Universidad Central "Marta Abreu" de Las Villas

Facultad de Matemática, Física y Computación

Licenciatura en Ciencia de la Computación

TRABAJO DE DIPLOMA

Herramienta de perfilado de metadatos en formato MARC 21

Autora: Mirielis Borges Zamora

Tutora: MSc. Lisandra Díaz de la Paz

Santa Clara, Cuba 2016

AGRADECIMIENTOS

- Le agradezco principalmente a mi tutora Lisandra, a Juan Luis y a Yaisel, por brindarme su ayuda incansable, su dedicación y paciencia.
- Le agradezco a mi familia por guiarme en la vida y permitirme estudiar en la universidad.
- Le agradezco y dedico mi tesis a mi novio Ruddy por acompañarme, ayudarme por 3 años, porque cada vez que tuve un problema estuvo allí y sin él esta tesis no sería posible.
- Le agradezco a mis amigos por apoyarme en los momentos difíciles y reír conmigo en los momentos de alegría.

RESUMEN

La catalogación de registros bibliográficos es un aspecto importante para la recuperación de la información. El ABCD se utiliza fundamentalmente en las bibliotecas y centros de documentación de las universidades del país ya que está dirigido a la gestión integrada de procesos de bibliotecas y operación automatizada. ABCD trabaja fundamentalmente con el formato MARC 21 para datos bibliográficos. En el módulo "Catalogación" la inserción de los metadatos es realizada por varios bibliotecarios de forma manual, se comprobó que esta última constituye una de las causas fundamentales que propician la aparición de problemas de calidad en los metadatos. Debido al impacto desfavorable que tienen los metadatos de baja calidad en la recuperación de la información y la carencia de una herramienta que detecte automáticamente anomalías en los campos del formato MARC 21 se realiza la presente investigación. Se establece una correspondencia entre las características de las herramientas de perfilado de metadatos y los errores más comunes en los campos mínimos encontrados en el módulo "Catalogación" mediante la cual se analizan las funciones principales con las que debe contar la herramienta de perfilado.

Palabras claves: calidad de metadatos, herramienta de perfilado, anomalías a nivel de instancia, catalogación de registros bibliográficos.

ABSTRACT

The cataloguing of bibliographic records is an important aspect for the recuperation of the information. The ABCD utilizes itself fundamentally at libraries and centers of documentation of universities of the country since he is guided to the steps integrated of processes of libraries and automated operation. ABCD works fundamentally with the format MARC 21 for bibliographic data. In the module Catalogación the insertion of metadata of information is sold off by several librarians of manual form, it was checked than this last constitutes one of the fundamental causes that propitiate the appearing of problems of quality in metadata of information. Due to the unfavorable impact that have the low-quality metadata of information in the recuperation of the information and the scarcity of a tool that MARC detect automatically anomalies at the Marc 21 format are accomplished present it investigation. A correspondency among the characteristics of the tools becomes established of in profile of metadata of information and commonnest errors at minimal fields found in the Catalogación module through which they analyze principal functions that the profile tool must have.

Keywords: Quality of metadata of information, profiling tool, level anomalies of instance, cataloguing of bibliographic records.

Tabla de contenidos

INTRODUCCIÓN *1*

CAPÍTULO I: ASPECTOS TEÓRICOS SOBRE CALIDAD DE METADATOS *5*

1.1. Catalogación de registros bibliográficos **5**

1.2. Metadatos **6**

1.2.1. Principales características 7

1.2.2. Calidad de los metadatos 10

1.2.3. Impacto de baja calidad en los metadatos 12

1.3. Formatos **18**

1.3.1. Estándar MARC 21 19

1.4. Perfilado de datos **25**

1.4.1. Herramientas de perfilado de datos 30

1.5. Auditoría de calidad de los datos **33**

1.6. Conclusiones parciales **34**

CAPÍTULO II: ESQUEMA DE ERRORES DE CALIDAD DE METADATOS *35*

2.1. Problemas de calidad de metadatos **35**

2.1.1. Nivel de instancia 36

2.1.2. Taxonomías de errores 38

2.2. Características de las herramientas de perfilado **42**

2.3. Correspondencia establecida **43**

2.4. Causas que conducen a los problemas de calidad **44**

2.5. Conclusiones parciales **46**

CAPÍTULO III: HERRAMIENTA DE PERFILADO DE METADATOS "PMMarc" *47*

3.1. Diseño de la herramienta de perfilado de datos **47**

3.1.1. Diagrama de casos de usos 47

3.1.2. Diagrama de clases 48

3.2. Perfilado de un conjunto de datos **50**

3.2.1. Visualización de los resultados por registros 50

3.3. Perfilado de múltiples conjuntos de datos **54**

3.4. Complejidad computacional de un algoritmo **56**

3.5. Campos mínimos más propensos a presentar errores de completitud **61**

3.6. Conclusiones parciales **62**

CONCLUSIONES *63*

RECOMENDACIONES *64*

REFERENCIAS BIBLIOGRÁFICAS *65*

ANEXOS 67

INTRODUCCIÓN

La recuperación de la información es uno de los procesos más usuales e importantes que realizan los profesionales de la información. Específicamente en las bibliotecas y centros de documentación donde los procesos bibliotecarios tales como préstamos, adquisiciones y catalogación de recursos bibliográficos se hacen de manera automática, la calidad con la que son descritos dichos recursos a través de los metadatos afecta directamente la recuperación de la información (Beall, 2006).

Uno de los sistemas que ha alcanzado la popularidad para tales fines lo constituye la suite para la Automatización de Bibliotecas y Centros de Documentación (ABCD), diseñada por el proyecto BIREME y financiado por el proyecto VLIR (Fernández and Lenzo, 2010)

ABCD está dirigido a la gestión integrada de procesos de bibliotecas y operación automatizada haciendo uso de una conexión directa a Internet aunque se puede utilizar sin conexión. Permite el manejo de cualquier estructura bibliográfica y no bibliográfica. Este paquete de software existe como un conjunto de módulos relativamente independientes, cuenta con cinco módulos principales que ofrecen las funciones de catalogación, préstamos, adquisiciones, registro de publicaciones periódicas (SeCSWeb), y administración de préstamos avanzados (EmpWeb). ABCD viene pre-configurado para trabajar con varios estándares bibliográficos tales como: UNIMARC, CEPAL, AGRIS, MARC 21. (Andreu Alvarez, 2015, Veloso Lara, 2012)

Por decisión del Ministerio de Educación Superior (MES), en las bibliotecas y centros de documentación cubanos se trabaja con el ABCD sin conexión directa a Internet. Esto implica que cada biblioteca cataloga de manera independiente sus registros bibliográficos y para ello deben utilizar el formato MARC 21. Una de las razones por las cuales se selecciona MARC 21 es porque constituye uno de los formatos más utilizados en países como Bélgica y España (Federico Medrano et al., 2012).

El sistema ABCD se utiliza fundamentalmente en las bibliotecas y centros de documentación de las universidades del país. A pesar de las ventajas que trae consigo su uso, la suite no está exenta de errores que pueden afectar la calidad de los procesos bibliotecarios mencionados anteriormente.

En Andreu Alvarez (2015) se realizó un análisis descriptivo acerca de los problemas presentes en los módulos del ABCD, a partir del procesamiento de una encuesta piloto aplicada a 16 especialistas que pertenecen al Centro de Documentación e Información Científico-Técnica (CDICT), que reside en el área académica de la Universidad Central "Martha Abreu" de Las Villas (UCLV). Algunos de los problemas que se mencionaron son los errores ortográficos y tipográficos, registros duplicados y campos nulos o vacíos. En dicho trabajo se ilustró a través de gráficos los principales problemas de calidad presentes en cada módulo y las causas que conducen a los mismos, de acuerdo a la opinión de los expertos encuestados. También se determinó que las dos causas fundamentales que permiten la aparición de errores son la entrada manual de los datos y la incorrecta aplicación de las reglas de catalogación angloamericanas, siendo el módulo de catalogación el más afectado y a su vez el más utilizado por los especialistas. Además, se comprobó que el sistema ABCD no cuenta con ninguna herramienta que permita detectar errores en los metadatos escritos en formato MARC 21, solamente en escasos casos los especialistas chequean algunos aspectos de calidad y lo hacen de manera manual.

Por consiguiente, en el presente trabajo se plantean los siguientes objetivos.

Objetivo General

Desarrollar una herramienta de perfilado de metadatos para detectar anomalías a nivel de instancia en los campos mínimos del formato MARC 21 y determinar el fondo real por tipo de material de los catálogos de las bibliotecas del país que utilizan dicho formato, mediante la recopilación de estadísticas descriptivas sobre los registros bibliográficos almacenados.

Objetivos específicos

1. Establecer una correspondencia entre un esquema de errores a nivel de instancia y las características inherentes de las herramientas de perfilado de datos.
2. Implementar una herramienta de perfilado de metadatos que detecte errores en campos del formato MARC 21 por tipo de material.
3. Determinar los campos mínimos del formato MARC 21 por tipo de material que están más propensos a presentar errores de completitud.

Preguntas de investigación

1. ¿Qué características comunes se pueden establecer entre un esquema de errores a nivel de instancia y las características inherentes de las herramientas de perfilado de datos?
2. ¿Cómo diseñar una herramienta de perfilado de metadatos que sea capaz de detectar errores por tipo de material en campos del formato MARC 21?
3. ¿Cuáles son los campos mínimos por tipo de material que están más propensos a presentar errores de completitud?

Justificación de la investigación

El impacto desfavorable que tienen los metadatos de baja calidad en la recuperación de la información y la carencia de una herramienta que detecte automáticamente anomalías en los campos que componen el formato MARC 21 y determine el fondo real por tipo de material en los catálogos de las bibliotecas y centros de documentación del país, constituyen las razones que motivan la presente investigación.

El trabajo es viable, pues se dispone de los recursos necesarios para llevarlo a cabo.

Estructura de la tesis

La tesis se encuentra estructurada en tres capítulos de la siguiente manera:

En el capítulo I se realiza una revisión bibliográfica sobre aspectos de calidad de metadatos, donde se exponen los principales conceptos teóricos que sirven de base para llevar a cabo la presente investigación. Sobre el tema se abordan aspectos importantes tales como la catalogación de registros bibliográficos, sus principales características y el impacto desfavorable que tienen los metadatos de baja calidad en la recuperación de la información. También se mencionan los formatos para describir los registros bibliográficos, haciéndose énfasis en MARC 21. Además, se presentan aspectos importantes sobre auditoría y perfilado de datos, y se muestra una tabla resumen con varias de las herramientas de perfilado de datos disponibles en internet.

En el capítulo II se investigan diferentes taxonomías buscando errores que se pueden encontrar en los metadatos almacenados en el módulo Catalogación del ABCD, tales como

anomalías a nivel de instancia en único y en múltiple registros, además de analizar diferentes errores en bases de datos, metadatos y registros bibliográficos. A continuación se exponen aspectos importantes presentes en las herramientas de perfilado y se establece la correspondencia entre los problemas más frecuentes en las taxonomías y las características inherentes de las herramientas de perfilado. Finalmente, se exponen las principales causas que conducen a los errores existentes para cada campo de los registros de nivel mínimo tomando en cuenta los tipos de materiales existentes en la suite ABCD para el formato MARC 21.

En el capítulo III se implementa la herramienta PMMarc, la cual responde, a través de funciones, las preguntas de calidad de metadatos que caracterizan a las herramientas de perfilado. Se visualizan la propuesta computacional, los diagramas de casos de uso y de clases, además se muestran los resultados por registro del perfilado de un conjunto de datos y para múltiples conjuntos de datos. Finalmente, se hallan los campos mínimos más propensos a errores de completitud a través del análisis de varios conjuntos de datos utilizando la herramienta de perfilado de metadatos PMMarc.

CAPÍTULO I: ASPECTOS TEÓRICOS SOBRE CALIDAD DE METADATOS

En el presente capítulo se analizan los principales aspectos relacionados con la calidad de los metadatos. Se exponen conceptos relacionados con los metadatos de investigaciones previas, registros de catalogación bibliográfica y se explica el impacto que tienen los metadatos de baja calidad en la información almacenada. Todos estos conceptos se ponen en práctica al utilizar la herramienta ABCD, con formato MARC 21. Además, se analizan las principales etapas de la gestión de la calidad de los metadatos, aspectos importantes sobre auditoría y perfilado de datos. Por último, se exponen las principales características de las herramientas de perfilado de datos a través de una tabla resumen.

1.1.Catalogación de registros bibliográficos

Un registro bibliográfico queda constituido por un conjunto de datos formales que identifican y permiten localizar un documento, es una representación del documento original y por tanto un documento secundario, constituido por la referencia bibliográfica, los puntos de acceso y la signatura topográfica. Por consiguiente facilita la recuperación del documento catalogado en un sistema de información (Veloso Lara, 2012).

Catalogación es un conjunto de operaciones bibliotecarias necesarias para la descripción completa de un documento y la asignación de una signatura topográfica. Las operaciones incluyen la catalogación por materias, además de la elección y redacción de los puntos de acceso principal y secundario. El resultado de la catalogación es un registro o asiento bibliográfico. Este formato usualmente se clasifica en estructurado o semiestructurado en dependencia de si posee una estructura fija a modo de tablas que se encuentran organizadas en bases de datos relacionales (estructurado), o si se usan normas complejas que permiten decidir dinámicamente el proceso siguiente después de la captura de los datos (semi-estructurado) (Amaya de Heredia, 2004).

El modelo semiestructurado de datos es la representación de datos con una estructura flexible. Algunos registros pueden tener campos faltantes, los otros pueden tener campos adicionales, algunos pueden tener dos o más ocurrencias del mismo campo. El tipo de datos es también

flexible: Puede ser un valor atómico o eso puede ser otro registro o colección. Además, las colecciones pueden ser heterogéneas. Pueden contener registros con estructuras diferentes (Ramalho, 2010).

En las bibliotecas de la UCLV junto a varias del país se utiliza el sistema ABCD, el cual cuenta con un módulo para la catalogación de los registros bibliográficos. En el mismo, los metadatos bibliográficos de los diferentes tipos de materiales son almacenados en bases de datos bibliográficas (Veloso Lara, 2012).

Las bases de datos bibliográficas son conjuntos de información estructurada en registros que describen documentos. Se encuentran almacenadas en un soporte electrónico legible desde un ordenador. Su finalidad es organizar dichos registros permitiendo la búsqueda y recuperación de la información. Hay dos tipos de bases de datos, las referenciales (ofrecen los datos identificativos de los documentos y, a veces, un resumen o abstract), y a texto completo (incluyen el texto completo de los documentos). Ambas son de vital importancia para ofrecer un mejor servicio de búsqueda y recuperación en el fondo documental de cualquier institución informativa. Puede ser una base de datos que contenga datos sobre libros y otras fuentes de información de una biblioteca (por ejemplo un catálogo de biblioteca) o incluso, un índice bibliográfico del contenido de un conjunto de revistas y otras publicaciones científicas como artículos científicos, actas de conferencias y congresos, capítulos de libros, etc. (Veloso Lara, 2012).

Específicamente en el caso de las bibliotecas de Cuba, se utiliza la base de datos NoSQL CDS/ISIS orientada a documentos. La familia de las bases de datos ISIS posee modelo semiestructurado de datos, está formada por registros y estos a su vez están compuestos por áreas de información conocidas como campos, los que identifican el autor, el título, el resumen, entre otros. Además, permite reconocer los campos de búsqueda y recuperación según su diseño, e identificar las características de los registros (Andreu Alvarez, 2015, Ramalho, 2010). Dicha información es almacenada en los metadatos de forma que constituyan una representación ideal del recurso bibliográfico.

1.2.Metadatos

Los metadatos son datos que describen otros datos denominados recursos. El concepto de metadatos es análogo al uso de índices para localizar objetos en vez de datos. Ejemplos de metadatos son el autor, título, tema, editorial y toda la información que sea necesaria para localizar tanto a libros como a los otros tipos de materiales. Los metadatos facilitan el flujo de trabajo convirtiendo datos automáticamente de un formato a otro. Para lo cual es necesario que los metadatos describan contenido y estructura de los datos (Bray, 1998).

Los metadatos describen el contenido, la calidad, el formato y otras características que lleva asociadas un recurso, constituyendo un mecanismo para caracterizar datos y servicios de forma que usuarios (y aplicaciones) puedan localizarlos y acceder a ellos (Sánchez Maganto et al., 2008).

Los metadatos buscan fomentar la reusabilidad de datos sin tener que recurrir al equipo humano que se encargó de su creación inicial. Aunque la creación de metadatos parezca un coste adicional, el valor de los datos a la larga es dependiente de su documentación. Permiten a los usuarios utilizar los datos de un modo más eficiente, determinando si serán de utilidad para ellos. Facilitan el acceso a los datos, su adquisición y una mejor utilización de los datos logrando una interoperabilidad de la información cuando esta procede de fuentes diversas. Los metadatos ayudan al usuario u organización que los recibe en el procesamiento, interpretación, y almacenamiento de los datos en repositorios internos (Sánchez Maganto et al., 2008).

1.2.1. Principales características

Según G. Stamou et al. (2006) los metadatos se clasifican de acuerdo a tres criterios:

- Contenido*:* se pueden separar los metadatos que describen el recurso mismo de los que describen el contenido del recurso. Es posible subdividir estos dos grupos más veces, por ejemplo para separar los metadatos que describen el sentido del contenido de los que describen la estructura del contenido o los que describen el recurso mismo de los que describen el ciclo vital del recurso. Subdividir metadatos por su contenido es lo más común.
- Variabilidad*:* se pueden distinguir metadatos mutables e inmutables. Los inmutables no cambian, sin importa la parte del recurso que se ve, por ejemplo el nombre de un fichero. Los mutables difieren de parte a parte, por ejemplo el contenido de un vídeo (J. R. Smith, 2006).

- Función*:* los datos se pueden dividir de acuerdo a su función en datos subsimbólicos, simbólicos o lógicos. Los subsimbólicos no contienen información sobre su significado. Los simbólicos describen datos subsimbólicos, es decir añaden sentido. Los datos lógicos describen cómo los datos simbólicos pueden ser usados para deducir conclusiones lógicas, es decir añaden comprensión.

Ciclo de vida de los metadatos

Según J. R. Smith (2006) el ciclo de vida de los metadatos comprende las fases de creación, manipulación y destrucción.

