

I Encuentro Nacional de Catalogadores

PONENCIA

Software de conversión de bases de datos CDS/Isis catalogadas en formato BIBUN a MARC21”

Por Marcos G. Lopez, Asesor Informático de la Fac. de Ciencia y Tecnología de la Univ. Autónoma de Entre Ríos (marcosg.lopez@gmail.com)

Sandra Bib. Emilce Osuna, Bibliotecaria a cargo de Procesos Técnicos de la Fac. de Ciencias Económicas de la Univ. Nac. de Entre Ríos (emilce_osuna@hotmail.com)

Resumen:

Software Open Source liberado bajo licencia MIT que permite migrar registros CDS/Isis catalogados en formato BIBUN a Marc21, contemplando asientos monográficos (tesis; libros; CD; DVD; VIDEOS, etc.)

Se trata de un esfuerzo inter-universitario entre la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos y la Facultad de Ciencias Económicas de la Universidad Nacional de Entre Ríos.

El eje de este desarrollo ha sido la combinación de conocimientos de dos disciplinas: las bibliotecológicas con aportes específicos sobre formatos y normas de catalogación y las informáticas aplicadas a sistemas de bibliotecas.

El producto es un conjunto de programas (scripts) que permiten primeramente realizar un control de calidad e integridad de los datos, informando errores y advertencias previas al proceso final de conversión.

La programación está realizada en lenguaje Wxis, destacándose la claridad para comprender la codificación y la posibilidad de ajustar la misma a políticas de procesamiento particulares de cada biblioteca. El mayor esfuerzo para entender el trabajo radica en la capacidad de interpretar el “lenguaje de formateo CISIS”.

El objetivo de presentar públicamente este desarrollo, es beneficiar a otras bibliotecas que se encuentran en los mismos procesos de migración de sus catálogos.

El código y el avance del proyecto estará dispuesto gratuitamente en el repositorio público

<http://code.google.com/p/bibun2marc>

Generalidades

En esta ponencia se presenta la experiencia en la conversión de las bases de datos bibliográficas de la Fac. de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos y la Fac. de Ciencias Económicas de la Universidad Nacional de Entre Ríos, quienes han señalado el Encuentro como el espacio oportuno para ofrecer gratuitamente el trabajo realizado.

La presentación describe brevemente los procesos informáticos llevados a cabo para convertir bases CDS/ISIS catalogadas en formato BIBUN llevándolas al formato MARC21.

Objetivo

La propuesta del trabajo consiste en socializar un desarrollo que aspira a satisfacer no solo las necesidades inmediatas de las instituciones involucradas sino además permitir que otras también se vean beneficiadas. A través de la licencia MIT se brinda gratuitamente el código fuente del software empleado para la conversión, el cual se dispondrá libremente en el repositorio público <http://code.google.com/p/bibun2marc>. El Encuentro es el evento elegido para dar a conocer este trabajo a partir del cual el código quedará disponible.

Características

Dado que las bases de origen y destino (BIBUN y MARC21) trabajan con estructura CDS/ISIS, se utilizó para la programación lógica de la conversión el software wwwisis (<http://bvsmodelo.bvsalud.org/php/level.php?lang=es&component=28&item=2>), el cual permitió definir estructuras independientes para el tratamiento de los grupos (centenas) de campos Marc21. Mediante la codificación de programas se fueron tratando cada centena, (v1XX.xis; v2XX.xis, etc.) las cuales se integraron a un programa o script principal (bibun2marc.xis).

Tipo de Material:

Inicialmente el programa cubre la conversión de asientos monográficos tales como tesis; libros; CD; DVD; VIDEOS, etc. aunque también se prevé la incorporación de análisis de texto entre otros.

Proceso de conversión

La conversión a Marc21 se inició como un ejercicio práctico abordando directamente la transformación de los campos, sin considerar el estado ni la calidad de carga de la base BIBUN. Este trabajo se caracterizó por su metodología empírica, la cual evidenció aspectos que se tuvieron que re-considerar a medida que se presentaban los inconvenientes, los que fueron planteados y sorteados tratando siempre de lograr un trabajo coherente entre tiempo/calidad.

