuación: (i) dentro del artículo no aparecen los términos de búsqueda definidos, (ii) el artículo no relaciona calidad de datos con los analistas y/o expertos del negocio.

La segunda etapa de la RSL, el **Desarrollo de la revisión** considera la realización de la revisión a partir de lo establecido en la planificación. En la Tabla 4.3 se muestra un resumen de los resultados de la revisión. En esta tabla, en la primera columna, se muestra las fuentes de búsqueda. En la segunda columna se indica el número de resultados obtenidos por cada fuente al utilizar las combinaciones de términos definidos en la Tabla 4.2. En la tercera columna se presenta el número de trabajos analizados por cada fuente. En este punto es importante indicar que en cada fuente se revisaron hasta los 100 primeros resultados arrojados por cada combinación de términos. Finalmente, en la cuarta columna se indica el número de trabajos seleccionados.

La tercera etapa corresponde a la **Publicación de resultados** con la idea de difundir los resultados de la investigación.

Fuente	Resultados de todas las combinaciones	Analizados	Seleccionados
Google Académico	213.547	800	22
Scopus	89	89	1
Springer Link	57	57	1
SciELO	675	430	6
Science Direct	924	678	1
Totales	215.292	2054	31

Tabla 4.3: Resumen de los resultados de búsqueda para la RSL

4.3. Revisión Sistemática de la Web

Para el desarrollo de la Revisión Sistemática en la Web durante la *Planificación de la Revisión*, y en base al objetivo e interrogante de investigación planteados, se definió el protocolo de revisión. En este se consideraron como fuentes de búsqueda: Bptrends, Information-management, Global-excellence, Globaldataexcellence y Google.

Se definieron los términos de búsqueda y sus combinaciones a utilizar en las fuentes de búsqueda, que se muestran en la tabla 4.1 y 4.2, con la salvedad que en los sitios cercanos a las personas del negocio (bptrends, information-management, global-excellence, globaldataexcellence) se omitieron los términos Business y Negocios. Luego se establecieron los siguientes criterios de inclusión (i) dentro de la página Web o documento de esta se hace referencia a problemas o requisitos de calidad de datos, (ii) en la página Web o documento se muestran dimensiones de calidad de datos en el contexto de las personas de negocio, y como criterio de exclusión que la información encontrada no relaciona de ninguna manera la calidad de datos y su definición en el contexto de las personas de negocio.

En la fase correspondiente al *Desarrollo de la Revisión*, se ejecutó la planificación realizada, a continuación en la Tabla 4.4 se muestra un resumen de los resultados de la revisión. En la primera columna de la tabla se muestran las fuentes, en la segunda columna se indica el número de resultados obtenidos en cada fuente con las combinaciones de términos definidos en la Tabla 4.2. En la tercera columna se presenta el número de documentos (o páginas web) analizados por cada combinación de términos en cada fuente. Cuando hubo resultados suficientes, se analizaron hasta los primeros 50 resultados de cada búsqueda. En la cuarta y última columna se indica el número de documentos seleccionados.

La tercera etapa corresponde a la *Publicación de resultados* con la idea de difundir los resultados de la investigación.

Fuente	Resultados de todas las combinaciones	Analizados	Seleccionados
bptrends.com (BPTrends, 2015)	46	46	5
information-management.com (SourceMedia, 2015)	1793	177	12
global-excellence.com (Global Excellence, 2015)	12	12	0
globaldataexcellence.com (Global Data Excellence, 2015)	59	59	3
google.com (Google.com, 2015)	11820000	400	39
Totales	11821910	694	59

Tabla 4.4: Resumen de los resultados de búsqueda para la RSW

4.4. Resultados de la revisión

Como resultado de la RSL y RSW se encontraron 97 problemas de DQ, que luego de ser analizados, y para un mejor manejo, fueron clasificados en 18 tipos. Entre los problemas de DQ identificados; la inexactitud, falta de completitud e inconsistencia fueron los más recurrentes. También, se obtuvo un listado de un total de 457 características de DQ, pesquisadas tanto en la RSL como en la RSW. Después de analizar las características identificadas en cada documento, se eliminaron las repetidas y se fusionaron aquellas que tenían el mismo significado y/o en su definición eran parte de otras, obteniendo un conjunto final de 36 características de DQ.

La lista de los principales tipos de problemas de DQ encontrados como resultado de la revisión se muestra en la Tabla 4.5. En dicha tabla se muestra en la primera columna el tipo de problema, la segunda columna se listan algunos de los problemas de DQ encontrados en la revisión y que fueron clasificados en el mismo tipo. En la tercera columna se muestra la o las fuentes desde donde se obtuvo y, finalmente, en la última columna se indica la cantidad de documentos encontrados (en la RSW una misma fuente podía arrojar más de un documento).

La lista de las características de DQ encontradas, con las cuales las personas del área del negocio se expresan sobre la calidad de sus datos e información se pueden apreciar en las tablas 4.6, 4.7 y 4.8. En la primera columna de estas tablas se presentan las características, luego en la segunda columna se muestran los términos relacionados con la característica definida, en la tercera columna se muestran las fuentes de donde fueron obtenidas las características, en la cuarta columna se presenta una definición de las características que engloba a su vez los términos relacionados y en la quinta columna se muestra el número de documentos en donde fue encontrada la característica.

En las tablas 4.6, 4.7 y 4.8, que muestra las características de DQ encontradas, se puede apreciar que 13 de las 15 características de DQ definidas en la norma ISO/IEC 25012 (ISO/IEC, 2008)

Tipo de problemas	Problemas de DQ identificados en la revisión	Fuentes	Núm. de doc.
Ambigüedad en los datos	Datos ambiguos, Datos sin sentido, Diferentes reglas culturales asociados a los datos	(Google.com, 2015; Redman, 1998).	10
Asimetría en los datos registrados	Problema de asimetría de información, Información asimétrica, Asimetría en la información	(Guardia y Fuentes, 2004; Hernández Ortega et al., 2008).	2
Datos duplicados	Registros de clientes duplicados, Números de contrato por duplicado en su almacén de datos, Entradas duplicadas, La duplicación de la información pre-existente, Superposición de datos dentro de las áreas funcionales, Datos redundantes	(Google.com, 2015; SourceMedia, 2015; Walsh y Walsh, 2005).	7
Datos imprecisos	Información pobre o imprecisa, Datos Pobre, Información débil, Información defectuosa	(Black, 2011; Google.com, 2015; Redman, 1998).	4
Datos mal escritos	Errores tipográficos, Un error de ortografía en cascada, Errores de ortografía habituales	(Google.com, 2015).	3
Datos no accesibles	Carencia en el acceso a los datos corporativos, No tener acceso a la información en verdadero tiempo real es perjudicial, Problemas de calidad, fiabilidad y capacidad de acceso a la información, Datos difíciles de acceder, Inaccesibilidad a los datos, Datos no disponibles	(Clarke, 2013; Google.com, 2015; Redman, 1998).	7
Datos no actualizados	Datos desactualizados, Envejecimiento de los datos, Datos obsoletos, Números que cambian de significado a través del tiempo	(Google.com, 2015; SourceMedia, 2015).	4
Datos no adecuados	La información debe ser apropiada al problema, Disponer de datos adecuados, El despilfarro de datos, Cumplir los requerimientos del problema, Datos inapropiados, Información fuera del objetivo	(Google.com, 2015; Howard, 2007; SourceMedia, 2015).	6
Datos no válidos, erróneos	Datos no válidos e inexactos, Los datos incorrectos, Información errónea, Registros de clientes con 99/99/99 como fecha de creación, Información imperfecta, Oportunidades perdidas debido a los datos de facturación inexactos, Información defectuosa	(Black, 2011; Global Data Excellence, 2015; Google.com, 2015; Redman, 1998; SourceMedia, 2015).	21
Datos no veraces	Información engañosa, Información fraudulenta	(Black, 2011)	1
Datos poco confiables	Información falsa, La libertad de prejuicios, Datos poco confiables debido a su fuente	(Google.com, 2015; Redman, 1998).	4
Falta de completitud de los datos	Encontrar datos sucios o faltantes, Datos que no se encuentran, Datos perdidos, Insuficiente información, La información incompleta e inexacta Datos incompletos, Datos faltantes, Potenciales campos no utilizados, Los campos no utilizados	(BPTrends, 2015; Google.com, 2015; Redman, 1998; Walsh y Walsh, 2005).	15
Falta de compromiso al registrar datos	Los datos correctos no son importantes para el usuario, que los registra, Las reglas han cambiado, pero los usuarios no son conscientes de ello	(Google.com, 2015).	2
Falta de estandarización en los datos	Los datos no estándar, Datos fuera del intervalo aceptable, Formatos de variables, Valores diferentes en materia de género, Datos fuera de rango, Datos que son excepciones de la regla, Datos mal clasificados, Las tecnologías más antiguas no son compatibles, por ejemplo las listas desplegables, Mala integración de datos	(Google.com, 2015).	9
Falta de integridad en los datos	La pérdida de integridad de datos, Problemas de integridad de datos, Información sesgada	(Clarke, 2013; SourceMedia, 2015).	2
Falta de procedimientos de validación de los datos	La falta de validación de datos en sistemas ERP	(Google.com, 2015).	1
Inconsistencia en los datos	Datos inconsistentes con la realidad, Datos inconsistentes entre sí, Sistemas con datos incongruentes, Información contradictoria, Conflictos en los datos, Desajustes de la definición de los datos	(Google.com, 2015; SourceMedia, 2015; Talalweh, 2010).	12
Pérdida de datos	Pérdida de información, Pérdida de Datos	(Clarke, 2013; Redman, 1998).	2

Tabla 4.5: Lista de problemas de Calidad de Datos

están presentes en la lista (marcadas con un “*”), siendo portabilidad y recuperabilidad las únicas dos características que no fueron encontradas. Es interesante que estas 2 características,

ambas de la perspectiva dependiente del sistema, no fueran encontradas. Creemos que esto es consecuente con el enfoque de búsqueda utilizado, que es tendiente a identificar la terminología usada por los expertos y/o analistas del negocio sobre DQ y no por los expertos en TI.

4.5. Validación de los resultados

En esta sección se muestra la especificación del estudio realizado para la validación de la terminología de DQ siguiendo las actividades indicadas en el método expuesto en la Sección 2.1.3 de Capítulo 2 de las metodologías de trabajo.

El estudio se basa en la aplicación de una encuesta a personas del área de los negocios que estuvieran encargadas de tomar decisiones dentro de las organizaciones y que de alguna manera trabajen con datos.

Aunque tanto la literatura consultada como los sitios Web revisados estaban restringidos al ámbito de negocio y no de TI, decidimos validar la terminología de DQ obtenida con personas del negocio. Para esto, se aplicó una encuesta a personas del área de los negocios, que manejan datos de su organización y que en base a ellos deben tomar decisiones.

A continuación se describe con mayor detalle cómo se preparó y aplicó la encuesta, comenzando con los objetivos.

4.5.1. Objetivos del estudio

Los objetivos establecidos en el estudio fueron los siguientes:

- Validar el uso de las características de DQ en el contexto del negocio, mediante una consulta a los expertos y/o analistas de negocio.
- Validar la tipología de problemas de DQ en el contexto del negocio, mediante una consulta a los expertos y/o analistas de negocio.

4.5.2. Diseño del instrumento

Para la construcción de la encuesta se consideró el diseño observacional o descriptivo, que es utilizado cuando se desea recolectar información sin cambiar el entorno, con una interacción de tipo transversal, lo que implica intervenir sólo una vez con grupos de personas. En este caso, se quiere conocer la opinión de un grupo de expertos y/o analistas de negocio sobre la utilización de la terminología de DQ identificada.

4.5.3. Desarrollo del instrumento

En concreto, tras el desarrollo de la identificación de la terminología, se obtuvo un total de 18 tipos de problemas y 36 características de DQ. Con la idea de no hacer un cuestionario

muy extenso ni tedioso para los encuestados, se crearon dos formularios, en cada uno de estos se incluyeron 9 tipos de problemas y 18 características de DQ en forma de preguntas cerradas para su validación. Además, a cada formulario se le agregaron 5 preguntas demográficas (para identificar el sector industrial, la organización y el nivel de experiencia del encuestado) y 1 pregunta abierta (para rescatar la terminología no considerada en el instrumento). Las preguntas cerradas fueron medidas con una escala de Likert de 5 niveles. El instrumento se puede apreciar en el Apéndice A.

4.5.4. Validación del instrumento

Previo a la aplicación de la encuesta se realizó una prueba piloto con 2 participantes. Como resultado del análisis de los comentarios obtenidos, se modificó levemente la forma en la que se planteaban las preguntas, con el fin de hacerlas más claras para los encuestados.

4.5.5. Selección de los participantes

La selección de los participantes en el estudio se realizó utilizando el método no probabilístico de muestreo por conveniencia (Kahn et al., 2002) y el muestreo bola de nieve o snowball (Goodman, 1961), estos tipos de muestreos se eligieron debido al tipo de estudio que se desarrolla y a su facilidad de aplicación.

4.5.6. Aplicación de la encuesta

El instrumento fue aplicado mediante el envío de un correo electrónico a los participantes donde se describía el estudio y se les entregaba un hipervínculo, donde ellos podrían participar respondiendo el formulario online. Las instrucciones formaban parte del formulario y el tiempo para responder la encuesta fue ilimitado. Se recolectaron 38 respuestas, 19 de ellas corresponden a los sujetos que respondieron el formulario A y las 19 restantes a los que respondieron el formulario B, estas fueron respondidas en forma completa por lo que no hubo necesidad de objetar ninguna de ellas.

4.6. Análisis de los resultados del estudio

En esta sección se realiza un análisis de las respuestas entregadas por los encuestados. Primero se definen los criterios para la validación de la terminología. Después se realiza un análisis demográfico de los participantes en el estudio. Luego se realiza un análisis estadístico de las

respuestas referente a las características y a los problemas de DQ y, finalmente, se presentan las conclusiones del estudio.

4.6.1. Criterios de validación

En primera instancia se debe considerar que la terminología a validar se obtuvo de foros Web, sitios Web y literatura formal cercana al área del negocio, por lo cual se puede establecer que está previamente validada como del área de negocio. Teniendo en cuenta lo anterior, se establecieron los siguientes criterios para validar cada uno de los tipos de problemas y cada una de las características de DQ encontradas:

- Sobre un 55 % de los participantes indica que usa el término consultado (tipo de problema o característica) desde el nivel 3 (“ocasionalmente”) hasta el nivel 5 (“siempre”). .
- El valor de la media de las respuestas a las preguntas en la escala de Likert debe ser al menos mayor a 2.5.
- El valor de la moda de las respuestas a las preguntas en la escala de Likert debe ser al menos 2.

4.6.2. Análisis demográfico

En esta subsección se realiza un análisis de los participantes del estudio. En la Tabla 4.9 se puede apreciar algunas de las medidas descriptivas (media, mediana, moda y desviación estándar) sobre aspectos referentes a la experiencia de los encuestados. En concreto de los años de experiencia laboral y los años que desempeñan en su cargo actual.

Con respecto a los años de experiencia laboral de los sujetos participantes del estudio se puede decir que el promedio es de 21.3 años con una variación de 10.7 años (los valores son altamente dispersos) y la mitad de los sujetos tiene 20 años de experiencia y la cantidad de años de experiencia más frecuente es de 20 años.

Ahora teniendo en consideración los años que llevan los sujetos desempeñando el cargo actual, es preciso señalar que en promedio es de 9.6 años con una variación de 9.7 años (los valores son altamente dispersos). La mitad de los sujetos llevan 4 años en su cargo actual y la cantidad de años desempeñando el cargo actual más frecuente es 4 años.

En la Tabla 4.10 se muestra un análisis de los sectores industriales en los que trabajan el grupo de participantes del estudio. De esta tabla se puede concluir que los participantes pertenecen a variados sectores industriales (11 diferentes) concentrando los porcentajes más altos, en el sector de salud y de educación.

4.6.3. Análisis estadístico de los resultados referente a problemas de DQ

En la Tabla 4.11 se presenta un análisis de los resultados del estudio tendiente a validar la tipología de problemas de DQ identificados en la literatura. La primera columna de esta tabla contiene los tipos de problemas, luego la distribución porcentual de las respuestas según la escala de 5 niveles que se usó. Dicha escala permite señalar el grado de ajuste de acuerdo del encuestado con el tipo de problema (forma de expresarlo) y su definición. En la siguiente columna se muestra el porcentaje acumulado de los niveles 3 al 5. En las últimas tres columnas se muestra el resultado del cálculo de las medidas descriptivas: media, moda y la desviación estándar.

Una vez que se estudiaron y se interpretaron los resultados presentados en la Tabla 4.11 se pudo establecer que los 3 tipos de problemas (forma de expresarlos) más usados son: “datos no actualizados”, “datos duplicados” y “falta de completitud de los datos”. Los 3 menos usados son: “falta de integridad en los datos”, “falta de compromiso al registrar datos” y “datos no adecuados”. Teniendo en consideración los criterios de validación definidos previamente, y la Tabla 4.11 con el análisis de resultados, se puede determinar que todos los tipos de problemas son válidos.

4.6.4. Análisis estadístico referente a las características de DQ

La Tabla 4.12 muestra el análisis de las respuestas del estudio tendientes a validar el nivel de utilización de las características de DQ en su vocabulario habitual durante la realización de su trabajo. La primera columna de la tabla contiene una característica de DQ identificada, seguida de la distribución porcentual de las respuestas según la escala de 5 niveles que se usó. Dicha escala permite señalar el nivel de uso de cada característica de DQ encontrada. En la siguiente columna se muestra el porcentaje acumulado de los niveles 3 al 5. Las últimas 3 columnas contienen el cálculo de medidas descriptivas: media, moda y la desviación estándar.

Una vez interpretada la información presentada en la Tabla 4.12 se puede establecer que las tres características de DQ que más utilizan los expertos del negocio son actualidad, aplicabilidad y exactitud. Las tres características que menos utilizan son Representación, Duplicidad y Reputación. Teniendo en consideración los criterios de validación definidos y la Tabla 4.11 con el análisis de resultados, se puede determinar que todas las características de DQ de negocio son validadas, excepto Representación (marcada con * en la Tabla 4.12) debido a que no cumple los criterios, porque entre las respuestas es más frecuente el nivel de utilización “rara vez”, además como media tiene un valor de 2.89 con una variación de 1.370, lo que quiere decir que en promedio tiene un nivel de utilización de “rara vez” u “ocasionalmente”.

4.7. Conclusiones del estudio

Al analizar las respuestas del instrumento se puede establecer que tanto los años de experiencia laboral como los años que los participantes desempeñan su cargo actual presentan valores elevados, lo que indica que poseen una gran experiencia. Además, la mayoría de los participantes pertenecen a dos sectores industriales, en concreto salud y educación.

Teniendo en consideración los objetivos establecidos en el estudio y dando respuesta al primero de ellos (“validar la utilización de las características de calidad de datos de negocio identificadas, mediante una consulta a los expertos y/o analistas de negocio”), se aplicó el estudio y se analizaron los resultados constatando que casi todas las características de DQ identificadas son utilizadas por expertos del negocio salvo Representación.