- Creación: se pueden crear metadatos manualmente, semiautomáticamente o automáticamente. El proceso manual puede ser muy laborioso, dependiente del formato usado y del volumen deseado, hasta un grado en el que los seres humanos no puedan superarlo. Por eso, el desarrollo de técnicas semiautomáticas o automáticas es altamente deseable. En la producción automática el software adquiere las informaciones que necesita sin ayuda externa. Aunque el desarrollo de algoritmos tan avanzados está siendo objeto de investigación actualmente, no es probable que la computadora vaya a ser capaz de extraer todos los metadatos automáticamente. En vez de ello, se considera la producción semiautomática más realista; donde un humano sostiene algoritmos autónomos con la aclaración de inseguridades o la proposición de informaciones que el software no puede extraer sin ayuda.
- Manipulación: si los datos cambian, los metadatos deben cambiar también. Aquí se hace la pregunta ¿quién va a adaptar los metadatos? Hay modificaciones que pueden ser manejadas sencilla y automáticamente, pero hay otras donde la intervención de un humano es indispensable. La metaproducción, el reciclaje de partes de recursos para crear otros recursos, demanda atención particular. La fusión de los metadatos afiliados no es trivial, especialmente si se trata de información con relevancia jurídica, como por ejemplo la gestión digital de derechos.
- Destrucción, para algunos casos es conveniente eliminar los metadatos junto con sus recursos, mientras que en otros es razonable conservar los metadatos, por ejemplo para supervisar cambios en un documento de texto.

Almacenamiento

Hay dos posibilidades para almacenar metadatos: depositarlos internamente, en el mismo documento que los datos, o depositarlos externamente, en su mismo recurso. Inicialmente, los metadatos se almacenaban internamente para facilitar la administración. Hoy, por lo general, se considera mejor opción la localización externa porque hace posible la concentración de metadatos para optimizar operaciones de búsqueda. Por el contrario, existe el problema de cómo se vincula un recurso con sus metadatos.

Desventajas del uso de metadatos

Según (Bultermann, 2004), algunos expertos critican fuertemente el uso de metadatos. Sus argumentos más sustanciosos son:

- Los metadatos son costosos y necesitan demasiado tiempo. Las empresas no van a producir metadatos porque no hay demanda y los usuarios privados no van a invertir tanto tiempo.
- Los metadatos son demasiado complicados. La gente no acepta los estándares porque no los comprende y no quiere aprenderlos.
- Los metadatos dependen del punto de vista y del contexto. No hay dos personas que añadan los mismos metadatos. Además, los mismos datos pueden ser interpretados de manera totalmente diferente, dependiendo del contexto.
- Los metadatos son ilimitados. Es posible adherir más y más metadatos útiles y no hay fin.
- Los metadatos son superfluos. Ya hay buscadores potentes para textos, y en el futuro la técnica *query by example* (búsqueda basada en un ejemplo) va a mejorarse, tanto para localizar imágenes como para música y vídeo.

Algunos estándares de metadatos están disponibles pero no se aplican: los críticos lo consideran una prueba de las carencias del concepto de metadatos.

Sin embargo, en el caso de la catalogación de registros bibliográficos cuyos recursos primarios (dígase libros, monografías, revistas impresas, etc.) se encuentran en un estante físico, todo el proceso de recuperación de información recae sobre los metadatos. He aquí la importancia de su uso y de hacer el proceso con la mayor calidad posible.

1.2.2. Calidad de los metadatos

Usualmente el concepto de calidad de datos está asociado a si los datos son adecuados para el propósito con el cual fueron creados. Los datos deben ser consistentes, completos y adecuados para su uso. Algunos expertos extienden el término a lo que también se denomina "calidad de metadatos" (Power Data, 2010, Wang and Strong, 1996).

"Los metadatos de alta calidad son los que respaldan los requerimientos funcionales del sistema que esté diseñado a soportar" lo que puede resumirse como "la calidad está relacionada con la aptitud para el propósito" (Federico Medrano et al., 2012).

La calidad de los metadatos se percibe como un tema importante para la operación y la interoperabilidad de los repositorios. La calidad debe ser medida de manera que se pueda apreciar la aptitud de los metadatos para cierta tarea. Cuando se dice medir la calidad no se evalúa la calidad del estándar de los metadatos o los vocabularios usados, ni la del recurso mismo. La calidad debe ser evaluada por la información introducida por los humanos o de forma automatizada acerca de un objeto, almacenada en registros. Esta información es importante ya que permite el acceso al objeto analizado (Hillmann, 2004).

Según Hillmann (2004) existen siete dimensiones que describen la calidad de los metadatos: completitud, exactitud, procedencia, conformidad a las expectativas, coherencia y consistencia lógica, actualidad y accesibilidad. Estas dimensiones son retomadas por otros trabajos como los expuestos por Ochoa y Duval Ochoa and Duval (2006), (2008).

- Completitud es el grado para el cual el registro de metadatos satisface toda la información necesaria para tener una representación ideal del objeto descrito.
- Exactitud es el grado para el cual los valores de los metadatos son correctos. Puede ser considerada como la distancia semántica entre la información que un usuario extrae del registro de metadatos y la información que el mismo usuario obtiene del documento mismo.

- Procedencia cubre la reputación que un registro de metadatos tiene en una comunidad. Por ejemplo, un usuario puede confiar más en metadatos generados por un experto de metadatos que él conoce, que los metadatos generados por un software, aunque lo metadatos generados automáticamente puedan ser mejor. La procedencia está relacionada con la percepción subjetiva que el usuario tiene acerca del origen de los metadatos.
- La conformidad a las expectativas mide el grado en el cual el registro de metadatos satisface los requisitos de una comunidad de usuarios. Existen varios parámetros que afectan esta calidad tales como el vocabulario, los campos necesarios para realizar una tarea especificada por el usuario (la búsqueda, la evaluación, la integración, etc.) y la cantidad de información para describir el objeto.
- La coherencia y consistencia lógica es el grado para el cual un registro de metadatos corresponde a una definición estándar y los valores usados en los campos se correlacionan positivamente entre ellos.
- La actualidad guarda relación con el grado para el cual un registro de metadatos permanece actual entre cierta comunidad. Si es necesaria la información sobre determinado tema escrito en los últimos cinco años, entonces, todos los libros que no cumplen con esta condición por ser más antiguos, pueden considerarse obsoletos.
- La accesibilidad es el grado para el cual los metadatos son accesibles, ambos en términos de la accesibilidad cognitiva así como también la accesibilidad física / lógica. La accesibilidad cognitiva es medida por cuán fácil es para un usuario entender la información contenida en el registro de metadatos. El bibliotecario revisa varias características: Los errores de ortografía, la conformidad con el vocabulario, incluyen una evaluación humana de la dificultad del texto (Ochoa and Duval, 2006).

La exactitud, completitud y consistencia son los criterios más comúnmente utilizados a la hora de medir la calidad de los metadatos. Es común emitir un criterio final cualitativo del grado de calidad que poseen los datos, se habla de alta, media y baja calidad. Siendo este último el criterio más analizado debido al impacto negativo que suele tener en la eficacia de los procesos de gestión de datos e incluso en la reputación de las organizaciones que los manejan (Ochoa and Duval, 2006).

1.2.3. Impacto de baja calidad en los metadatos

Conocer el momento en que se producen los errores ayuda a prevenirlos. Según (Power Data, 2010), estos suelen ocurrir en tres ocasiones fundamentales: entrada de datos, por agregación de datos externos y migraciones:

- Entrada de datos: la mayor fuente de errores son las entradas de información manual. Producido por ruido en la comunicación, errores tipográficos o equivocaciones, o por otros factores externos, como por ejemplo cuando la persona que está entrando los datos de un contacto, desconoce su código postal y no tiene herramientas para localizar esta codificación, por tanto se deja en blanco ese ítem.
- Datos externos: frecuentemente se incorporan datos externos de forma automática, en los repositorios de datos, sin tomar las precauciones oportunas, y esto provoca que se generen multitud de problemas de calidad de datos tales como registros duplicados.
- Migraciones: ocurren errores cuando se realiza una migración de metadatos sin prever los cambios en la información, una de las muchas consecuencias es la ausencia de calidad de datos, existencia de valores obsoletos o en un formato distinto al esperado en el nuevo sistema, además de los registros duplicados.

La baja calidad de los metadatos datos influye negativamente en los procesos que se basan en ellos y además provoca daños y desconfianza. Algunos ejemplos planteados por Power Data (2010) de los perjuicios que ocasiona son:

- Impide la inteligencia de negocios**:** metadatos incorrectos conducen a informes erróneos, defectos en los análisis y las decisiones que se basan en información incorrecta pueden ser perjudiciales para la organización.
- Costes de gestión**:** discrepancias entre aplicaciones pueden requerir trabajos de reconciliación de registros. Muchas empresas realizan procesos de cuadre, corrección y eliminación de duplicados con desarrollos manuales que tienen un gran coste y aportan poco valor.
- Daña la relación con el cliente: imposibilidad de ofrecer un buen servicio y un trato personalizado. El cliente suele ser el primer afectado ante un problema de calidad de

datos. Con frecuencia se realizan labores correctivas tras una reclamación del cliente cuando el daño ya está hecho.

Las bibliotecas digitales suelen tener muchos errores en los metadatos. Su estudio es muy importante porque ellos pueden bloquear el acceso a los documentos digitales y las bibliotecas pueden contener datos inexactos o incompletos necesarios para brindar información.

Según Beall (2006), algunos errores comunes son los tipográficos, de escaneo y errores de conversión. A los cuales (Rosie, 2014) añade los errores de codificación de caracteres, los elementos duplicados y los datos ausentes.

Errores tipográficos

Beall (2006) afirma que los errores tipográficos pueden ser clasificados de la siguiente forma:

1. Errores de omisión de letras (omt para omit)
2. Errores de inserción de letras (asnd para and)
3. Errores de sustitución de letras (cliar para clear)
4. Errores de transposición de letras (iknd para kind)

Un argumento que busca minimizar los efectos de los errores tipográficos sugiere que una palabra mal escrita en un documento puede aparecer luego escrita correctamente, y que no debe existir la posibilidad de que ese error impida el acceso al documento; mientras que por el contrario, los errores en los metadatos que se asociaron con un propósito pueden impedir el acceso a ese documento, puesto que los metadatos son los sustitutos de un objeto entero y se utilizan para brindar información importante del archivo. Para algunos objetos digitales como imágenes y documentos cuyo texto no es indexado, los metadatos son el único acceso por lo que se hace necesario la calidad de los metadatos guardados como elemento crucial para la recuperación precisa. En términos de bibliotecas digitales, probablemente los errores tipográficos más serios son los que ocurren en la URL. Ellos pueden bloquear completamente el acceso al documento ocasionando la pérdida de su información, aunque con el correcto manejo de la biblioteca digital los errores en URL pueden ser evitados o incluso arreglados (Beall, 2006).

De acuerdo a lo expresado en (Rosie, 2014), existen tres causas principales para los errores tipográficos, usualmente causado por los humanos:

- Enfocando la atención en la cantidad en vez de la calidad debido a presión/tareas de trabajo.
- La falta de interés.
- Ver lo que usted piensa que ve (SWYTYS, por sus siglas en inglés) en vez de lo que está realmente allí.

En la Figura 1 se muestran ejemplos de SWYTYS, generalmente se trata de almacenar los datos que son semejantes a los existentes en realidad, debido a no prestar la atención necesaria a la información. Las palabras de la izquierda son los datos que debieron ser almacenados y en su lugar se escribieron los de la derecha.

special reference	specific reference
private	pirate
conservation	conversation
hypertensive	hypersensitive
isoform	inform
perceptions	perspectives
economic aspects	economic analysis
aspects	concepts
biochemical	biomechanical
experiential	experimental

Figura 1 SWYTYS Fuente: (Rosie, 2014)

Errores de escaneo

Los errores de escaneos ocurren cuando un libro en copia física se desea convertir a digital. Algunos errores tipográficos pueden ser encontrados en documentos digitales que fueron creados a partir del escaneo de un libro en copia física. Aunque generalmente las palabras incorrectas pueden ser encontradas en el contenido del texto también pueden estar en el título, el autor o incluso en el año de publicación causando la pérdida de información. Aunque los errores en algunos metadatos como por ejemplo el número de páginas de un artículo no tienen

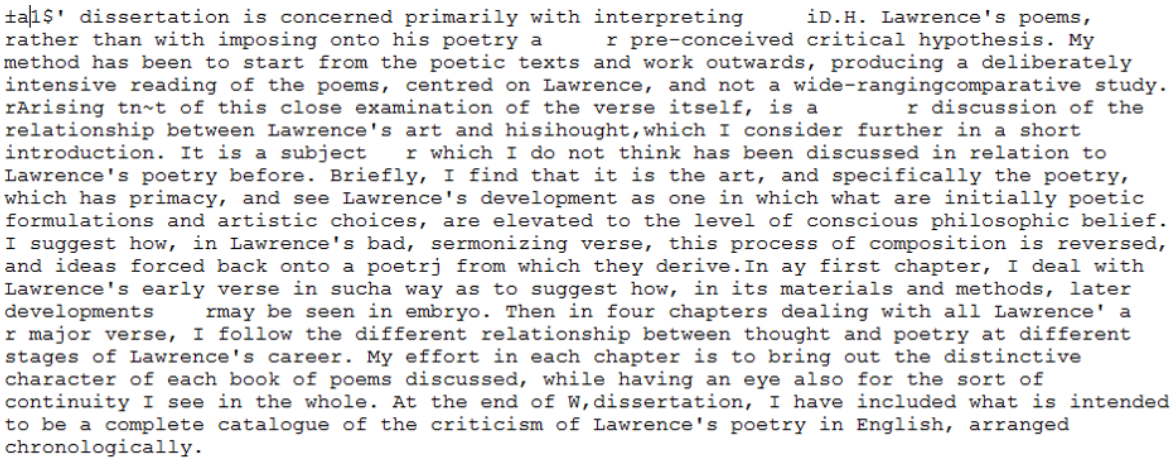
gran influencia en la recuperación del artículo en sí. En cambio otros, como el caso del año de la creación del artículo muestra que tan reciente es la investigación y le suma importancia al contenido.

Los errores que se encuentran en la versión original física no pueden ser arreglados en versión digital ya que esas son copias de página originales de publicación.

Otros problemas se manifiestan a partir de la conversión de un archivo digital de un formato a otro. La incorrecta conversión de caracteres afecta la indexación y la recuperación del documento ya que no existe forma de recuperar la información una vez que se ha modificado el original.

Los errores de escaneo ocurren por la máquina utilizada, ver Figura 2.

- El escaneo de textos es imperfecto.
- Los caracteres no son usualmente reconocidos.
- Palabras/oraciones pueden ser concatenadas.
- Revisado por humanos trae como consecuencia los SWYTYS.



±a1$' dissertation is concerned primarily with interpreting iD.H. Lawrence's poems, rather than with imposing onto his poetry a r pre-conceived critical hypothesis. My method has been to start from the poetic texts and work outwards, producing a deliberately intensive reading of the poems, centred on Lawrence, and not a wide-rangingcomparative study. rArising tn~t of this close examination of the verse itself, is a r discussion of the relationship between Lawrence's art and hisihought,which I consider further in a short introduction. It is a subject r which I do not think has been discussed in relation to Lawrence's poetry before. Briefly, I find that it is the art, and specifically the poetry, which has primacy, and see Lawrence's development as one in which what are initially poetic formulations and artistic choices, are elevated to the level of conscious philosophic belief. I suggest how, in Lawrence's bad, sermonizing verse, this process of composition is reversed, and ideas forced back onto a poetrj from which they derive.In ay first chapter, I deal with Lawrence's early verse in sucha way as to suggest how, in its materials and methods, later developments rmay be seen in embryo. Then in four chapters dealing with all Lawrence' a r major verse, I follow the different relationship between thought and poetry at different stages of Lawrence's career. My effort in each chapter is to bring out the distinctive character of each book of poems discussed, while having an eye also for the sort of continuity I see in the whole. At the end of W,dissertation, I have included what is intended to be a complete catalogue of the criticism of Lawrence's poetry in English, arranged chronologically.

Figura 2 Errores de escaneo en el abstract Fuente: (Rosie, 2014)

Errores de conversión de datos

Los errores de conversión son aquellos que ocurren cuando los metadatos son convertidos de un formato (como MARC 21) a otro (como Dublin Core Qualified) y no se tienen en cuenta una serie de incompatibilidades, por ejemplo:

- no exista compatibilidad entre ambos formatos, o sea no son interoperables,

- la cantidad de campos de un formato respecto al otro es muy diferente,
- los tipos de datos a introducir en los campos de cada formato sean totalmente diferentes e incompatibles.

Por tanto, forzar la conversión de un formato a otro puede ocasionar pérdida de información relevante.

Codificación de caracteres

Según (Rosie, 2014), los errores de codificación existen cuando el software no es capaz de reconocer un conjunto de caracteres identificándolos como extraños. El uso de caracteres extraños puede causar la falla del software. En la Figura 3 se muestra como el software identifica la tilde del nombre del autor como un carácter extraño. Esto también suele suceder con palabras de otros idiomas como japonés o árabe.

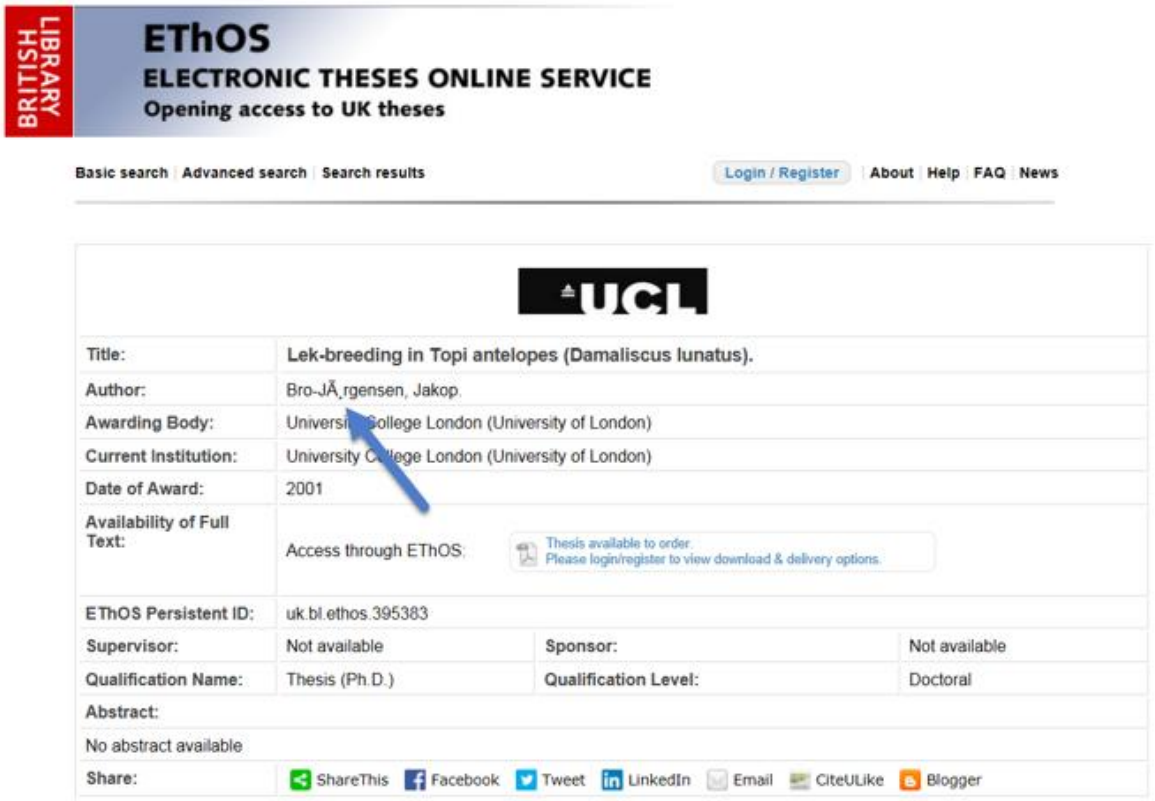


Figura 3 Nombre del autor con tilde. Fuente:(Rosie, 2014)

El caso del carácter especial (tilde diacrítica) se reconoce como la letra + acento. Un software debe ser capaz de unirlos correctamente. Si el software no puede agregar el acento y lo omite, la diferencia entre los registros afecta la recuperación de metadatos.