La revisión de la integridad, consistencia y acato a su formato de la base BIBUN se plantea como un ejercicio posterior y previo a la conversión; asumiendo inicialmente que la misma se encuentra en un estado hipotéticamente ideal, a fin de abordar directamente la práctica tomando variados casos a fin de testear y comprobar el procedimiento de conversión.

Puntos de acceso:

El primer grupo que se convirtió fue el 1XX, analizando inicialmente los campos existentes en el registro BIBUN, es decir, que si por ej. se encontraba un autor personal (v28) se asumía la creación de un v100 en Marc21. Esto fue inmediatamente desestimado, aunque la mayoría de las obras catalogadas se ajusten a este criterio hubo que considerar otros casos donde por ej. la obra se identificaba por su título a pesar de tener indicado un autor personal, o si la misma surgía de una jornada/reunión y se encontraban autores corporativos y personales. Se replanteó el análisis considerando más preciso el campo v9 de BIBUN (Índice de procesamiento), el cual define el punto de acceso principal de la obra. Se tuvieron en cuenta los siguientes valores:

PE: Autor personal

IN: Autor institucional

TI: Título

RE: Reunión/Jornada/Congreso, etc.

Dados estos valores se consideraba cuál de los campos Marc21 de la centena 1XX/2XX se debían crear.

Aquí quedo apuntado como tarea posterior la necesidad de revisar y controlar el estado del campo v9 de BIBUN, ya que el mismo ahora tiene un papel preponderante al momento de definir el acceso principal de la obra.

Caracteres a ignorar en la indización:

Otro de los procesos postergados que se dejaron para la etapa final, fue sobre los indicadores que especifican los caracteres a ignorar en la indización de los campos. En aquellos casos se optó simplemente por indicar con un signo “?” el indicador que luego debería convertirse para especificar la cantidad de caracteres obviados.

Se conformó una base de datos (car_ignora) con un solo campo para especificar los términos que se deben ignorar. El procedimiento de transformación del indicador “?”, analizaba entonces la primer palabra del campo y de encontrarse en la base de datos (car_ignora) convertía el mismo por su longitud (más una posición correspondiente al espacio). Este proceso se realiza en un punto previo la grabación del registro realizándose de una sola vez para todos los campos.

Lógica de la conversión:

Básicamente el proceso de conversión consiste en estructurar un campo global v9999 el cual tendrá tantas líneas (u ocurrencias) como campos Marc21 surjan del registro BIBUN, cada una de las cuales está formateada de acuerdo al comando PROC (lenguaje formateo Isis) que posteriormente permitirá crear los diferentes campos que conformarán el registro convertido en la base MARC21.

Ejemplo del campo v9999 con formato PROC:

```
a100~1#^aChilkowski, Oscar Daniel~  
a245~0?^aContabilidad general : ^bconceptualización teórica elemental^c / Oscar  
Daniel Chilkowski~
```

Este caso crearía dos campos, v100 y v245. Este último previamente debe convertir el segundo indicador “?” por el valor 0.

Definiciones predefinidas:

Se tomaron algunas consideraciones en forma de facto, a fin de simplificar y agilizar el proceso de conversión, buscando el mejor coeficiente tiempo/calidad. Estos casos , entre otros que se encuentran documentado en el programa fuente, fueron:

- *Nombre personal en forma invertida*
El primer indicador del campo v100 y v700 se asume siempre en 1, indicando de esta manera que el nombre personal es tratado en orden invertido (Apellido, Nombre)
- *Nombre corporativo en forma directa*
El segundo indicador del campo v110 y v710 se asume siempre en 2, indicando de esta manera que el nombre institucional es ingresado en forma directa
- *Título uniforme*
El primer indicador del campo v240 se asume siempre en 1, indicando de esta manera que se visualiza el título uniforme
- *Mención de responsabilidad*
Se omite la posibilidad que exista más de tres ocurrencias del campo v29 (autor institucional). Esta situación se apunta para verificar esta condición antes del proceso de conversión
- *Títulos uniforme (acceso principal), paralelos, alternativos o variantes del título*
No se considera en la migración el tratamiento de títulos uniformes, paralelos, alternativos o variantes del mismo, esta posibilidad debe analizarse posteriormente a la conversión para revisar y atender estos casos.

