Preprocessing von E-Media Metadaten

Aufbereiten der 360 MARC von ProQuest/SerialSolutions des IDS Basel Bern für den Import in Swissbib/CBS

Stand: Oktober 2016 / Andres von Arx

Inhalt

Datenquellen

E-Medien werden von der Fachstelle E-Media (UB Basel) bzw. E-Library (UB Bern) im ERM von Serial Solutions verwaltet. Siehe: http://www.swissbib.org/wiki/index.php?title=E-Book-Import

Instanzen im ProQuest Intota ERM:

Betreut von E-Media Basel	
BS	UB Basel
FREE	Open Access und Freie Ressourcen
Betreut von E-Library Bern	
BE	UB Bern
BBZ	Berner Bildungszentrum Pflege
EHB	Eidgenössisches Hochschulinstitut für Berufsbildung, Bern

ProQuest stellt zweiwöchentlich Zip-Archive mit dem Gesamtbestand der Metadaten im Format ISO 2790 zum Download bereit, ebenso Zip-Archive mit dem Delta gegenüber der letzten Version (CSV-Dateien für neue, geänderte und gelöschte Records).

Preprocessing

Die Metadaten werden auf Seiten IDSBB von der UB Basel konvertiert, normalisert, gemerged und für Swissbib bereitgestellt.

Programmaufruf

Wann?	Zweiwöchentlich, nachdem alle Mails von ProQuest eingetroffen sind, dass die 360MARC Daten bereit stehen. ACHTUNG: Das Skript sollte nicht über einen Datumswechsel (d.h. über Mitternacht hinaus) laufen.
Wie?	Zur Zeit manuell TO DO: Automatisierung (Cronjob), Logging mit Email (vgl. ub-edoc)
Was?	Auf ub-catmandu: `$ /opt/scripts/e-books/bin/make-idsbb-emedia.sh`
Laufzeit	Ca. 30 Minuten

Output

Für die Weiterbearbeitung in Swissbib werden zwei Dateien bereitgestellt:

sersol-idsbb-emedia-updates.xml.gz

Inhalt	Enthält alle neuen und geänderten Metadaten
Format	MARC21 XML Datei im Format für den Import ins CBS
Upload	harvester@sb-coa1.swissbib.unibas.ch:/swissbib/harvesting/incomingSersol

sersol-idsbb-emedia-deletions.txt

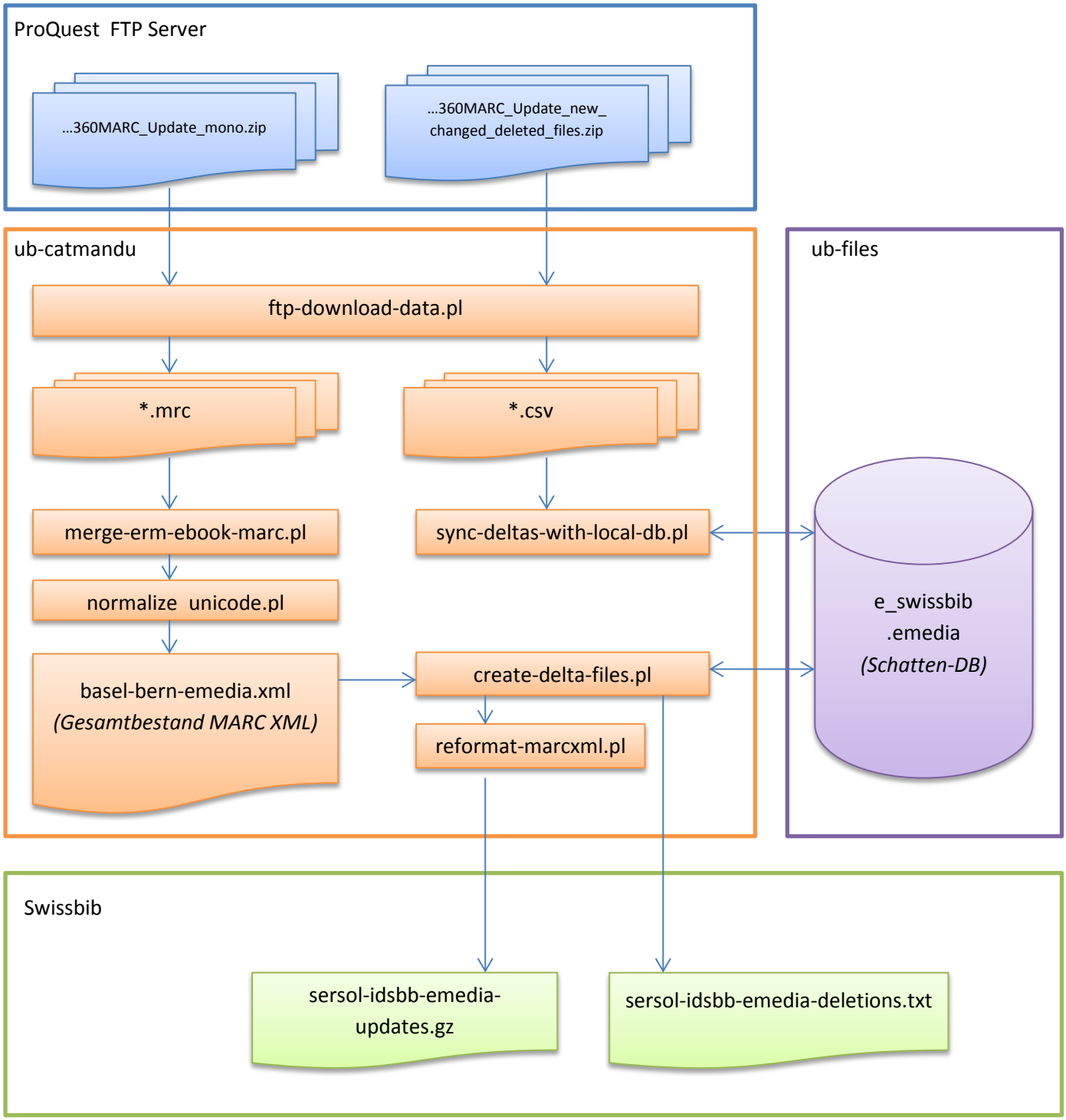
Inhalt	Serial-Solutions IDs (Feld 001) der Metadaten, die in allen Instanzen des ERM gelöscht sind.
Format	Unix Textdatei. Die Zeile 1 enthält einen Datums-Stempel.
Upload	harvester@sb-coa1.swissbib.unibas.ch:/swissbib/harvesting/oaiDeletes