Para dar cumplimiento al segundo objetivo (“validar la tipología de problemas de calidad de datos de negocio identificada, mediante una consulta a los expertos y/o analistas de negocio”), se aplicó el estudio y al analizar los resultados, llegando a que se corroboraron todos los tipos de problemas ya que estos cumplieron los criterios de validación erigidos.

Además, las características de calidad de datos fueron ampliamente validadas excepto representación ya que no cumplió los criterios establecidos para su validación. Con la idea de usar terminología ampliamente validada se decidió no incluir esta característica en la ontología.

4.8. Análisis de la terminología

En esta sección se presenta el análisis realizado de la terminología de calidad de datos, teniendo en consideración los tipos de problemas y las características de DQ de negocio. Concretamente, se analizaron las relaciones entre ellas, estas relaciones se puede apreciar en las Tablas 4.13, 4.14 y 4.15, en donde la primera columna de estas muestra cada uno de los tipos de problemas de calidad de datos de negocio (BDQ), en la segunda columna el tipo de relación, en la tercera la característica BDQ con la que se relaciona y por último, en la cuarta columna las justificaciones de las relaciones encontradas basadas en el estudio de las definiciones de cada tipo de problema BDQ y característica BDQ.

Al analizar las características BDQ y las características de calidad de datos de tecnologías de información (ITDQ) se encontraron las relaciones entre ellas se presentan en las Tablas 4.16, 4.17 y 4.18, en donde la primera columna presenta cada una de las características BDQ, en la segunda columna, el tipo de relación, en la tercera la característica ITDQ con la que se relaciona y, por último, en la cuarta columna las justificaciones de las relaciones encontradas basadas en el estudio de las definiciones de cada una de las características.

Este análisis servirá de ayuda para el desarrollo de la ontología, ya que proveerá de las diversas relaciones que existen entre los conceptos de calidad datos estudiados.

Características	Términos asociados	Fuente	Descripción	Núm. de doc.
Accesibilidad*	Accesibilidad, Acceso, Recuperabilidad, Puntualidad	(Berente et al., 2009; Borek et al., 2014; Eppler y Muenzenmayer, 2002; Falge et al., 2012; Google.com, 2015; Lillrank, 2003; SourceMedia, 2015; Torres y Rojas, 2008; Wang, 1998).	Medida en que los datos están disponibles en un momento dado en el tiempo.	33
Actualidad*	Actual, Actualidad, Actualizada, Actualizar, Caducado, No obsoleto	(Global Data Excellence, 2015; Google.com, 2015; Liy et al., 2013; Padilla).	Trata sobre la actualización de los datos y su vigencia.	9
Adecuación	Adecuación representacional, Atracción, Apropiaada, Cantidad apropiada, Cantidad apropiada de los datos, Cantidad de datos, Suficiencia	(Aranda et al., 2013; Google.com, 2015; Lillrank, 2003; Liy et al., 2013; Redman, 1998; Torres y Rojas, 2008; Wang, 1998).	Grado en el cual el dato o la información se presenta de manera concisa, flexible y organizada, considerando los objetivos del usuario para ayudarlo a alcanzar las metas que ha especificado.	10
Amplitud	Alcance, Amplitud, Cobertura	(Google.com, 2015; Redman, 1998; Torres).	La información que se presenta cubre los temas en un grado o alcance que es satisfactorio para el usuario.	4
Aplicabilidad	Aplicabilidad	(Google.com, 2015)	Es la característica de la información que le permite ser usada directamente por el usuario en un contexto específico y de cómo la información se encuentra organizada para acciones.	1
Complejidad*	Completa, Complejidad, Completo, Exhaustividad, Existencia, Especificación	(Berente et al., 2009; BPTrends, 2015; Burguete, 2004; Calderón Amaya y Rodríguez Monroy, 2012; Global Data Excellence, 2015; Google.com, 2015; Redman, 1998; SourceMedia, 2015; Tee et al., 2007; Torres).	Grado en que los valores necesarios de los atributos están disponibles.	27
Comprensibilidad*	Comprensibilidad	(Google.com, 2015; Redman, 1998; Torres y Rojas, 2008).	Medida en que los datos se entienden fácilmente.	5
Confiabilidad	Fiabilidad, Confiabilidad, Confiante, Confianza	(Google.com, 2015; Gorla et al., 2010; Redman, 1998; Tee et al., 2007; Torres y Rojas, 2008).	Grado de concordancia entre las diferentes medidas tomadas en condiciones similares.	9
Confidencialidad*	Confidencialidad	(Google.com, 2015).	Acceso restringido a la información de calidad.	1
Conformidad*	Conformidad	(Google.com, 2015; SourceMedia, 2015).	Grado de ajuste de los valores a los formatos estándares establecidos.	7
Consistencia*	Coherencia, Semántica, Consistencia, Consistente, Representación Concisa, Representación	(Global Data Excellence, 2015; Google.com, 2015; Guerra-García et al., 2011; SourceMedia, 2015; Wang, 1998).	Grado en que la información de diferentes conjuntos de datos se puede corroborar.	36
Credibilidad*	Credibilidad, Creíble	(Berente et al., 2009; Black, 2011; Google.com, 2015; Wang, 1998).	Medida en que se considera los datos como verdadero y creíble.	8
Disponibilidad*	Disponibilidad, Disponible	(Burguete, 2004; Google.com, 2015; Torres)	Grado en que los datos están actualizados y disponibles para el uso.	7
Duplicidad	Duplicación, Duplicados, Duplicidad	(Google.com, 2015; Torres).	Grado en el que hay ocurrencias redundantes o registros de un mismo objeto del mundo real o evento.	9
Eficacia	Eficacia	(Google.com, 2015).	La información es eficaz si satisface las necesidades del consumidor de la información que utiliza la información para una tarea específica.	1

Tabla 4.6: Lista de características de Calidad de Datos (1/3)

Características	Términos asociados	Fuente	Descripción	Núm. de doc.
Eficiencia*	Eficiencia	(Google.com, 2015; Redman, 1998).	Grado en que la información satisface las necesidades del consumidor de la información, la obtiene y utiliza de una manera fácil.	2
Estandarización	Estandarización, Formato, Formato útil	(Correa y Cruz, 2005; Google.com, 2015; Redman, 1998; SourceMedia, 2015).	Grado en que los elementos de dato están definidos y se comprenden sus valores.	5
Estructurada	Estructura	(Google.com, 2015; SourceMedia, 2015).	Formato correcto para utilizar.	2
Exactitud*	Claridad, Correcta, Exacta, Exactitud, Precisión*	(Bagchi et al., 2006; Borek et al., 2014; Calderón Amaya y Rodríguez Monroy, 2012; Eppler y Muenzenmayer, 2002; Global Data Excellence, 2015; Google.com, 2015; Lee et al., 2011; Lillrank, 2003; Redman, 1998; Tee et al., 2007; Torres).	Medida en que los datos se representan correctamente, con valores estándar y reales.	50
Integridad	Derivación de integridad, Integralidad, Integridad	(BPTrends, 2015; Eppler y Muenzenmayer, 2002; Global Data Excellence, 2015; Google.com, 2015; Lillrank, 2003; SourceMedia, 2015; Wang, 1998).	Grado en que todos los datos necesarios están presente.	35
Interpretabilidad	Interpretabilidad	(Berente et al., 2009; Google.com, 2015; Lillrank, 2003; Redman, 1998; Wang, 1998).	Medida en que los datos son apropiados en idioma, símbolos y unidades, y las definiciones son claras.	7
Libre de errores	Libre de error, Libre de errores, Sin errores	(Berente et al., 2009; Google.com, 2015).	Medida en que los datos son correctos y confiables	3
Mantenibilidad	Conservación, Mantenibilidad	(Eppler y Muenzenmayer, 2002; Torres).	Grado en que los datos pueden ser accedidos usados, conservados, actualizados, mantenidos y manipulados.	2
Objetividad	Objetividad	(Berente et al., 2009; Google.com, 2015; Lillrank, 2003; Wang, 1998)	Medida en que los datos es imparcial y sin prejuicios, implica un enfoque que asegurar que la información es exacta, fiable, completa y clara.	10
Operatividad	Operatividad	(Torres y Rojas, 2008).	Calidad intrínseca de la información.	1
Oportunidad	Oportunidad, Oportunidades, Oportuna, Oportuno, Pertinente	(Calderón Amaya y Rodríguez Monroy, 2012; Google.com, 2015; Lillrank, 2003; Talalweh, 2010; Tee et al., 2007; Torres y Rojas, 2008; Torres).	Grado en que los datos están actualizados y disponibles para el uso.	15

Tabla 4.7: Lista de características de Calidad de Datos (2/3)

Características	Términos asociados	Fuente	Descripción	Núm. de doc.
Relevancia	Relevancia, Relevante, Importancia	(Falge et al., 2012; Google.com, 2015; Lillrank, 2003; Redman, 1998; SourceMedia, 2015; Talalweh, 2010; Tee et al., 2007; Torres; Wang, 1998).	Medida en que los datos son aplicables y pertinentes para la tarea en cuestión.	16
Representación	Representación, Representación coherente	(Berente et al., 2009; Wang, 1998).	La característica de calidad de los datos que se ocupa de la forma, el patrón, la legibilidad y utilidad de los datos para su uso previsto.	2
Reputación	Reputación, Percepción	(Falge et al., 2012; Google.com, 2015; Lillrank, 2003).	Medida en que los datos están muy bien considerado en términos de su origen o contenido.	8
Seguridad	Seguridad	(Falge et al., 2012; Google.com, 2015; Lillrank, 2003).	Medida en que el acceso a los datos es restringido adecuadamente para mantener su seguridad.	8
Sincronización	Sincronización	(Google.com, 2015; Torres)	Medida de la equivalencia de la información almacenada para hacer los datos equivalentes .	2
Trazabilidad*	Trazabilidad	(Eppler y Muenzenmayer, 2002).	(Sin definición encontrada)	1
Usabilidad	Facilidad de comprensión, Facilidad de localización, Facilidad de manipulación, Facilidad de operación, Facilidad de uso, Usabilidad, Presentación, Presentación adecuada, Interactividad	(Berente et al., 2009; Eppler y Muenzenmayer, 2002; Google.com, 2015; Lillrank, 2003; SourceMedia, 2015; Torres).	Grado en que la presentación de la información es directamente utilizable y eficaz para su propósito.	14
Utilidad	Útil, Utilidad	(Google.com, 2015; Mosquera et al., 2001; Redman, 1998).	Utilidad implica la utilidad y disponibilidad de la información para su uso previsto.	6
Valor añadido	Valor añadido	(Aranda et al., 2013; Google.com, 2015; Lillrank, 2003; Wang, 1998).	Grado en el cual el dato o la información son beneficiosos y proveen ventajas como producto de su utilización.	5
Veracidad	Real, Relación Validez, Veraz, Verídica, Veracidad, Fundamentos	(Guzmán et al., 2012a; Lij et al., 2013; Padilla; Torres y Rojas, 2008; Torres).	Grado en que los datos se ajustan a su definición, los valores de dominio y las reglas de negocio.	6

Tabla 4.8: Lista de características de Calidad de Datos (3/3)

Experiencia /Medidas descriptivas	Media	Mediana	Moda	Desviación estándar
Años en su cargo	9.5	4	4	9.7
Años de experiencia	21.3	20	20	10.7

Tabla 4.9: Análisis respecto de la experiencia de los encuestados

Sector industrial	Frecuencia	Porcentaje (%)
Agricultura	2	5,3
Alimentación	2	5,3
Banca	2	5,3
Construcción	2	5,3
Educacional	10	26,3
Energía	3	7,9
Industrial	1	2,6
Financiero	1	2,6
Público	2	5,2
Salud	12	31,6
Servicio	1	2,6
Total	38	100

Tabla 4.10: Análisis del sector industrial al que pertenecen los encuestados

Tipo de problema	% de respuestas según escala de Likert					Σ nivel 3 al 5	Media	Moda	Desviación
	5 Siempre	4 Casi Siempre	3 Ocasionalmente	2 Rara vez	1 Nunca				
Falta de integridad en los datos	5.3	21	31.6	36.8	5.3	57.9	2.84	2	1.015
Datos no validos erróneos	21.1	36.8	10.5	21.1	10.5	68.4	3.37	4	1.342
Perdida de datos	31.6	21	15.8	26.3	5.3	68.4	3.47	5	1.349
Datos mal escritos	15.8	31.6	15.8	36.8	0	63.2	3.26	2	1.147
Datos duplicados	26.3	26.3	21.1	21	5.3	73.7	3.47	5	1.264
Falta de estandarización en los datos	15.8	36.8	5.3	26.3	15.8	57.9	3.11	4	1.41
Datos no veraces	21.1	26.3	15.8	15.7	21.1	63.2	3.11	4	1.487
Datos no accesibles	10.5	36.9	15.8	26.3	10.5	63.2	3.11	4	1.243
Asimetría en los datos registrados	5.3	31.5	26.4	31.5	5.3	63.2	3	4	1.054
Falta de completitud de los datos	15.79	31.58	31.58	15.79	5.26	78.95	3.37	4	1.116
Falta de procedimientos de validación de datos	31.58	5.26	31.58	15.79	15.79	68.42	3.21	5	1.475
Falta de compromiso al registrar datos	26.32	21.05	10.52	36.85	5.26	57.89	3.26	2	1.368
Inconsistencia en los datos	26.32	26.31	21.05	15.79	10.53	73.68	3.42	5	1.346
Datos imprecisos	15.79	36.84	10.53	31.58	5.26	63.16	3.26	4	1.24
Datos no actualizados	31.58	26.31	10.53	26.32	5.26	68.42	3.53	5	1.349
Ambigüedad en los datos	15.79	26.32	31.57	15.79	10.53	73.68	3.21	3	1.228
Datos poco confiables	26.32	10.52	31.58	15.79	15.79	68.42	3.16	3	1.425
Datos no adecuados	15.79	21.05	21.05	42.11	0	57.89	3.11	2	1.15

Tabla 4.11: Análisis de los resultados del estudio de los tipos de problemas de calidad de datos

Característica	% de respuestas según escala de Likert					Σ nivel 3 al 5	Media	Moda	Desviación
	5 Siempre	4 Casi Siempre	3 Ocasionalmente	2 Rara vez	1 Nunca				
Actualidad	47.4	31.5	15.8	5.3	0	94.7	4.21	5	0.918
Adecuación	15.8	57.9	21	5.3	0	94.7	3.79	4	0.918
Amplitud	15.8	57.9	10.5	15.8	0	84.2	3.74	4	0.933
Aplicabilidad	47.4	36.8	10.5	5.3	0	94.7	4.26	5	0.872
Completitud	26.3	52.6	21.1	0	0	100	4.05	4	0.705
Comprensibilidad	21.1	47.3	26.3	5.3	0	94.7	3.79	4	0.976
Confiabilidad	21.1	57.8	5.3	10.5	5.3	84.2	3.79	4	1.084
Confidencialidad	15.8	36,8	21.1	10.5	15.8	73.7	3.26	4	1.327
Consistencia	10.5	36.9	21	21.1	10.5	68.4	3.16	4	1.214
Credibilidad	36.8	47.4	10.5	5.3	0	94.7	4.11	4	0.994
Disponibilidad	21.1	47.3	5.3	5.2	21.1	73.7	3.42	4	1.465
Duplicidad	26.3	15.8	31.6	21	5.3	73.7	3.37	3	1.257
Eficacia	5.3	57.9	21	10.5	5.3	84.2	3.47	4	0.964
Estandarización	31.6	57.9	10.5	0	0	100	4.21	4	0.631
Estructuración	15.8	57.9	21	5.3	0	94.7	3.84	4	0.765
Exactitud	42.1	42.1	10.5	5.3	0	94.7	4.21	5	0.855
Integridad	15.8	47.4	36.8	0	0	100	3.79	4	0.713
Interpretabilidad	21.1	57.8	21.1	0	0	100	4	4	0.667
Accesibilidad	10.53	36.84	15.79	31.58	5.26	63.16	3.16	4	1.167
Conformidad	21.05	31.58	31.58	15.79	0	84.21	3.58	4	1.017
Eficiencia	26.32	26.31	26.32	15.79	5.26	78.95	3.53	5	1.219
Libre de errores	15.79	31.58	36.84	10.53	5.26	84.21	3.42	3	1.071
Mantenibilidad	15.79	42.1	31.58	10.53	0	89.47	3.63	4	0.895
Objetividad	26.32	42.1	26.32	0	5.26	94.74	3.84	4	1.015
Operatividad	31.58	57.89	10.53	0	0	100	4.21	4	0.631
Oportunidad	26.32	31.57	21.06	15.79	5.26	78.95	3.58	4	1.216
Relevancia	26.32	31.57	26.32	10.53	5.26	84.21	3.63	4	1.165
Representación*	15.79	21.05	15.79	31.58	15.79	52.63	2.89	2	1.37
Reputación	10.53	26.31	47.37	0	15.79	84.21	3.16	3	1.167
Seguridad	36.84	26.32	10.52	15.79	10.53	73.68	3.63	5	1.422
Sincronización	31.58	21.05	31.58	10.53	5.26	84.21	3.63	5	1.212
Trazabilidad	31.58	26.31	31.58	5.27	5.26	89.47	3.74	5	1.147
Usabilidad	21.05	42.11	21.05	10.53	5.26	84.21	3.63	4	1.116
Utilidad	31.58	36.84	21.05	5.27	5.26	89.47	3.84	4	1.119
Valor añadido	21.05	31.58	21.05	5.27	21.05	73.68	3.26	4	1.447
Veracidad	26,32	31.57	21.06	15.79	5.26	78.95	3.58	4	1.216

Tabla 4.12: Análisis de los resultados del estudio de las características de calidad de datos

Tipos de problema BDQ	Tipo de relación	Característica de BDQ	Justificación
Falta de integridad en los datos	Está relacionado con	Integridad	Debido a que falta de integridad en los datos en su descripción hace mención a problemas con datos incompletos al igual que la definición de integridad.
	Está relacionado con	Libre de errores	Debido a que falta de integridad en los datos en su descripción hace mención a problemas con datos incorrectos al igual que la definición de libre de errores.
Datos no validos y erróneos	Está relacionado con	Libre de errores	Debido a que datos no valido y erróneos hace referencia a problemas con datos incorrectos y la definición de libre de errores hace referencia a datos libres de defectos.
	Está relacionado con	Adecuación	Porque datos no validos y erróneos en su descripción hace referencia a la representación correcta de la realidad y la definición de adecuación hace alusión a la presentación concisa flexible y organizada de los datos.
Perdida de datos	Está relacionado con	Completitud	Debido a que perdida de datos en su descripción menciona problemas con la ausencia de datos y la definición de completitud esta relacionada con la presencia de los datos necesarios.
Datos mal escritos	Está relacionado con	Interpretabilidad	Porque datos mal escritos en su descripción hace referencia a errores de ortografía y símbolos inadecuados al igual que la definición de interpretabilidad.
Datos duplicados	Está relacionado con	Duplicidad	Debido a que datos en su descripción hace alusión a problemas con la repetición de datos en diversas partes cuando no es necesario al igual que la definición de duplicidad.
Falta de estandarización en los datos	Está relacionado con	Estandarización	Debido a que falta de estandarización en los datos hace mención a problemas con los datos que no se ajustan a los estándares y estandarización en su definición hace alusión a los datos definidos de forma uniforme.
Datos no veraces	Está relacionado con	Veracidad	Porque datos no veraces en su descripción hace referencia a problemas con la los datos irreales o engañosos y la definición de veracidad se hace referencia al grado en que los datos se ajustan con su definición, los valores del dominio y las reglas del negocio .
Datos no accesibles	Está relacionado con	Accesibilidad	Porque datos no accesibles esta relacionado con problemas con no poder acceder a los datos cuando estos se necesitan.
	Está relacionado con	Disponibilidad	Porque datos no accesibles esta relacionado con problemas con de la disponibilidad de los datos.