Duplicados

Los duplicados ocurren por errores de los seres humanos y de las máquinas, la falta de estándares en la representación de caracteres y palabras, en la falta de identificadores únicos como ISBN, DOI. Los duplicados aumentan el flujo de trabajo automático interno (ejemplo: leer y recuperar automáticamente de registros) cuando no es confiable la llave correspondiente.

- Eliminar o prevenir los duplicados es difícil debido a los SWYTYS tanto en los datos entrados como en los recuperados.
- El esfuerzo desaprovechado recreando los mismos datos y luego eliminándolos.
- Causa la digitalización del mismo objeto múltiples veces.
- La incorrecta representación del objeto.
- Consumo de tiempo y costo para remover el duplicado.
- Crea problemas para compartir y mantener múltiples conjuntos de datos.

Razones por la que no coinciden los títulos:

- Subtítulo incluido o no incluido.
- Artículo inicial incluido o no incluido.
- Artículos o conjunciones perdidas, e incluso diferentes (de, para, en, uno, pero, etc.).
- Diferentes representaciones de palabras, letras, números o símbolos.
- Uso de singular o plural
- Deletreo incorrecto, términos técnicos no familiares.
- SWYTYS.
- Incluir números de parte o volumen en el título.

Coincidencias autor y título

La lectura automática usualmente incluye la correspondencia del autor y el título entre los registros para garantizar la unicidad y prevenir coincidencias incorrectas con títulos

genéricos. En el nombre del autor no existen efectos de SWYTYS, los autores no coinciden debido a:

- Iniciales vs nombre completos.
- Segundo nombre o inicial omitido.
- Acento omitido o incluido.
- Selección incorrecta de apellidos donde existen más de dos elementos (ejemplo: Jakub Tadeusz Flasz).

Datos ausentes

Los datos o valores ausentes ocurren cuando los datos no son almacenados en el campo que le corresponde. Ocurre comúnmente y puede tener un gran efecto en las conclusiones arribadas a partir de los datos. Es la falta de respuesta, cuando la información no es obtenida, a veces los valores perdidos son causados por el investigador, por ejemplo, cuando un conjunto de datos es terminado incorrectamente o existen errores en los datos guardados. La presencia de valores perdidos (información ausente o faltante) no puede ser ignorada en el análisis de datos.

Los errores en los metadatos no se deben solo a los problemas mencionados anteriormente sino también debido a la variedad de estructuras, calidad, esquemas y formatos, convirtiéndolos los metadatos en inconsistentes, pocos fiables e inutilizables. Por lo que es necesario escoger el formato adecuado, teniendo en cuenta la estructura que deben poseer los datos y el propósito del sistema.

1.3.Formatos

Existe gran variedad de formatos y estándares que se utilizan para representar los metadatos. Entre ellos los más reconocidos en la actualidad son los siguientes: ID3, *Dublin Core*, *Learning Object Metadata* (LOM), *Resource Description Framework* (RDF), RDF *Schema* y MARC con su versión MARC 21 para registros bibliográficos. En este último es el que se centra la presente investigación.

1.3.1. Estándar MARC 21

MARC es el acrónimo de *Machine Readable Cataloging* o Catalogación legible por máquina, significa que una máquina puede leer e interpretar los datos contenidos en un registro catalográfico en función de una serie de elementos que actúan como "señaladores" de los datos. MARC es un estándar de transmisión de metadatos. Además de algunos campos fijos definidos por el estándar MARC y sus versiones, el contenido real que un catalogador coloca en cada registro MARC es usualmente regido y definido por otros estándares aparte de MARC. Las Reglas de Catalogación Anglo-Americanas (RCAA) (Amaya de Heredia, 2004), por ejemplo, definen cómo deben ser catalogadas las características físicas de los libros y de otros materiales. Las cabeceras LCSH (*Library of Congress Subject Headings*) proporcionan una lista de términos autorizados para cada materia, que se usan para describir el contenido principal de un material.

Cada campo en un registro MARC proporciona información acerca de un ítem que el registro está describiendo y hace uso de un código numérico simple de tres dígitos (000-999) a fin de identificar cada campo en el registro. El estándar bibliográfico define, por ejemplo, el campo 100 como el autor primario de una obra, 245 como el título y el 260 es usado para información sobre el editor (Hernández Agustí et al., 2015).

Los campos con valor por encima de 008 son posteriormente divididos en subcampos usando una designación de una sola letra o número. El campo 260, por ejemplo, es dividido en el subcampo 'a' para designar el lugar de publicación, 'b' para el nombre del editor, y 'c' para la fecha de publicación

Los registros bibliográficos MARC (Hernández Agustí et al., 2015) se distinguen de todos los otros tipos de registros MARC mediante códigos específicos en la Cabecera/06 (Tipo de registro) que identifican a los siguientes tipos de registros bibliográficos según Garcia B. (2008):

*Material impreso textual
*Material manuscrito textual
*Archivo de computador
*Material cartográfico
*Material cartográfico manuscrito
*Grabación sonora no musical
*Grabación sonora musical
*Medio proyectable
*Gráfico bidimensional no proyectable
*Artefacto tridimensional y objetos naturales

*Música impresa con notación	*Conjunto
*Música manuscrita	*Materiales mixtos

Los datos constituyen la información bibliográfica contenida en el registro, definida por normas ajenas al formato, sistema de clasificación, tesauro de materias, listas de códigos y otras convenciones usadas por la organización que crea un registro. Esa información son los metadatos necesarios para acceder al registro los cuales caracterizan los tipos de materiales ya que son específicos para cada uno, aunque hay metadatos que se utilizan para la mayoría de los registros bibliográficos, siendo por su utilidad los de mayor importancia. Una de las versiones de MARC es el formato MARC 21 para registros de datos bibliográficos.

MARC 21 es una estructura para registros de datos bibliográficos que se aplica a una base de datos creada con el software que se considere más conveniente. Este formato fue creado por un grupo de expertos de la Biblioteca del Congreso (EE.UU) liderados por Henriette Avram en el año 1965 (Congress, 2008b). Está compuesto por catálogos de bibliotecas usualmente llamados registros para cada uno de los tipos de materiales, la información es almacenada tanto en campos como subcampos de forma organizada, permitiendo el acceso rápido a los datos catalogados. Los formatos o familias de formatos MARC 21 son de las siguientes cinco clases desarrolladas para fines específicos con formato para datos: bibliográficos, de autoridad, de existencias, de clasificación y de la comunidad.

- Formato para datos bibliográficos: codifica los datos para describir, recuperar y controlar los diferentes tipos de materiales bibliográficos; por ejemplo: libros, publicaciones seriadas, recursos electrónicos, mapas, música, materiales visuales y materiales mixtos.
- Formato para datos de autoridad: codifica la información contenida en registros de autoridad de nombres, materias y series.
- Formato para datos de existencias: codifica información específica, como la signatura topográfica, cantidad de ejemplares y/o volúmenes, instituciones que posee un ítem, etc.

- Formato para datos de clasificación: codifica información contenida en un sistema de clasificación, por ejemplo: CDD, CDU o cualquier otro sistema que se desee automatizar.
- Formato para datos de la comunidad: codifica la descripción de recursos no bibliográficos que responden a las necesidades de información de una comunidad, como individuos, organizaciones, programas o servicios, eventos y otros recursos que pueden ayudar a los usuarios a conseguir la información que necesitan.

En correspondencia con los objetivos del trabajo, a partir de este momento solo se hace alusión al formato MARC 21 para registros bibliográficos. Este ha sido diseñado para servir como portador de la información bibliográfica relativa a: materiales textuales impresos y manuscritos, archivos de computador, mapas, música, recursos continuos, materiales visuales y materiales mixtos (OCLC, 2016). La información bibliográfica comúnmente incluye: títulos, nombres, tópicos, notas, datos de publicación e información sobre la descripción física de un ítem. El formato bibliográfico contiene elementos de información para los siguientes tipos de materiales (Hernández Agustí et al., 2015):

- Libros: materiales textuales cuya naturaleza es monográfica.
- Recursos continuos: materiales textuales con un patrón de publicación recurrente, Anteriormente al 2002 se hacía referencia a los recursos continuos como publicaciones seriadas ej: publicaciones periódicas, diarios, anuarios.
- Archivos de computador: utilizado para programas lógicos de computador, datos numéricos, multimedia para computador, sistemas o servicios en línea. Otras clases de recursos electrónicos se codifican de acuerdo a su aspecto más significativo. Los materiales pueden ser de naturaleza monográfica o seriada.
- Mapas: todo tipo de materiales cartográficos, incluyendo mapas planos y globos impresos, manuscritos, electrónicos y en microformas.
- Música: impresa, electrónica, en microformas y música impresa con notación.
- Grabaciones sonoras: grabaciones sonoras no musicales y grabaciones sonoras musicales.
- Materiales visuales: medios proyectables, gráficos bidimensionales, artefacto tridimensionales u objetos naturales; y, conjuntos.

- Materiales mixtos: principalmente colecciones de archivo y manuscritos con una mezcla de formas de materiales. Los materiales pueden ser de naturaleza monográfica o seriada.

Designación del contenido

- Etiquetas: números de tres dígitos usados para identificar los campos que componen un registro y al tipo de dato que le sigue, por ejemplo el campo título se identifica con la etiqueta 245.
- Indicadores: dos caracteres que se usan al comienzo de los campos variable, contienen valores que interpretan la información sobre el tipo o función de los datos en el campo; identificados independientemente.
- Códigos de subcampos: entendemos como subcampos a cada elemento que compone un campo. Los códigos de subcampo son de dos caracteres que preceden a cada elemento para identificarlo y señalar su localización dentro de un registro. Un código de subcampo está compuesto por: un delimitador, representado por el símbolo $ y de un identificador o carácter alfanumérico o numérico.

En la Figura 4 se ilustra la estructura de un registro con formato MARC 21.

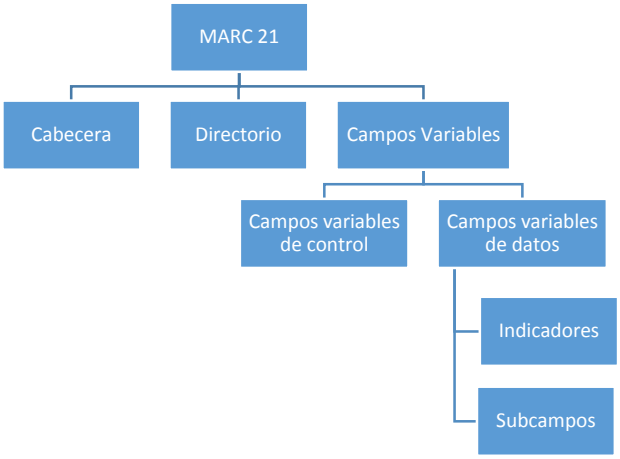


Figura 4 Estructura de un registro con formato MARC 21. Fuente: (García Mendoza et al., 2015)

Especificada en la norma ISO 2709 (ANSI Z39.2) que establece que las partes que componen un registro mostradas en la **¡Error! No se encuentra el origen de la referencia.** son:

- Líder o cabecera: son las primeras 24 posiciones (00-23) de cada registro y suministra información para el procesamiento del registro.
- Directorio: es un índice generado e interpretado por la computadora, sigue inmediatamente a la cabecera en la posición 24 y remite a la localización de los campos de datos de un registro. Es una serie de entrada de posiciones de 12 caracteres que indican la etiqueta, longitud y la posición de comienzo del primer carácter de cada campo variable.
- Campos variables:
    - Campos variables de control: campos etiquetados 00X y no contienen indicadores ni códigos de subcampos.
    - Campos variables de datos: campos etiquetados 01X-8XX y pueden tener definidos los indicadores o no, todos tienen el subcampo $a.

Formato MARC 21 Completo y Conciso para datos bibliográficos (Garcia B. , 2008)

- Cabecera /Directorio
- 00X -- Campos de Control
- 01X-09X -- Números y Códigos
- 1XX -- Entradas Principales
- 2XX -- Títulos, Ediciones, Imprentas, etc.
- 3XX -- Descripciones Físicas
- 4XX -- Declaraciones de Serie
- 5XX -- Notas
- 6XX -- Entradas de la Materia
- 7XX – Entradas de Secundarias
- 8XX -- Entradas de Secundarias de Serie, etc.

Los registros bibliográficos de nivel mínimo son aquellos que cumplen con ciertos requisitos, proveen suficiente información para identificar un ítem bibliográfico y generan una descripción básica (Congress, 2008a). Los registros bibliográficos deben incluir más elementos que los identificados como requisitos mínimos, pero se considera que los mínimos son los más relevantes, ver Anexo 1. En la Tabla 1 se señala con una cruz (X) aquellos campos expuestos en la columna denominada “Campos” que están presentes en cada tipo de

material. Resaltándose en azul los campos que están presentes en todos los tipos de materiales (245, 260 y 300) y en gris aquel donde solamente no está presente un solo tipo de material, tal es el caso del campo 100.

Tabla 1 Campos de nivel mínimo

Campos	Libro	Archivo de computadora	Música	Mapa	Archivo seriado	Material Mixto
028 (Número de publicación)			X			
100 (Autor)	X		X	X	X	X
245 (Título)	X	X	X	X	X	X
256 (Características del archivo)		X				
260 (Lugar de publicación)	**X**	**X**	**X**	**X**	**X**	**X**
300 (Descripción física)	X	X	X	X	X	X
310 (Frecuencia)					X	
500 (Nota general)				X	X	X
650 (Temática)			X	X		X
710 (Entrada secundaria del autor corporativo)					X	

1.4.Perfilado de datos

En la Figura 5 se muestra el ciclo de vida de la calidad de datos propuesto en (Matt Casters, 2010). El primer paso es el perfilado de datos, luego es la validación y el manejo de errores, cuando esos resultados son almacenados en una auditoría separada o tablas, esa información puede ser utilizada para reportar o analizar los valores. El paso final es corregir los datos tomados en el lugar base de la búsqueda de validación y auditoría.

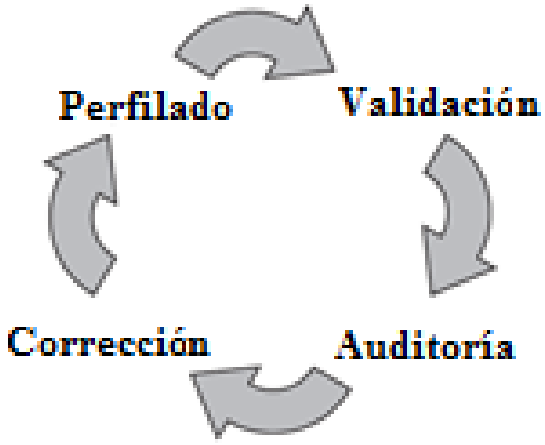


Figura 5 Ciclo de vida de la calidad de datos. Fuente:(Matt Casters, 2010)

Como se ilustra en la Figura 6 la gestión de la calidad de los datos está compuesta por dos procesos fundamentales:

- Exploración de datos
- Calidad de datos

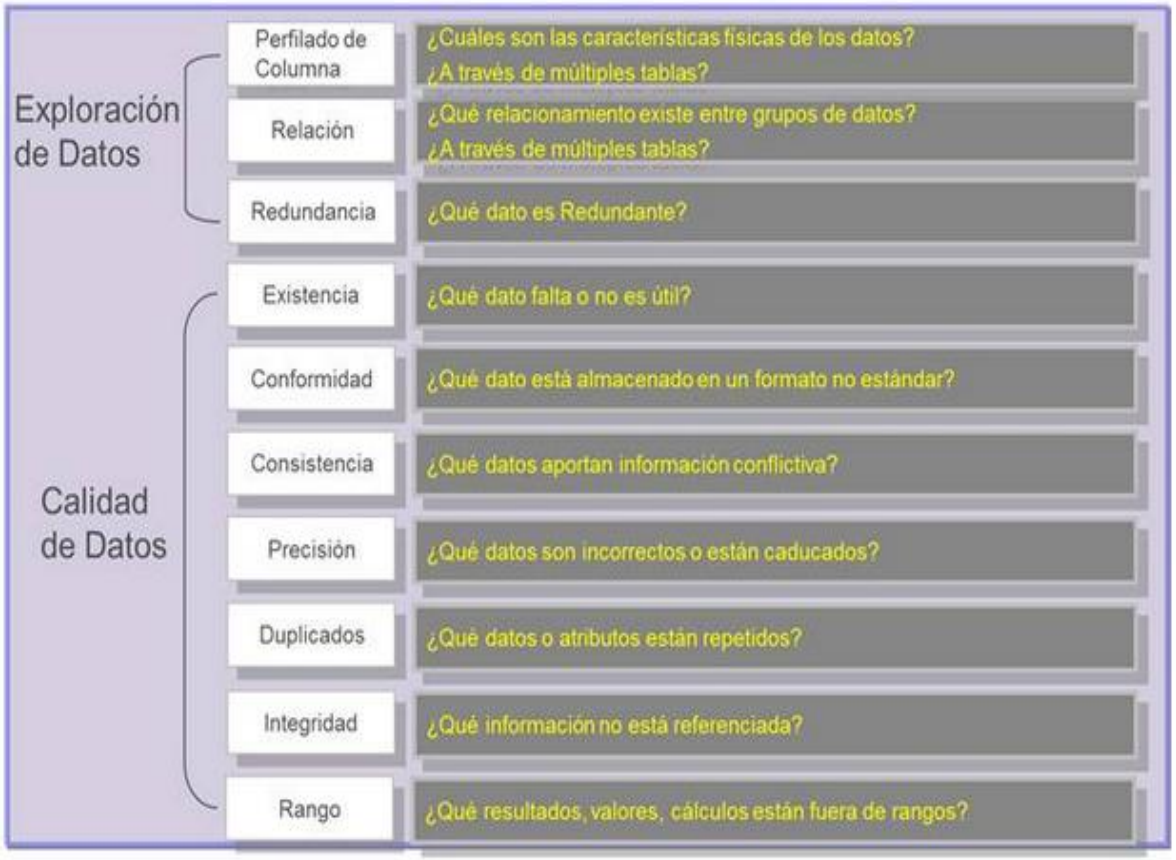


Figura 6 Procesos de la gestión de calidad de datos. Fuente: (Power Data, 2010)

La fase de perfilado realiza la exploración de los datos y busca un valor que representa la calidad de lo información, obteniendo las respuestas a las preguntas descubre los errores para seguidamente corregir y eliminarlos.

