Herramienta CompMARC para la medición de la completitud de registros bibliográficos en formato MARC 21

Juan Luis García-Mendoza[1], Lisandra Díaz-De la Paz[2], Luisa González-González[3], Yaisel Nuñez-Arcia[4], Amed Abel Leiva-Mederos[5]

1 Universidad Central "Marta Abreu" de Las Villas, juanluis@uclv.cu

2 Universidad Central "Marta Abreu" de Las Villas, ldp@uclv.edu.cu

3 Universidad Central "Marta Abreu" de Las Villas, luisagon@uclv.edu.cu

4 Universidad Central "Marta Abreu" de Las Villas, ynunes@uclv.cu

5 Universidad Central "Marta Abreu" de Las Villas, amed@uclv.edu.cu

RESUMEN

MARC 21 constituye uno de los estándares más utilizados para la catalogación de registros bibliográficos. Según los resultados del procesamiento de encuestas aplicadas a especialistas en Ciencias de la Información de la Universidad Central "Marta Abreu" de Las Villas, uno de los principales problemas de calidad de datos que presentan los registros bibliográficos en este formato es la incompletitud de sus datos. Por consiguiente, el presente trabajo tiene como objetivo medir la dimensión de calidad de datos completitud de registros bibliográficos en formato MARC 21. En el proceso de medición de la completitud se utilizaron dos métricas propuestas en la literatura para metadatos. Como principal resultado se implementó la herramienta CompMARC que utiliza ambas métricas y determina el grado de completitud de estos registros a partir de los umbrales propuestos en este trabajo.

Palabras claves: calidad de datos, completitud, MARC 21, métrica, registros bibliográficos

ABSTRACT

MARC 21 is one of the most used for cataloging bibliographic records standards. According to the results of the processing of surveys of specialists in Information Sciences from the Central University "Marta Abreu" of Las Villas, one of the leading data quality problems that present bibliographic records in this format is the incompleteness of data. Therefore, this study aims to measure the completeness data quality dimension of bibliographic records in MARC 21 format. In the process of measuring the completeness two metrics proposed in the literature for metadata were used. As the main result CompMARC tool that uses both metrics and determines the degree of completeness of these records from the thresholds proposed in this paper was implemented.

Keywords: data quality, completeness, MARC 21, metric, bibliographic records.

1. INTRODUCCIÓN

En la actualidad la catalogación de registros bibliográficos es una tarea fundamental que sirve de soporte a diversos procesos bibliotecarios. Con la informatización y la automatización de las bibliotecas han surgido varios formatos y estándares para la catalogación de estos registros (Garrido Arilla, 1996), uno de ellos es el formato MARC 21 (acrónimo de *Machine Readable Cataloging*) (Moreno & Brascher, 2007).

El formato MARC 21 es un modelo de metadatos y constituye una norma utilizada para la representación e intercambio de datos bibliográficos, de autoridad, de existencias, de clasificación y de información de interés para la comunidad. Un subconjunto del formato MARC 21 completo[1] lo constituye el formato MARC 21 para datos bibliográficos. Este último formato se utiliza para almacenar información bibliográfica, materiales textuales impresos y manuscritos, archivos de computador, mapas, música, recursos continuos, materiales visuales y materiales mixtos.

Los registros bibliográficos en formato MARC 21 contienen los elementos de datos esenciales que se necesitan para crear descripciones bibliográficas de información de los ítems. Estos registros deben incluir campos necesarios de acuerdo a su tipo para que presenten una completitud mínima y "proporcionen información suficiente para identificar un elemento bibliográfico y generar una descripción bibliográfica básica"[2].

Según encuestas aplicadas a 16 especialistas en Ciencias de la Información de la Universidad Central "Marta Abreu" de Las Villas (UCLV), uno de los principales problemas de calidad de datos que se presentan en la catalogación de registros bibliográficos es la incompletitud (Abreu-Álvarez, 2015). Lo anterior constituye la motivación fundamental que justifica el principal objetivo del presente trabajo: medir la dimensión de calidad de datos completitud de registros bibliográficos en formato MARC 21.

2. METODOS

Según Furrie (2003), la estructura del formato MARC 21 está formada por tres componentes principales: cabecera, directorios y campos variables (ver Figura 1).

[1]Formato Bibliográfico MARC 21 LITE (Oficina de Desarrollo de Redes y Normas MARC) http://www.loc.gov/marc/bibliographic/litespa/elbdspa.html
[2] Appendix C - Minimal Level Record Examples. http://www.loc.gov/marc/bibliographic/bdapndxc.html.