Tabla 4.13: Relaciones entre tipos de problemas DQ y características DQ de negocio (1/3)

Tipos de problema BDQ	Tipo de relación	Característica de BDQ	Justificación
Falta de procedimientos de validación de los datos	Está relacionado con	Exactitud	Debido a que falta de procedimientos de validación de los datos esta relacionado problemas de ausencia de validación al registrar los datos y la definición de exactitud corresponde al grado en que los dato son reales.
	Está relacionado con	Completitud	Debido a que falta de procedimientos de validación de los datos esta relacionado con problemas de ausencia de validación al registrar los datos y la definición de completitud hace referencia a que todos los datos que se necesitan estén registrado.
Falta de compromiso al registrar datos	Está relacionado con	Completitud	Porque falta de compromiso al registrar datos esta relacionado con problemas en formularios con datos faltantes y la definición de completitud hace referencia a que todos los datos que se necesitan estén registrado.
	Está relacionado con	Libre de errores	Debido a que falta de compromiso al registrar datos esta relacionado con problemas con formularios con datos incorrectos y la definición de libre de errores hace referencia a datos libres de defectos.
Inconsistencia en los datos	Está relacionado con	Consistencia	Porque inconsistencia en los datos esta relacionado con problemas de datos no coherentes y la definición de consistencia hacer referencia a el nivel de coherencia de los datos.
Datos imprecisos	Está relacionado con	Exactitud	Porque datos imprecisos esta vinculado con problemas con la falta de precisión en los datos y la definición de exactitud hace referencia a la representación correcta de los datos.
	Está relacionado con	Completitud	Debido a que datos imprecisos en su descripción hace referencia a problemas con datos poco detallados y las definición de completitud corresponde al grado en que todos los datos necesarios están presentes.
Datos no actualizados	Está relacionado con	Actualidad	Porque datos no actualizados se vincula con problemas con datos obsoletos que no representan la realidad actual y la definición de actualidad hace referencia al nivel de actualización y vigencia de los datos.
Ambigüedad en los datos	Está relacionado con	Interpretabilidad	Porque ambigüedad en los datos por su descripción se vincula con problemas con datos que tienen significados diferentes para distintas personas.
Datos poco confiables	Está relacionado con	Confiability	Debido a que datos poco confiables se vinculan con problemas con los datos en termino de su confiabilidad.
	Está relacionado con	Trazabilidad	Porque datos poco confiables se vincula con problemas de no poder verificar los datos y la definición de trazabilidad hace referencia al seguimiento de los datos .

Tabla 4.14: Relaciones entre tipos de problemas DQ y características DQ de negocio (2/3)

Tipos de problema BDQ	Tipo de relación	Característica de BDQ	Justificación
Datos no adecuados	Está relacionado con	Adecuación	Debido a que datos no adecuados esta vinculado con problemas de la mala presentación de los datos y datos inadecuados.
Asimetría en los datos registrados	Está relacionado con	Consistencia	Porque asimetría en los datos registrados se vincula con problemas de diferencias de entre los datos registrados y la definición de consistencia corresponde al grado son coherentes y no contradictorios.
Falta de completitud de los datos	Está relacionado con	Compleitud	Porque falta de completitud de los datos se vincula con problemas con datos que no están completos.

Tabla 4.15: Relaciones entre tipos de problemas DQ y características DQ de negocio (3/3)

Característica BDQ	Tipo de relación	Característica ITDQ	Justificación
Actualidad	se relaciona con	Actualidad	Debido a que ambas características en su definición hacen referencia a la vigencia de los datos.
Adecuación	se relaciona con	Comprensibilidad	Debido a que adecuación se vincula con el grado en que se representan de manera concisa y organizada los datos. Y comprensibilidad se define como el grado en que los datos pueden ser leídos e interpretados.
Amplitud	se relaciona con	Eficiencia	Debido a que amplitud se asimila como el grado en que los datos cubren satisfactoriamente las necesidades de los usuarios. Y eficiencia esta vinculado con el grado de desempeño de los datos.
	se relaciona con	Comprensibilidad	Debido a que amplitud se define como el alcance en que los datos cubren satisfactoriamente las necesidades de los usuarios. Y comprensibilidad hace referencia a los atributos que permiten que los datos sean interpretados.
Aplicabilidad	se relaciona con	Eficiencia	Debido a que aplicabilidad se interpreta como el nivel en que los datos son directamente utilizables. Y eficiencia se asimila como el grado de desempeño de los datos.
Completitud	se relaciona con	Completitud	Debido a que ambas características hacen referencia en su definición al en nivel en que todos los datos necesarios están presentes.
Comprensibilidad	se relaciona con	Comprensibilidad	Debido a que en ambas características se detallan como el nivel en que los datos son interpretados y apropiados.
Confiabilidad	se relaciona con	Consistencia	Debido a que la definición de confiabilidad se vincula al grado de concordancia entre los datos. Y consistencia hace relación con el grado en que los datos son libres de contradicción y coherentes.
Confidencialidad	se relaciona con	Confidencialidad	Debido a que ambas características se relacionan con el debido control de acceso a los datos.
Consistencia	se relaciona con	Consistencia	Debido a que ambas características en su definición se vinculan con la coherencia entre los datos.
Credibilidad	se relaciona con	Credibilidad	Debido a que ambas características en su definición hacen referencia a el nivel en que los datos son verdaderos y creíbles.
Disponibilidad	se relaciona con	Disponibilidad	Debido a que la definición de ambas características corresponde a la obtención de los datos por los usuarios autorizados.
Duplicidad	se relaciona con	Consistencia	Debido a que duplicidad se define como el nivel en que los datos se repiten entre sistemas cuando no es necesario. Y consistencia tiene vinculación al grado en que los datos son coherentes, lo cual podría ser afectado por los datos repetidos.
Eficacia	se relaciona con	Eficiencia	Debido a que ambas características se relacionan con el nivel de desempeño de los datos para satisfacer necesidades.

Tabla 4.16: Relaciones entre características de DQ de negocio y IT (1/3)

Característica BDQ	Tipo de relación	Característica ITDQ	Justificación
Estandarización	se relaciona con	Conformidad	Debido a que estandarización se vincula al grado en que los datos están definidos de forma uniforme. Y conformidad hace referencia al nivel en que los datos se adhieren a las normas y convenciones.
Estructuración	se relaciona con	Conformidad	Debido a que estructuración se asimila como el nivel en que los datos poseen estructura y el formato correcto.
Exactitud	se relaciona con	Exactitud	Debido a que ambas características se definen como al nivel en que los datos representan el valor correcto y son reales.
Integridad	se relaciona con	Comprensibilidad	Debido a que integridad se define como el nivel en que todos los datos están presentes y de forma apropiada. Y comprensibilidad se vincula con la representación apropiada de los datos, para ser leídos e interpretados.
	se relaciona con	Complejidad	Debido a que integridad se relaciona con la presencia de todos los datos necesarios al igual que complejidad.
Interpretabilidad	se relaciona con	Comprensibilidad	Debido a que interpretabilidad se vincula al nivel en que los datos son apropiados en idioma, símbolos y unidades , al igual que comprensibilidad.
Accesibilidad	se relaciona con	Accesibilidad	Debido a que ambas características son relacionadas con el acceso a los datos.
Conformidad	se relaciona con	Conformidad	Debido a que ambas características en función con su definición se relacionan con el ajuste de los datos con normas o formatos.
Eficiencia	se relaciona con	Eficiencia	Debido a que ambas características se vinculan con el nivel de desempeño de los datos para satisfacer necesidades.
Libre de errores	se relaciona con	Exactitud	Debido a que ambas características se vinculan con el nivel en que los datos están correctos.
Mantenibilidad	se relaciona con	Recuperabilidad	Debido a que mantenibilidad se define como el grado en que los datos mantenidos y conservados, al igual que recuperabilidad.
Objetividad	se relaciona con	Credibilidad	Debido a que objetividad se define como el nivel en los datos son imparciales y sin perjuicio. Y credibilidad se relaciona al grado en que los datos son verdaderos.
Operatividad	se relaciona con	Eficiencia	Debido a que operatividad se vinculan con la cercanía de los datos como la realidad para su óptima utilización. Y eficiencia hace referencia al nivel de desempeño de los datos.
Oportunidad	se relaciona con	Disponibilidad	Debido a que oportunidad y disponibilidad se vinculan con el grado en que los datos pueden ser recuperados (estén disponibles).
	se relaciona con	Actualidad	Debido a que oportunidad y actualidad se relacionan con el nivel en que los datos están actualizados.
Relevancia	se relaciona con	Eficiencia	Debido a que relevancia se define como el grado en que los datos son útiles y pertinentes. Y eficiencia se relaciona el desempeño de los datos.

Tabla 4.17: Relaciones entre características de DQ de negocio y IT (2/3)

Característica BDQ	Tipo de relación	Característica ITDQ	Justificación
Reputación	se relaciona con	Credibilidad	Debido a que reputación se define al nivel en que los datos están bien considerados. Y credibilidad se vincula con el grado en que los datos son verdaderos y creíbles.
Seguridad	se relaciona con	Confidencialidad	Debido a que seguridad y confidencialidad se vinculan con el adecuado control de acceso a los datos.
Sincronización	se relaciona con	Consistencia	Debido a que sincronización se define como el nivel de equivalencia entre los datos almacenados en diferentes sistemas. Y consistencia se define como el grado en que los datos son libres de contradicciones y coherentes.
Trazabilidad	se relaciona con	Trazabilidad	Debido a que en ambas características se refieren a seguimiento o auditoría de los datos.
Usabilidad	se relaciona con	Eficiencia	Debido a que usabilidad se define como el nivel en que la presentación de los datos es útil y eficaz. Y eficiencia hace referencia al nivel de desempeño de los datos.
Utilidad	se relaciona con	Eficiencia	Debido a que utilidad se define como en nivel en que los datos son beneficiosos para un uso previstos. Y eficiencia esta vinculado con el grado de desempeño de los datos.
Valor añadido	se relaciona con	Eficiencia	Debido a que valor añadido se define como en nivel en que los datos proveen beneficios y ventajas producto de su utilización. Y eficiencia esta vinculado con el grado de desempeño de los datos.
Veracidad	se relaciona con	Credibilidad	Debido a que la definición de veracidad hace referencia al grado en que los datos se ajustan a su definición, los valores de dominio y las reglas de negocio. Y la definición de exactitud que corresponde al grado en que los datos representan correctamente el verdadero valor de los atributos deseados.

Tabla 4.18: Relaciones entre características de DQ de negocio y IT (3/3)

Capítulo 5

Desarrollo de la Ontología

Como herramienta para plasmar el conocimiento adquirido respecto a la terminología de calidad de datos existen principalmente tres opciones: un meta modelo, un modelo conceptual y una ontología (Calero et al., 2006).

Un meta modelo busca mejorar la rigurosidad de los modelos y es definido como un “modelo de modelos” (Calero et al., 2006).

Los modelos conceptuales permiten representar entidades y las relaciones entre ellas en forma muy estructurada (Tablas, clases de objetos, etc.) para un sistema en concreto, por ello en términos generales son contenedores de datos (Calero et al., 2006).

Por otra parte, una ontología busca mejorar la rigurosidad de los modelos para representar conocimiento, estas se definen como “una especificación explícita y formal sobre una conceptualización compartida”(Gruber, 1993), en términos generales es una herramienta para la representación de conocimiento, por ello se utilizó como herramienta una ontología para la representación del conocimiento adquirido.

La metodología definida para el desarrollo de la ontología es METHONTOLOGY la cual se describió en la Sección 5 del Capítulo 2 y establece 5 actividades que se pueden apreciar en la Fig. 5.1 y se describen a continuación.



Fig. 5.1: Actividades de la metodología METHONTOLOGY (Corcho et al., 2005)

- **La especificación**, que permite determinar por qué se construye la ontología, cuál será su uso, y quiénes serán sus usuarios finales.
- **La conceptualización**, que se encarga de organizar y convertir una percepción informal del dominio en una especificación semi-formal, para lo cual utiliza un conjunto de representaciones intermedias, basadas en notaciones tabulares y gráficas, que pueden ser fácilmente comprendidas por los expertos de dominio y los desarrolladores de ontologías. El resultado de esta actividad es el modelo conceptual de la ontología.
- **La formalización**, que se encarga de la transformación de dicho modelo conceptual en un modelo formal o semi-computable.
- **La implementación**, donde se construyen modelos computables en un lenguaje de ontologías (RDF Schema, OWL, etc.). La mayor parte de las herramientas de ontologías permiten llevar a cabo esta actividad de manera automática.
- **Finalmente el mantenimiento**, que se encarga de la actualización y/o corrección de la ontología, en caso necesario.

Este capítulo se estructura de la siguiente manera, comienza con la especificación de la ontología en la Sección 5.1. Luego en la Sección 5.2 se describe paso a paso como se realizó la conceptualización de la ontología. Después en la Sección 5.3 se presenta la formalización e implementación de la ontología. Finalmente en la Sección 5.5 se analiza la vinculación de la ontología con el método BPiDQ.

5.1. Especificación de la Ontología

En esta sección se define en términos generales la ontología a desarrollar, determinando exactamente por qué se construye la ontología, cuál será su uso, y quiénes serán sus usuarios finales.

5.1.1. ¿Por qué se construye la ontología?

La ontología se construye para definir la terminología de DQ utilizada en el ámbito de negocio y establecer su equivalencia en el ámbito de las tecnologías de información. Con la idea de facilitar la especificación de requisitos de calidad de datos para la construcción de un sistema de información.

5.1.2. ¿Cuál será su uso?

Proveer un lenguaje para que la gente de negocio exprese requisitos DQ del área de negocio en forma de tipos de problemas y/o características de DQ. Además de mapear requisitos de DQ expresados por los expertos y/o analistas de negocio a requisitos de DQ expresados con terminología utilizada en el área de las TI.

Alguna de las preguntas de competencia que debe responder la ontología son las siguientes.

- ¿Qué problemas o características de DQ existen?
- ¿Qué característica de DQ está vinculada a un problema de DQ?
- ¿Qué características de DQ son utilizadas en el área de negocios?
- ¿Qué características de DQ son utilizadas en el área de las TI?
- ¿Cómo se relaciona una característica de DQ del área de negocios con una característica de DQ de las TI?

5.1.3. ¿Quiénes serán sus usuarios finales?

Los usuarios finales son los analistas de negocio y los analistas de IT que estén involucrados con el modelado de procesos consientes de la calidad de datos con la ayuda del método BPiDQ (Rodríguez y Caro, 2012).

5.2. Conceptualización de la Ontología

La conceptualización se encarga de organizar y convertir una percepción informal del dominio en una especificación semi-formal, para lo cual utiliza un conjunto de representaciones intermedias, basadas en notaciones tabulares y gráficas, las tareas que se debe desarrollar, se pueden visualizar en la figura 5.2.



Fig. 5.2: Tareas de la metodología METHONTOLOGY, extraído de (Corcho et al., 2005)

5.2.1. Tarea 1: Construcción de un glosario de términos

Para comenzar con el desarrollo de la conceptualización se creó un glosario con los términos relevantes del dominio estudiado (conceptos, instancias, atributos, relaciones entre conceptos, etc.). En las Tablas 5.1, 5.2, 5.3, 5.4, 5.5 y 5.6 se puede apreciar el glosario de términos realizado. Estas tablas poseen 3 columnas, la primera tiene el nombre del término, la segunda muestra la descripción en detalle del término, y en la tercera el tipo de término (constante, concepto, atributo de instancia, relación).