Log von make-idsbb-emedia.sh

```
------------------------------------------------
Download and preprocess 360MARC data for swissbib
------------------------------------------------
START: Mit Jun 29 09:30:05 CEST 2016
* download and extract data
ftp: contact ftp.serialssolutions.com
ftp: downloading 1UB_360MARC_Update_mono.zip...
ftp: downloading BB3_360MARC_Update_mono.zip...
ftp: downloading EH1_360MARC_Update_mono.zip...
ftp: downloading BB3_360MARC_Update_new_changed_deleted_files.zip...
ftp: downloading EH1_360MARC_Update_new_changed_deleted_files.zip...
ftp: downloading 1UB_360MARC_Update_new_changed_deleted_files.zip...
ftp: downloading 3UB_360MARC_Update_mono.zip...
ftp: downloading 2UB_360MARC_Update_mono.zip...
ftp: downloading 2UB_360MARC_Update_new_changed_deleted_files.zip...
ftp: downloading 3UB_360MARC_Update_mono_new_changed_deleted_files.zip...
ftp: download complete
ftp: extracting monographs full
ftp: extracting monographs delta
* merge all ERM instances
* [please be patient for about 30 minutes...]
merge: storing data for BE in memory
merge: storing data for FREE in memory
merge: storing data for EHB in memory
merge: storing data for BBZ in memory
merge: merging data into BS
merge: merging data into BE
merge: merging data into FREE
merge: merging data into EHB
merge: merging data into BBZ
* synchronizing MySQL database
delta: reading BS_delta/NewMonographRecords.csv
delta: reading BS_delta/ChangeMonographRecords.csv
delta: reading BS_delta/DeleteMonographRecords.csv
delta: reading BE_delta/NewMonographRecords.csv
delta: reading BE_delta/ChangeMonographRecords.csv
delta: reading BE_delta/DeleteMonographRecords.csv
delta: reading EHB_delta/ChangeMonographRecords.csv
* writing delta files
* [please be patient for about 15 minutes...]
delta: writing deletion list
delta: writing updates xml
delta: done
* reformatting updates xml
* gzipping xml
* writing stats
END: Mit Jun 29 10:00:06 CEST 2016
```

Datenfluss Preprocessing

(vereinfacht, ohne Zwischenschritte)



Wenn in den *.csv Dateien neue oder geänderte Dateien gemeldet wurden, wird die entsprechende MARC-XML Datei aus dem Gesamtbestand extrahiert und in den Mutationsmeldungen geliefert.