- *Acceso principal por Titulo*
Si la obra tiene acceso principal el título y además tiene definido autores corporativos y personales, en la mención de responsabilidad solo se consideran los primeros
- *Autores colaboradores*
Se omite la puntuación (y función) para los otros autores que colaboran en la obra.
- *Notas*
Las notas no se discriminan y se consideran todas como “Notas Generales”, enviándose todas al campo v500
- *Series*
Si la obra pertenece a una Publicación Seriada, el título de la misma se considera como un punto de acceso, creándose el campo v440
- *Términos temáticos*
Las palabras claves no se discriminan y se migran al campo v653 como “Término clave sin controlar”
- *Cabecera (leader):*
Solo se definen los siguientes valores/posiciones:
 - Estado del registro (Pos. 5): Se asume en “n” (Nuevo registro)
 - Tipo de registro (Pos. 6):
Evalúa el campo “soporte del documento” (v8) y considera CD y DVD para los materiales digitales, asumiendo que estos soportes NO SON MUSICALES (Pos. 6=m)
Las cintas (v8=CINTMAG) también se asumen como NO MUSICALES, es decir grabaciones habladas (Pos. 6=i)
En el resto de los casos se asume una monografía (Pos. 6=a)
 - Nivel bibliográfico (Pos. 7):
Analiza si es una colección o en su defecto asume una monografía “m”
 - Tipo de control (Pos. 8): Se asume en “#” (No especificado)
 - Código del esquema de caracteres (Pos. 9): Se asume en “#” (MARC-8)
 - Nivel de codificación (Pos. 17): Se asume en “1” (Nivel completo, sin examinar el material)
 - Forma de la catalogación descriptiva (Pos. 18): Se asume en “a” (RCAA2)
 - Registro ligado o relacionado (Pos. 19): Se asume en “#” (No requiere)
- *Campo 008:*
Se tienen las siguientes consideraciones:
 - La fecha de creación (Pos. 00-05) se toma del campo fecha de alta (v3^a)
 - Se asume fecha múltiple (Pos. 06) si la fecha de edición contiene guiones y su longitud es mayor a 4 posiciones (se omiten corchetes, “?”, y “c”)
 - La/s fecha/s de publicación se toman del campo edición (v45)
 - El lugar de publicación se obtiene de la primer ocurrencia del campo país de edición (v48), convirtiendo su valor según la definición de la base auxiliar “cod_lugar”, la cual define la correlación entre los códigos definidos entre BIBUN y MARC21
 - El tipo de ilustración se limita a reflejar el valor “a####” sin especificar las variantes
 - La audiencia no se especifica “#”
 - La forma del ítem solo considera “electrónico” o en su defecto “material impreso”
 - El contenido solo considera los valores:
 - 'BIBLIOG'

- 'CATALOGO'
 - 'VOCABULA'
 - 'ENCICLOP'
 - 'MANUAL'
 - 'NORMALEG'
 - 'NORMATEC'
 - 'TESIS'
- Publicación gubernamental (Pos. 28) se considera solo si el tipo de documento (v7) contiene “NORMALEG” (no codificada “[”)
 - Publicación de conferencia (Pos. 29), se considera solo si el índice de procesamiento (v9) contiene “RE” (no codificada “[”)
 - No se consideran los homenajes (Pos. 30)
 - No se consideran los índices (Pos. 31)
 - No se consideran las obras literarias (Pos. 33)
 - No se consideran las biografías (Pos. 34)
 - El código de idioma se obtiene de la primera ocurrencia del campo idioma (v50), convirtiendo su valor según la definición de la base auxiliar “cod_idoma”, la cual define la correlación entre los códigos definidos entre BIBUN y MARC21
 - Las posiciones 38 (Registro modificado) y 39 (Fuente de catalogación) no se codifican “#”

Procesos pre-conversión

Una vez logrado el programa para convertir los registros se tuvo que considerar la revisión de aquellas situaciones que escapaban a las consideraciones de facto establecidas, además de analizar ciertas inconsistencia en la carga.