Cuando se lleva a cabo la gestión de la calidad de los datos hay que tratar de evitar los errores más frecuentes, que pueden llevar al fracaso del proyecto y, lo que es más importante, extender su impacto negativo en las acciones futuras de la empresa. Hay que prestar una especial atención a:

- La fase de perfilado: es el punto de partida y sólo se lleva a cabo una vez. Por perfilado de datos se entiende el análisis de los datos de los sistemas para entender su contenido, estructura, calidad y dependencias. (Soto, 2008)
- Diversidad de fuentes: la variedad de orígenes de datos puede hacer caer en riesgos de duplicados en un mismo sistema, suelen venir motivados por errores humanos.

- Datos incompletos: la completitud es uno de los atributos de la calidad del dato que mayor importancia tienen en términos de análisis. La detección de datos carentes de ella es imprescindible para garantizar la consistencia de la información con que se trabajará.

El perfilado de datos permite localizar, medir, monitorizar y reportar problemas de calidad de datos. El perfilado no debe ser sólo el inicio de un proyecto de calidad de datos, es un proyecto en sí. Debe permanecer durante los procesos posteriores e incluso después para la monitorización (Matt Casters, 2010).

Existen dos tipos de perfilado:

- Perfilado de estructura.
- Perfilado de contenido, también llamado análisis de datos.

Perfilado de estructura

El perfilado de estructura consiste en el análisis de los datos sin tener en cuenta su significado. Se analiza la información desde un punto de vista estructural. Por ejemplo, un dato que contiene el nombre "Juan J Gómez" no tiene en cuenta si el nombre es válido, simplemente lo analiza y lo identifica como una cadena de caracteres de longitud 12.

El análisis se realiza de forma semi-automática y masiva. Las soluciones especializadas en este tipo de perfilado pueden analizar cientos de tablas sin apenas necesidad de parametrización (Power Data, 2010).

Tipos de análisis del perfilado de estructura:

- Perfilado de columnas: análisis de atributos que puede tener una columna de una tabla: tipo de datos, longitud, número de nulos, número de valores únicos, frecuencias de valores, patrones de caracteres, máximos, mínimos, medias, etc.
- Perfilado de dependencias: análisis de columnas dependientes de otras en una misma tabla. Típicamente usado para validación de claves primarias y/o candidatas.
- Perfilado de redundancias: búsqueda de relaciones entre las tablas. Generalmente usado para validación de claves foráneas y/o *joins*. Análisis de valores "huérfanos", es decir, valores de una tabla referenciada no existentes en una tabla de referencia ("*Child Orphan*"), o viceversa ("*Parent Orphan*") (Power Data, 2010).

Perfilado de contenido

El perfilado de contenido consiste en el análisis de la información contenida en los datos desde un punto de vista sintáctico y semántico. Por ejemplo, un dato que contiene el nombre "Juan J Gómez", podría indicar que contiene un nombre válido, una letra que podría ser una inicial de un segundo nombre y un apellido válido.

Este tipo análisis requiere una configuración para cada campo para aplicar las reglas de negocio pertinentes en cada dato. Se deben combinar componentes específicos para tratamientos de cadenas, diccionarios o listas de valores e incluso diccionarios de patrones válidos. Con el perfilado de contenido también puede obtenerse las mismas conclusiones que el perfilado de columnas (es decir: tipo de datos, longitud, número de nulos, número de valores únicos, frecuencias de valores, patrones de caracteres, máximos, mínimos, medias, etc.), sin embargo sería necesario una configuración para cada campo. Generalmente el perfilado de contenido se usa sobre un número limitado de atributos (10-50).

Por lo general se pueden distinguir seis categorías principales de indicadores de calidad de datos (Power Data, 2010):

- Existencia: datos omitidos (nulos, en blanco), o que contienen información no útil (valores por defecto, como "N/D").
- Conformidad: adecuación del formato de un dato con respecto a un estándar establecido. Por ejemplo, que los teléfonos españoles estén todos en el formato de nueve dígitos, sin prefijo internacional, ni guiones, ni espacios.
- Consistencia: nivel de coherencia entre dos o más conjuntos de datos. Por ejemplo, una columna indica que el cliente es varón y tiene por nombre "Luisa".
- Precisión: existencia de datos incorrectos u obsoletos. Generalmente se comprueba comparando con datos de referencia. Por ejemplo: fecha de nacimiento "1/3/1890"
- Duplicados: datos referentes a la misma entidad, se consideran repetidos y no aportan información añadida. Por ejemplo: una misma persona que aparece dos o más veces en la misma tabla.

- Integridad: problemas tanto de integridad referencial (claves primarias y foráneas), como de datos relacionados no unidos por un campo común. Por ejemplo, empresas del mismo grupo de empresas no relacionadas entre sí en una base de datos.

Una vez identificados los registros y los modelos de datos a través del perfilado, se debe realizar el tratamiento de excepciones y de errores. Este punto consiste en analizar los informes del perfilado de datos y detectar todos aquellos formatos, registros nulos, contenidos o errores de integridad referencial que han aparecido y definir, junto con los responsables de los datos.

Para el perfilado de datos es fundamental una solución de reportes para la visualización de los resultados.

Se requieren las siguientes funcionalidades:

- Reportes: informes simples de conteo de registros por los indicadores de calidad de datos (exploración al detalle).
- Tableros de mando (*scorecarding*): son resúmenes de resultados de todos los indicadores de calidad de datos. Iconos de tendencia, códigos semafóricos según cumplimiento con unos objetivos fijados. Valores de ejecuciones previas.
- Evolución histórica: nivel de un indicador de calidad de datos a lo largo del tiempo para su monitorización.

Esta fase proporciona un conocimiento minucioso de la información de origen que debe servir para establecer pautas sobre su posterior normalización, depuración, mejora y tratamiento. Se trabaja a todos los niveles (columna, tabla, modelo). En esta fase se debe identificar qué fuentes pueden ocasionar problemas por cuestiones estructurales para su descarte o modificación con el propósito de evitar problemas posteriores.

Una vez que los datos son recolectados deben ser entendidos; por ello un análisis de los mismos debería ser capaz de responder preguntas como: ¿qué campos de los metadatos están realmente presentes y en qué porcentaje del total de registros, el grado de normalización, y la consistencia entre los valores reales de esos campos?, o si es posible detectar patrones de los contenidos en tales metadatos (Federico Medrano et al., 2012).

En la actualidad, para dar solución a todas las interrogantes planteadas anteriormente, se utilizan las herramientas de perfilado de datos.

1.4.1. Herramientas de perfilado de datos

Una herramienta de perfilado es la encargada de obtener informes que permitan apreciar la estructura y formato de cada uno de los registros, número de registros para cada tipo de formato, comprobación de integridad referencial y dependencias entre tablas.

Las actuales herramientas de perfilado de datos proveen una interfaz que facilita la visualización de las relaciones entre los datos y automatizan considerablemente la creación, ejecución y reutilización de reglas de validación y limpieza. Estas herramientas, como es lógico, deben disponer de conectores para acceder a todo tipo de fuentes de datos y todo tipo de plataformas, considerando diferentes infraestructuras tanto locales, remotas como la nube informática (*cloud*). Así mismo, deben minimizar el movimiento de datos.

Las herramientas de generación de perfiles varían en la arquitectura que usan para analizar los datos y en el entorno de trabajo que proveen para el equipo que genera el perfil de estos datos. Según TrilliumSoftware (2008) usualmente existen dos opciones de arquitectura, la de perfil basado en consultas y la basada en un repositorio para la generación de perfiles de datos.

Perfil basado en consultas

Algunas tecnologías de generación de perfiles implican tener destreza con las técnicas de SQL que se ejecutan contra sistemas de origen o contra una copia instantánea de los datos de origen. Aunque esto genera información útil sobre los datos, tiene numerosas limitaciones:

- Riesgos de rendimiento: Las consultas generan presión sobre los sistemas, desacelerando las operaciones, a veces de modo significativo. Cuando se necesita información adicional o si los usuarios quieren ver los datos reales, se ejecuta una segunda consulta creando más presión sobre el sistema. Las organizaciones reducen este riesgo haciendo una copia de los datos pero para ello se necesita replicar el ambiente completo, tanto en los sistemas de hardware como de software, lo cual puede ser costoso y llevar tiempo.
- Riesgos de trazabilidad: Los datos en los sistemas de producción cambian constantemente. Las estadísticas y los metadatos capturados de los perfiles generados en base a consultas corren el riesgo de quedar desactualizados de inmediato.

- Riesgos de integridad: Es complicado adquirir conocimientos integrales utilizando análisis basados en consultas. Las consultas se basan en hipótesis, y el objetivo es confirmar y cuantificar las expectativas sobre lo que está mal y bien en los datos. Dado esto, es fácil pasar por alto problemas que uno aún no está informado.
- La generación de perfiles por consulta: es valiosa cuando se desea controlar los datos de producción bajo ciertas condiciones. Pero no es la mejor forma de analizar grandes volúmenes de datos en preparación para integraciones y migraciones de datos a gran escala.

Repositorio para la generación de perfiles de datos

Otras tecnologías generan los perfiles de datos como parte de un proceso programado y almacenan los resultados en un repositorio de perfiles. Los resultados guardados pueden incluir contenido, como estadísticas resumidas, metadatos, patrones, claves, relaciones y valores de datos. Estos resultados pueden seguir siendo analizados por los usuarios o bien ser guardados para realizar más adelante análisis de tendencias. Los repositorios de perfiles que permiten a los usuarios explorar la información y ver los valores de los datos originales en el contexto de registros de origen son los que ofrecen más versatilidad y estabilidad para las audiencias no técnicas. Proveen una vista común y amplia de los datos del sistema de origen desde el cual se pueden basar decisiones trazables.

Consideraciones de volumen: En el caso de que las tablas o los archivos ingresen al rango de centenares de millones de registros, se debería considerar una estrategia de repositorio de perfiles. Con volúmenes de esta magnitud, la mejor estrategia podría ser una mezcla entre procesos de generación de perfiles programados los fines de semana y de una generación de perfiles enfocado, sin pugnas, basado en consultas.

Entorno de trabajo: Interfaz gráfica

Puesto que los usuarios pueden no estar familiarizados con las estructuras y tecnologías de bases de datos, es importante hallar una herramienta que provea una interfaz de usuario (GUI, por sus siglas en inglés) intuitiva y fácil de aprender. Las características de seguridad apropiadas también deberían ser parte del entorno de trabajo, para garantizar que el acceso a campos o registros restringidos se pueda permitir o negar, eso es lo planteado por TrilliumSoftware (2008).

La herramienta de perfilado de datos debe responder las siguientes preguntas (Matt Casters, 2010):

- Número de valores distintos: ¿Cuántos registro poseen campos de valor único?
- Número de null y valores vacíos: ¿Cuántos registros no tienen valores o tienen valores vacíos?
- Mayor y menor valor: No solo para valores numéricos, también para textos.
- Suma, media y desviación estándar: Cálculos y distribución en valores numéricos.
- Patrones y longitud de cadenas: ¿Los valores están correctamente almacenados?
- Número de palabras en total, con letras mayúsculas, minúsculas y combinadas: ¿Cuál es el total de palabras en un registro, y las palabras en mayúsculas, minúsculas o combinadas?
- Frecuencia: ¿Cuáles son los valores mínimos y máximo de una cantidad determinada de elementos en un registro?

En la Tabla 2 se muestran una serie de aspectos que permiten comparar las herramientas de perfilado, en especial se toman como referente las herramientas MARCEdit y MARCXIMIL por ser las dos que permiten realizar perfiles sobre datos bibliográficos y son de código abierto. Ambas herramientas hacen un perfilado por columna y detectan registros duplicados, aunque MARCEdit permite hacer análisis del contenido de los datos y otras validaciones. Sin embargo no son capaces de realizar análisis según el grado de importancia de los campos por tipo de material, donde los campos de mayor relevancia son los campos mínimos, luego los que forman parte de la versión completa y por último el resto de los campos.

Tabla 2 Aspectos fundamentales de las herramientas de perfilado de datos

Herramientas	**Profiler**	**MARCEdit (Reese, 2010)**	**MARCReport (Rosie, 2014)**	**MARCXIMIL (Borel, 2011b) (Borel, 2011a)**	**DataCleaner (Matt Casters, 2010)**
Licencia de software	Código abierto	Código abierto	Propietario	Código abierto	Código abierto

Fuentes de datos	Postgres MySQL JDBC (Base de datos)	MARC MRK MRK8 MARCXML Documento de texto	MARC MARCXML	MARCXML	Postgres MySQL JDBC (Base de datos)
Tipo de perfilado	Perfilado por columna	Perfilado por columna	Perfilado por columna	Perfilado por columna	Perfilado por columna Perfilado de dependencia Perfilado de redundancias
Registros bibliográficos	NO	SI	SI	SI	NO
Duplicados	X	X	X	X	
Análisis del contenido de los datos	X	X	X		
Validaciones		X	X		X
Distribución de los datos	X				X
Frecuencias	X	X	X		X
Análisis de los campos y patrones	X	X	X		X
Identificar campos comunes y estandarizados	X		X		X
Reportes	X	X			

1.5.Auditoría de calidad de los datos

La auditoría de calidad de los datos es el tercer paso en el ciclo de vida de la calidad de datos.

En esta fase se debe realizar:

- Análisis y medición de los problemas de calidad de datos
- Conclusiones de los problemas de calidad de datos

- Propuesta de soluciones
- Análisis del retorno de la inversión (ROI, *Return On Investment*)

Programar auditorías de datos con regularidad es una excelente manera de asegurarse de que los datos continúen cumpliendo las expectativas y además resalta las áreas de calidad que se han descuidado o las nuevas áreas de problemas que se han manifestado. Para facilitar este proceso, muchas organizaciones aprovechan la tecnología utilizada para la evaluación de riesgos, las mediciones de referencia, el diseño y análisis de los sistemas de origen y las pruebas de aceptación del usuario (Harte, 2008). Tanto la fase de auditoría como la de perfilado tienen como objetivo analizar la calidad de los datos y plantear estrategias para la corrección de los errores y mejora de la información.

1.6.Conclusiones parciales

En el presente capítulo se hizo una revisión de la bibliografía. Primeramente se plantearon aspectos sobre conceptos importantes, como lo es la catalogación de registros bibliográficos. Seguidamente se expuso ideas sobre los metadatos, tales como sus principales características y el impacto causado por la baja calidad en los metadatos. Luego se analizó la estructura y funcionalidad del formato MARC 21, además de la importancia que representan las etapas de perfilado y auditoría de la calidad de los datos. Las herramientas mostradas en la Tabla 2 Aspectos fundamentales de las herramientas de perfilado de datostienen sus ventajas pero no se pudo escoger una de ellas porque no daban respuestas a la información necesitada. De ellas solo tres utilizan formato MARC 21 para registros bibliográficos, MARCXIMIL solamente se especializa en búsqueda de duplicados. MARCEdit no es capaz de identificar campos comunes y estandarizados. MARCReport identifica errores y es capaz de arreglarlos pero no realiza un análisis tan detallado a nivel de campo y subcampo. No diferencia los campos por su importancia y no realiza análisis teniendo en cuenta las características únicas de cada tipo de material, por lo que las estadísticas mostradas en los reportes son generales y no permite comparaciones entre conjuntos de datos, entonces se puede decir que las herramientas analizadas se encuentran incompletas con relación a la necesidad de detectar los errores en los registro bibliográficos para MARC 21 en el módulo Catalogación de la suite ABCD.

CAPÍTULO II: ESQUEMA DE ERRORES DE CALIDAD DE METADATOS

En este capítulo se analizan los principales errores que se pueden encontrar en los metadatos almacenados en el módulo Catalogación del ABCD a partir de la investigación de los problemas a nivel de instancia, luego se estudian varias taxonomías buscando los errores que se encuentran en los campos, con los cuales se construye un esquema de errores. Además se presentan las características que debe cumplir una herramienta de perfilado, con el objetivo de establecer un correspondencia entre el esquema de errores y dichas características. Finalmente, se exponen las principales causas que conducen a los errores existentes para cada campo de los registros de nivel mínimo tomando en cuenta los tipos de materiales existentes en la suite ABCD para el formato MARC 21.

2.1. Problemas de calidad de metadatos

De acuerdo con lo expresado por López Porrero (2011), "…*los datos que contienen anomalías son denominados datos erróneos o sucios*". Las anomalías (errores o problemas) en los datos se clasifican en correspondencia con diferentes criterios, algunas de estas clasificaciones son:

- Anomalías sintácticas, semánticas o de contexto.
- Anomalías a nivel de esquemas o a nivel de instancias.

Estas clasificaciones también se pueden extender al contexto de los metadatos, en los cuales influye el hecho de si estas ocurren en un único registro o si se producen por la integración de múltiples registros.

Cuando se trata de un único registro los problemas a nivel de esquema de datos pueden ser motivados por problemas de un mal diseño del esquema, como por ejemplo problemas de no unicidad, o violaciones de la integridad referencial. En el caso de múltiples registros éstos pueden ser ocasionados por la heterogeneidad de los modelos y diseño de los esquemas, como son, por ejemplo, los conflictos estructurales y de nombres. Estos problemas a nivel del esquema se reflejan también al nivel de las instancias, y pueden ser resueltos a través de un

mejoramiento en el diseño de los esquemas (evolución de esquemas), y por la traslación e integración de los esquemas. Por otra parte, los problemas a nivel de instancias se refieren a errores e inconsistencias en los datos que no son visibles a nivel de esquema (López Porrero, 2011). La presente investigación se centra solamente en los problemas a nivel de instancia existentes en los campos de cada registro.