Wenn in den *.csv Dateien Löschungen gemeldet wurden, wird in der Schatten-DB geprüft, ob alle Holdings gelöscht sind. Wenn ja, wird die ID des Records in der Löschtabelle eingetragen und der Datensatz aus der Schatten-Datenbank entfernt. Wenn nein, wird ebenfalls eine MARC-XML als Mutationsmeldung geliefert.

Die Schattendatenbank

Um die Hinzufügungen und Löschungen der verschiedenen Instanzen im Griff zu behalten, wird eine MySQL-Tabelle aller IDs in allen Instanzen gepflegt.

Produktion	e_swissbib.emedia @ ub-filesvm
Test	e_swissbib.emedia @ ub-filesqm

Die Datei erhält für jede ID eine Holding-Flag für jede Instanz im ERM plus eine Flag, die beim Export der Datenlieferung gesetzt wird.

Field	Type	Default	Bedeutung
ssid	varchar(16)	NULL	ID im ERM von SerialSolutions
HolBS	tinyint(1)	0	Holdings UB Basel vorhanden?
HolBE	tinyint(1)	0	Holdings UB Bern vorhanden?
HolBBZ	tinyint(1)	0	Holdings BBZ vorhanden?
HolEHB	tinyint(1)	0	Holdings EHB vorhanden?
HolFREE	tinyint(1)	0	Holdings FREE vorhanden?
MARC	tinyint(1)	0	MARC Metadaten geliefert und exportiert?
modified	date	NULL	Datum des letzten Deltas

Daten und Verzeichnisse

Programme und Daten des Preprocessing liegen auf ub-catmandu

/opt/scripts/e-books/bin	Programs + libraries
/opt/scripts/e-books/bin/devel	Helpers + progs for development
/opt/data/e-books/data	Output data and log files
/opt/data/e-books/download	Temporary download files

Einzelne Skripte

ftp-download-data.pl

Das Skript holt die aktuellen 360 MARC-Dateien als Zipfiles per FTP von Serial Solutions per FTP.

Output:

- In /opt/data/e-books/data: je eine *mrc Datei des Gesamtbestands Monografien pro ERM-Instanz sowie die konkatenierten RunSummaries (= Statistiken).
- In /opt/data/e-books/download: je ein Unterverzeichnis mit den Delta-Dateien pro ERM Instanz

Achtung: Das Skript löscht gnadenlos vorhandene *zip und *mrc Dateien und alle Unterverzeichnisse im Ordner /opt/data/e-books/download.

merge-erm-ebook-marc.pl

```
NAME
    merge-erm-ebook-marc.pl - merge holdings and links from 360MARC data
    files

SYNOPSIS
     perl merge-erm-ebook-marc.pl

DESCRIPTION
    Input sind je eine 360MARC Dateien von Serial Solutions fuer die
    verschiedenen Instanzen des Intota-ERM (UB Basel, UB Bern, BBZ Bern, EHB
    Bern, FREE fuer freie Ressourcen [geplant]). Input-Format ist ISO 2709
    (MARC).

    Output ist eine Datei im Format MARC21 XML ohne Dubletten und mit
    gemergten Holdings und URL-Links.

    Bei einigen Feldern wird die Interpunktion am Ende der Unterfelder
    entfernt: Felder: 100 110 111 245 260 264 300 600 610 611 630 650 651
    653 655 700 710 711 Entfernt werden: .,:;/ (inkl. vorausgehende Spatien)
    Nicht entfernt werden z.B.: ?()[]! Punkt nach "Dr.", "Jr.", "Inc." und
    nach Grossbuchstaben

    Im Feld 520 $b werden HTML-Markups und -Entities entfernt.
    Im Feld 949 werden Unterfelder normalisiert

  Laufzeit
    merging der E-Books von BE/BS/BBZ/EHB @ catmandu: 30 Minuten (380'080
    records)

CAVEAT
    In den Inputdateien vom 11.09.2014 wurden dublette Records festgestellt
    (Records mit derselben ID, laut Stichproben inhaltlich identisch). Die
    Fehler wurden Serial Solutions gemeldet. Sie sind mittlerweile offenbar
    behoben.

AUTHOR
    andres.vonarx@unibas.ch

HISTORY
     07.03.2013 Testversion
     13.11.2013 generiere Parkfeld 950 fuer 856/ava
     10.12.2013 multiple Parkfelder, fixe Interpunktion/ava
     17.12.2013 weitere fixes.../ava
     23.09.2014 ignoriere dublette BS records
     16.04.2015 workflow fuer frei zugaengliche E-Book Pakete
     02.09.2015 erweitert auf 3. Datenquelle BBZ Bern, Dateinamen hardcoded.
     22.01.2016 erweitert auf 4. Datenquelle EHB Bern, Dateinamen hardcoded.
     19.05.2016 - rewrite fuer beliebige Anzahl Sets, verhindert dublette
                "ssib.." IDs in Records.
                - Der spezielle Workflow fuer FREE-Pakete wird eliminiert.
                - Feld 898 wird geloescht.
                - Feld 950 wird nicht mehr generiert
                - Feld 949 wird gepatcht
                - Interpunktion fixen zusaetzlich fuer Feld 264
```