Nombre	Descripción	Tipo
Ámbitos	Los ámbitos corresponden a las diversos sectores que se involucran cuando se desarrolla un software.	Concepto
Expertos	Personas hábiles o con mucha experiencia en un ámbito.	Concepto
Expertos de IT	Personas hábiles o con mucha experiencia en las tecnologías de la información.	Concepto
Expertos de negocio	Personas hábiles o con mucha experiencia en negocios.	Concepto
Negocio	Corresponde a una actividad o ocupación tendiente a la obtención de beneficios.	Concepto
IT	Las tecnologías de la información son todo lo relacionado con los aspectos de la gestión y tratamiento de la información, dentro de una gran organización o empresa.	Concepto
Soluciones IT	Corresponde a la soluciones apoyadas por tecnología de la información para la gestión y tratamiento de la información organización.	Concepto
Requisitos solución IT	Corresponde a una necesidad documentada sobre el contenido, forma o funcionalidad de una solución apoyada por las tecnologías de la información.	Concepto
Requisitos funcionales	Corresponde una descripción de lo que una solución IT debe hacer.	Concepto
Requisitos no funcionales	Corresponde a una especificación de algo sobre la propia solución y de cómo debe realizar sus funciones.	Concepto
Calidad de datos	La calidad de datos corresponde al grado en que las características de los datos satisfacen las necesidades implícitas establecidas, cuando son utilizados en condiciones específicas.	Concepto
Tipos de problemas DQ	Los tipos de problemas DQ son una clasificación de problemas de calidad de datos que surgen en las organizaciones.	Concepto
Tipos de problemas BDQ	Los tipos de problemas DQ son una clasificación de problemas de calidad de datos que surgen en las organizaciones estipulados con terminología del área de negocio.	Concepto
Características DQ	Las característica DQ corresponden a los atributos que deben poseer los datos para ser utilizados por sistemas y por seres humanos.	Concepto
Características BDQ	Las característica DQ corresponden a los atributos que deben poseer los datos para ser utilizados por sistemas y por seres humanos estipulados con terminología del área de negocio.	Concepto
Características ITDQ	Las característica DQ corresponden a los atributos que deben poseer los datos para ser utilizados por sistemas y por seres humanos estipulados con terminología de las tecnologías de la información.	Concepto
Características ITDQ inherente	Las características ITDQ inherentes son las que tienen el potencial intrínseco para satisfacer las necesidades establecidas y necesarias cuando los datos son utilizados bajo condiciones específicas.	Concepto
Características ITDQ dependientes del sistema	Las características ITDQ dependientes del sistema corresponden a las vinculadas con el grado con el que la calidad de datos es alcanzada y preservada a través de un sistema informático cuando los datos son utilizados bajo condiciones específicas.	Concepto

Tabla 5.1: Glosario de términos (1/6)

Nombre	Descripción	Tipo
DQ desde el punto de vista de IT	Corresponde a la calidad de datos vista desde la perspectiva de los expertos en tecnologías de la información.	Concepto
DQ desde el punto de vista de negocio	Corresponde a la calidad de datos vista desde la perspectiva de los expertos en negocio.	Concepto
Actualidad	Grado de actualización y vigencia de los datos.	Instancia
Adecuación	Grado en el cual los datos se presentan de manera concisa, flexible y organizada, considerando los objetivos de quienes los necesitan para ayudarlo a alcanzar sus metas.	Instancia
Amplitud	Grado de alcance en que los datos cubren satisfactoriamente las necesidades de los usuarios.	Instancia
Aplicabilidad	Grado en que los datos permiten ser directamente utilizados por un trabajador en un contexto específico.	Instancia
Eficacia	Grado en que los datos satisfacen las necesidades de quienes los necesitan para el desarrollo de una tarea específica.	Instancia
Compleitud	Grado en que todos los datos necesarios están presentes para el desarrollo de una tarea.	Instancia
Comprensibilidad	Grado en que los datos se entienden fácilmente.	Instancia
Confiabilidad	Grado de concordancia entre diferentes datos registrados en condiciones similares.	Instancia
Confidencialidad	Grado en que se establece control al acceso a los datos de una organización.	Instancia
Consistencia	Grado en que los datos de diferentes conjuntos de datos son coherentes y no contradictorios.	Instancia
Credibilidad	Grado en que se consideran los datos como verdaderos y creíbles.	Instancia
Disponibilidad	Grado en que los datos se pueden obtener cuando se necesitan para un determinado uso.	Instancia
Duplicidad	Grado de existencia de datos repetidos en más de un sistema de información de una organización, cuando esto no es necesario.	Instancia
Estandarización	Grado en que los datos están definidos de manera uniforme y se comprenden sus valores.	Instancia
Estructuración	Grado en que los datos poseen la estructura y el formato correcto para utilizarlos.	Instancia
Exactitud	Grado en que los datos son reales y se representan correctamente en concordancia con los valores estándares definidos.	Instancia
Integridad	Grado en que todos los datos necesarios están presentes y de manera apropiada.	Instancia
Interpretabilidad	Grado en que los datos son apropiados en idioma, símbolos y unidades, y las definiciones son claras.	Instancia

Tabla 5.2: Glosario de términos (2/6)

Nombre	Descripción	Tipo
Libre de errores	Grado en que los datos son correctos y libres de defectos.	Instancia
Mantenibilidad	Grado en que los datos pueden ser accedidos, usados, conservados, actualizados, mantenidos y manipulados.	Instancia
Objetividad	Grado en que los datos son imparciales y sin prejuicios. Esto implica aplicar un enfoque que permita asegurar que los datos sean exactos, fiables, completos y claros.	Instancia
Operatividad	Grado de conformidad expresada por un trabajador respecto a la correspondencia, cercanía de los datos con la realidad y la utilidad, en concordancia con sus necesidades.	Instancia
Oportunidad	Grado en que los datos están actualizados y disponibles en el momento que se requiere su uso.	Instancia
Relevancia	Grado en que los datos son útiles y pertinentes para el desarrollo de una tarea determinada.	Instancia
Reputación	Grado en que los datos están bien considerados en términos de su origen y contenido.	Instancia
Eficiencia	Grado en que los datos satisfacen las necesidades de quien los necesita, teniendo en consideración el acceso y el fácil uso de esta.	Instancia
Seguridad	Grado en que el acceso a los datos es restringido adecuadamente para mantener su seguridad.	Instancia
Sincronización	Grado de equivalencia entre los datos almacenados en un sistema principal y los sistemas locales de una organización.	Instancia
Trazabilidad	Grado en que se puede hacer un seguimiento del origen de los datos disponibles.	Instancia
Usabilidad	Grado en que la presentación de los datos es útil y/o eficaz para su propósito.	Instancia
Utilidad	Grado en que los datos son beneficiosos y están disponibles para un uso previsto.	Instancia
Valor añadido	Grado en que los datos proveen beneficios y ventajas como producto de su utilización.	Instancia
Veracidad	Grado en que los datos se ajustan a su definición, a los valores de dominio y a las reglas de negocio.	Instancia
Accesibilidad	Grado en que los datos están disponibles en un momento dado en el tiempo para su uso.	Instancia
Conformidad	Grado en que los datos se ajustan a los formatos estándares establecidos por una organización.	Instancia
Exactitud	El grado en el cual el dato tiene atributos que representan el valor correcto de un concepto o evento en un contexto específico de uso.	Instancia
Compleitud	El grado en el cual el dato asociado a una entidad tiene valores para todos los atributos esperados e instancias de entidad relacionadas, de acuerdo a un contexto específico de uso.	Instancia

Tabla 5.3: Glosario de términos (3/6)

Nombre	Descripción	Tipo
Consistencia	El grado en el cual el dato tiene atributos libres de contradicción y son coherentes con otros datos en un contexto específico de uso.	Instancia
Credibilidad	El grado en el cual el dato tiene atributos considerados como verdaderos y creíbles por usuarios en un contexto específico de uso.	Instancia
Actualidad	El grado en el cual el dato tiene los atributos que son del período correcto en un contexto específico de uso.	Instancia
Accesibilidad	El grado en el cual se puede acceder al dato en un contexto específico de uso.	Instancia
Conformidad	El grado en el cual el dato tiene atributos que se adhieren a normas, convenciones o regulaciones vigentes y reglas relacionadas con la calidad de datos en un contexto específico de uso.	Instancia
Confidencialidad	El grado en el cual el dato tiene los atributos que aseguran que este es sólo accesible e interpretable por usuarios autorizados en un contexto específico de uso.	Instancia
Eficiencia	El grado en el cual el dato tiene los atributos que pueden ser procesados y proporciona los niveles esperados de funcionamiento (desempeño) usando las cantidades y los tipos de recursos apropiados en un contexto específico de uso.	Instancia
Precisión	El grado en el cual el dato tiene atributos que son exactos o que proporcionan discernimiento en un contexto específico de uso.	Instancia
Trazabilidad	El grado en el cual el dato tiene atributos que proporcionan un rastro de auditoría de acceso a los datos y de cualquier cambio hecho a los datos en un contexto específico de uso.	Instancia
Comprensibilidad	El grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios, y es expresado en lenguajes apropiados, símbolos y unidades, en un contexto específico de uso.	Instancia
Disponibilidad	El grado en el cual el dato tiene atributos que le permiten ser recuperados por usuarios autorizados y/o aplicaciones en un contexto específico de uso.	Instancia
Portabilidad	El grado en el cual los datos tienen atributos que les permiten ser instalados, substituidos o movidos de un sistema a otro conservando la calidad existente, en un contexto específico de uso.	Instancia
Recuperabilidad	El grado en el cual el dato tiene atributos que le permiten mantener y conservar un nivel especificado de operaciones y calidad, aún en caso de falla, en un contexto específico de uso.	Instancia
Falta de integridad en los datos	Corresponde a problemas relacionados con los datos registrados cuando estos están incompletos y/o incorrectos. Por ejemplo, la dirección de entrega de un producto sin el número de la casa o departamento.	Instancia

Tabla 5.4: Glosario de términos (4/6)

Nombre	Descripción	Tipo
Datos no válidos, erróneos	Los datos no son correctos o no representan de forma correcta la realidad. Esto implica costos adicionales para recapturar o actualizar los datos y/o pérdida de oportunidades de negocio por el manejo de datos incorrectos (Datos erróneos de clientes, Datos de facturación erróneos, entre otros.)	Instancia
Pérdida de datos	La ausencia de los datos que se guardaron en un sistema de información en un momento dado, cuando se necesitan pensando que deberían estar guardados.	Instancia
Datos no adecuados	Mala presentación de los datos para su uso, los cuales pueden estar con un nivel de detalle excesivo o ser inapropiados.	Instancia
Datos duplicados	La organización posee datos repetidos en diferentes partes, lo que puede llevar a interpretaciones erróneas e inconsistencias (por ejemplo, un mismo cliente con direcciones diferentes).	Instancia
Falta de estandarización en los datos	Los datos no se ajustan a los estándares requeridos por la organización. No son registrados siempre con el mismo formato (números telefónicos, direcciones), no se ajustan a los rangos de valores válidos (edades, género), entre otros.	Instancia
Datos no veraces	Corresponde a datos irreales o engañosos.	Instancia
Datos no accesibles	Un trabajador no puede acceder a los datos que necesita para el desarrollo de una tarea, ya sea porque estos no estén disponibles o porque la forma de obtenerlos sea engorrosa, causando demoras en el desarrollo de dicha tarea.	Instancia
Asimetría en los datos registrados	Corresponde a la diferencia que existe entre los datos registrados de algo particular, por dos o más trabajadores de áreas diferentes, ejemplo un encargado de bodega y un vendedor tienen datos diferente sobre una transacción.	Instancia
Falta de completitud de los datos	Se refiere a que algunos datos en la organización no están completos, por ejemplo, una ficha médica que no contiene enfermedades preexistentes, una factura donde no se registran todos los ítems comprados. Lo anterior, provoca que datos necesarios para la organización no estén disponibles para su uso, generando problemas en la toma de decisiones.	Instancia
Datos mal escritos	El registro de los datos se puede realizar con errores tales como faltas ortográficas, intercambio de letras, símbolos inadecuados, etc. Este tipo de problemas puede ocasionar pérdida de ingresos, ineficiencia de un proceso, incumplimiento de normas o regulaciones, entre otros.	Instancia
Falta de procedimientos de validación de los datos	El registro de datos (en sistemas de información), sin las respectivas validaciones, ocasiona problemas de calidad de datos, por ejemplo el ingreso de un RUT erróneo por no validar el dígito verificador.	Instancia

Tabla 5.5: Glosario de términos (5/6)

Nombre	Descripción	Tipo
Falta de compromiso al registrar datos	En una organización algunos trabajadores no están conscientes de la importancia de los datos que están registrando, e ingresan documentos y/o formularios con datos faltantes y/o incorrectos.	Instancia
Inconsistencia en los datos	Los datos no son coherentes, por ejemplo, en las organizaciones para hablar de un mismo tema (servicio, producto, recurso) en las distintas áreas funcionales se utilizan términos distintos.	Instancia
Datos imprecisos	Los datos registrados en los sistemas de las organizaciones se encuentran poco detallados o faltos de precisión, con respecto a la realidad, provocando una menor productividad, mayores costos, menor satisfacción de los clientes finales o pérdida de ingresos.	Instancia
Datos no actualizados	Los datos que tienen disponibles los trabajadores están obsoletos y no representan la realidad actual, por ejemplo el correo electrónico o el número telefónico de un proveedor que está registrado no corresponde al actual.	Instancia
Ambigüedad en los datos	Los datos ambiguos son aquellos que pueden tener significados diferentes para las distintas personas que los utilizan, perdiendo el sentido de lo que se quería representar con ellos.	Instancia
Datos poco confiables	Corresponde a datos que provienen de fuentes o cuyo origen es poco confiable o no verificable.	Instancia

Tabla 5.6: Glosario de términos (6/6)

5.2.2. Tarea 2: Taxonomía de términos

La taxonomía de términos se desarrolló en base a los términos tipo conceptos del glosario que se pueden apreciar en la Tabla 5.1 y 5.2, utilizando las relaciones establecidas por la metodología METHONTOLOGY (subclase-de, descomposición-disjunta, descomposición-exhaustiva, partición de un concepto). La taxonomía de conceptos definida para esta conceptualización se puede apreciar en la la Fig. 5.3.

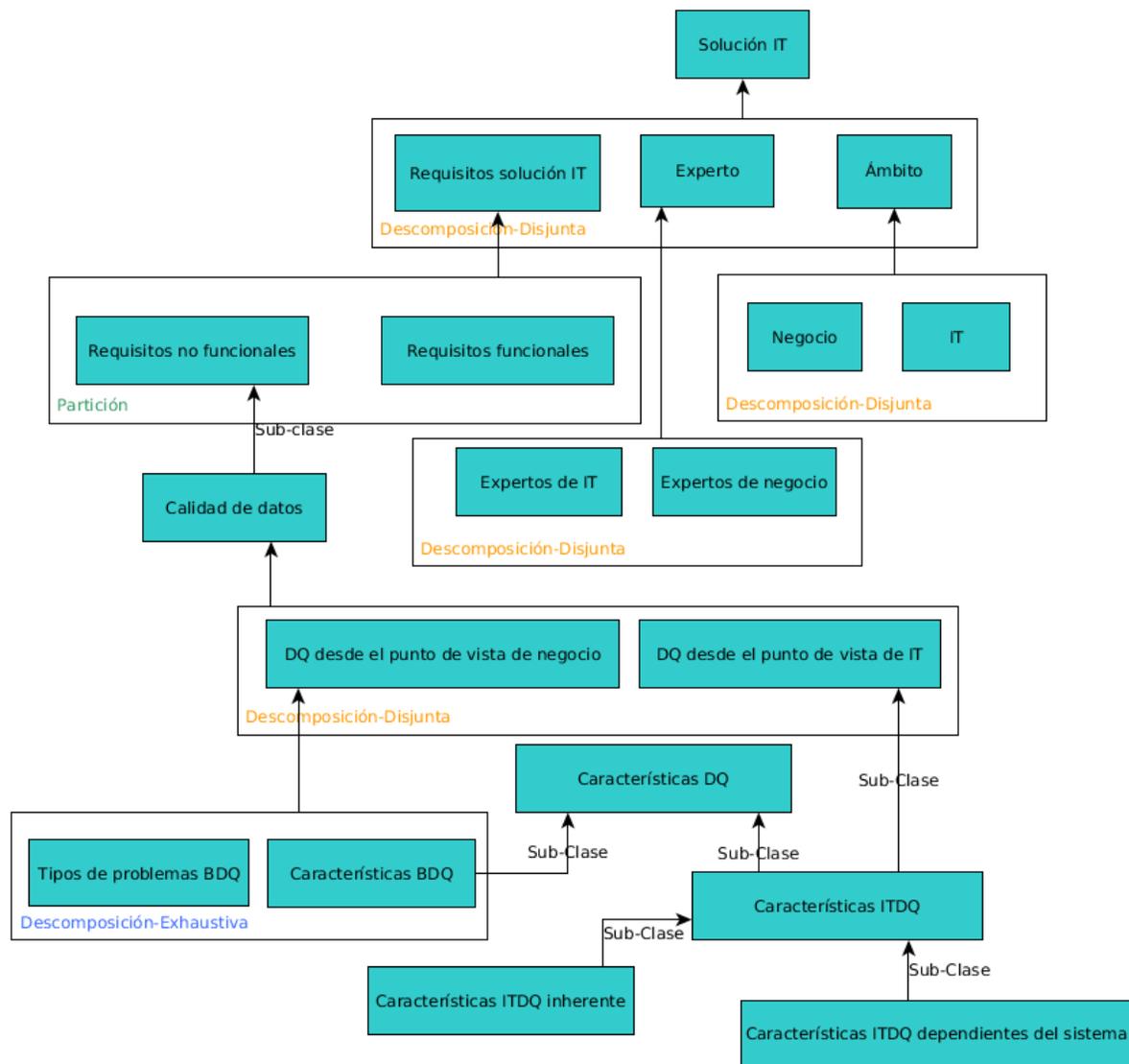


Fig. 5.3: Taxonomía de Conceptos

5.2.3. Tarea 3: Diagramas de relaciones binarias

En base a la taxonomía (ver Fig. 5.3) se construyó un diagrama de relaciones binarias ad hoc. Ésto con el objetivo de establecer las relaciones ad hoc existentes entre conceptos de la taxonomía. Los diagramas de relaciones binarias de esta conceptualización se presentan en las Fig. 5.4.

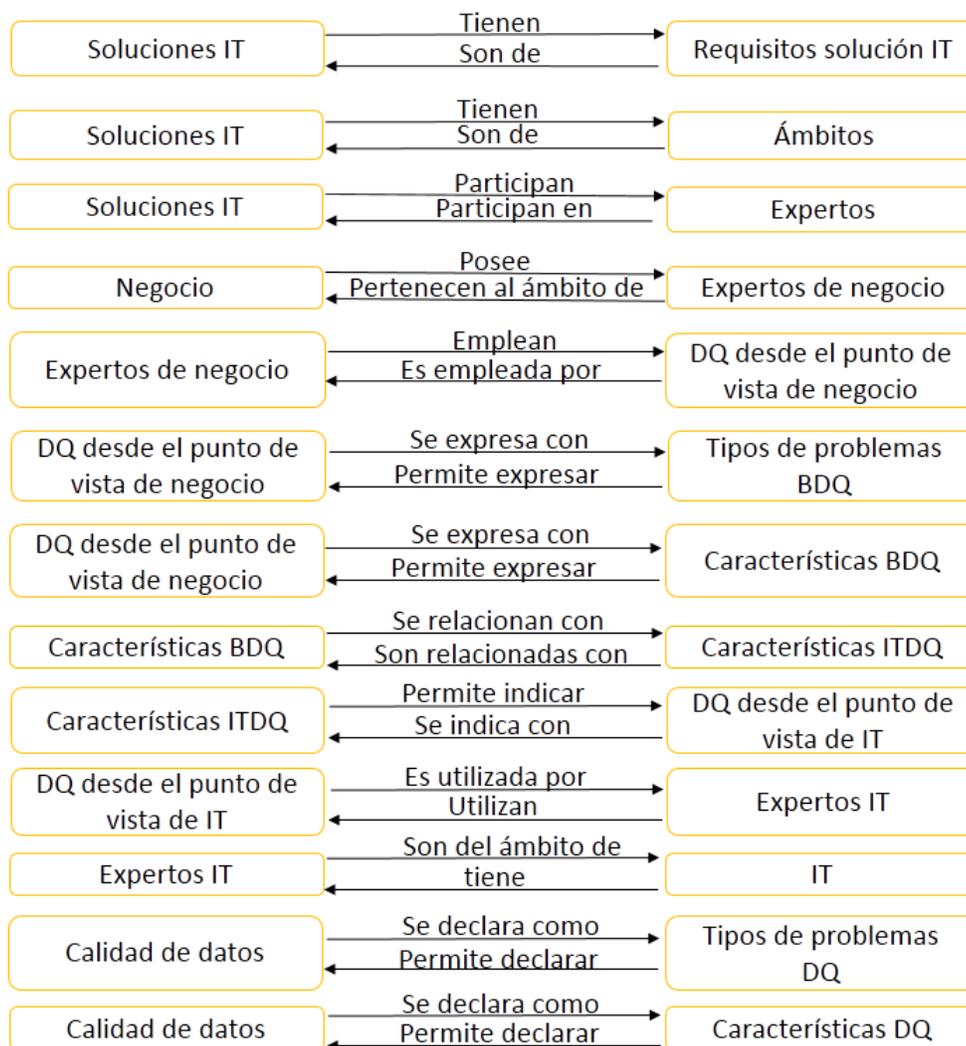


Fig. 5.4: Diagramas de relaciones binarias ad hoc

5.2.4. Tarea 4: Diccionario de conceptos

En la Tabla 5.7 se puede ver el diccionario de conceptos que contiene todos los conceptos del dominio, sus relaciones, sus instancias, sus atributos de clase y de instancia.