2.1.1. Nivel de instancia

Tanto los datos estructurados como semi-estructurados pueden presentar errores a nivel de instancia. Los problemas de datos a nivel de instancia se pueden dividir en problemas de un único registro y problemas de múltiples registros (Nuñez Arcia et al., 2016).

2.1.1.1. Único registro

Los problemas con los datos explicados por Nuñez Arcia et al. (2016) en un registro individual o simple se refieren a uno o varios campos de un único registro.

- Falta de datos en un campo no nulo: atributos que son rellenados con algún valor ficticio; por ejemplo, un número de año de publicación 9999 es un valor indefinido utilizado para superar la restricción no nula.
- Datos erróneos: datos que son válidos, pero no se ajustan a la entidad real; un ejemplo de datos erróneos es "Gente Nueva" que indica la editorial de un registro bibliográfico cuando este en realidad es "Samuel Feijó".
- Errores ortográficos: palabras mal escritas en campos de las fuentes de datos; por ejemplo el nombre de autor escrito como "Ernesto Rodrigues" en lugar de "Ernesto Rodríguez".
- Valores implícitos: existencia de datos extraños en algún campo de datos; un ejemplo común de valores implícitos es la inserción de un título en un campo de nombre, por ejemplo, "Licenciado Ernesto Rodríguez".
- Valores de campos perdidos: datos que son almacenados en el campo equivocado. Por ejemplo, el valor "50" ubicado en el campo referente al ISBN, cuando este realmente indica el número de páginas de una monografía.

- Datos ambiguos: datos que pueden ser interpretados en más de una manera, con diferentes significados. La existencia de datos ambiguos puede ocurrir debido a la presencia de abreviatura: por ejemplo: el nombre abreviado "A. Rodríguez" puede ser ampliado de diferentes formas, como: "Alberto Rodríguez", "Antonio Rodríguez" etc.

2.1.1.2. Múltiples registros

Los problemas con los datos en múltiples registros no pueden ser detectados considerando cada registro por separado, ya que como el nombre lo indica, se está haciendo referencia a más de un registro (Nuñez Arcia et al., 2016).

- Registros duplicados: son aquellos que no contienen información contradictoria; por ejemplo, los siguientes registros de empleados se consideran duplicados: Empleado1 (Nombre = "Padrón, Juan Nicolás", Título: "Héroes y padres", Parte restante del título = "Una lectura de Ilíada y Odisea"); Empleado2 (Nombre = "Juan Nicolás Padrón", Título: "Héroes y padres. Una lectura de Ilíada y Odisea", Parte restante del título = "").
- Contradicción de registros: son aquellos que contienen información contradictoria; por ejemplo, si se consideran los registros Empleado1 y Empleado2, pero con la siguiente información: Empleado1 (Nombre = "Padrón, Juan Nicolás", Título: "Héroes y padres"; Empleado2 (Nombre = "Padrón, Juan Nicolás", Título: "Héroes y padres. Una lectura de Ilíada y Odisea").
- Datos no estandarizados: diferentes registros que no utilizan las mismas representaciones de datos, invalidando así su comparación:
  - Transposición de palabras: en un solo campo, las palabras pueden aparecer en diferente orden; por ejemplo, los nombres "Carlos Rodríguez" y "García, Julio" no utilizan la misma regla de ordenación.
  - Formato de codificación: uso de diferentes formatos de codificación, por ejemplo, ASCII, UTF-8, entre otros.

- Formato de representación: diferentes representaciones de un formato para la misma información; por ejemplo el formato de moneda puede ser $ 10.5, 10.5 $, etc.
- Unidad de medida: diferentes unidades utilizadas en registros distintos, por ejemplo las distancias dadas en cm y pulgadas.

2.1.2. Taxonomías de errores

En los registros bibliográficos se pueden encontrar disímiles anomalías tales como: datos no estandarizados, errores ortográficos, tipográficos y de codificación. A continuación se muestran varias taxonomías de errores propuestas por diferentes autores con el propósito de determinar cuáles de los problemas más comunes se encuentran en los campos de los registros bibliográficos de nivel mínimo. Entiéndase por problemas comunes aquellos que están presentes en la mayoría de las clasificaciones expuestas.

Las taxonomías analizadas son: la propuesta por Beall (2006) en el contexto de las bibliotecas digitales, la presentada por López Porrero (2011) en el contexto de las bases de datos y la expuesta por Rosie (2014) sobre los errores en los metadatos. Además se añade a esta comparación la lista de los problemas más relevantes encontrados en el módulo Catalogación del ABCD implantado en la UCLV, producto del procesamiento estadístico de la encuesta piloto aplicada a 16 bibliotecarios realizado por Andreu Alvarez (2015) y las anomalías a nivel de instancia para un único registro y para múltiples registros expuestas en los epígrafes 2.1.1.1 y 2.1.1.2 respectivamente.

Taxonomía de errores en bibliotecas digitales (Beall, 2006):

- Errores tipográficos: palabras mal escritas en los campos debido a la cercanía de las letras teclado.
- Errores de escaneo: errores ocurridos al escanear un libro en copia física para digitalizarlo, se pueden encontrar errores tipográficos, palabras incorrectas con símbolos extraños, además de información ilegible, etc.
- Errores de conversión de datos: ocurren cuando los datos son convertidos de un formato a otro.

Taxonomía de errores en bases de datos relacionales (López Porrero, 2011):

- Datos incompletos
  - Registros que faltan (tablas vacías): se definen las estructuras de las tablas y no se almacena la información, en el caso de los registros bibliográficos ocurre cuando el registro se encuentra vacío.
  - Campos que faltan: cuando existen muchos campos pero se utilizan unos pocos mientras que el resto de los campos permanecen vacíos.
- Datos incorrectos
  - Códigos incorrectos: se utiliza un código que no aparece en el codificador asociado a ese atributo.
  - Registros duplicados: información repetida en varios registros de datos que corresponde a un mismo objeto.
  - Información incorrecta entrada al sistema: es el error cometido en el proceso de entrada, por fallas tipográficas o de otro tipo.
  - Carné de identidad incorrecto: los errores vienen dados por: no contar con 11 dígitos, no seguir al inicio el patrón: año, mes, día.
- Datos incomprensibles
  - Múltiples campos dentro de uno: en un campo se escriben valores de atributos diferentes, separados por coma o algún otro signo o símbolo.
- Datos inconsistentes
  - Uso y significado inconsistente de diversos códigos: cuando los codificadores cambian, si no se actualizan los codificadores antiguos, se pueden encontrar datos que hacen referencia a estos últimos y otros a los nuevos, lo cual provoca inconsistencias.
  - Uso inconsistente de nulos, vacíos y espacios: el valor NULL se utiliza para indicar la ausencia de información, es muy común que en su lugar se utilicen cadenas vacías, espacios, valor cero y valores por defecto.
  - Falta de integridad referencial: no existe una correspondencia entre una llave extranjera y la llave primaria que refiere.

Taxonomía de errores en los metadatos (Rosie, 2014):

- Codificación de caracteres: los errores de codificación existen cuando el software no es capaz de reconocer un conjunto de caracteres extranjeros, los identifica como extraños, considerándolos caracteres inválidos (por ejemplo MARC-8 vs UTF-8).
- Errores tipográficos: errores existentes en las palabras incorrectamente escritas entradas de forma manual.
- Escaneo de textos: errores al escanear un libro de texto en formato física para llevarlo a digital.
- Datos ausentes: ocurren cuando los datos no son almacenados en el campo correspondiente. También se conocen como datos o valores perdidos.
- Falta de estandarización: Usualmente ocurre en campos como el nombre calificativo, nombre del autor o representaciones de palabras (o símbolos). Es un error causado porque los metadatos pueden escribirse en diferentes posiciones como por ejemplo: el autor se puede escribir de varias formas: nombre apellido1 apellido2 o apellido1 apellido2, nombre.
- Falta de identificadores: ocurre cuando el libro no contiene un valor capaz de identificarlo únicamente, por ejemplo libros que no tengan el campo ISBN.
- Duplicados: los duplicados ocurren por errores de falta de estándar en la representación de caracteres y palabras, falta de identificadores como ISBN.

Según Andreu Alvarez (2015) los problemas más relevantes que se encontraron en el módulo Catalogación son:

- Errores ortográficos.
- Duplicados.
- Datos con ruido, incorrectos: sucede cuando los datos describen información incorrecta, por ejemplo, un año con cinco cifras, valores fuera de rangos o simplemente incorrectos cuando no se corresponde con el valor real.
- Valores nulos o vacíos: un valor nulo tiene el significado general de un valor perdido, o sea, un valor que existe en el mundo real, pero no está disponible en una colección de datos. Un valor puede faltar ya sea porque existe pero no se conoce, porque no existe, o porque no se sabe si existe.

- Datos no estandarizados: diferentes registros que no utilizan las mismas representaciones de datos, invalidando así su comparación.
- Inexactitud en los datos: son datos que al ser obtenidos de los registros no se corresponden con la información del documento real.
- Datos desfasados temporalmente: los datos que violan una restricción temporal, que especifica el instante de tiempo o intervalo en el que los datos son válidos.
- Conflictos de nombres: son los nombres que al ser comparados se dice que no son iguales incluso siendo el mismo porque se encuentran traspuestos o con diferente estándar.
- Valores ficticios en los campos: ocurre cuando se inserta un dato fingido (que no es real) en un campo determinado sin violar las reglas sintácticas de dicho campo. Por ejemplo cuando al describir un libro se inserta en el nombre del autor un nombre diferente al del autor real.
- Inconsistencias entre diferentes copias del mismo dato: definidos como contradicciones entre duplicados.
- Tipo de datos incorrecto: sucede cuando el campo debe almacenar un valor numérico y en vez de eso guarda una cadena.
- Valores implícitos.
- Datos ambiguos.

Los errores que no han sido definidos se mencionaron anteriormente en las otras taxonomías por lo que su significado sigue siendo el mismo. A continuación se realiza un análisis de los errores encontrados en las taxonomías definiendo de esta forma los errores más comunes que se pueden encontrar en el módulo de Catalogación para el formato MARC 21.

En la Tabla 3 se señala con una cruz (X) aquellos errores en expuestos en la columna denominada “Tipos de errores” que están presentes las taxonomías ((Beall, 2006), (López Porrero, 2011), (Rosie, 2014), (Andreu Alvarez, 2015), (Nuñez Arcia et al., 2016)) analizadas anteriormente. Resaltándose en azul los errores existentes en cuatro de las cinco taxonomías siendo éstos los errores más comunes y en gris aquellos presentes en tres taxonomías, denominándose como errores comunes.

Tabla 3 Tipos de errores

Tipos de errores	**Bibliotecas digitales** (Beall, 2006)	**Bases de datos** (López Porrero, 2011)	**Metadatos** (Rosie, 2014)	**Módulo Catalogación** (Andreu Alvarez, 2015)	**Nivel de instancia** (Nuñez Arcia et al., 2016)
Valores ausentes		X	X	X	X
Datos con ruido o incorrectos				X	X
Errores ortográficos		X		X	X
Errores tipográficos	X	X	X		X
Valores implícitos		X		X	X
Datos ambiguos				X	X
Registros duplicados		X	X	X	X
Contradicción de registros				X	X
Datos no estandarizados			X	X	X
Transposición de palabras				X	X
Errores de formato de codificación		X	X		X
Diferente formato de representación					X
Diferente unidad de medida					X
Errores de escaneo de texto	X		X		
Falta de identificadores			X		
Errores de conversión de datos	X				
Falta de integridad referencial		X			
Tipo de datos incorrecto				X	
Datos desfasados temporalmente				X	
Violaciones de restricciones de integridad				X	
Valores ficticios en los campos		X		X	X

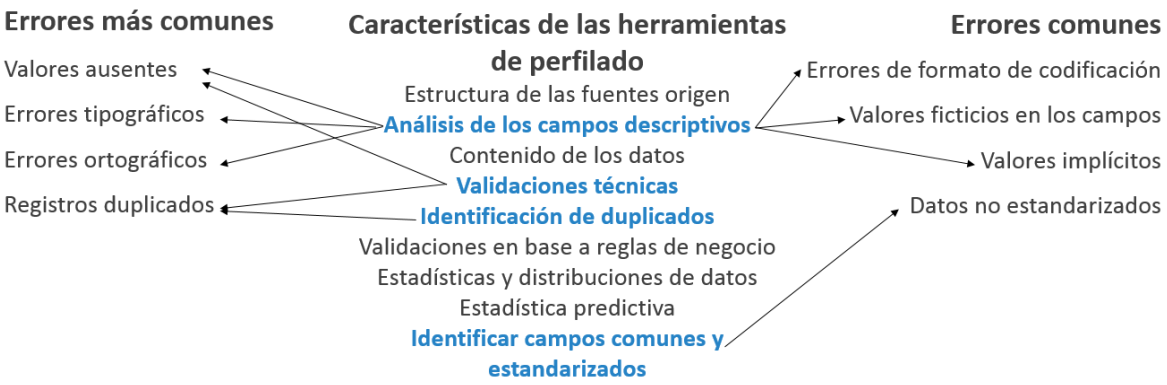
2.2. Características de las herramientas de perfilado

Durante la fase de perfilado, es importante no sólo la documentación de los datos, sino la detección de errores y su clasificación (Vidal Gil, 2013). Al realizar el perfilado de los datos se debe analizar:

- La estructura de las fuentes origen: tipo de fuente (fichero *host*, Excel, tabla BBDD, XML, etc.) tipo de estructura (bloque fijo, bloque variable, etc.), formatos de los campo (tipo de datos, longitud, etc.) y nivel de granularidad.
- Contenido de los datos: tipo de información que contiene cada fuente y campo. Identificar qué información es realmente relevante para el análisis.
- Validaciones técnicas: integridad, duplicidades, valores obligatorios, nulos, etc…
- Validaciones en base a reglas de negocio: por ejemplo, rangos de valores límites, valores no posibles, umbrales (edad < 0, código postal de más de n dígitos, número de cuenta corriente de menos de n dígitos, etc.).
- Identificación de duplicados: no sólo por validación técnica de clave idéntica, sino de posibles duplicados entre registros que contienen campos con diferentes valores, pero que por error de grabación pueden ser el mismo.
- Estadísticas y distribuciones de datos: con respecto a una columna consiste en efectuar conteos, agregaciones, valores máximos, mínimos, medio, desviaciones, frecuencias, distribución de los valores, rangos de valores, número de valores distintos, ratio de nulos, distribución de longitudes, etc…
- Estadística predictiva: en base a las estadísticas realizadas, obtener el valor más probable, más frecuente, construir tablas de frecuencia en base a las existentes.
- Análisis de los campos descriptivos: cualquier tipo de error que deba ser corregido o normalizado, por ejemplo descripciones incompletas, errores ortográficos, confusiones fonéticas, caracteres extraños y descripciones en diferentes idiomas. También se reconocen patrones.
- Identificar campos comunes y estandarizables, que deben seguir un patrón establecido, para tratamiento y normalización estándar. Ejemplo: nacionalidad, nombres, direcciones postales, teléfono, e-mail, código divisa, código cuenta corriente, código entidad financiera, etc…

2.3. Correspondencia establecida

El siguiente diagrama muestra la correspondencia establecida entre los errores más comunes encontrados en las taxonomías estudiadas y las características inherentes de las herramientas de perfilado.



De la correspondencia establecida se concluye que la herramienta de perfilado debe cumplir principalmente con las validaciones técnicas, identificación de duplicados, el análisis de los campos descriptivos, además de determinar los campos comunes y estandarizados, entonces la herramienta de perfilado debe ser capaz de detectar los errores más comunes en los datos por lo que los cubren en un 100%.

2.4. Causas que conducen a los problemas de calidad

En (Andreu Alvarez, 2015) se exponen las principales causas asociadas a los problemas existentes en el módulo Catalogación, dentro de las que sobresalen las siguientes:

- Incorrecta aplicación de las reglas de catalogación angloamericanas: son reglas establecidas por expertos, reconocidas a nivel internacional que deben cumplir los registros catalogados.
- Entrada manual de datos: introducir datos en el sistema de forma manual.
- Migración de datos: transferencia de materiales digitales de un origen de datos a otro.
- Información redundante: información duplicada en las bases de datos del sistema.
- Metadatos incompletos: carencia de información sobre los datos bibliográficos (metadatos) en los documentos catalogados.
- Desactualizaciones de datos: falta de actualización en todas las réplicas de datos del sistema (Ejemplo: los diferentes ejemplares de un libro determinado).
- Cambios no capturados: se han producido diferentes cambios en la biblioteca, pero estos no han sido capturados en el sistema.

- Procesamiento de datos: cambio en los programas responsables del procesamiento regular de datos (recolección y manipulación de elementos de datos para producir información significativa).
- Actualizaciones del sistema: actualizaciones en el software del sistema cada cierto tiempo.
- Automatización de procesos: con el progreso de la tecnología, cada vez más tareas son automatizadas sin intervención humana.
- Purga de datos: eliminación rutinaria de datos antiguos del sistema para dar paso a datos más nuevos.
- Diferentes formatos de codificación: falta de actualización en todas las réplicas de datos del sistema (Ejemplo: los diferentes ejemplares de un libro determinado).
- Integración de datos: proceso de combinar datos que residen en diferentes fuentes y permitirle al usuario final tener una vista unificada de todos sus datos.
- Manejo de mucha información: el manejo de tanta información dificulta tener una visión global de la información almacenada.
- Carencia de una herramienta de perfilado: la no existencia de una herramienta que muestra reportes sobre los registros y sus tipos de materiales, además que identifique los errores existentes.

La entrada manual de datos y la carencia de una herramienta de perfilado son las principales causas que conducen a los errores existentes. El trabajo con miles de registros causa que ciertos errores no sean encontrados fácilmente por ejemplo: registros duplicados o contradicciones o registros incompletos, además de información incorrecta, los campos que deben ser estandarizados, incluso errores de codificación. Otra de las causas es el desconocimiento de los aspectos específicos del estándar MARC 21 debido a la complejidad del formato, tales como los campos que se deben utilizar por tipo de material, específicamente los definidos como mínimos, además de ignorar las reglas de catalogación angloamericanas, aunque los datos no sean incorrectos sintácticamente, si lo son semánticamente, puesto que el objeto no ha sido descrito con la información adecuada.