normalize_unicode.pl

Dieses Skript verwandelt UTF-8 in einem "decomposed" Format (Grundbuchstabe plus Diakritikum) in die "precomposed" Form von Unicode, d.h. in die "Normalization Form D (NFD), Canonical Decomposition". Fehlerhafte UTF-8-Zeichen werden geloescht.

sync-deltas-with-local-db.pl

```
NAME
    sync-deltas-with-local-db.pl

SYNOPSIS
     perl sync-deltas-with-local-db.pl

DESCRIPTION
    Pro ERM-Instanz wird eine CSV-Datei mit den Mutationen geliefert:
    DeleteMonographRecords.csv ChangeMonographRecords.csv
    NewMonographRecords.csv

    Die Mutationen werden in der lokalen DB e_swissbib.emedia @ ub-filesvm
    nachgetragen. Fuer jede Aenderung wird ein Timestamp gesetzt.

    Zusaetzlich werden die RunSummary.txt Dateien gemerged nach
    */data/Mono-Delta-RunSummary.txt

    Laufzeit: < 1 Minute

AUTHOR
    andres.vonarx@unibas.ch

HISTORY
     08.06.2016 beta / ava
```

create-delta-files.pl

```
NAME
    create-delta-files.pl

SYNOPSIS
     perl create-delta-files.pl

DESCRIPTION
    The program produces a MARC21 XML file with new and updated E-media
    records from all ProQuest ERM, ready for import into Swissbib. It also
    produces a list of E-media IDs for which no holdings exist anymore;
    these records should be deleted from Swissbib.

  Data sources
    basel-bern-emedia.xml
    A merge of the complete 360MARC files of all ProQuest ERM instances.
    Holdings and specific URLs are merged. Converted to MARC21 XML format.

    e_swissbib.emedia @ ub-filesvm
    The local MySQL database e_swissbib contains a table emedia. There we
    keep track of all the messages regarding added, new or deleted files.

  Output
    basel-bern-emedia-updates.xml
    MARC21 XML file for import into Swissbib. Must be reformatted, gzipped
    and uploaded to Swissbib host.

    basel-bern-emedia-deletions.txt
    List of IDSs to be deleted. Note: first line contains date stamp

  Program flow
    Step 1: The program queries the emedia table to find IDs whithout any
    holdings. It adds these IDs to the deletion list and removes the record
    from the table.

    Step 2:The program iterates over basel-bern-emedia.xml. It checks the
    records ID against the emedia table. It will set the MARC flag for the
```

```
[…]
    record and output the record to basel-bern-emedia-updates.xml if one of
    two conditions are met: (1) The *modified* field is set to the current
    date, signaling a recent addition, update or deletion. (2) The MARC flag
    is not set, signaling that the record has been reported as addition in
    an earlier run but the MARC record has not been delivered.

  Run time
    15 minutes (ub-catmandu, 380'000 records total, 2016-06-09)

  CAVEAT
    ProQuest/SerialSolutions produces the 360MARC data and the lists for
    new/changed/deleted records asynchronously. Therefore it is possible the
    records will be reported as new before the 360MARC record has been
    delivered.

    This results in messy statistics. We also have to keep track of whether
    we actually could deliver a delta record to Swissbib or not.

    An example:
    [. . .]

AUTHOR
    andres.vonarx@unibas.ch

HISTORY
     09.06.2016 beta / ava
```

Logfiles

RunSummary-Mono.txt

Konkatenierte RunSummary.txt von SerialSolution für alle Instanzen der MARC Datenlieferungen

RunSummary-Mono-Delta.txt

Konkatenierte RunSummary.txt von SerialSolution für alle Instanzen der Delta Dateilieferungen

statistik.txt

Gesamtstatistik von make-idsbb-emedia.sh

Siehe auch:

http://www.swissbib.org/wiki/index.php?title=Members:OaiDeleteRecordsCompiledSingleIds#Bereitstellung_der_summon_ebooks_f.C3.BCr_den_Import_in_den_CBS_Datenhub