Nombre del concepto	Instancias	Atributos de clase	Atributos de Instancias	Relaciones
Ámbitos	-	-	-	-
Expertos	-	-	-	-
Expertos de IT	-	-	-	-
Expertos de negocio	-	-	-	-
Negocio	-	-	-	-
IT	-	-	-	-
Soluciones IT	-	-	-	-
Requisitos solución IT	-	-	-	-
Requisitos funcionales	-	-	-	-
Requisitos no funcionales	-	-	-	-
Calidad de datos	-	-	-	-
Tipos de problemas DQ	-	-	-	-
Tipos de problemas BDQ	Falta de integridad en los datos, Datos no válidos, erróneos, Pérdida de datos, Datos no adecuados, Datos duplicados, Falta de estandarización en los datos, Datos no veraces, Datos no accesibles, Asimetría en los datos registrados, Falta de completitud de los datos, Datos mal escritos, Falta de procedimientos de validación de los datos, Falta de compromiso al registrar datos, Inconsistencia en los datos, Datos imprecisos, Datos no actualizados, Ambigüedad en los datos, Datos poco confiables.	-	nombre, descripción.	se vinculan con
Características DQ	-	-	-	-
Características BDQ	Actualidad, Adecuación, Amplitud, Aplicabilidad, Eficacia, Completitud, Comprensibilidad, Confiabilidad, Confidencialidad, Consistencia, Credibilidad, Disponibilidad, Duplicidad, Estandarización, Estructuración, Exactitud, Integridad, Interpretabilidad, Libre de errores, Mantenibilidad, Objetividad, Operatividad, Oportunidad, Relevancia, Reputación, Eficiencia, Seguridad, Sincronización, Trazabilidad, Usabilidad, Utilidad, Valor añadido, Veracidad, Accesibilidad, Conformidad.	-	nombre, definición.	se relacionan con
Características ITDQ	Exactitud, Completitud, Consistencia, Credibilidad, Actualidad, Accesibilidad, Conformidad, Confidencialidad, Eficiencia, Precisión, Trazabilidad, Comprensibilidad, Disponibilidad, Portabilidad, Recuperabilidad.	-	nombre, definición.	-
Características ITDQ inherente	Precisión, Completitud, Consistencia, Credibilidad, Actualidad, Accesibilidad, Conformidad, Confidencialidad, Eficiencia, Exactitud, Trazabilidad, Comprensibilidad.	-	nombre, definición.	-
Características ITDQ dependientes del sistema	Accesibilidad, Conformidad, Confidencialidad, Eficiencia, Exactitud, Trazabilidad, Comprensibilidad, Disponibilidad, Portabilidad, Recuperabilidad.	-	nombre, definición.	-
DQ desde el punto de vista de IT	-	-	-	-
DQ desde el punto de vista de negocio	-	-	-	-

Tabla 5.7: Diccionario de conceptos de la Ontología

5.2.5. Tarea 5: Describir las relaciones binarias ad hoc

Corresponde a la tarea de describir en detalle todas las relaciones binarias ad hoc identificadas en el diagrama de relaciones binarias e incluidas en el diccionario de conceptos. La descripción de las relaciones binarias ad hoc de esta conceptualización se muestran en la Tabla 5.8.

Nombre de la Relación	Concepto origen	Cardinalidad máxima	Concepto destino	Relación inversa
emplean	Expertos de negocio	1	DQ desde el punto de vista de negocio	es empleada por
participan	Soluciones IT	N	Expertos	participan en
posee	Negocio	N	Expertos de negocio	pertenecen al ámbito de
se declara como	Calidad de datos	N	Tipos de problemas de DQ, Características DQ.	permiten declarar
se expresa con	DQ desde el punto de vista de negocio	N	Tipos de problemas de BDQ, Características BDQ	permiten expresar
se indica con	DQ desde el punto de vista de IT	N	Características ITDQ	permite indicar
se relacionan con	Características BDQ	N	Características ITDQ	son relacionadas con
se vinculan con	Tipos de problemas de DQ	N	Características BDQ	son vinculadas con
tiene	IT	N	Expertos de IT	son del ámbito de
tienen	Soluciones IT	N	Ámbitos, Requisitos solución IT	son de
utilizan	Expertos de IT	1	DQ desde el punto de vista de IT	es utilizada por

Tabla 5.8: Descripción de las relaciones binarias ad hoc

5.2.6. Tarea 6: Describir los atributos de instancia

Corresponde a la tarea de describir en detalle todos los atributos de instancia incluidos en el diccionario de conceptos.

Cada fila de la tabla de atributos de instancia contiene la descripción detallada de un atributo de instancia. Por cada atributo de instancia, el desarrollador de la ontología debe especificar su nombre, el concepto al que el atributo pertenece (los atributos son locales a los conceptos),

su tipo de valor, su rango de valores (en el caso de atributos numéricos), y sus cardinalidades mínima y máxima. En la Tabla 5.9 se describen los atributos de instancia.

Nombre del atributo de instancia	Concepto	Tipo de valores	Rango de valores	Cardinalidad
nombre	Tipos de problemas BDQ, Características BDQ, Características ITDQ.	Cadena de caracteres	-	[1,1]
descripción	Tipos de problemas BDQ	Cadena de caracteres	-	[1,1]
definición	Características BDQ, Características ITDQ	Cadena de caracteres	-	[1,1]

Tabla 5.9: Descripción de los atributos de instancia

5.2.7. Tareas de la 7 a la 10

La tarea 7 corresponde a la descripción de los atributos de clase, esta tarea se realizó, pero no se logró identificar alguno. La tarea 8 corresponde a la descripción de las constantes, esta tarea se realizó, al analizar el conocimiento no se logró encontrar constantes. La tarea 9 corresponde a la identificación de axiomas formales, se revisó la tarea y se buscaron axiomas, pero no se logró identificar ninguna como parte del conocimiento. La tarea 10 corresponde a la definición de reglas, se analizó el conocimiento a modelar en profundidad, pero no se logró identificar alguna regla.

5.2.8. Tarea 11: Describir instancias

En esta tarea, para cada instancia, se define: su nombre, el nombre del concepto al que pertenece y los valores de sus atributos de instancia, si se conocen. En las Tablas 5.10, 5.11, 5.12, 5.13, 5.14 y 5.15 se presentan todas las descripciones de instancia.

Nombre de la instancia	Nombre del concepto	Atributo	Valores
exactitud	Características ITDQ	nombre	Exactitud
		definición	El grado en el cual el dato tiene atributos que representan el valor correcto de un concepto o evento en un contexto específico de uso.
completitud	Características ITDQ	nombre	Completitud
		definición	El grado en el cual el dato asociado a una entidad tiene valores para todos los atributos esperados e instancias de entidad relacionadas, de acuerdo a un contexto específico de uso.
consistencia	Características ITDQ	nombre	Consistencia
		definición	El grado en el cual el dato tiene atributos libres de contradicción y son coherentes con otros datos en un contexto específico de uso.
credibilidad	Características ITDQ	nombre	Credibilidad
		definición	El grado en el cual el dato tiene atributos considerados como verdaderos y creíbles por usuarios en un contexto específico de uso.
actualidad	Características ITDQ	nombre	Actualidad
		definición	El grado en el cual el dato tiene los atributos que son del período correcto en un contexto específico de uso.
accesibilidad	Características ITDQ	nombre	Accesibilidad
		definición	El grado en el cual se puede acceder al dato en un contexto específico de uso.
conformidad	Características ITDQ	nombre	Conformidad
		definición	El grado en el cual el dato tiene atributos que se adhieren a normas, convenciones o regulaciones vigentes y reglas relacionadas con la calidad de datos en un contexto específico de uso.
confidencialidad	Características ITDQ	nombre	Confidencialidad
		definición	El grado en el cual el dato tiene los atributos que aseguran que este es sólo accesible e interpretable por usuarios autorizados en un contexto específico de uso.
eficiencia	Características ITDQ	nombre	Eficiencia
		definición	El grado en el cual el dato tiene los atributos que pueden ser procesados y proporciona los niveles esperados de funcionamiento (desempeño) usando las cantidades y los tipos de recursos apropiados en un contexto específico de uso.
precisión	Características ITDQ	nombre	Precisión
		definición	El grado en el cual el dato tiene atributos que son exactos o que proporcionan discernimiento en un contexto específico de uso.
trazabilidad	Características ITDQ	nombre	Trazabilidad
		definición	El grado en el cual el dato tiene atributos que proporcionan un rastro de auditoría de acceso a los datos y de cualquier cambio hecho a los datos en un contexto específico de uso.
comprensibilidad	Características ITDQ	nombre	Comprensibilidad
		definición	El grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios, y es expresado en lenguajes apropiados, símbolos y unidades en un contexto específico de uso.
disponibilidad	Características ITDQ	nombre	Disponibilidad
		definición	El grado en el cual el dato tiene atributos que le permiten ser recuperados por usuarios autorizados y/o aplicaciones en un contexto específico de uso.

Tabla 5.10: Descripción de instancias(1/6)

Nombre de la instancia	Nombre del concepto	Atributo	Valores
portabilidad	Características ITDQ	nombre	Portabilidad
		definición	El grado en el cual los datos tienen atributos que les permiten ser instalados, substituidos o movidos de un sistema a otro conservando la calidad existente, en un contexto específico de uso.
recuperabilidad	Características ITDQ	nombre	Recuperabilidad
		definición	El grado en el cual el dato tiene atributos que le permiten mantener y conservar un nivel especificado de operaciones y calidad, aún en caso de falla, en un contexto específico de uso.
Actualidad	Características BDQ	nombre	Actualidad
		definición	Grado de actualización y vigencia de los datos.
Adecuación	Características BDQ	nombre	Adecuación
		definición	Grado en el cual los datos se presentan de manera concisa, flexible y organizada, considerando los objetivos de quienes los necesitan para ayudarlo a alcanzar sus metas.
Amplitud	Características BDQ	nombre	Amplitud
		definición	Grado de alcance en que los datos cubren satisfactoriamente las necesidades de los usuarios.
Aplicabilidad	Características BDQ	nombre	Aplicabilidad
		definición	Grado en que los datos permiten ser directamente utilizados por un trabajador en un contexto específico.
Eficacia	Características BDQ	nombre	Eficacia
		definición	Grado en que los datos satisfacen las necesidades de quienes los necesitan para el desarrollo de una tarea específica.
Compleitud	Características BDQ	nombre	Compleitud
		definición	Grado en que todos los datos necesarios están presentes para el desarrollo de una tarea.
Comprensibilidad	Características BDQ	nombre	Comprensibilidad
		definición	Grado en que los datos se entienden fácilmente.
Confiabilidad	Características BDQ	nombre	Confiabilidad
		definición	Grado de concordancia entre diferentes datos registrados en condiciones similares.
Confidencialidad	Características BDQ	nombre	Confidencialidad
		definición	Grado en que se establece control al acceso a los datos de una organización.
Consistencia	Características BDQ	nombre	Consistencia
		definición	Grado en que los datos de diferentes conjuntos de datos son coherentes y no contradictorios.
Credibilidad	Características BDQ	nombre	Credibilidad
		definición	Grado en que se consideran los datos como verdaderos y creíbles.
Disponibilidad	Características BDQ	nombre	Disponibilidad
		definición	Grado en que los datos se pueden obtener cuando se necesitan para un determinado uso.
Duplicidad	Características BDQ	nombre	Duplicidad
		definición	Grado de existencia de datos repetidos en más de un sistema de información de una organización, cuando esto no es necesario.
Estandarización	Características BDQ	nombre	Estandarización
		definición	Grado en que los datos están definidos de manera uniforme y se comprenden sus valores.

Tabla 5.11: Descripción de instancias(2/6)

Nombre de la instancia	Nombre del concepto	Atributo	Valores
Estructuración	Características BDQ	nombre	Estructuración
		definición	Grado en que los datos poseen la estructura y el formato correcto para utilizarlos.
Exactitud	Características BDQ	nombre	Exactitud
		definición	Grado en que los datos son reales y se representan correctamente en concordancia con los valores estándares definidos.
Integridad	Características BDQ	nombre	Integridad
		definición	Grado en que todos los datos necesarios están presentes y de manera apropiada.
Interpretabilidad	Características BDQ	nombre	Interpretabilidad
		definición	Grado en que los datos son apropiados en idioma, símbolos y unidades, y las definiciones son claras.
Libre de errores	Características BDQ	nombre	Libre de errores
		definición	Grado en que los datos son correctos y libres de defectos.
Mantenibilidad	Características BDQ	nombre	Mantenibilidad
		definición	Grado en que los datos pueden ser accedidos, usados, conservados, actualizados, mantenidos y manipulados.
Objetividad	Características BDQ	nombre	Objetividad
		definición	Grado en que los datos son imparciales y sin prejuicios. Esto implica aplicar un enfoque que permita asegurar que los datos sean exactos, fiables, completos y claros.
Operatividad	Características BDQ	nombre	Operatividad
		definición	Grado de conformidad expresada por un trabajador respecto a la correspondencia, cercanía de los datos con la realidad y la utilidad, en concordancia con sus necesidades.
Oportunidad	Características BDQ	nombre	Oportunidad
		definición	Grado en que los datos están actualizados y disponibles en el momento que se requiere su uso.
Relevancia	Características BDQ	nombre	Relevancia
		definición	Grado en que los datos son útiles y pertinentes para el desarrollo de una tarea determinada.
Reputación	Características BDQ	nombre	Reputación
		definición	Grado en que los datos están bien considerados en términos de su origen y contenido.
Eficiencia	Características BDQ	nombre	Eficiencia
		definición	Grado en que los datos satisfacen las necesidades de quien los necesita, teniendo en consideración el acceso y el fácil uso de esta.
Seguridad	Características BDQ	nombre	Seguridad
		definición	Grado en que el acceso a los datos es restringido adecuadamente para mantener su seguridad.
Sincronización	Características BDQ	nombre	Sincronización
		definición	Grado de equivalencia entre los datos almacenados en un sistema principal y los sistemas locales de una organización.
Trazabilidad	Características BDQ	nombre	Trazabilidad
		definición	Grado en que se puede hacer un seguimiento del origen de los datos disponibles.
Usabilidad	Características BDQ	nombre	Usabilidad
		definición	Grado en que la presentación de los datos es útil y/o eficaz para su propósito.

Tabla 5.12: Descripción de instancias(3/6)

Nombre de la instancia	Nombre del concepto	Atributo	Valores
Utilidad	Características BDQ	nombre	Utilidad
		definición	Grado en que los datos son beneficiosos y están disponibles para un uso previsto.
Valor añadido	Características BDQ	nombre	Valor añadido
		definición	Grado en que los datos proveen beneficios y ventajas como producto de su utilización.
Veracidad	Características BDQ	nombre	Veracidad
		definición	Grado en que los datos se ajustan a su definición, a los valores de dominio y a las reglas de negocio.
Accesibilidad	Características BDQ	nombre	Accesibilidad
		definición	Grado en que los datos están disponibles en un momento dado en el tiempo para su uso.
Conformidad	Características BDQ	nombre	Conformidad
		definición	Grado en que los datos se ajustan a los formatos estándares establecidos por una organización.
Falta de integridad en los datos	Tipos de problemas BDQ	nombre	Falta de integridad en los datos
		descripción	Corresponde a problemas relacionados con los datos registrados cuando estos están incompletos y/o incorrectos. Por ejemplo, la dirección de entrega de un producto sin el número de la casa o departamento.
Datos no válidos, erróneos	Tipos de problemas BDQ	nombre	Datos no válidos, erróneos
		descripción	Los datos no son correctos o no representan de forma correcta la realidad. Esto implica costos adicionales para recapturar o actualizar los datos y/o pérdida de oportunidades de negocio por el manejo de datos incorrectos (Datos erróneos de clientes, Datos de facturación erróneos, entre otros.)
Pérdida de datos	Tipos de problemas BDQ	nombre	Pérdida de datos
		descripción	La ausencia de los datos que se guardaron en un sistema de información en un momento dado, cuando se necesitan pensando que deberían estar guardados.
Datos no adecuados	Tipos de problemas BDQ	nombre	Datos no adecuados
		descripción	Mala presentación de los datos para su uso, los cuales pueden estar con un nivel de detalle excesivo o ser inapropiados.
Datos duplicados	Tipos de problemas BDQ	nombre	Datos duplicados
		descripción	La organización posee datos repetidos en diferentes partes, lo que puede llevar a interpretaciones erróneas e inconsistencias (por ejemplo, un mismo cliente con direcciones diferentes).
Falta de estandarización en los datos	Tipos de problemas BDQ	nombre	Falta de estandarización en los datos
		descripción	Los datos no se ajustan a los estándares requeridos por la organización. No son registrados siempre con el mismo formato (números telefónicos, direcciones), no se ajustan a los rangos de valores válidos (edades, género), entre otros.
Datos no veraces	Tipos de problemas BDQ	nombre	Datos no veraces
		descripción	Corresponde a datos irreales o engañosos.

Tabla 5.13: Descripción de instancias(4/6)

Nombre de la instancia	Nombre del concepto	Atributo	Valores
Datos no accesibles	Tipos de problemas BDQ	nombre	Datos no accesibles
		descripción	Un trabajador no puede acceder a los datos que necesita para el desarrollo de una tarea, ya sea porque estos no estén disponibles o porque la forma de obtenerlos sea engorrosa, causando demoras en el desarrollo de dicha tarea.
Asimetría en los datos registrados	Tipos de problemas BDQ	nombre	Asimetría en los datos registrados
		descripción	Corresponde a la diferencia que existe entre los datos registrados de algo particular, por dos o más trabajadores de áreas diferentes, ejemplo un encargado de bodega y un vendedor tienen datos diferentes sobre una transacción.
Falta de completitud de los datos	Tipos de problemas BDQ	nombre	Falta de completitud de los datos
		descripción	Se refiere a que algunos datos en la organización no están completos, por ejemplo, una ficha médica que no contiene enfermedades preexistentes, una factura donde no se registran todos los ítems comprados. Lo anterior, provoca que datos necesarios para la organización no estén disponibles para su uso, generando problemas en la toma de decisiones.
Datos mal escritos	Tipos de problemas BDQ	nombre	Datos mal escritos
		descripción	El registro de los datos se puede realizar con errores tales como faltas ortográficas, intercambio de letras, símbolos inadecuados, etc. Este tipo de problemas puede ocasionar pérdida de ingresos, ineficiencia de un proceso, incumplimiento de normas o regulaciones, entre otros.
Falta de procedimientos de validación de los datos	Tipos de problemas BDQ	nombre	Falta de procedimientos de validación de los datos
		descripción	El registro de datos (en sistemas de información), sin las respectivas validaciones, ocasiona problemas de calidad de datos, por ejemplo el ingreso de un RUT erróneo por no validar el dígito verificador.
Falta de compromiso al registrar datos	Tipos de problemas BDQ	nombre	Falta de compromiso al registrar datos
		descripción	En una organización algunos trabajadores no están conscientes de la importancia de los datos que están registrando, e ingresan documentos y/o formularios con datos faltantes y/o incorrectos.
Inconsistencia en los datos	Tipos de problemas BDQ	nombre	Inconsistencia en los datos
		descripción	Los datos no son coherentes, por ejemplo, en las organizaciones para hablar de un mismo tema (servicio, producto, recurso) en las distintas áreas funcionales se utilizan términos distintos.
Datos imprecisos	Tipos de problemas BDQ	nombre	Datos imprecisos
		descripción	Los datos registrados en los sistemas de las organizaciones se encuentran poco detallados o faltos de precisión, con respecto a la realidad, provocando una menor productividad, mayores costos, menor satisfacción de los clientes finales o pérdida de ingresos.