Así mismo y debido a que el formato BIBUN contempla registros relacionados madre-hijos, omitiendo en estos últimos la información que está previamente descrita, se dispara un procedimiento que consolida estos datos, completando aquellos datos que fueron obviados, de manera que el proceso de conversión asuma que el registro se encuentra con toda la información descriptiva.

Consolidación de datos:

Este programa genera una base de datos (bibun_preconversion) que reúne toda la información diseminada que un registro pueda tener con otros vinculados. En esta etapa se analiza la relación madre-hijo de los registros, informando e interrumpiendo la consolidación al encontrar un error.

IMPORTANTE: El procedimiento de consolidación utilizado se ajusta a una variante del formato BIBUN que se implementó para relacionar los registros vinculados, definiendo un nuevo caso para el campo v5: NIVEL DE DESCRIPCIÓN BIBLIOGRÁFICA, el nivel MadreHijo. Quedando entonces tres casos posibles:

- Registro Madre (v5='m')
- Se utiliza cuando el documento es una obra completa en sí misma editada en una o varias partes físicas. Ej.: informe, trabajo monográfico, curso, en uno o más volúmenes, obra en línea a texto completo, etc.

- Registro MatreHijo (v5='mx#####')
Se utiliza cuando el documento es el primer ejemplar de cada parte física (volumen o tomo) para aquellas ediciones que están divididas en varias partes. En este caso, al igual que en el siguiente, se describen sólo aquellos datos que difieren del registro madre, cuyo número de registro (MFN) está indicado y representado en el ejemplo por #####
- Registro Hijo (v5='x#####')
Este código, que no responde al nivel descripción bibliográfica, es utilizado para establecer la relación entre el registro hijo y el registro madre. Se utiliza en los registros hijos. Está constituido por una x y el número de acceso del registro madre.

Si bien esta adaptación es particular, el programa está pensado para que se le realicen las correcciones necesarias a fin de ajustar el criterio propio de catalogación adoptado por una institución alineándolo con el empleado por el proceso de migración, a fin de que los registros obtenidos puedan interpretarse debidamente por el programa de conversión.

La base creada finalmente contiene solo registros de nivel de descripción bibliográfica (campo v5 de BIBUN) que corresponde a "m" (monográfico) compaginando la información de los registros vinculados.

Control de datos

Luego de obtener la base consolidada de registros se diseñó un programa para controlar los datos obtenidos. El informe que se emite detalla dos tipos de mensajes: Error y Advertencia, interpretando que sólo los primeros son condicionales y deben atenderse para el proceso final de conversión. Se contemplaron las siguientes situaciones:

- Error E2, inconsistencia de v9. Los valores encontrados no corresponden a ninguno de los siguientes: PE, IN, TI, RE; o no existe coherencia entre el valor de v9 (índice de procesamiento) y el punto de acceso principal, es decir por ej. si v9='PE' debe existir al menos un v28.
- Error E3, código de país no definido
- Error E4, código de idioma no definido
- Error E5, fecha fuera de rango o no definida
- Error E6, no se ha indicado una fecha de edición
- Advertencia A2, no se puede representar debidamente al editor/lugar
- Advertencia A3, no se pudo validar la fecha de edición
- Advertencia A4, no se encontró nro. de inventario
- Advertencia A5, no se encontró la clasificación/librística
- Advertencia A6, la secuencia de ejemplares es incorrecta

Resultado final

Una vez controlados los datos el proceso final de conversión genera una base compatible con el formato MARC21, en la que se pueden lograr los siguientes campos:

Cabecera:

La cabecera está diseminada entre los campos v9XX, donde XX representa la posición del dato dentro de la misma

Campos:

005, Fecha de Última Modificación
008, Información general
020, ISBN
040, Fuente de Catalogación
100, Nombre Personal
110, Nombre Corporativo
111, Nombre de Reunión
240, Título Uniforme
245, Título Propiamente Dicho
250, Mención de Edición
260, Publicación
300, Descripción Física
440, Mención de Serie
500, Nota General
504, Nota de Bibliografía
653, Término Indizado No Controlado
700, Entrada Secundaria, Nombre Personal
710, Entrada Secundaria, Nombre Corporativo
800, Asiento Secundario de Serie, Nombre Personal
810, Asiento Secundario de Serie, Nombre Corporativo