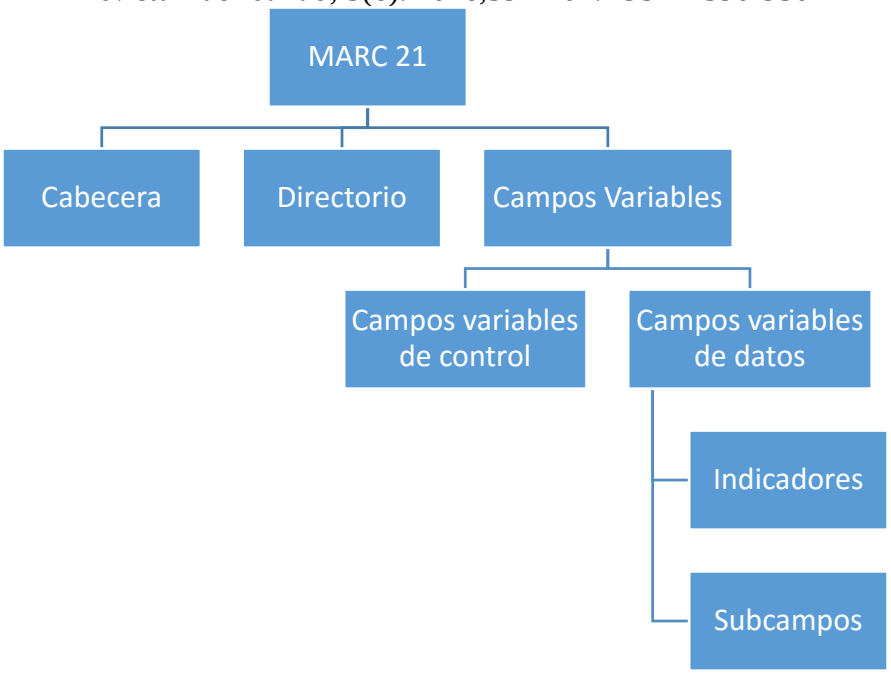

Figura 1. Componentes de un registro MARC 21 (Fuente: Elaboración propia).

La dimensión de calidad de datos completitud se encuentra entre las dimensiones agrupadas en la categoría contextual de acuerdo al marco de trabajo propuesto por Wang and Strong (1996). Existen varias definiciones de completitud de acuerdo al contexto en el cual es aplicada dicha dimensión (Liu & Chi, 2002; Pipino, Lee, & Wang, 2002; Sivogolovko, 2011; Wand & Wang, 1996). En el contexto de las bases de datos relacionales Batini, Cappiello, Francalanci, and Maurino (2009) definen la completitud como "el grado en que una colección de datos dada incluye datos que describen el conjunto correspondiente de objetos del mundo real".

En el contexto de los metadatos los autores Ochoa y Duval (2006a, 2006b) definen la completitud como "el grado en el cual un registro de metadatos almacena toda la información necesaria para tener una representación global del objeto descrito".

Debido a que el formato MARC 21 es un modelo de metadatos, en el presente trabajo se utiliza la definición de completitud en el contexto de los metadatos dada por Ochoa y Duval (2006a, 2006b).

Además se tienen en cuenta las siguientes consideraciones:

- Un campo variable de un registro en formato MARC 21 está completo si:
  - su valor no es la cadena vacía, para el caso de los campos variables de control.

- si presenta los subcampos mínimos, para el caso de los de los campos variables de datos necesarios para que el registro presente una completitud mínima.
- si al menos está presente el subcampo a, para el resto de los campos variables de datos.

- Ante la presencia de múltiples instancias se asume que un campo es completo si al menos una de sus instancias es completa (Ochoa & Duval, 2006a, 2006b).

En el proceso de medición se utilizaron dos métricas propuestas por Ochoa y Duval (2006a, 2006b) para medir completitud en metadatos, las cuales se pueden extender al formato MARC 21. La primera de ellas se muestra en la ecuación 1.

$$C = \frac{\sum_{i=1}^{N} P(i)}{N} \qquad (1)$$

Donde *N* representa el número total de campos del formato, $P(i)$ toma valor 1 si el i-ésimo campo está completo y 0 en otro caso. En lo adelante esta ecuación se denomina métrica 1.

A la métrica 1 Ochoa y Duval (2006a, 2006b) le introducen un factor de peso para cuando todos los campos no tengan la misma importancia. Esta modificación se muestra en la ecuación 2 y constituye la segunda métrica propuesta por ambos autores.

$$C = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i} \qquad (2)$$

Donde α_i representa el grado de importancia o peso del campo i-ésimo y tanto P(i) como *N* significan lo mismo que en la métrica 1. En lo adelante la ecuación anterior se denomina métrica 2.