Tabla 5.14: Descripción de instancias(5/6)

Nombre de la instancia	Nombre del concepto	Atributo	Valores
Datos no actualizados	Tipos de problemas BDQ	nombre	Datos no actualizados
		descripción	Los datos que tienen disponibles los trabajadores están obsoletos y no representan la realidad actual, por ejemplo el correo electrónico o el número telefónico de un proveedor que está registrado no corresponde al actual.
Ambigüedad en los datos	Tipos de problemas BDQ	nombre	Ambigüedad en los datos
		descripción	Los datos ambiguos son aquellos que pueden tener significados diferentes para las distintas personas que los utilizan, perdiendo el sentido de lo que se quería representar con ellos.
Datos poco confiables	Tipos de problemas BDQ	nombre	Datos poco confiables
		descripción	Corresponde a datos que provienen de fuentes o cuyo origen es poco confiable o no verificable.

Tabla 5.15: Descripción de instancias(6/6)

5.3. Implementación de la Ontología

En esta sección se expondrá en detalle cómo se realizó la formalización e implementación de la ontología.

5.3.1. Ambiente de desarrollo

Para el desarrollo de la ontología se utilizó el software Protégé versión 5.0 que es una herramienta para el desarrollo de ontologías y sistemas basados en el conocimiento creada en la Universidad de Stanford. Protégé está desarrollada en Java y funciona adecuadamente bajo el sistema operativo Linux (en concreto con la distribución Elementary OS Freya). Las aplicaciones desarrolladas con Protégé son empleadas en resolución de problemas y toma de decisiones en dominios particulares. La herramienta Protégé emplea una interfaz gráfica de usuario que facilita la creación de ontologías. En la Fig. 5.5 se puede apreciar la interfaz de la herramienta Protégé (Research Stanford Center for Biomedical Informatics, 2015).

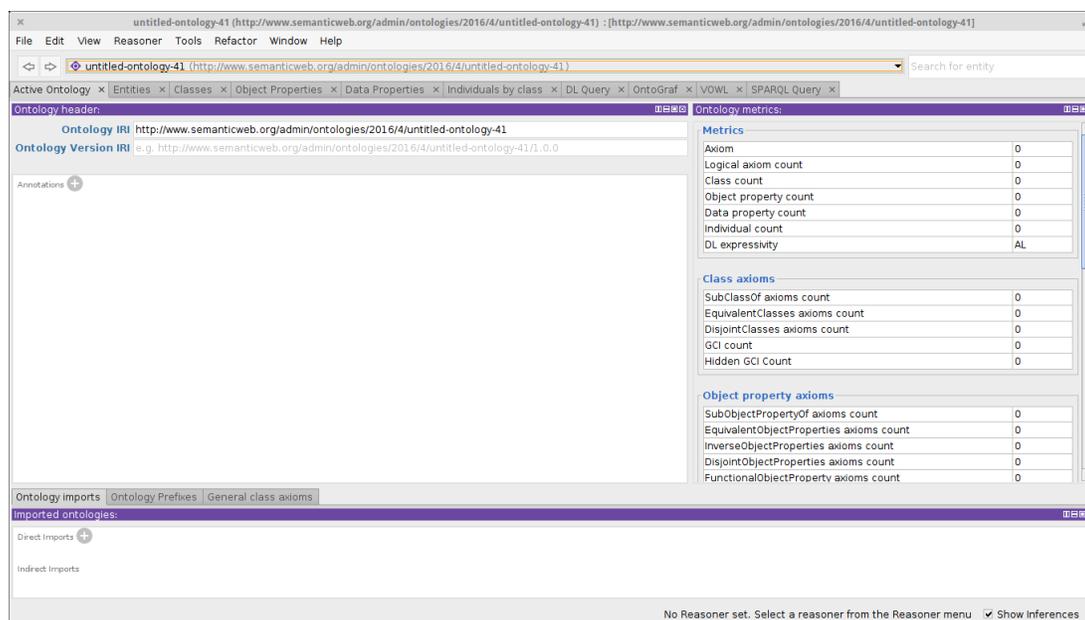


Fig. 5.5: Interfaz gráfica de Protégé

5.3.2. Desarrollo de la ontología

En base a la conceptualización descrita en la Sección 5.2 con la ayuda de la metodología METHONTOLOGY (Juristo et al., 1997) se comienza con la implementación de la ontología con la herramienta Protégé. Las actividades que se realizaron para desarrollar la ontología se presenta a continuación.

Primero se creó una nueva ontología en el software Protégé con el lenguaje OWL. Además, se escribieron los comentarios esenciales de la ontología como lo son una descripción, el número de versión, entre otras cosas, como que se puede observar en la Fig 5.6.

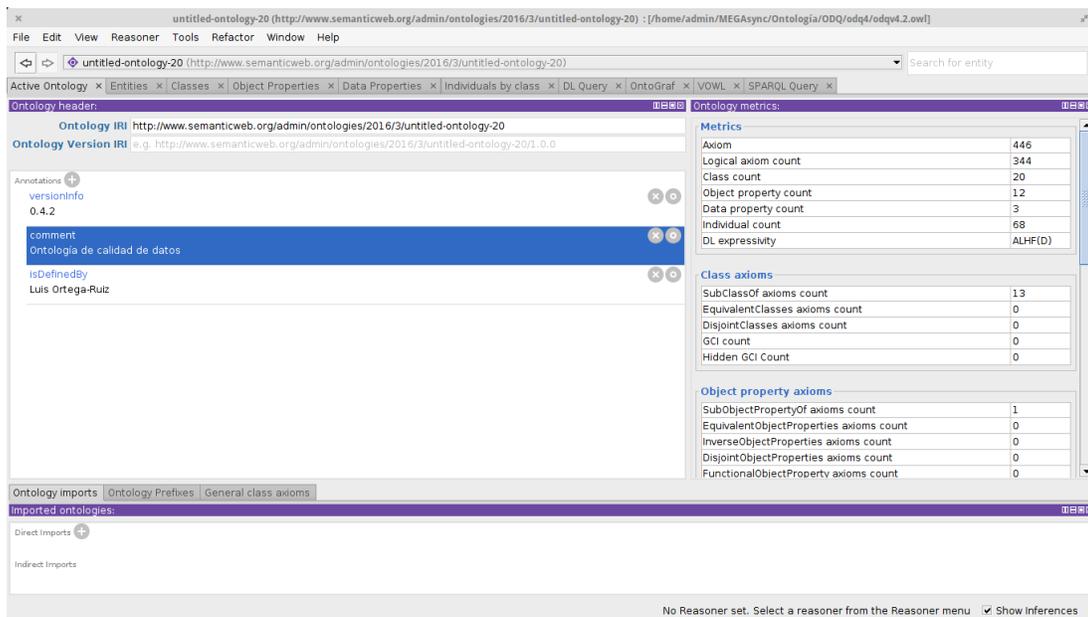


Fig. 5.6: Captura de la creación y configuración de la ontología

Luego se crearon las clases y la jerarquía de clases de la ontología en base a la taxonomía desarrollada en la conceptualización que se puede ver en la Fig. 5.3. En la Fig. 5.7 se puede apreciar cómo se realizó la jerarquía de clases con el software Protégé, donde cada concepto definido en la taxonomía corresponde a una clase.

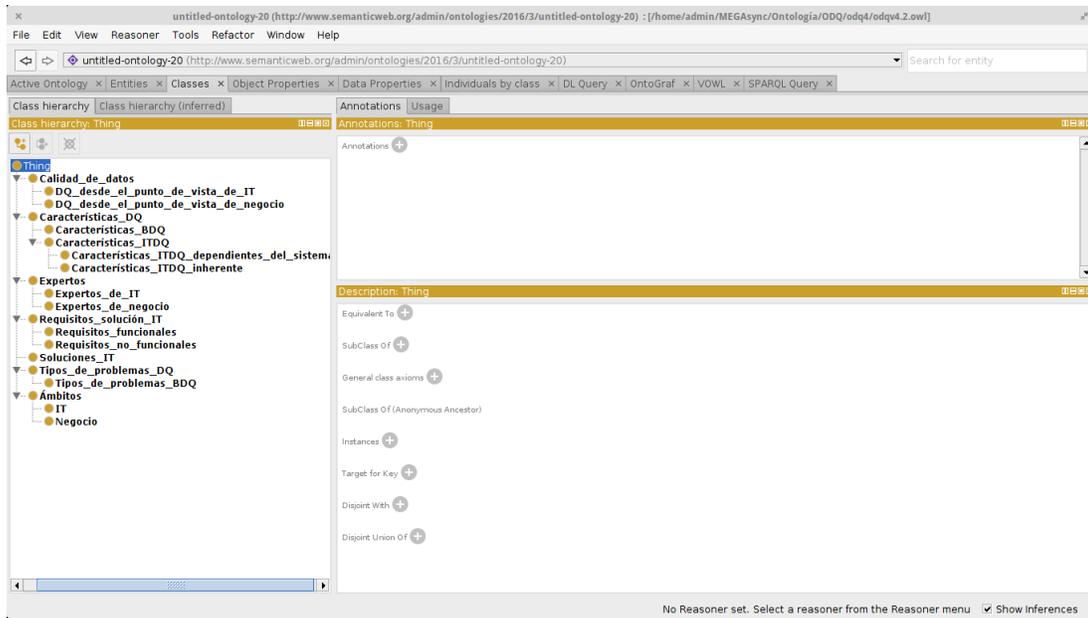


Fig. 5.7: Captura de la jerarquía de clases

Una vez creada la jerarquía de clases se establecieron las propiedades de objeto siguiendo lo establecido en la Tabla 5.8 de relaciones ad hoc de la conceptualización. En la Fig. 5.8 se puede apreciar cómo se crearon en el software Protégé.

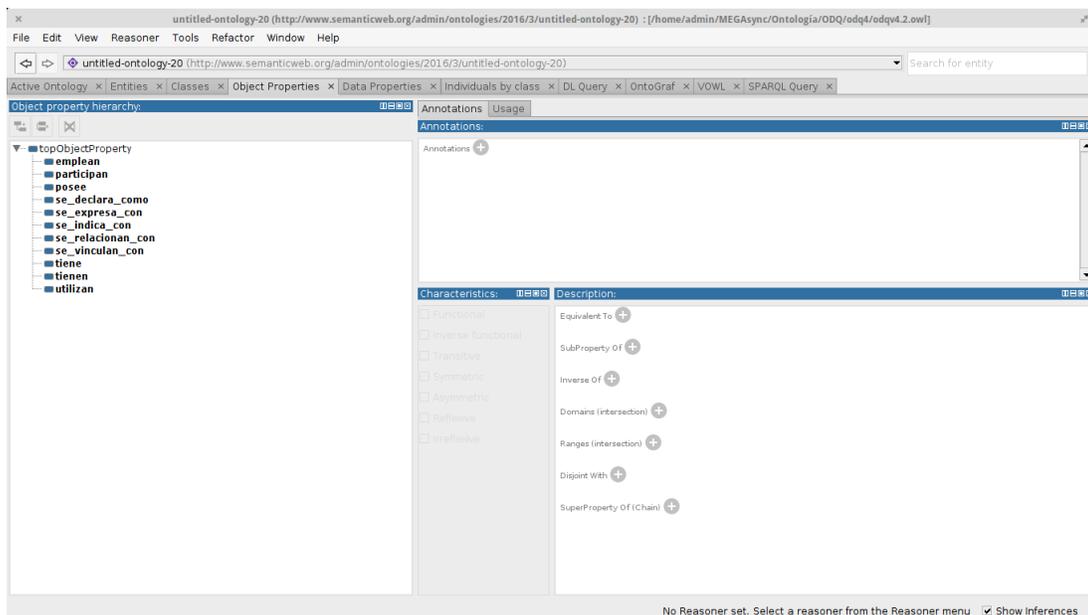


Fig. 5.8: Captura de la definición de las propiedades de objeto

También se definieron las propiedades de datos con lo establecido en la Tabla 5.9 que describen atributos de instancia. En la Fig. 5.9 se puede apreciar cómo se establecieron en el software.

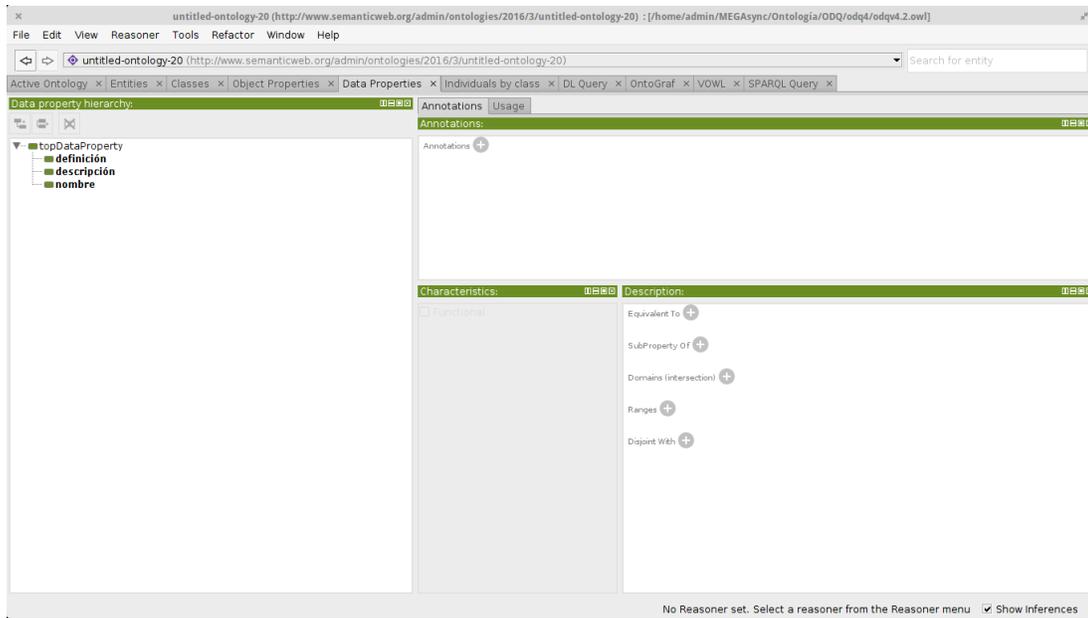


Fig. 5.9: Captura de la definición de las propiedades de dato

Finalmente se agregaron las instancias, tarea que consiste en traspasar la información que se presenta en las Tablas 5.10, 5.11, 5.12, 5.13, 5.14 y 5.15 que describe las instancias de la conceptualización de la ontología creada con el software Protégé, en la Fig. 5.10 se puede ver alguna de las instancias agregadas en el software Protégé.

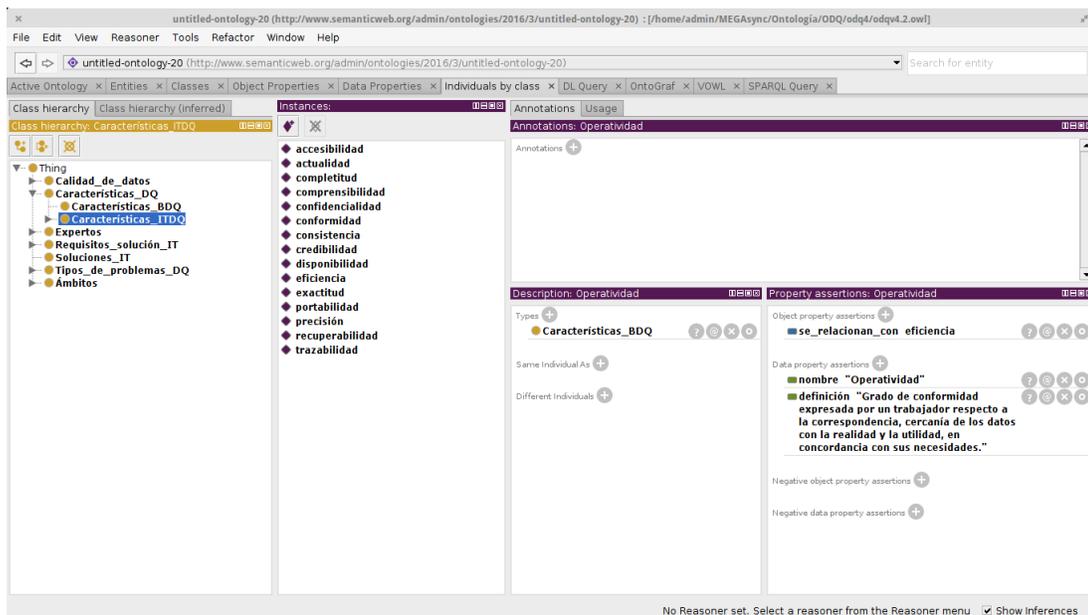


Fig. 5.10: Captura de las instancias agregadas a la ontología

La ontología creada con sus clase e instancias se puede observar en la Fig. 5.11.