2.5. Conclusiones parciales

En este capítulo se estudiaron los principales errores de la calidad de los metadatos y las causas de su aparición. Algunos de estos errores no tienen gran efectos como los errores ortográficos en campos como el resumen, mientras que en el título del libro o en el nombre del autor puede causar la pérdida de información importante. Se estudiaron las principales características que deben poseer las herramientas de perfilado de datos. Se estableció una correspondencia entre los errores más comunes encontrados en los metadatos y las características de perfilado obteniéndose como resultado las principales características con las que debe contar la herramienta para que sea capaz de encontrar los errores más comunes. Se analizaron las principales causas de los problemas de calidad para el módulo Catalogación entre las que se encontraron la carencia de una herramienta de perfilado y la entrada manual de los datos.

CAPÍTULO III: HERRAMIENTA DE PERFILADO DE METADATOS “PMMarc”

En el presente capítulo se implementa la herramienta de perfilado de metadatos PMMarc (**P**erfilado de **M**etadatos de **MARC** 21). Se visualiza la propuesta computacional, los diagramas de casos de uso y de clases, además se muestran los resultados del perfilado de un conjunto de datos y para múltiples conjuntos de datos. Finalmente, se hallan los campos mínimos más propensos a errores de completitud a través del análisis de varios conjuntos de datos utilizando la herramienta de perfilado de metadatos PMMarc.

3.1. Diseño de la herramienta de perfilado de datos

La herramienta de perfilado de datos a crear debe cumplir con las principales características analizadas en el capítulo II para que se capaz de encontrar los errores más frecuentes en el módulo Catalogación de la suite ABCD para el formato MARC 21, ellas son:

- Validaciones técnicas
- Validaciones en base a reglas de negocio
- Análisis de los campos descriptivos
- Identificar campos comunes y estandarizados

3.1.1. Diagrama de casos de usos

Los modelos de casos de uso dirigen todo el proceso de desarrollo de un programa ya que constituyen el punto de partida para llevar a cabo la mayoría de las actividades: el análisis, diseño y prueba del software. Este modelo se realiza identificando cada actor de la como los posibles usuarios para los cuales está realizado el mismo. Un caso de uso es una descripción de las acciones de un sistema desde el punto de vista del usuario. (Schmuller, 2000).
PMMarc tiene como actor principal al usuario que desea identificar los errores existentes en sus catálogos, por lo que la herramienta debe proveer una interfaz amigable y fácil de utilizar. Los usuarios a través de la herramienta pueden buscar en los registros los campos de nivel mínimo con valores ausentes, descubrir la cantidad de registros existente por cada tipo de

material, calcular la frecuencia de los campos más utilizados, encontrar patrones incorrectos en el año de publicación, visualizar la información existente en cada registro y comparar conjuntos de datos según el total de valores ausentes en los campos mínimos. En la Figura 7 se presenta el diagrama de casos de uso de la herramienta:

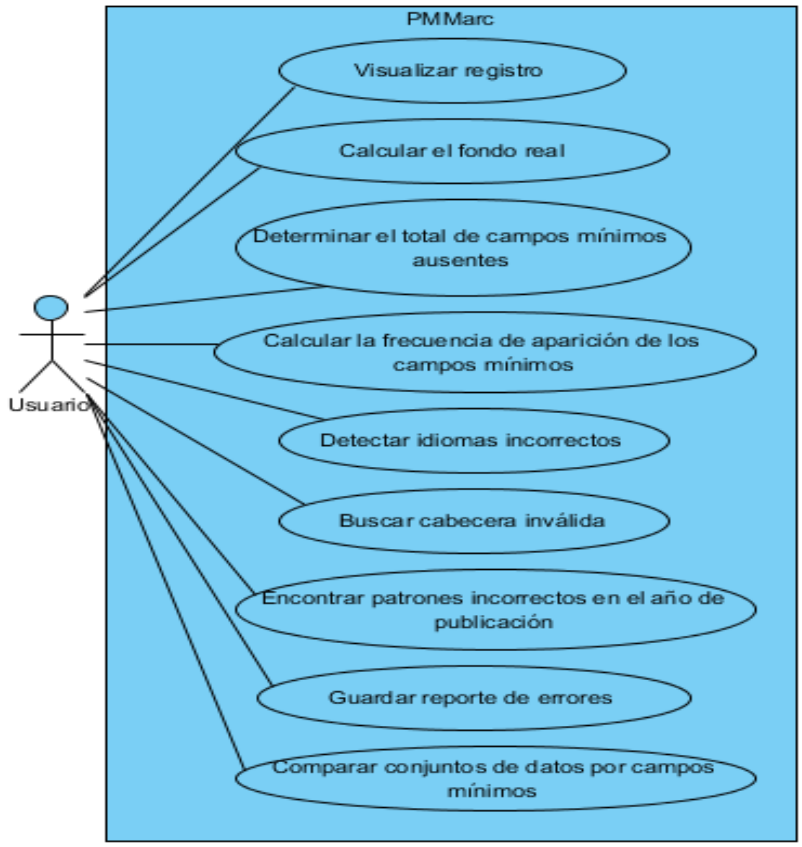


Figura 7 Diagrama de casos de uso Fuente: elaboración propia

3.1.2. Diagrama de clases

Un diagrama de clases muestra la manera en que las clases se relacionan entre sí. Diagrama de clases es un diagrama de objetos que describe las clases en forma de esquema representando gráficamente las interrelaciones, atributos y operaciones. El área superior de cada clase contiene el nombre, el área central contiene los atributos, y el área inferior las operaciones o acciones (Schmuller, 2000, Rumbaugt, 2000).

La herramienta cuenta con 9 clases PMMarc, Arbol, Analisis, Grafico, Registro, Idioma, CantMaterial, AInvalido y Reporte mostradas en la Figura 8 Diagrama de clases Fuente:

elaboración propia. Las mismas están divididas en tres paquetes, uno para las clases que trabajan con los componentes de la vista, otro para las clases encargadas de trabajar específicamente con el formato MARC 21, además de un paquete que trabaja con reportes.

El paquete "Visual" está compuesto por las siguientes clases: PMMarc, Arbol, Analisis, Grafico. PMMarc es la clase principal que maneja de las vistas del proyecto. Arbol posee las funciones que construyen el árbol donde se visualiza la estructura de cada registro con los valores de sus campos. Por otra parte, la clase Soporte escribe los valores en las tablas mostradas por PMMarc como la cantidad de registros por tipo de material o el total de registros con campos nulos. La clase Analisis busca los valores ausentes en los campos y crea los gráficos mostrados por PMMarc.

El segundo paquete está compuesto por las clases: Registro, Idioma, CantMaterial y AInvalido. Registro posee los funciones que obtienen la información almacenada en el *.*mrc*, a medida que se lee el formato MARC 21 se guardan los registros en un objeto, luego cada función trabaja con el objeto en vez de acceder directamente al documento. La clase Idioma recorre los registros buscando idiomas almacenados en el campo 041 subcampo $a, que no se correspondan con los códigos MARC de dos y tres caracteres o con los códigos correspondientes al ISO 639-2 de idiomas de tres letras, ISO 639-3 de idiomas de dos letras. AInvalido determina el patrón de cada año de publicación almacenado en el campo 260 subcampo $c y luego si el valor no es válido como por ejemplo: el año tiene 5 dígitos o posee otros caracteres (. , ; [] () a, etc) es mostrado en una lista por PMMarc permitiendo que el usuario obtenga el número de registro que guarda ese patrón. Por último la clase CantMaterial es una de las más importantes porque permite buscar los registros pertenecientes a cada tipo de material y devuelve su total, además de la cantidad de campos mínimos y la cantidad de campos ausentes.

El paquete Reporte solo contiene la clase con el mismo nombre utilizada para mostrar y guardar los reportes de errores encontrados en los catálogos, incluyendo la fecha actual, por lo que se pueden utilizar para comparar y encontrar mejorías con respecto a los errores existentes anteriormente.

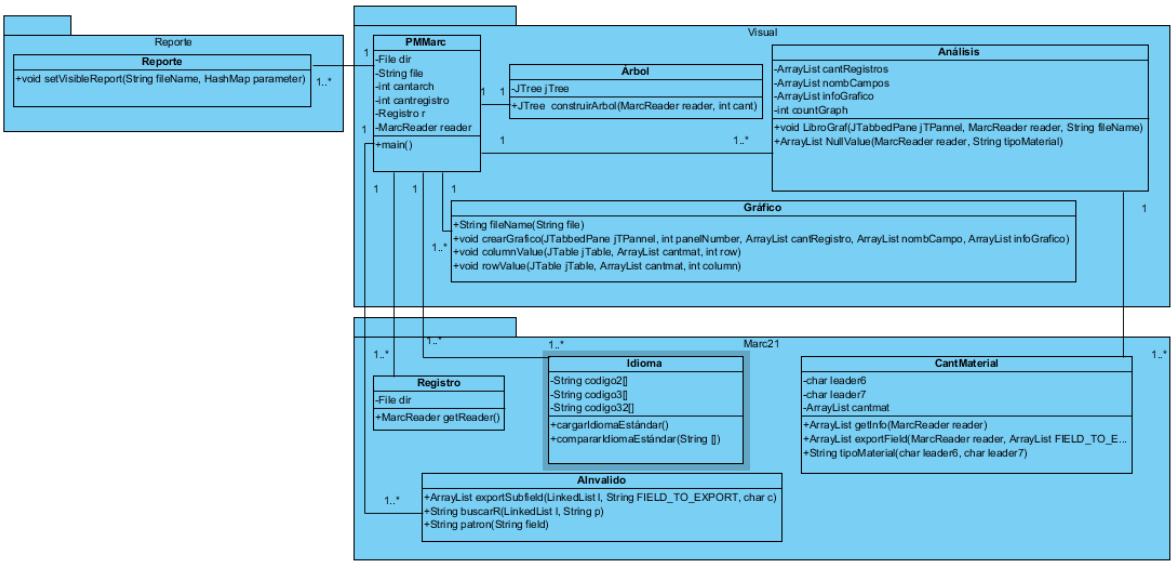


Figura 8 Diagrama de clases Fuente: elaboración propia

3.2. Perfilado de un conjunto de datos

El perfilado de un conjunto de datos permite detectar los errores existentes en un solo conjunto de datos sin establecer comparaciones con otros. Busca los problemas existentes en cada registro de forma independiente, utiliza funciones que determinan el tipo de material y los errores en los campos mínimos mostrándolos en tablas o gráficos. La herramienta visualiza de forma general los problemas más comunes en los campos mínimos específicamente para un único conjunto de datos.

3.2.1. Visualización de los resultados por registros

Cuando se inicia el sistema, se selecciona un archivo con extensión *mrc*, luego automáticamente se crea un componente JTree encargado de mostrar toda la información correspondiente a cada registro del conjunto de datos como se muestra en la Figura 9. Primeramente encontraremos la cabecera, que es el número que identifica al registro de forma única. La posición 6 y 7 definen el tipo de material (música, mapa, libro, etc.). Luego se muestran las etiquetas de los campos (100, 245,650…) y sus valores. Los subcampos son identificados por $letra. Las flechas obtienen la información

correspondiente a los registros de uno en uno. El botón “Ir a” recorre todo el conjunto de datos hasta llegar al número de registro especificado en el JTextField.

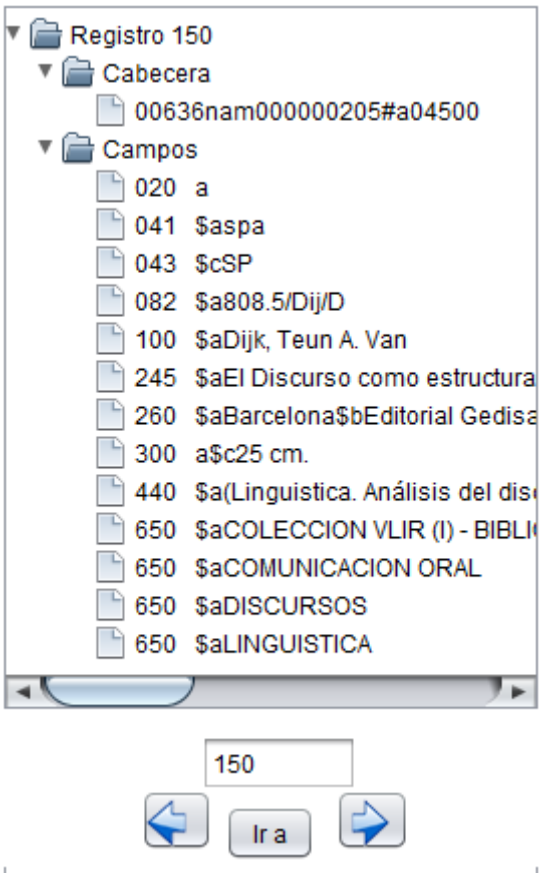


Figura 9 Árbol de registros Fuente: Elaboración propia

Cuando se realiza el perfilado de un conjunto de datos uno de los aspectos a considerar son los valores más importantes almacenados en los campos. La opción “Contar campos” calcula la cantidad de registros que poseen los campos mínimos, por ejemplo: existen 15741 registros en el archivo UCLV(2) que poseen el campo autor(100) como se muestra en la Figura 10.

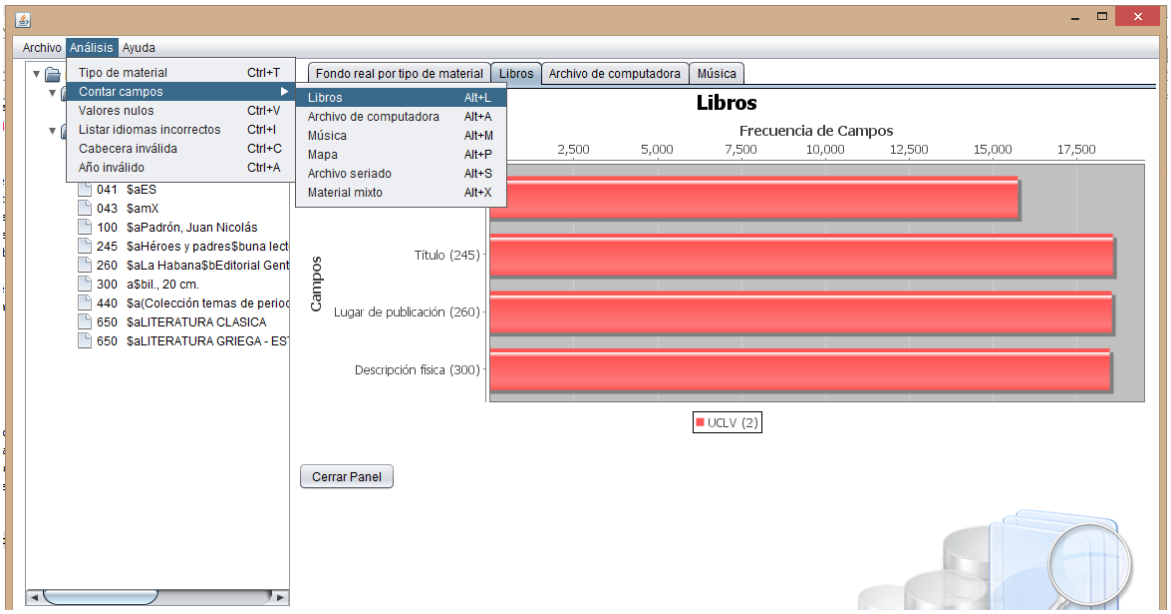


Figura 10 Cantidad de registros que poseen los campos mínimos Fuente: Elaboración propia

En la Figura 11 se muestra uno de los análisis más importantes de la herramienta, busca la cantidad de campos mínimos ausentes existentes por tipo de material que deben poseer valor para que el registro esté completo, por ejemplo: existen 63 libros que no poseen Título por lo que cuando se realice una búsqueda en otro software utilizando el campo 245, ninguno de los 63 libros aparece como resultado.

Registros con campos nulos

Campos	Libros	Archivo de co...	Música	Mapa	Archivo seri...	Material mixto	Otros	Total
Número de publicación (028)	18646	43	7	8	2	0	37	18745
Autor (100)	2905	42	0	0	0	0	16	2964
Título (245)	63	0	0	0	0	0	8	71
Características del archivo (256)	18646	43	7	8	2	0	37	18745
Lugar de publicación (260)	98	0	0	0	1	0	7	106
Descripción física (300)	160	1	0	0	0	0	12	174
Frecuencia de publicación (310)	18645	43	7	8	2	0	37	18744
Nota general (500)	12475	40	6	1	1	0	31	12556
Temática (650)	180	0	0	0	1	0	11	193
Nombre corporativo (710)	15734	42	7	4	2	0	34	15825

Figura 11 Registros con campos nulos Fuente: Elaboración propia

La opción “Listar idioma incorrecto” del menú Análisis busca los lenguajes incorrectos almacenados en la herramienta y que no se encuentran en los códigos MARC para idiomas de 2 y 3 caracteres o los listados en el ISO 639 también para idiomas de dos y tres caracteres. Luego al presionar el botón “Listar registros”, se encuentran todos los registros que poseen el idioma seleccionado como se muestra en la Figura 12 Lista de idiomas incorrectos Fuente: Elaboración propia

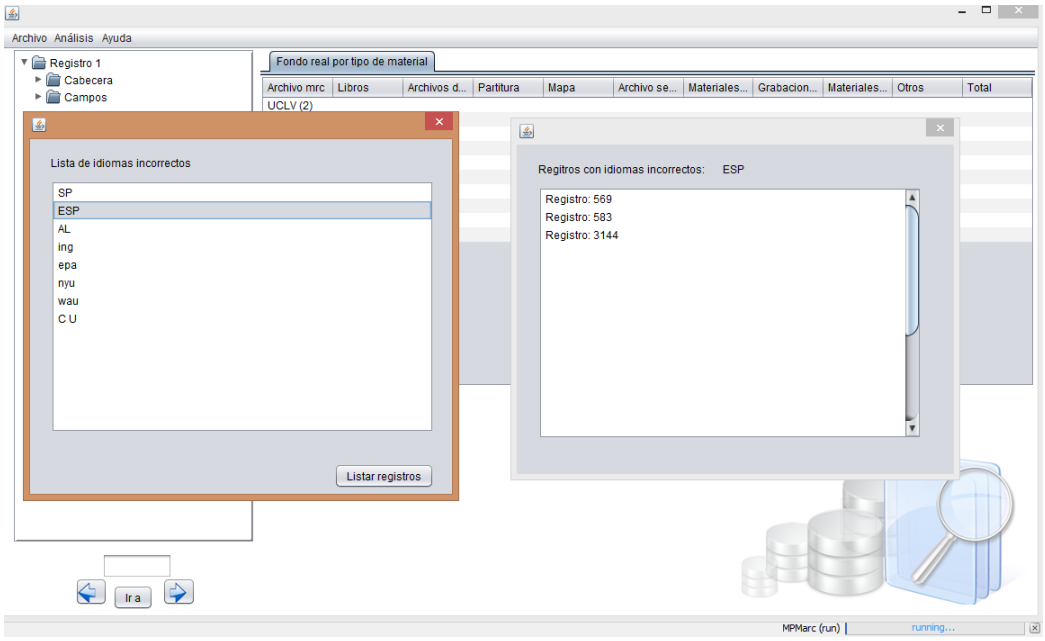


Figura 12 Lista de idiomas incorrectos Fuente: Elaboración propia

Para obtener los patrones de los años de edición incorrectos entrados de forma manual la herramienta cuenta con la opción “Año inválido” del menú “Análisis”. Se recorren todos los registros obteniendo los patrones de los años que deben ser numéricos pero son almacenados como cadenas con diferentes caracteres (. , “ ; [] ¿?) causando que el valor no pueda ser comparado. Se muestran todos los patrones diferentes. El botón “Listar registros”, busca los registros que poseen el patrón seleccionado en la lista. Ver Figura 13.