En ambas métricas se garantiza que el máximo valor que puede tomar la métrica es 1 (cuando todos los campos contienen información) y el valor mínimo es 0 (cuando ningún campo contiene información). Además cuando existe más de una instancia de algún campo, este se considera completo si al menos una de sus instancias es completa (Ochoa & Duval, 2006a, 2006b).

Revista Publicando, 3(6). 2016,397-407. ISSN 1390-930

En el desarrollo del presente trabajo se utilizan las métricas 1 y 2 para el desarrollo de una aplicación para la medición de la dimensión de calidad de datos completitud en bases de datos con formato MARC 21.

3. RESULTADOS

La herramienta que se obtiene, nombrada como CompMARC permite la medición de la completitud en bases de datos en formato MARC 21 utilizando las métricas 1 y 2. La aplicación CompMARC mide la completitud de los siguientes tipos de registros bibliográficos:

- Libro: Código a en la posición seis de la cabecera.
- Archivo de computadora: Código m en la posición seis de la cabecera.
- Música impresa con notación: Código c en la posición seis de la cabecera.
- Material cartográfico: Código e en la posición seis de la cabecera.
- Materiales mixtos: Código p en la posición seis de la cabecera.

En el caso de la métrica 1 la aplicación brinda la posibilidad de definir un umbral y tres opciones para definir la cantidad de campos (N):

- El valor 999 que representa el total de etiquetas del formato MARC 21.
- Calcular el total de campos a partir de las diferentes etiquetas utilizadas en la base de datos de entrada.
- Un valor introducido por el especialista.

En el caso de la métrica 2 no es necesario pasar ningún parámetro. Esto es debido a la propia definición de la métrica y su implementación en la aplicación. Los pesos de cada campo y el umbral se definen a partir de los campos necesarios que necesita el registro para que presente completitud mínima.

Para determinar los pesos de cada campo necesario para que el registro presente una completitud mínima, se debe tener en cuenta las siguientes consideraciones:

- A cada uno de ellos se le asigna el mismo peso.
- El peso de cada uno de estos campos analizados de manera independiente, tiene que ser mayor que la suma de los pesos de los otros campos (los que no se consideran necesarios para que el registro presente una completitud mínima).

Además, se cumple que la sumatoria de los pesos de todos los campos siempre es uno.

Lo anteriormente expuesto permite determinar si un registro presenta completitud mínima o no a partir de los pesos de sus campos.

Asimismo se puede considerar como umbral la suma de los pesos de los campos necesarios. Si un registro contiene un campo necesario incompleto o que esté ausente aunque contenga el resto de los campos, no sobrepasa el valor definido como umbral.

Por ejemplo un registro bibliográfico de tipo archivo de computadora debe tener los campos 001, 003, 005, 008, 040, 245, 256, 260, 300 y 538 para que presente completitud mínima (ver Figura 2). Un peso que puede utilizarse para cada uno de los campos anteriores es 0.091. La suma de todos estos pesos es 0.91, quedando solo 0.09 para los restantes campos no necesarios (0.091>0.09). Se considera como peor caso que a un registro de este tipo le falte un campo necesario y presente el resto de los campos. Aún así la completitud para este peor caso no sobrepasa el valor 0.91, por lo cual se considera este valor como un umbral factible para este tipo de registros.

Leader/00-23	*****nmm##22*****7a#4500
001	<control number>
003	<control number identifier>
005	19930729093320.6
008/00-39	930729s1989####azu####|###b||######eng#d
040	##$a[organization code]$c[organization code]
245	00$aPC nations$h[electronic resource] :$bflags and national anthems of 175 countries.
256	##$aComputer program.
260	##$aTempe, Ariz. :$bPC Globe,$cc1989.
300	##$a2 computer disks :$bcol. ;$c5 1/4 in.
538	##$aSystem requirements: IBM/Tandy; EGA or VGA monitor only.

Figura 2. Campos necesarios para que un archivo de computadora presente completitud mínima.

En la Tabla 1 se muestran los umbrales definidos en la herramienta CompMARC para la métrica 2.

Tabla 1. Pesos y umbrales utilizados en la herramienta CompMARC para la métrica 2.

Tipo	Campos necesarios	Umbral	Peso de cada campo necesario	Suma de los pesos del resto de los campos
Libro	9	0.91	0.1011111111111111	0.09
Archivo de computadora	10	0.91	0.091	0.09
Música impresa con notación	12	0.93	0.0775	0.07
Material cartográfico	15	0.94	0.0626666666666667	0.06
Materiales mixtos	9	0.91	0.1011111111111111	0.09

Para el proceso de medición de la completitud se utilizan dos bases de datos con formato MARC 21, las cuales sirven de entrada a la herramienta CompMARC. La primera es una base de datos de la Universidad de Cambridge3 y la segunda es una base de datos pública de la Universidad de Michigan4 que se encuentra bajo la licencia Creative Commons CC0. En lo adelante estas base de datos se denominan como BD_CAMB y BD_UMICH. La base de datos BD_CAMB contiene 1 350 737 de registros bibliográficos, de los cuales 1 341 717 tiene un código permitido por la aplicación CompMARC en la posición seis de la cabecera. En tanto, la base de datos BD_UMICH presenta 1 327 753 de registros bibliográficos, de los cuales 1 291 200 son válidos para la herramienta CompMARC.