La primera pregunta de competencia es “¿Qué problemas o características de DQ existen?”, su traducción al lenguaje de consultas SPARQL es la siguiente:

```
PREFIX odq: <http://www.odeq.quesi.cl/odq.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?Características_y_tipos_de_problemas
WHERE {
  { ?Características_y_tipos_de_problemas rdf:type odq:Características_BDQ }
  UNION { ?Características_y_tipos_de_problemas rdf:type odq:Características_ITDQ }
  UNION { ?Características_y_tipos_de_problemas rdf:type odq:Tipos_de_problemas_BDQ }
}
```

La segunda pregunta de competencia es “¿Qué característica DQ está vinculada a un problema de DQ?”, su traducción al lenguaje de consultas SPARQL es la siguiente:

```
PREFIX odq: <http://www.odeq.quesi.cl/odq.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?Tipos_de_problemas_BDQ ?Características_BDQ
WHERE {
  ?Características_BDQ rdf:type odq:Características_BDQ.
  ?Tipos_de_problemas_BDQ rdf:type odq:Tipos_de_problemas_BDQ.
  ?Tipos_de_problemas_BDQ odq:se_vinculan_con ?Características_BDQ
}
```

La tercera pregunta de competencia es “¿Qué características de DQ son utilizadas en el área de negocios?”, su traducción al lenguaje de consultas SPARQL es la siguiente:

```
PREFIX odq: <http://www.odeq.quesi.cl/odq.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?Características_BDQ
WHERE { ?Características_BDQ rdf:type odq:Características_BDQ
}
```

La cuarta pregunta de competencia es “¿Qué características de DQ son utilizadas en el área de las TI?”, su traducción al lenguaje de consultas SPARQL es la siguiente:

```
PREFIX odq: <http://www.odeq.quesi.cl/odq.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?Características_ITDQ
WHERE {
  ?Características_ITDQ rdf:type odq:Características_ITDQ
}
```

La quinta pregunta de competencia es “¿Cómo se relaciona una característica de DQ del área de negocios con una característica de DQ de las TI?” su traducción al lenguaje de consultas SPARQL es la siguiente:

```
PREFIX odq: <http://www.odeq.quesi.cl/odq.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?Características_BDQ ?Características_ITDQ
WHERE {
?Características_BDQ rdf:type odq:Características_BDQ.
?Características_ITDQ rdf:type odq:Características_ITDQ.
?Características_BDQ odq:se_relacionan_con ?Características_ITDQ
}
```

La ontología implementada con la herramienta Protégé siguiendo la metodología METHONTOLOGY es capaz de entregar respuestas a todas las consultas de competencia establecidas durante la especificación con éxito.

5.4. Validación de la ontología

El desarrollo de la ontología fue guiado por la metodología METHONTOLOGY, esta metodología asimila el desarrollo de una ontología como un desarrollo en términos de la ingeniería de software. Por lo cual, para la construcción de la ontología se realizaron varias iteraciones, comprobando en cada una de estas la capacidad de la ontología para dar respuesta a las preguntas de competencia definidas por un panel de expertos. En la Fig. 5.13 se muestra la ontología resultante de la primera y última interacción.

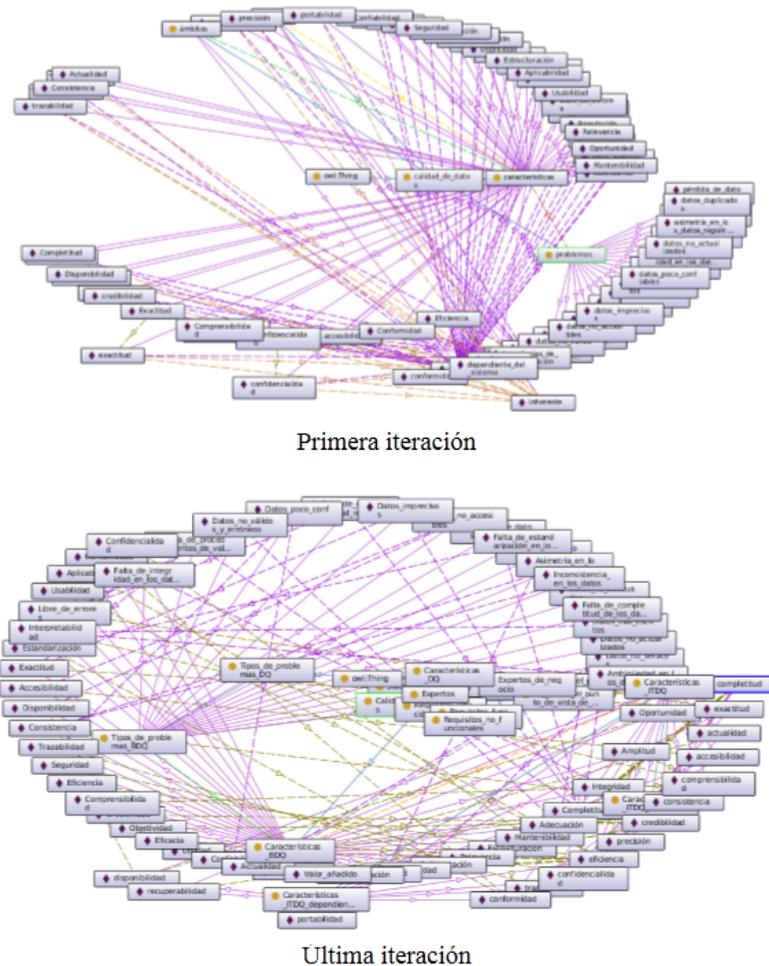


Fig. 5.13: Ontología resultante de la primera y última interacción.

5.5. Vinculación de la ontología con el método BPIDQ

El método BPIDQ consta de 4 etapas. La etapa 1 es de modelado de procesos de negocio consciente de la calidad de datos, la etapa 2 es de especificación de requisitos de calidad de datos, la etapa 3 es de análisis y mejora de BP relacionado con la calidad de datos y la etapa 4 que corresponde a la generación de diagramas de casos de uso. Como se observa en la Tabla 5.16 en el método se utiliza como herramienta el BPMN 2.0 extendido con dqBP en las tres primeras etapas y en las dos primeras etapas participan expertos de negocio y analistas de BP. Es en estas etapas donde la ontología juega su papel otorgando terminología de DQ utilizada por expertos de negocio (en forma de tipos de problemas o características de calidad de datos de negocio) y la relación de esta con la terminología de DQ del área de las TI (características de DQ). De esta manera al utilizarla el BPMN 2.0 extendido con dqBP con la ontología permite facilitar la especificación de requisitos de calidad de datos para la construcción de un sistema de información.

	BPIDQ-S1	BPIDQ-S2	BPIDQ-S3	BPIDQ-S4
Entradas	Modelo BP	Modelo de BP con DQ-Flags	Modelo BP con DQ-Flags - Detalle de requisitos DQ-Flags	Modelo BP con actividades DQ agregadas
Trabajadores	Experto del Negocio/Analista BP	Analista BP - Experto DQ	Diseñador BP - Experto DQ	Analista de sistemas - Experto DQ
Herramientas	BPMN 2.0 extendido con dqBP	BPMN 2.0 extendido con dqBP	BPMN 2.0 extendido con dqBP - Repositorio Actividades DQ	Casos de uso UML - Repositorio de Casos de uso DQ
Salidas	Modelo de BP con DQ-Flags	Modelo de BP con DQ-Flags - Detalle de requisitos DQ	Modelo BP con actividades DQ agregadas	Diagramas de casos de uso del sistema

Tabla 5.16: Resumen de las etapas de BPIDQ (extraído de (Rodríguez y Caro, 2012))

Capítulo 6

Conclusiones

Este capítulo se estructura de la siguiente manera, en la Sección 6.1 se muestran los objetivos propuestos y se detalla como estos se cumplieron. Después, en la Sección 6.2 se dan a conocer los principales aportes. Luego, en la Sección 6.3 presenta el contraste de resultados y, finalmente, en la Sección 6.4 se establecen las líneas de investigación futuras abiertas por el desarrollo de esta tesis.

6.1. Análisis del cumplimiento de objetivos

El objetivo general de esta tesis es: “Crear una ontología que provea un modelo conceptual, que facilite la comunicación, respecto de la especificación de requisitos de calidad de datos en los procesos de negocio, entre los expertos de negocio y los expertos en TI” y los objetivos específicos se presentan a continuación con su análisis del cumplimiento.

El objetivo específico número 1. “Estudiar la literatura para establecer la terminología utilizada por los expertos y/o analistas de negocio para expresar problemas y/o requisitos de DQ en el contexto de los BP.” Para la Identificación de la terminología de calidad de datos se definió un método el cual se a presentado en el Capítulo 2, el cual se ejecutó obteniendo una lista de características y otra de problemas de calidad de datos como se puede evidenciar en el Capítulo 4.

El objetivo específico número 2. “Identificar conceptos y el vocabulario utilizado para describir requisitos y/o problemas de DQ en los BP, por parte de los expertos de negocio, mediante entrevistas.” Al estudiar la literatura para establecer la terminología utilizada se obtuvo un conjunto de características y problemas de calidad de datos descritos con vocabulario utilizado en el área de los negocios. Luego, esta terminología se utilizó para elaborar un estudio que permitiera validarla ante expertos y/o analistas de negocio como se puede observar en el Capítulo 4.

El objetivo específico número 3. “Definir una metodología y una herramienta para el desarrollo de la ontología a realizar.” Se estudiaron y analizaron las metodologías existentes para el desarrollo de ontologías como se evidencia en el Capítulo 3 y se optó por utilizar METHONTOLOGY por su nivel de documentación. Además, se estudiaron y analizaron las herramientas para la implementación de ontologías como se evidencia en el Capítulo 3 y se optó por utilizar Protége 5.0 por su nivel de documentación y compatibilidad.

El objetivo específico número 4. “Desarrollar una ontología de calidad de datos para el contexto de los procesos de negocio.” Se desarrolló la ontología siguiendo la metodología METHONTOLOGY partiendo con la conceptualización y, luego, la implementación que se desarrolló con la herramienta Protégé. Todo el desarrollo de este trabajo se documenta en el Capítulo 5.

6.2. Principales aportes

Los principales aportes de esta tesis son:

- Obtención de una lista de características y una tipología de problemas de calidad de datos expresadas con el lenguaje de las personas del negocio.
- Una ontología desarrollada en base a la terminología de calidad de datos identificada en el área de negocio y la definida en el estándar ISO/IEC25012.

6.3. Contrates de resultados

Como resultado de esta tesis se ha generado una publicación:

- Luis Ortega-Ruiz, Angélica Caro, Alfonso Rodríguez. Hacia una Ontología de Calidad de Datos, en el Ámbito del Negocio, para el Modelado de Procesos. SCCC Track - Workshop on Business Process Management, Santiago (Chile), 11 de Noviembre del 2015.

6.4. Líneas de trabajo futuro

Las principales líneas de investigación abiertas como resultado de esta tesis son las siguientes:

- Integrar la ontología desarrollada con la herramienta DQ BPMN Modeler que apoya al método BPiDQ con la idea de proveer la terminología identificada.
- Especificación de una ontología de DQ en el contexto de los procesos de negocio que permita establecer un lenguaje común entre los expertos de negocio y los expertos en TI respecto a los requisitos de seguridad.
- Especificación de una ontología para la evaluación de la calidad de datos mediante la definición de métricas.

Apéndice A

Instrumento

Terminología de Calidad de Datos en los Negocios

(Formulario)

El presente instrumento tiene como finalidad verificar la terminología utilizada por expertos del negocio para expresar tanto los problemas que tienen con sus datos, como la forma en que caracterizan la calidad de los datos.

Solicitamos a usted contestar el siguiente cuestionario, teniendo presente que la información sólo será utilizada con fines académicos y es de carácter estrictamente reservado.

Junto con el cuestionario, también le solicitamos llenar un consentimiento informado en el que Ud. nos autoriza a utilizar la información que nos proporcione.

Desde ya agradecemos su participación en este estudio.

Información referente a su filiación y experiencia profesional

Nombre de su organización	
Sector industrial (marque el correspondiente con una X) Retail Telecomunicaciones Energía Educacional Salud Alimentación Otro:
Cargo que desempeña en su organización:	
Años que desempeña en su cargo:	
Años de experiencia profesional:	

1. Problemas de Calidad de Datos.

A continuación se le presentará una lista de descripciones de problemas de Calidad de Datos que pueden ocurrir en su organización y una denominación de dicho problema. Por favor indicar su opinión respecto del nivel de conformidad de la descripción del problema con su denominación. Marque con una X el nivel de ajuste, siendo: (1) ninguno, (2) poco, (3) medio, (4) alto y (5) perfecto.

Descripción del problema de Calidad de datos	Denominación del problema de Calidad de Datos	Nivel de conformidad				
		1	2	3	4	5
Corresponde a problemas relacionados con los datos registrados cuando estos están incompletos y/o incorrectos. Por ejemplo, la dirección de entrega de un producto sin el número de la casa o departamento.	Falta de integridad en los datos					
Los datos no son correctos o no representan de forma correcta la realidad. Esto implica costos adicionales para recapturar o actualizar los datos y/o pérdida de oportunidades de negocio por el manejo de datos incorrectos (Datos erróneos de clientes, Datos de facturación erróneos, entre otros.)	Datos no válidos, erróneos					
La ausencia de los datos que se guardaron en un sistema de información en un momento dado, cuando se necesitan pensando que deberían estar guardados.	Pérdida de datos					
Mala presentación de los datos para su uso, los cuales pueden estar con un nivel de detalle excesivo o ser inapropiados.	Datos no adecuados					
La organización posee datos repetidos en diferentes partes, lo que puede llevar a interpretaciones erróneas e inconsistencias (por ejemplo, un mismo cliente con direcciones diferentes).	Datos duplicados					
Los datos no se ajustan a los estándares requeridos por la organización. No son registrados siempre con el mismo formato (números telefónicos, direcciones), no se ajustan a los rangos de valores válidos (edades, género), entre otros.	Falta de estandarización en los datos					
Un trabajador no puede acceder a los datos que necesita para el desarrollo de una tarea, ya sea porque estos no estén disponibles o porque la forma de obtenerlos sea engorrosa, causando demoras en el desarrollo de dicha tarea.	Datos no accesibles					
Corresponde a la diferencia que existe entre los datos registrados de algo particular, por dos o más trabajadores de áreas diferentes, ejemplo un encargado de bodega y un vendedor tienen datos diferente sobre una transacción.	Asimetría en los datos registrados					
Corresponde a datos que provienen de fuentes o cuyo origen es poco confiable o no verificable.	Datos poco confiables					

Descripción del problema de Calidad de datos	Denominación del problema de Calidad de Datos	Nivel de conformidad				
		1	2	3	4	5
Corresponde a datos irreales o engañosos.	Datos no veraces					
Se refiere a que algunos datos en la organización no están completos, por ejemplo, una ficha médica que no contiene enfermedades preexistentes, una factura donde no se registran todos los ítems comprados. Lo anterior, provoca que datos necesarios para la organización no estén disponibles para su uso, generando problemas en la toma de decisiones.	Falta de completitud de los datos					
El registro de los datos se puede realizar con errores tales como faltas ortográficas, intercambio de letras, símbolos inadecuados, etc. Este tipo de problemas puede ocasionar pérdida de ingresos, ineficiencia de un proceso, incumplimiento de normas o regulaciones, entre otros.	Datos mal escritos					
El registro de datos (en sistemas de información), sin las respectivas validaciones, ocasiona problemas de calidad de datos, por ejemplo el ingreso de un RUT erróneo por no validar el dígito verificador.	Falta de procedimientos de validación de los datos					
En una organización algunos trabajadores no están conscientes de la importancia de los datos que están registrando, e ingresan documentos y/o formularios con datos faltantes y/o incorrectos.	Falta de compromiso al registrar datos					
Los datos no son coherentes, por ejemplo, en las organizaciones para hablar de un mismo tema (servicio, producto, recurso) en las distintas áreas funcionales se utilizan términos distintos.	Inconsistencia en los datos					
Los datos registrados en los sistemas de las organizaciones se encuentran poco detallados o faltos de precisión, con respecto a la realidad, provocando una menor productividad, mayores costos, menor satisfacción de los clientes finales o pérdida de ingresos.	Datos imprecisos					
Los datos que tienen disponibles los trabajadores están obsoletos y no representan la realidad actual, por ejemplo el correo electrónico o el número telefónico de un proveedor que está registrado no corresponde al actual.	Datos no actualizados					
Los datos ambiguos son aquellos que pueden tener significados diferentes para las distintas personas que los utilizan, perdiendo el sentido de lo que se quería representar con ellos.	Ambigüedad en los datos					

2. Caracterización de la Calidad de Datos.

La siguiente tabla contiene una lista de características o atributos de Calidad de Datos acompañadas de su definición. Indique marcando con una X la medida en que dicha denominación es usada o podría ser usada en su organización, considerando los siguientes niveles de uso: (1) nunca, (2) rara vez, (3) ocasionalmente, (4) casi siempre y (5) siempre.

Características de Calidad de Datos	Definición	Nivel de utilización				
		1	2	3	4	5
Accesibilidad	Grado en que los datos están disponibles en un momento dado en el tiempo para su uso					
Actualidad	Grado de actualización y vigencia de los datos.					
Adecuación	Grado en el cual los datos se presentan de manera concisa, flexible y organizada, considerando los objetivos de quienes los necesitan para ayudarlo a alcanzar sus metas.					
Amplitud	Grado o alcance en que los datos cubren satisfactoriamente las necesidades de los usuarios.					
Aplicabilidad	Grado en que los datos permiten ser directamente utilizados por un trabajador en un contexto específico.					
Eficacia	Grado en que los datos satisfacen las necesidades de quienes los necesitan para el desarrollo de una tarea específica.					
Complejidad	Grado en que todos los datos necesarios están presentes para el desarrollo de una tarea.					
Comprensibilidad	Grado en que los datos se entienden fácilmente.					
Confiabilidad	Grado de concordancia entre diferentes datos registrados en condiciones similares.					
Confidencialidad	Grado en que se establece control al acceso a los datos de una organización.					
Conformidad	Grado en que los datos se ajustan a los formatos estándares establecidos por una organización.					
Consistencia	Grado en que los datos de diferentes conjuntos de datos son coherentes y no contradictorios.					
Credibilidad	Grado en que se consideran los datos como verdaderos y creíbles.					
Disponibilidad	Grado en que los datos se pueden obtener cuando se necesitan para un determinado uso.					
Duplicidad	Grado de existencia de datos repetidos en mas de un sistema de información de una organización, cuando esto no es necesario.					
Estandarización	Grado en que los datos están definidos de manera uniforme y se comprenden sus valores.					

Características de Calidad de Datos	Definición	Nivel de utilización				
		1	2	3	4	5
Estructuración	Grado en que los datos poseen la estructura y el formato correcto para utilizarlos.					
Exactitud	Grado en que los datos son reales y se representan correctamente en concordancia con los valores estándares definidos.					
Integridad	Grado en que todos los datos necesarios están presentes y de manera apropiada.					
Interpretabilidad	Grado en que los datos son apropiados en idioma, símbolos y unidades, y las definiciones son claras.					
Libre de errores	Grado en que los datos son correctos y libres de defectos.					
Mantenibilidad	Grado en que los datos pueden ser accedidos, usados, conservados, actualizados, mantenidos y manipulados.					
Objetividad	Grado que los datos son imparciales y sin prejuicios. Esto implica aplicar un enfoque que permita asegurar que los datos sean exactos, fiables, completos y claros.					
Operatividad	Grado de conformidad expresada por un trabajador respecto a la correspondencia, cercanía de los datos con la realidad y la utilidad, en concordancia con sus necesidades.					
Oportunidad	Grado en que los datos están actualizados y disponibles en el momento que se requiere su uso.					
Relevancia	Grado en que los datos son útiles y pertinentes para el desarrollo de una tarea determinada.					
Representación	Grado en que los datos poseen la forma, legibilidad y utilidad adecuada para un uso previsto.					
Reputación	Grado en que los datos están bien considerados en términos de su origen y contenido.					
Eficiencia	Grado en que los datos satisfacen las necesidades de quien los necesita, teniendo en consideración el acceso y el fácil uso de esta.					
Seguridad	Grado en que el acceso a los datos es restringido adecuadamente para mantener su seguridad.					
Sincronización	Grado de equivalencia entre los datos almacenados en un sistema principal y los sistemas locales de una organización.					
Trazabilidad	Grado en que se puede hacer un seguimiento del origen de los datos disponibles.					
Usabilidad	Grado en que la presentación de los datos es útil y/o eficaz para su propósito.					
Utilidad	Grado en que los datos son beneficiosos y están disponibles para un uso previsto.					