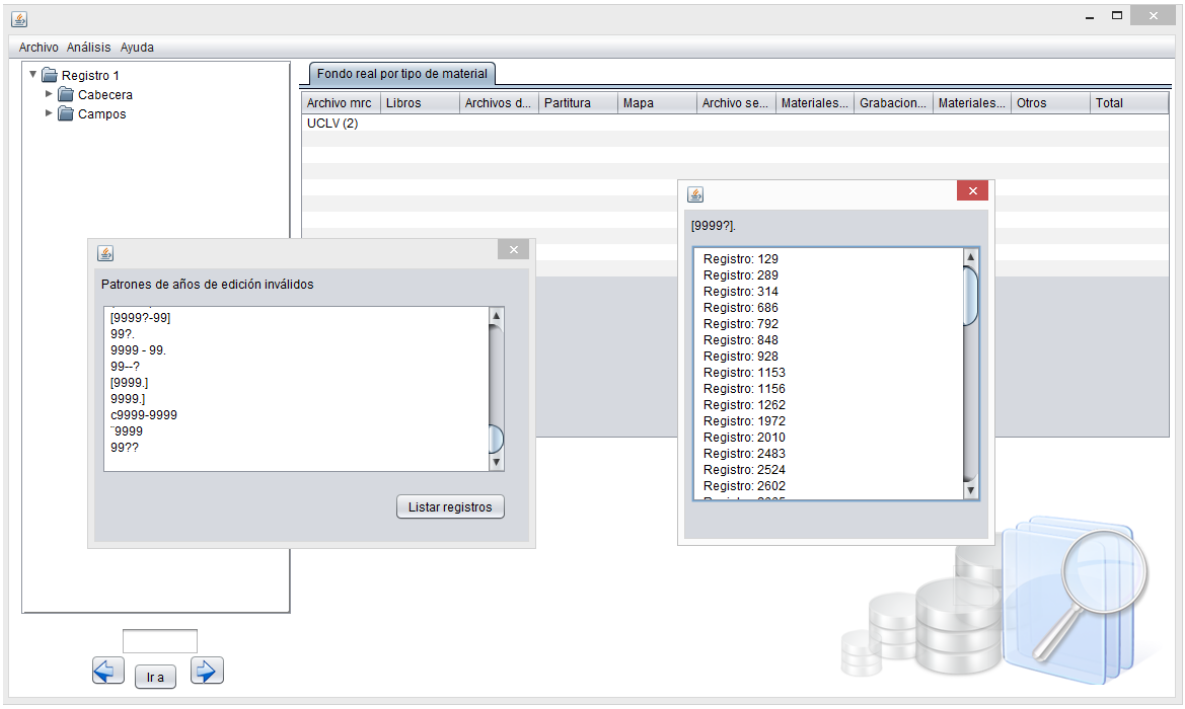


Figura 13 Patrones de años de edición inválidos Fuente: Elaboración propia

3.3. Perfilado de múltiples conjuntos de datos

El perfilado de múltiples conjuntos de datos permite que la información almacenada en los registros se compare con respecto a la cantidad de registros por tipo de material y a la frecuencia de los campos mínimos ausentes.

Para obtener la cantidad de registros por tipo de material para el último archivo *mrc* leído, la herramienta cuenta con la opción "Tipo de material" del menú Comparar. La tabla muestra las estadísticas de hasta un total de 20 archivos, ver la Figura 14. La clase CantMaterial determina el total de registros ya sea de tipo música, mapa, libro, etc. Todos los registros que no se encuentren en esas categorías serán contados como "Otros", los cuales usualmente poseen encabezados inválidos.

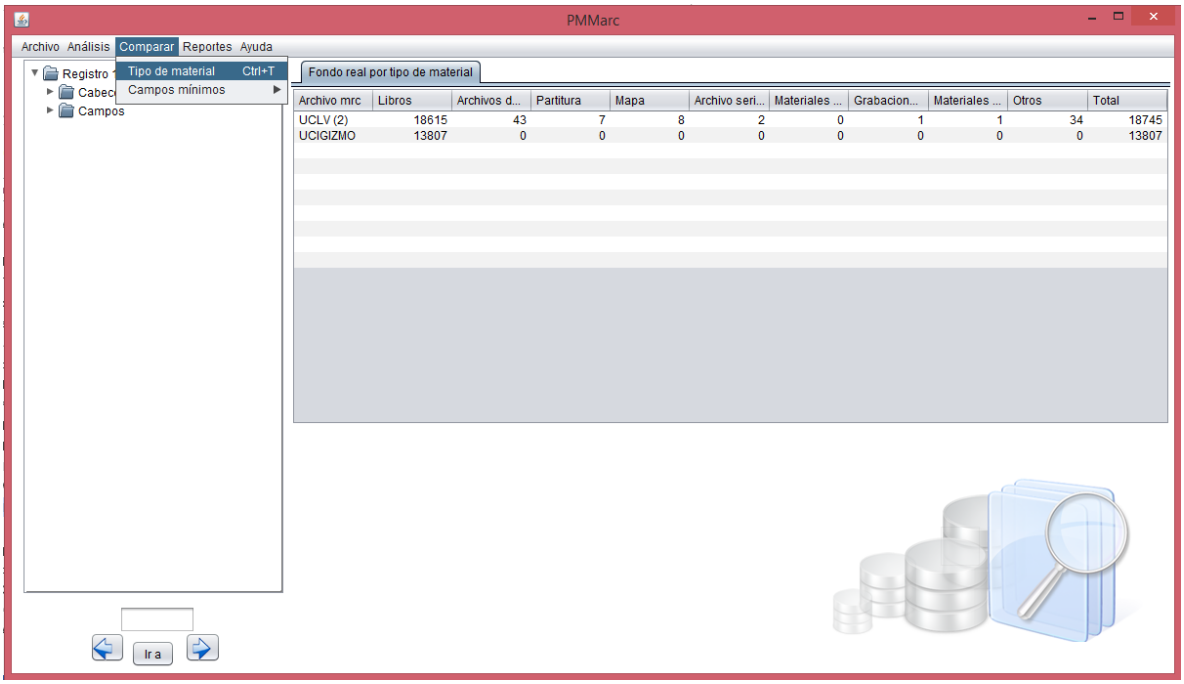


Figura 14 Fondo real por tipo de material Fuente: Elaboración propia

PMMarc muestra un gráfico que permite visualizar el total de campos mínimos ausentes en los conjuntos de datos seleccionados, ver Figura 15, por ejemplo: se observa que en los catálogos de la UCLV el campos mínimos con más problemas de completitud es el autor (100), mientras que para la UCI el campo que guarda la descripción física (300) es el que más problemas de completitud presenta.

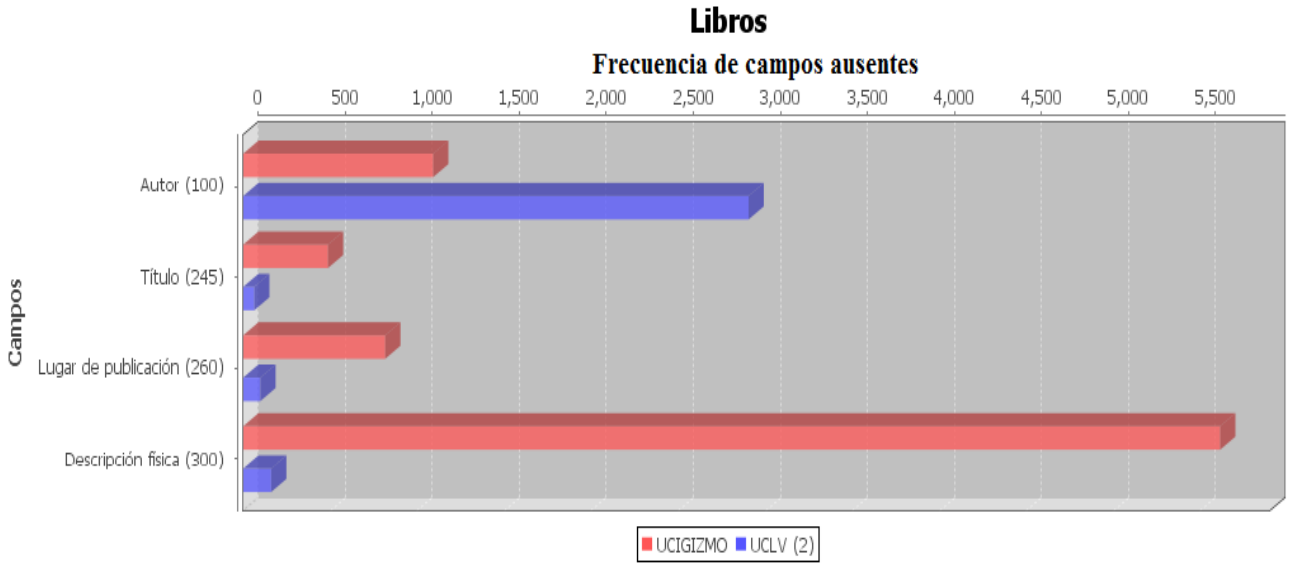


Figura 15 Frecuencia de campos ausentes en libros Fuente: Elaboración propia

3.4. Complejidad computacional de un algoritmo

Para calcular la complejidad computacional se utiliza la notación O. Tomando en cuenta los órdenes de complejidad que establecen un orden entre los algoritmos sobre la base de su eficiencia, siendo los algoritmos más eficientes aquellos cuyo orden de complejidad se encuentra más a la izquierda en esta relación. Se utilizan también las propiedades de la notación asintótica O y las reglas para la determinación de la complejidad de un algoritmo (Allen Weiss, 2002).

$O(1) \subset O(\log n) \subset O(n^a) \subset O(n^a (\log n)^b) \subset O(n^c) \subset O(d^n) \subset O(n!) \subset O(n^n)$

con $a > 0$; $b \geq 1$; $c > a$; $d > 1$

Propiedades de la notación asintótica O:

- $O(f(n) + g(n)) = O(\max\{f(n), g(n)\})$
- $O(f(n)).O(g(n)) = O(f(n).g(n))$
- $O(cf(n)) = O(f(n))$
- $O(f(n)) = O(g(n))$ si, y solo si, $f(n) \in O(g(n))$ y $g(n) \in O(f(n))$
- $O(f(n)) \subset O(g(n))$ si, y solo si, $f(n) \in O(g(n))$ pero $g(n) \notin O(f(n))$
- Si $f(n) \in O(g(n))$ y $g(n) \in O(h(n))$ entonces $f(n) \in O(h(n))$

Reglas para la determinación de la complejidad de un algoritmo:

Determinación del parámetro entero que servirá como medida de la dimensión de las instancias.

La complejidad temporal de las instrucciones de lectura, escritura, comparación, operaciones aritméticas y asignación se consideran que son un O(1).

La complejidad temporal de una sucesión de instrucciones está dada por la regla de la suma.

$C(S1;S2;\ldots;Sn)=C(S1)+C(S2)+..+C(Sn)$

La complejidad de la instrucción condicional es la suma de las complejidades de evaluar la condición y la complejidad de la instrucción a ejecutar. En el caso de una instrucción if-then-else, el costo de la instrucción a ejecutar es la complejidad de la instrucción de mayor costo de las dos posibles.

$C(IF) = C(\text{condición}) + C(\text{Sentencia}) = \max\{ C(\text{condición}), C(\text{Sentencia})\}$

C(IF-Else)=C(condición)+C(Sentencia)+C(Else)=max{C(condición),C(Sentencia), C(Else) }

La complejidad de un ciclo está dada por el producto del total de veces que se ejecuta el ciclo por la complejidad del cuerpo del ciclo.

Esto es porque una sentencia dentro de un ciclo se ejecuta tantas veces como se entre al ciclo, por lo que cada sentencia dentro del ciclo tiene una complejidad que es igual al producto de la sentencia por la cantidad de veces que se ejecuta.

C(ciclo)= C(cuerpo) * cantidad de iteraciones

Se debe tener en cuenta que cuando existen ciclos anidados el cálculo se realiza de adentro hacia fuera, desde el ciclo que está más adentro. Calculando la cantidad de veces que se entra al ciclo más adentro (muchas veces es la cantidad de veces que se ejecuta el de afuera por la cantidad de veces que se ejecuta el ciclo de adentro, pero no siempre es así, el de adentro si tiene cierta dependencia de la iteración del ciclo más afuera, en cada iteración no tiene por qué repetirse la misma cantidad de veces.

La complejidad de una instrucción donde se llama a otro método es igual a la complejidad del método que se está llamando.

Para el caso del algoritmo getInfo el peor caso es cuando se recorren todos los registros, con un total de 18745 para el conjunto de datos de la UCLV, además si los registro no pertenecen a un tipo de material por lo que se debe comprobar todas las condiciones del *case*, lo cual aumenta la complejidad.

```
public static ArrayList<Integer> getInfo(LinkedList l) throws FileNotFoundException { //O(n)
 //Son asignaciones por lo que C(asig)= O(1)
                        int index;
                        int libros = 0;
                        int archivoComputadora = 0;
                        int musica = 0;
                        int mapa = 0;
```

```
int serial = 0;
int materialMixto = 0;
int materialVisual = 0;
int grabacionSonora = 0;
int otros = 0;
int pos = 0;

while (pos < l.size()) {  //18745*C(cuerpo) = 18745*O(1)=O(n)
        // C(asig)= O(1)
        Record record = (Record) l.get(pos);
         leader6 = record.getLeader().getTypeOfRecord();
         leader7 = record.getLeader().toString().charAt(7);

          switch (tipoMaterial(leader6, leader7)) { //C(switch)=
max(C(tipoMat), C(case)...) = O(1)
            case "libro":
               libros++;
               break;
            case "archivoComputadora":
               archivoComputadora++;
               break;
            case "musica":
               musica++;
               break;
            case "mapa":
               mapa++;
               break;
            case "serial":
               serial++;
               break;
            case "materialMixto":
```

```
                    materialMixto++;
                    break;
                case "materialVisual":
                    materialVisual++;
                    break;
                case "grabacionSonora":
                    grabacionSonora++;
                    break;
                case "otro":
                    otros++;
                    break;
            }
        pos++;
    }

  int total = mapa + materialMixto + libros + archivoComputadora + musica + serial + materialVisual + grabacionSonora + otros;
        // C(add)=O(1)
            cantmat.add(0, libros);
            cantmat.add(1, archivoComputadora);
            cantmat.add(2, musica);
            cantmat.add(3, mapa);
            cantmat.add(4, serial);
            cantmat.add(5, materialMixto);
            cantmat.add(6, grabacionSonora);
            cantmat.add(7, materialVisual);
            cantmat.add(8, otros);
            cantmat.add(9, l.size());

            return cantmat;
        }
```

```
public static String tipoMaterial(char leader6, char leader7) {    //O(1)    C(IF) = max(C(<Condicion>),C(<SentenciaIf>),C(<SentenciaElse>)) = O(1)
        String flag = "otro";

        if ((leader6 == 'a' &&( leader7 == 'm'|| leader7 == 'a'  || leader7 == 'c' || leader7 == 'd')) || (leader6 == 't' && (leader7 == 'm'|| leader7 == 'a'  || leader7 == 'c' || leader7 == 'd'))) { //max(C(<Condicion>),C(asig)) = max(O(1), O(1) = O(1))
            flag = "libro";
        } else
            if (leader6 == 'm' && (leader7 == 'a' ||leader7 == 'b' || leader7 == 'c' ||leader7 == 'd' ||leader7 == 'i' ||leader7 == 'm' ||leader7 == 's')) {
            flag = "archivoComputadora";
        } else
              if ((leader6 == 'c' && (leader7 == 'a' ||leader7 == 'b' || leader7 == 'c' ||leader7 == 'd' ||leader7 == 'i' ||leader7 == 'm' ||leader7 == 's')) || (leader6 == 'd' && ((leader7 == 'm'|| leader7 == 'a'  || leader7 == 'c' || leader7 == 'd')))) {
            flag = "musica";
        } else
            if ((leader6 == 'e' && (leader7 == 'a' ||leader7 == 'b' || leader7 == 'c' ||leader7 == 'd' ||leader7 == 'i' ||leader7 == 'm' ||leader7 == 's'))||((leader6 == 'f' && (leader7 == 'a' || leader7 == 'c' ||leader7 == 'd' ||leader7 == 'm')))) { // a, b, c, d, i, m, s
            flag = "mapa";
        } else
            if (leader6 == 'a' && (leader7 == 's' || leader7 == 'b' || leader7 == 'i')) {
            flag = "serial";
        } else if (leader6 == 'p' && (leader7 == 'c' || leader7 == 'd' )) {
            flag = "materialMixto";
        } else
```

```
        if ((leader6 == 'g' || leader6 == 'k' || leader6 == 'r' || leader6 == 'o') && (leader7 == 'a'
||leader7 == 'b' || leader7 == 'c' ||leader7 == 'd' ||leader7 == 'i' ||leader7 == 'm' ||leader7 == 's'))
{
         flag = "materialVisual";
      } else
           if ((leader6 == 'i' || leader6 == 'j')&& (leader7 == 'a' ||leader7 == 'b' || leader7 == 'c'
||leader7 == 'd' ||leader7 == 'i' ||leader7 == 'm' ||leader7 == 's') ){
        flag = "grabacionSonora";
      }


      return flag;
   }
```

3.5. Campos mínimos más propensos a presentar errores de completitud

Se realiza el análisis de los campos mínimos por tipo de material de los conjuntos de datos de la UCLV y la UCI. Para los libros el resultado obtenido por la herramienta se muestra en la Figura 15. El campo autor y el título son uno de los campos más importantes de cada registro ya que las búsquedas en los software, se realizan sobre estos valores, un libro que no posea estos campos lo más probable es que no pueda ser accedido el 90% de las veces.
Los campos mínimos más propensos a presentar errores de incompletitud por tipo de material son los siguientes:

- Libros: Autor (100), descripción física (300).
- Archivo de computadora: Características del archivo (265).
- Música: Número de publicación (028).
- Mapa: Nota general (500).
- Archivo seriado: Frecuencia de publicación (310), entrada secundaria de nombre corporativo (710).