La Tabla 2 muestra los resultados obtenidos de la medición de la dimensión de calidad de datos completitud en las bases de datos BD_CAMB y BD_UMICH con la herramienta CompMARC. En el caso de la métrica 1 se utiliza como umbral 0.03. Además el resultado de calcular el total de campos utilizados da como resultado 194 campos en BD_CAMB y 235 en BD_UMICH. Lo anterior significa que un registro bibliográfico debe contener 30 campos completos cuando se utiliza como total 999 y seis y ocho campos completos para las bases de datos BD_CAMB y BD_UMICH respectivamente cuando el total de campos es calculado.

Tabla 2. Resultados de la medición de la dimensión de calidad de datos completitud

	BD_CAMB		BD_UMICH	
	> umbral	< umbral	> umbral	< umbral

[3] http://data.lib.cam.ac.uk/data/cambridge.mrc.gz
[4] http://www.lib.umich.edu/files/open-access-marc/umich_created_20140827.marc.gz

Sin Pesos	999	**25**	1 350 712	**33**	1 324 720
	Calculado	1 084 501	**266 236**	1 278 346	**46 407**
Con Pesos		**0**	1 341 717	**108 083**	1 183 117

La tabla anterior muestra como resultados significativos que solo 25 (0.002 %) y 33 (0.003 %) registros bibliográficos pertenecientes a las bases de datos BD_CAMB y BD_UMICH respectivamente contienen 30 o más campos completos. Además 266 236 (19,8 %) registros de la base de datos BD_CAMB presentan menos de seis campos completos. En el caso de BD_UMICH 46 407 (3,57 %) registros tienen menos de ocho campos completos. Por último, en el caso de la métrica 2 ningún registro de BD_CAMB contiene todos los campos necesarios para cada tipo de registro y solo 108 083 (8,37 %) registros de BD_UMICH los contiene.

Se exponen los resultados obtenidos en la investigación.

4. CONCLUSIONES

En el presente trabajo se midió la completitud de los registros bibliográficos en formato MARC 21 presentes en las bases de datos BD_CAMB y BD_UMICH a través de la herramienta CompMARC. En ambos casos se utilizaron dos métricas para metadatos tomadas de la literatura y se determinó el grado de completitud de estos registros a partir de los umbrales propuestos en este trabajo, lo cual permitió corroborar que la incompletitud de los registros bibliográficos constituye un problema latente que incide directamente en la calidad de los datos. Además, se establecieron las consideraciones a tener en cuenta a la hora de otorgar los pesos de cada campo necesario para determinar la completitud mínima de un registro. En trabajos futuros se debe continuar trabajando en el mejoramiento de la herramienta CompMARC y de los umbrales para ambas métricas, extender la medición hacia otras dimensiones de calidad de datos y aplicar técnicas de limpieza de datos para completar los valores ausentes e incompletos.

5. REFERENCIAS BIBLIOGRÁFICAS

Abreu-Álvarez, Y. (2015). *Análisis de la calidad de datos en fuentes de la suite ABCD.* (Tesis de Grado), Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Cuba.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR), 41*(3), 16.

Furrie, B. (2003). Conociendo MARC Bibliográfico: Catalogación Legible por Máquina. Retrieved 1 de Septiembre, 2015, from http://www.loc.gov/marc/umbspa/umbspa.html

Garrido Arilla, M. R. (1996). Tendencias que presenta la catalogación automatizada hoy. *Revista general de información y documentación, 6*(2), 51.

Liu, L., & Chi, L. (2002). *Evolutionary data quality.* Paper presented at the Proceedings of the 7th international conference on information quality (IQ).

Moreno, F. P., & Brascher, M. (2007). MARC, MARCXML e FRBR: relações encontradas na literatura. *Informação & Sociedade: Estudos, 17*(3).

Ochoa, X., & Duval, E. (2006a). *Quality Metrics for learning object Metadata.* Paper presented at the Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006.

Ochoa, X., & Duval, E. (2006b). Towards Automatic Evaluation of Metadata Quality in Digital Repositories. *Lecture Notes in Computer Science, 4231*, 372-381.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218.

Sivogolovko, E. (2011). Evaluation of impact of data quality on clustering with syntactic cluster validity methods: Technical report, Christian-Albrechts University.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5-33.