Características de Calidad de Datos	Definición	Nivel de utilización				
		1	2	3	4	5
Valor añadido	Grado en que los datos proveen beneficios y ventajas como producto de su utilización.					
Veracidad	Grado en que los datos se ajustan a su definición, a los valores de dominio y a las reglas de negocio.					

Si usted suele usar otros términos para expresar Calidad de Datos por favor indicarlos en la siguiente tabla.

Término	Significado

Bibliografía

Gabriela N Aranda, Nadina Martínez, Pamela Faraci, y Alejandra Cechich. Hacia un framework de evaluación de calidad de información en foros de discusión técnicos. 2013.

Gabriela N Aranda y Francisco Ruiz. Clasificación y ejemplos del uso de ontologías en ingeniería del software. En *XI Congreso Argentino de Ciencias de la Computación*. 2005.

Sugato Bagchi, Xue Bai, y Jayant Kalagnanam. Data quality management using business process modeling. En *Services Computing, 2006. SCC'06. IEEE International Conference on*, págs. 398–405. IEEE, 2006. ISBN 0769526705.

Nicholas Berente, Betty Vandebosch, y Benoit Aubert. Information flows and business process integration. *Business Process Management Journal*, 15(1):119–141, 2009.

Bernard S Black. Requisitos legales e institucionales para el establecimiento de un mercado de valores sólido. *revista del mercado de valores*, pág. 11, 2011.

Alexander Borek, Ajith Kumar Parlikad, Philip Woodall, y Maurizio Tomasella. A risk based model for quantifying the impact of information quality. *Computers in Industry*, 65(2):354–366, 2014. ISSN 0166-3615. doi:<http://dx.doi.org/10.1016/j.compind.2013.12.004>. URL <http://www.sciencedirect.com/science/article/pii/S0166361513002467>.

BPTrends. BPTrends BPM Analysis, Opinion and Insight. 2015.

José Luis Vézquez Burguete. Pasado, presente y futuro de las dimensiones pública y social en el desarrollo conceptual del marketing. *International Review on Public and Nonprofit Marketing*, 1(1):9–34, 2004.

José Luis Calderón Amaya y Carlos Rodríguez Monroy. Un enfoque gerencial de factores críticos para el éxito de los sistemas de información en la Pyme metalmecánica venezolana. *Contaduría y administración*, 57(1):79–102, 2012.

- Coral Calero, Francisco Ruiz, y Mario Piattini. *Ontologies for software engineering and software technology*. Springer Science & Business Media, 2006.
- Maybel Díaz Capote. Necesidad de una ontología de Ciencia y Tecnología para el Observatorio Tecnológico del ISPJAE. *Serie Científica*, 6(2), 2013.
- Carlos Carrascosa y Vicente Julián. Herramienta de creación de ontologías OntoEdit. 2005. URL [http://personales.upv.es/ccarrasc/doc/2003-2004/OntoEdit/presentacion\[_\].html/definiciononto.html](http://personales.upv.es/ccarrasc/doc/2003-2004/OntoEdit/presentacion[_].html/definiciononto.html).
- Roger Clarke. Data Risks in the Cloud. *Journal of theoretical and applied electronic commerce research*, 8:59–73, 2013. ISSN 0718-1876. URL [http://www.scielo.cl/scielo.php?script=sci\[_\].arttext{&}pid=S0718-18762013000300005{&}nrm=iso](http://www.scielo.cl/scielo.php?script=sci[_].arttext{&}pid=S0718-18762013000300005{&}nrm=iso).
- Jesús Contreras y Juan-Antonio Martínez-Comeche. *Ontologías: ontologías y recuperación de información*. 2007.
- Oscar Corcho, M Fernández-López, A Gómez-Pérez, y A López-Cima. Construcción de ontologías legales con la metodología METHONTOLOGY y la herramienta WebODE. En R Benjamins, P Casanovas, J Breuker, y A Gangemi, eds., *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, tomo 3369, págs. 142–157. Springer-Verlag, 2005. URL <http://oa.upm.es/5289/>.
- Oscar Corcho, Mariano Fernández-López, y Asunción Gómez-Pérez. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & knowledge engineering*, 46(1):41–64, 2003. ISSN 0169-023X.
- Patricio Ramírez Correa y Rosario García Cruz. Una investigación empírica sobre los factores que afectan el éxito de los sistemas ERP en Chile. 2005.
- John Domingue. Tadzebao and WebOnto: Discussing, browsing, and editing ontologies on the web. 1998.
- Walid el Abed. Data Governance: A Business Value-Driven Approach. En *La Gouvernance Des Données: Une Approche De Valeur Conduite Par Les Métiers*. 2009.
- Martin J Eppler y Peter Muenzenmayer. Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. En *IQ*, págs. 187–196. Citeseer, 2002.
- Clarissa Falge, Boris Otto, y H Osterle. Data quality requirements of collaborative business processes. En *System Science (HICSS), 2012 45th Hawaii International Conference on*, págs. 4316–4325. IEEE, 2012. ISBN 1457719258.

- Hector Florez Fernandez. Construcción de ontologías owl. *Vínculos*, 4(1):19–34, 2013. ISSN 2322-939X. URL <http://revistas.udistrital.edu.co/ojs/index.php/vinculos/article/view/4112>.
- Jakob Freund, Bernd Rucker, Bernhard Hitpass, y Others. *BPMN 2.0 Manual de referencia y guía práctica*. Bernhard Hitpass, 2013.
- Antonio Garcia. Instrumentos De Representación Del Conocimiento : Tesoros Versus Ontologías. *Anales de documentación: Univerisdad Rey Juan Carlos*, págs. 79–95, 2004. ISSN 1575-2437. doi:10.6018/1691.
- Francisco-Javier García-Marco. Ontologías y organización del conocimiento: retos y oportunidades para el profesional de la información. *El Profesional de la Información*, 16(6):541–550, 2007. ISSN 1386-6710. doi:10.3145/epi.2007.nov.01.
- Marcela Genero, JA Cruz-Lemus, y M Piattini. Métodos de investigación en ingeniería del software. *RaMa. Shull, Forrest*, 2014.
- John H Gennari, Mark A Musen, Ray W Ferguson, William E Grosso, Monica Crubézy, Henrik Eriksson, Natalya F Noy, y Samson W Tu. The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89 – 123, 2003. ISSN 1071-5819. doi:[http://dx.doi.org/10.1016/S1071-5819\(02\)00127-1](http://dx.doi.org/10.1016/S1071-5819(02)00127-1). URL <http://www.sciencedirect.com/science/article/pii/S1071581902001271>.
- Global Data Excellence. Global Data Excellence. 2015.
- Global Excellence. GLOBAL EXCELLENCE. 2015.
- Paul Glowalla y Ali Sunyaev. Process-Driven Data Quality Management Through Integration of Data Quality into Existing Process Models. *Business & Information Systems Engineering*, 5(6):433–448, 2013. doi:10.1007/s12599-013-0297-x. URL <http://dx.doi.org/10.1007/s12599-013-0297-x>.
- Ascuncion Gómez-Pérez, Mariano Fernández, y A de Vicente. Towards a method to conceptualize domain ontologies. 1996.
- Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, págs. 148–170, 1961.
- Google.com. Google. 2015. URL <https://www.google.com/>.
- Narasimhaiah Gorla, Toni M Somers, y Betty Wong. Organizational impact of system quality, information quality, and service quality. *The Journal of Strategic Information Systems*, 19(3):207–228, 2010. ISSN 0963-8687.
- Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993. ISSN 1042-8143.

- Michael Grüninger y Mark S Fox. Methodology for the Design and Evaluation of Ontologies. 1995.
- Francisco Montijano Guardia y Nuria Ruiz Fuentes. La naturaleza competitiva de la empresa banco-aseguradora. *Estudios de economía aplicada*, 22(2):377–378, 2004.
- César Guerra-García, Ismael Caballero, y Mario Piattini Velthius. A survey on how to manage specific data quality requirements during information system development. En *Evaluation of Novel Approaches to Software Engineering*, págs. 16–30. Springer, 2011. ISBN 3642233902.
- Jaime Guzmán, Mauricio López, y Ingrid Durley. Metodologías y métodos para la construcción de ontologías. En *Scientia et Technica*, tomo 2, págs. 133–140. Universidad Tecnológica de Pereira, 2012a. ISSN 0122-1701.
- Jaime Alberto Guzmán, Mauricio López, y Ingrid Durley. Metodologías y métodos para la construcción de ontologías Methodologies and methods for building ontologies . XVII:133–140, 2012b.
- Paul Harmon y Celia Wolf. Business Process Modeling Survey. *Business Process Trends* (<http://www.bptrends.com/>), 2011.
- Andreas Harth. This is a static HTML export of the Neon-Toolkit Wiki. 2014. URL <http://neon-toolkit.org/wiki/Main{ }Page.html>.
- Robert Heinrich, Alexander Kappe, y Barbara Paech. Modeling Quality Information within Business Process Models. En *Proceedings of the 4th SQMB Workshop, TUM-I1104*, págs. 4–13. 2011a.
- Robert Heinrich, Alexander Kappe, y Barbara Paech. Tool Support for the Comprehensive Modeling of Quality Information within Business Process Models. En *EMISA*, págs. 213–218. Citeseer, 2011b.
- Martin Hepp y Dumitru Roman. An ontology framework for semantic business process management. *Wirtschaftsinformatik Proceedings 2007*, pág. 27, 2007.
- Blanca Hernández Ortega, J Martínez Jiménez, y J DeHoyos. Calidad de la información Web en la banca electrónica. En *International Congress Marketing Trends. Venice*. 2008.
- Bernhard Hitpass. *BPM: Business Process Management Fundamentos y Conceptos de Implementación: Fundamentos y Conceptos de Implementación*. Bernhard Hitpass, 2012.
- Walter Howard. Data Quality Isn't Just a Data Management Problem. *Information Management*, 17(10):16, 2007.
- ISO/IEC-FDIS-25012. Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data Quality model. 2008.

- Dieter E Jenz. Business process ontologies: Speeding up business process implementation. *Jenz & Partner GmbH*, 2003.
- N Juristo, A Gómez, y J Pazos. Ingeniería del conocimiento: Construcción de sistemas expertos. 1997.
- Beverly K Kahn, Diane M Strong, y Richard Y Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.
- Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33:2004, 2004.
- Barbara D Klein. User perceptions of data quality: Internet and traditional text sources. *Journal of Computer Information Systems*, 41(4):9–15, 2001. ISSN 0887-4417.
- Huei Diana Lee, Wilson Jung, Adrieli Cristina da Silva, Luiz Henrique Dutra da Costa, Bianca Espindola, João José Fagundes, Feng Chung Wu, y Others. Prototype system to manage data on coloproctology surgery. *Journal of Coloproctology (Rio de Janeiro)*, 31(4):351–361, 2011.
- Douglas B. Lenat y R. V. Guha. Building large knowledge-based systems: Representation and inference in the CYC project. *Artificial Intelligence*, 61(1), 1993.
- Paul Lillrank. The quality of information. *International Journal of Quality & Reliability Management*, 20(6):691–703, 2003.
- Roger Liy, Rafael Montes, Richard Nakashima, y Others. *Propuesta de mejora para la cadena de suministro en el negocio de lubricantes*. Tesis Doctoral, Universidad Peruana de Ciencias Aplicadas-UPC. Escuela de Postgrado, 2013.
- Alexander Mosquera, Derlisiret Rincón, y María Gracia Romero. La organización basada en los sistemas de información. *Opción: Revista de Ciencias Humanas y Sociales*, (34):68–85, 2001.
- W S Nasution y Albarda. Improvement of business process in order to manage the quality of information. En *Proceedings - International Conference on ICT for Smart Society 2013: "Think Ecosystem Act Convergence"*, *ICISS 2013*, págs. 162–168. 2013. doi:10.1109/ICTSS.2013.6588084. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84884883134&partnerID=40&md5=ce6ec5c011494362193695e2e53e8321>.
- Matthias Nickles, Adam Pease, Andrea C Schalley, y Dietmar Zaefferer. Ontologies across disciplines. *Ontolinguistics-How Ontological Status Shapes the Linguistic Coding of Concepts*, págs. 23–67, 2007.
- Nf Noy y Dl McGuinness. Ontology development 101: A guide to creating your first ontology. *Development*, 32:1–25, 2001. ISSN 09333657. doi:10.1016/j.artmed.2004.01.014. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.5085&rep=>

[rep1{&}type=pdf\\$\delimiter"026E30F\\$nhhttp://liris.cnrs.fr/alain.mille/enseignements/Ecole{ }Centrale/Whatisanontologyandwhyweneedit.htm](http://liris.cnrs.fr/alain.mille/enseignements/Ecole{ }Centrale/Whatisanontologyandwhyweneedit.htm).

Martin H Ofner, Boris Otto, y Hubert Österle. Integrating a data quality perspective into business process management. *Business Process Management Journal*, 18(6):1036–1067, 2012. ISSN 14637154. URL [10.1108/14637151211283401http://search.ebscohost.com/login.aspx?direct=true{&}db=bth{&}AN=85895712{&}lang=es{&}site=ehost-live](http://search.ebscohost.com/login.aspx?direct=true{&}db=bth{&}AN=85895712{&}lang=es{&}site=ehost-live).

Paulo Oliveira, Fátima Rodrigues, Pedro Henriques, y Helena Galhardas. A taxonomy of data quality problems. En *2nd Int. Workshop on Data and Information Quality*, págs. 219–233. 2005a.

Paulo Oliveira, Fátima Rodrigues, y Pedro Rangel Henriques. A Formal Definition of Data Quality Problems. En *IQ*. 2005b.

OMG. OMG Unified Modeling Language (OMG UML). Inf. Téc. November, 2007. doi:10.1007/s002870050092.

OMG. Notation (BPMN) Version 2.0. *OMG Specification, Object Management Group*, 2011.

Inga Lillian Padilla. Planeacion De Los Recursos De La Empresa: ERP.

Rosalba Talavera Pereira y Yelitza Josefina Marcano Aular. Los lenguajes de representación semántica y su uso en la construcción de ontologías. *Revista de Ciencias Sociales*, 13(1), 2007. ISSN 1315-9518.

Juan Diego Pérez. Notaciones y lenguajes de procesos. Una visión global. *Departamento de sistemas y lenguajes informáticos, Universidad De Sevilla*, 2007.

Mario Piattini y Emilio Del Peso. Auditoría Informática. *Un enfoque práctico*, 2001.

Jan Recker. Opportunities and constraints: the current struggle with BPMN. *Business Process Management Journal*, 16(1):181–201, 2010.

Thomas C Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2):79–82, 1998.

Thomas C Redman. *Data driven: profiting from your most important business asset*. Harvard Business Press, 2008. ISBN 1422119122.

Research Stanford Center for Biomedical Informatics. Protégé PRODUCTOS. 2015. URL <http://protege.stanford.edu/products.php>.

Alfonso Rodríguez y Angélica Caro. Obteniendo Casos de Uso centrados en la Calidad de los Datos desde Procesos de Negocio descritos con BPMN. *Iberian Journal of Information Systems and Technologies*, (10):65–80, 2012. ISSN 1646-9895.

- Alfonso Rodríguez, Angélica Caro, Cinzia Cappiello, y Ismael Caballero. *A BPMN extension for including data quality requirements in business process modeling*. Paper presented at the 4th International Workshop on the Business Process Model and Notation, Vienna, Austria., 2012.
- Marco Rospocher, Chiara Ghidini, y Luciano Serafini. An ontology for the Business Process Modelling Notation. En *Formal Ontology in Information Systems: Proceedings of the Eighth International Conference (FOIS 2014)*, tomo 267, pág. 133. IOS Press, 2014. ISBN 1614994382.
- N Sánchez-Serrano, I Caballero, y F García. Extending BPMN to support the modeling of data quality issues. En *Proceedings of the 2009 International Conference on Information Quality, ICIQ 2009*. 2009. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84871539616&partnerID=40&md5=643a961fc514cefd4e4ea9c529f18696>.
- SourceMedia. information management. 2015. URL <http://www.information-management.com/>.
- Irena Spasic, Sophia Ananiadou, John McNaught, y Anand Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics*, 6(3):239–251, 2005. ISSN 1467-5463.
- Steffen Staab, Rudi Studer, Hans-Peter Schnurr, y York Sure. Knowledge processes and ontologies. *IEEE Intelligent systems*, (1):26–34, 2001. ISSN 1541-1672.
- Diane M Strong, Yang W Lee, y Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997. ISSN 0001-0782.
- Bill Swartout, R. Patil, Kevin Knight, y Tom Russ. Toward Distributed Use of Large-Scale Ontologies. *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, págs. 138–148, 1996. URL [#](http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Toward+Distributed+Use+of+Large-Scale+Ontologies)0.
- Mohammad Talalweh. La divulgación de información financiera en internet de las principales empresas españolas: variables económicas determinantes. *Gestión Joven*, (5):2, 2010.
- Sing What Tee, Paul L Bowen, Peta Doyle, y Fiona H Rohde. Factors influencing organizations to improve data quality in their information systems. *Accounting & Finance*, 47(2):335–355, 2007.
- Maritza Torres y Daniel Rojas. Review of data quality dimensions and applied methods in the evaluation of health information systems. *Enlace*, 5:25–44, 2008. ISSN 1690-7515.
- Verónica Evora Torres. Una propuesta metodológica para la gestión de la calidad de datos del Sistema Informativo del Banco Central de Cuba.

- Michael Uschold y Martin King. Towards a methodology for building ontologies. En *Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh.*, págs. 15–30. Citeseer, 1995.
- Raphael Volz. KAON Ontology Framework. 2015. URL <http://sourceforge.net/projects/kaon/?source=navbar>.
- Jan vom Brocke, Alessio Maria Braccini, Christian Sonnenberg, y Paolo Spagnoletti. Living IT infrastructures—an ontology-based approach to aligning IT infrastructure capacity and business needs. *International Journal of Accounting Information Systems*, 15(3):246–274, 2014.
- John G Walsh y Jeremy M Walsh. Method and system for enterprise business process management. 2005.
- Richard Y Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.
- Mathias Weske. *Business process management: concepts, languages, architectures*. Springer Science & Business Media, 2012. ISBN 364228616X.