3.6. Conclusiones parciales

En el presente capítulo se implementó la primera versión de la herramienta de perfilado de metadatos PMMarc para la detección de errores en los campos del formato MARC 21 por tipo de material. Para el diseño de la herramienta se analizaron las características que relacionadas con los errores más comunes presentes en el módulo Catalogación para determinar que funciones que mejoran la calidad de la información. Se determinaron los campos mínimos por tipo de material que se encuentran más propensos a presentar errores de completitud utilizando la herramienta de perfilado implementada.

CONCLUSIONES

- Se identificaron los errores tipográficos, ortográficos, valores ausentes, errores de formato de codificación, registros duplicados, valores ficticios en los campos, valores implícitos y los datos no estandarizados como los más comunes e importantes a partir de la comparación realizada entre diferentes taxonomías de errores expuestas en la literatura y los problemas más frecuentes que se pueden encontrar en el módulo Catalogación del ABCD.
- Se determinó que las herramientas de perfilado de datos cubren en un 100% los errores más comunes mediante la correspondencia establecida entre dichos errores y las características inherentes de las herramientas de perfilado de datos.
- Se implementó la primera versión de la herramienta de perfilado de metadatos denominada "PMMarc", la cual permite calcular el fondo real por tipo de material de los catálogos de las bibliotecas que utilizan el formato MARC 21, detectar errores y mostrar estadísticas descriptivas sobre los registros bibliográficos analizados.
- Se identificaron los campos autor (100), descripción física (300), características del archivo de computadora (265), número de publicación (028), nota general (500), frecuencia de publicación (310), entrada secundaria de nombre corporativo (710) como los campos mínimos del formato MARC 21 que están más propensos a presentar errores de completitud según los catálogos analizados por tipo de material mediante la herramienta PMMarc.

RECOMENDACIONES

Extender la herramienta PMMarc de manera que permita:

- Incorporar otros formatos de entrada y salida compatibles con MARC 21.
- Detectar errores no tratados en el presente trabajo.
- Extender el análisis de los errores al resto de los campos del formato MARC 21.
- Incluir nuevos estándares bibliográficos.

REFERENCIAS BIBLIOGRÁFICAS

ALLEN WEISS, M. 2002. *Estructuras de Datos en Java*.

AMAYA DE HEREDIA, M. 2004. Reglas de catalogación angloamericanas.

ANDREU ALVAREZ, Y. 2015. Análisis de la calidad de datos en fuentes de la suite ABCD.

BEALL, J. 2006. Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information,* 6**,** 20.

BOREL, A. K., JAN 2011a. MarcXimiL A deduplication tool for library records... and more.

BOREL, A. K., JAN 2011b. MarcXimiL – near-duplicates detection.

BRAY, T. 1998. RDF and Metadata.

BULTERMANN, D. C. A. 2004. Is It Time for a Moratorium on Metadata? : IEEE Computer Society Press.

CONGRESS, L. O. 2008a. *Appendix C - Minimal Level Record Examples* [Online]. Available: http://www.loc.gov/marc/bibliographic/.

CONGRESS, L. O. 2008b. *Marc 21* [Online]. Available: http://www.loc.gov/marc/marcspa.html.

FEDERICO MEDRANO, J., FIGUEROLA , C. G. & ALONSO BERROCAL, J. L. 2012. Repositorios digitales en España y calidad de los metadatos.

FERNÁNDEZ, G. & LENZO, N. 2010. Software ABCD (Automatización de Bibliotecas y Centros de Documentación). *1ra Jornada "Temas actuales en Bibliotecología". 1ra Jornada "Temas actuales en Bibliotecología".* Argentina.

G. STAMOU, J., OSSENBRUGGEN, J. & PAN, G. S. 2006. Multimedia Annotations on the Semantic Web. IEEE Computer Society Press.

GARCIA B. , A. 2008. Formato MARC 21 para registros bibliográficos.

GARCÍA MENDOZA, J. L., DÍAZ DE LA PAZ, L., GONZÁLEZ GONZÁLEZ, L. M., NUÑEZ ARCIA, Y. & LEIVA MEDEROS, A. A. 2015. Herramienta CompMARC para la medición de la completitud de registros bibliográficos en formato MARC 21. *III Taller Internacional de Bibliotecas Especializadas* La Habana, Cuba.

HARTE, H. 2008. Aspectos básicos sobre la calidad de datos. Trillium Software.

HERNÁNDEZ AGUSTÍ, M., SANTOS MUÑOZ, R. & TEJERO LÓPEZ, P. 2015. *Formato MARC 21 para registros bibliográficos*.

HILLMANN, D. I. B., THOMAS R. 2004. THE CONTINUUM OF METADATA QUALITY: DEFINING, EXPRESSING, EXPLOITING. 17.

J. R. SMITH, P. S. 2006. Metadata Standards Roundup. IEEE Computer Society Press.

LÓPEZ PORRERO, B. 2011. LIMPIEZA DE DATOS: REEMPLAZO DE VALORES AUSENTES Y ESTANDARIZACIÓN.

MATT CASTERS, R. B., JOS VAN DONGEN 2010. *Pentaho Kettle Solutions: Building Open Source ETL Solutions with,* Wiley Publishing, Inc., Indianapolis, Indiana.

NUÑEZ ARCIA, Y., DÍAZ DE LA PAZ, L. & GARCÍA MENDOZA, J. L. 2016. ALGORITMO PARA CORREGIR ANOMALÍAS A NIVEL DE INSTANCIA EN GRANDES VOLÚMENES DE DATOS UTILIZANDO MAPREDUCE. *III Conferencia Internacional en Ciencias Computacionales e Informáticas.* La Habana, Cuba.

OCLC. 2016. *Marcas comerciales y/o marcas de servicios nacionales e internacionales de OCLC y de sus miembros* [Online]. Available: file:///bibformats/es/introduction.html#BABDFJDA.

OCHOA, X. & DUVAL, E. 2006. Quality Metrics for Learning Object Metadata.

OCHOA, X. & DUVAL, E. 2008. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries,* 10**,** 67-91.

POWER DATA, E. E. G. D. D. 2010. La Calidad de los Datos: una radiografía completa.

RAMALHO, L. G. S. 2010. Implementing a modern API for CDS/ISIS, a classic semistructured NoSQL database.
REESE, T. 2010. *MarcEdit: A simplified metadata processing tool*.
ROSIE, H. 2014. Repository v. Catalogue, Top tips for thesis metadata management. EThOS Metadata Manager.
RUMBAUGT, J. 2000. *Modelaje y diseño orientado a objetos*.
SÁNCHEZ MAGANTO, A., NOGUERAS ISO, J. & BALLARI, D. 2008. Normas sobre metadatos. 123**,** 48-57.
SCHMULLER, J. 2000. *Aprendiendo UML en 24 horas*.
SOFTWARE, T. 2008. Conceptos básicos para la generación de perfiles de datos.
SOTO, D. 2008. *Integración y Calidad de Datos* [Online]. Available: http://integracionycalidad.blogspot.com.
VELOSO LARA, A. 2012. Evaluación del Sistema de Recuperación de Información de la herramienta de gestión bibliotecaria ABCD versión 1.0.6.
VIDAL GIL, J. 2013. *Recursos y análisis sobre herramientas de Business Intelligence* [Online]. Available: http://sasybi.blogspot.com/.
WANG, R. Y. & STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems***,** 5-33.

ANEXOS

Anexo 1: Registro de nivel mínimo - Libro

Los registros de libros se pueden identificar por el código a en Leader/06 como material impreso textual y como monografía por el código m en Leader/07. El registro contiene los siguientes campos como información básica: **100** (Autor), **245** (Título), **260** (Publicación, Distribución, etc. (Imprenta)), y **300** (Descripción Física).

Leader/00-23	*****cam##22*****7a#4500
001	<control number>
003	<control number identifier>
005	19921113071152.5
008/00-39	870102s1962####at############000#0#eng#d
040	##$a[organization code]$c[organization code]
100	1#$aKelleher, James M.
245	10$aRoman fever :$ba story of the struggle by the Catholic people of New South Wales to remedy the injustice caused by the Education Acts of 1866 and 1880--an observer's opinion of the causes which so far have rendered that struggle in vain /$cby James M. Kelleher.
260	##$aSurrey Hills, N.S.W. :$bCatholic Press Newspaper Co.,$c[1962?]
300	##$a139 p. ;$c16 cm.

Anexo 2: Campos mínimos que debe poseer un registro.

A continuación se utilizan 3 códigos diferentes para diferenciar los elementos que siempre deben estar presentes (M), los que se deben registrar si se tiene la información (A) y los datos opcionales (O).

		National Level	Minimal Level
LEADER		M	M
00-04	Record length	M	M
05	Record status	M	M
06	Type of record	M	M
07	Bibliographic level	M	M
08	Type of control	M	M
09	Character Coding Scheme	M	M
10	Indicator count	M	M
11	Subfield code count	M	M
12-16	Base address of data	M	M
17	Encoding level	M	M
18	Descriptive cataloging form	M	M
19	Multipart resource record level	M	M
20	Length of the length-of-field portion	M	M
21	Length of the starting-character-	M	M

	position portion		
22	Length of the implementation-defined portion	M	M
23	Undefined	M	M
DIRECTORY		M	M
00-02	Tag	M	M
03-06	Field length	M	M
07-11	Starting character position	M	M
001 CONTROL NUMBER		M	M
003 CONTROL NUMBER IDENTIFIER		M	M
005 DATE AND TIME OF LATEST TRANSACTION		M	M
008-FIXED-LENGTH DATA ELEMENTS		M	M
008-All material types			
00-05	Date entered on file or execution	M	M
35-37	Language	M	M
39	Cataloging source	M	M
008-Books			
008-Computer files			
008-Maps			
25	Type of cartographic material	M	M
008-Music			
20	Format of music	M	M
008-Continuing resources			
34	Entry convention	M	M
008-Visual materials			
33	Type of visual material	M	M
008-Mixed materials			
010 LIBRARY OF CONGRESS CONTROL NUMBER		A	A
$a	LC control number	A	A
$b	NUCMC control number	A	A
$z	Canceled or invalid LC control number	A	A
016 NATIONAL BIBLIOGRAPHIC AGENCY CONTROL NUMBER		A	A
$a	Record control number	A	A
$z	Canceled or invalid control number	A	A

$2	Source	A	A
020	**INTERNATIONAL STANDARD BOOK NUMBER**	A	A
$a	International Standard Book Number	A	A
$c	Terms of availability	A	A
$z	Canceled or invalid ISBN	A	A
$6	Linkage	A	A
022	**INTERNATIONAL STANDARD SERIAL NUMBER**	A	A
$a	International Standard Serial Number	A	A
$l	ISSN-L	A	A
$m	Canceled ISSN-L	A	A
$y	Incorrect ISSN	A	A
$z	Canceled ISSN	A	A
$2	Source	A	A
$6	Linkage	A	A
024	**OTHER STANDARD IDENTIFIER**	A	A
$a	Standard number or code	M	M
$c	Terms of availability	A	A
$d	Additional codes following the standard number or code	A	A
$z	Canceled or invalid standard number or code	A	A
$2	Source of number or code	A	A
$6	Linkage	A	A
028	**PUBLISHER NUMBER**	A	A
$a	Publisher number	M	M
$b	Source	M	M
$6	Linkage	A	A
034	**CODED CARTOGRAPHIC MATHEMATICAL DATA**	A	A
$a	Category of scale	M	M
$b	Constant ratio linear horizontal code	A	A
$c	Constant ratio linear vertical scale	A	A
$6	Linkage	A	A
040	**CATALOGING SOURCE**	M	M
$a	Original cataloging agency	A	A
$b	Language of cataloging	A	A
$c	Transcribing agency	M	M
$d	Modifying agency	A	A
$6	Linkage	A	A
042	**AUTHENTICATION CODE**	A	A
$a	Authentication code	M	M
066	**CHARACTER SET PRESENT**	A	A
$c	Alternate G0 or G1 character set	M	M
100	**MAIN ENTRY PERSONAL NAME**	A	A
$a	Personal name	M	M
$b	Numeration	A	A

$c	Titles and other words associated with a name	A	A
$d	Dates associated with a name	A	A
$q	Fuller form of name	A	A
$4	Relator code	O	A
$6	Linkage	A	A

110	**MAIN ENTRY CORPORATE NAME**	A	A
$a	Corporate name or jurisdiction name as entry element	M	M
$b	Subordinate unit	A	A
$c	Location of meeting	A	A
$d	Date of meeting or treaty signing	A	A
$g	Miscellaneous information	A	A
$k	Form subheading	A	A
$n	Number of part/section/meeting or standard number	A	A
$4	Relator code	O	A
$6	Linkage	A	A

111	**MAIN ENTRY MEETING NAME**	A	A
$a	Meeting name or jurisdiction name as entry element	M	M
$c	Location of meeting	A	A
$d	Date of meeting	A	A
$e	Subordinate unit	A	A
$n	Number of part/section/meeting	A	A
$4	Relator code	O	A
$6	Linkage	A	A

130	**MAIN ENTRY UNIFORM TITLE**	A	A
$a	Uniform title	M	M
$d	Date of treaty signing	A	A
$f	Date of a work	A	A
$k	Form subheading	A	A
$l	Language of a work	A	A
$m	Medium of performance for music	A	A
$n	Number of part/section of a work	A	A
$o	Arranged statement for music	A	A
$p	Name of part/section of a work	A	A
$r	Key for music	A	A
$s	Version or standard number	A	A
$6	Linkage	A	A

240	**UNIFORM TITLE**	A	A
$a	Uniform title	M	M
$d	Date of treaty signing	A	A
$f	Date of a work	A	A
$g	Miscellaneous information	A	A
$k	Form subheading	A	A
$l	Language of a work	A	A
$m	Medium of performance for music	A	A
$n	Number of part/section of a work	A	A
$o	Arranged statement for music	A	A
$p	Name of part/section of a work	A	A
$r	Key for music	A	A

$s	Version	A	A
$6	Linkage	A	A
245	**TITLE STATEMENT**	M	M
$a	Title	M	M
	[Subfield is A for Mixed materials.]		
$c	Statement of responsibility, etc.	A	A
$f	Inclusive dates	A	A
$g	Bulk dates	A	A
$k	Form	A	A
$n	Number of part/section of a work	A	A
$p	Name of part/section of a work	A	A
$s	Version	A	A
$6	Linkage	A	A
246	**VARYING FORM OF TITLE**	A	A
$a	Title proper/short title	M	M
$f	Date or sequential designation	A	A
$n	Number of part/section of a work	A	A
$p	Name of part/section of a work	A	A
$6	Linkage	A	A
250	**EDITION STATEMENT**	A	A
$a	Edition statement	M	M
$6	Linkage	A	A
254	**MUSICAL PRESENTATION STATEMENT**	A	A
$a	Musical presentation statement	M	M
$6	Linkage	A	A
255	**CARTOGRAPHIC MATHEMATICAL DATA**	A	A
$a	Statement of scale	M	M
$b	Statement of projection	A	A
$c	Statement of coordinates	A	A
$6	Linkage	A	A
256	**COMPUTER FILE CHARACTERISTICS**	A	A
$a	Computer file characteristics	M	M
$6	Linkage	A	A
257	**COUNTRY OF PRODUCING ENTITY**	A	A
$a	Country of producing entity	M	M
$6	Linkage	A	A
$2	Source	A	A
260	**PUBLICATION, DISTRIBUTION, ETC. (IMPRINT)**	A	A
	[Field is M for Serials.]		
$b	Name of publisher, distributor, etc. [Subfield is M for Serials.]	A	A
$c	Date of publication, distribution, etc. [Subfield is M for Maps.]	A	A
300	**PHYSICAL DESCRIPTION**	M	M
	[Field is A for Computer files.]		
$a	Extent	M	M
$f	Type of unit	A	A
$6	Linkage	A	A

362	**DATES OF PUBLICATION AND/OR SEQUENTIAL DESIGNATION**	A	A
	$a Dates of publication and/or sequential designation	M	M
	$6 Linkage	A	A
440	**SERIES STATEMENT/ADDED ENTRY TITLE**	A	A
	$a Title	M	M
	$n Number of part/section of a work	A	A
	$p Name of part/section of a work	A	A
	$v Volume/sequential designation	A	A
	$x International Standard Serial Number	A	A
	$6 Linkage	A	A
490	**SERIES STATEMENT**	A	A
	$a Series statement	M	M
	$v Volume/sequential designation	A	A
	$x International Standard Serial Number	A	A
	$6 Linkage	A	A
511	**PARTICIPANT OR PERFORMER NOTE**	A	A
	$a Participant or performer note	M	A
	$6 Linkage	A	A
533	**REPRODUCTION NOTE**	A	A
	$a Type of reproduction	M	M
	$b Place of reproduction	M	M
	$c Agency responsible for reproduction	A	A
	$d Date of reproduction	A	A
	$e Physical description of reproduction	A	A
	$f Series statement of reproduction	A	A
	$m Dates and/or sequential designation of issues reproduced	A	A
	$6 Linkage	A	A
773	**HOST ITEM ENTRY**	A	A
	$a Main entry heading	A	A
	$b Edition	A	A
	$r Report number	A	A
	$s Uniform title	A	A
	$t Title	A	A
	$u Standard Technical Report Number	A	A
	$x International Standard Serial Number	A	A
	$6 Linkage	A	A
780	**PRECEDING ENTRY**	A	A
	$a Main entry heading	A	A
	$b Edition	A	A
	$s Uniform title	A	A
	$t Title	A	A
	$x International Standard Serial Number	A	A
	$6 Linkage	A	A
785	**SUCCEEDING ENTRY**	A	A
	$a Main entry heading	A	A
	$b Edition	A	A
	$s Uniform title	A	A

$t Title	A	A
$x International Standard Serial Number	A	A
$6 Linkage	A	A

880 ALTERNATE GRAPHIC REPRESENTATION	A	A
$6 Linkage	M	M
$a-z Same as associated field	M	M
$0-5, 7-9 Same as associated field	M	M