This is an interesting metadata source. Can I import it into Koha?

KohaCon12, Edinburgh, 5-7 June 2012

Marijana Glavica <mglavica@ffzg.hr>
Dobrica Pavlinušić <dpavlin@rot13.org>

acer
RJEČNIK
HRVATS
ILI SRPSKOGA JEZIKA
JUGOSLAVENSKA
ZNANOSTI I UMJETNOSTI
U ZAGREBU 1904
hp scanjet 4070

Material

- 6000 scans of book front pages

- directories organised by person who did the scanning, and **location** of the books

- filenames - **inventory** number (having duplicates)

Task

- add metadata to scanned material
  - some books already catalogued somewhere else
  - not all sources have Z39.59

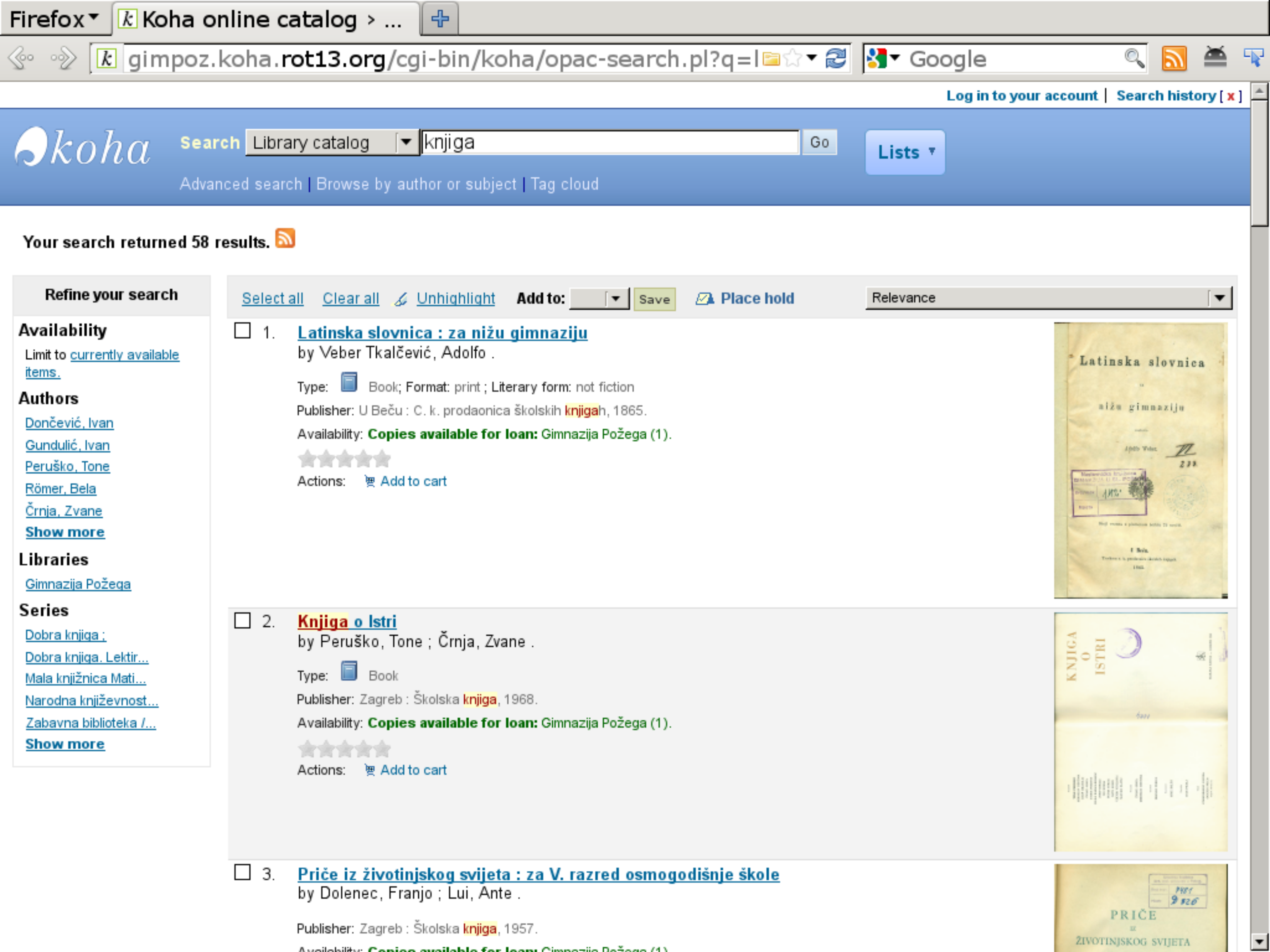
- upload images

- keep track of what is done
  - separate spreadsheet file?

Solution

- Create MARC records from file names - bibliographic and items
  - itemtype - not yet processed

- import MARC records and upload all images in Koha



Search: knjiga

Your search returned 58 results.

Refine your search

Availability

Limit to currently available items.

Authors

- Dončević, Ivan
- Gundulić, Ivan
- Peruško, Tone
- Römer, Bela
- Črnja, Zvane
- Show more

Libraries

- Gimnazija Požega

Series

- Dobra knjiga ;
- Dobra knjiga. Lektir...
- Mala knjižnica Mati...
- Narodna književnost...
- Zabavna biblioteka /...
- Show more

□ 1. **Latinska slovnica : za nižu gimnaziju**
by Veber Tkalčević, Adolfo .
Type: Book; Format: print ; Literary form: not fiction
Publisher: U Beču : C. k. prodaonica školskih knjigah, 1865.
Availability: **Copies available for loan:** Gimnazija Požega (1).
Actions: Add to cart

□ 2. **Knjiga o Istri**
by Peruško, Tone ; Črnja, Zvane .
Type: Book
Publisher: Zagreb : Školska knjiga, 1968.
Availability: **Copies available for loan:** Gimnazija Požega (1).
Actions: Add to cart

□ 3. **Priče iz životinjskog svijeta : za V. razred osmogodišnje škole**
by Dolenec, Franjo ; Lui, Ante .
Publisher: Zagreb : Školska knjiga, 1957.
Availability: **Copies available for loan:** Gimnazija Požega (1).

What is wrong with metadata?

http://www.catholicresearch.net/blog/2012/05/oai/

The harvested Dublin Core metadata was typical of OAI-PMH repositories: thin, a bit ambiguous, and somewhat inconsistent across repositories. -- *Eric Lease Morgan*

Europeana is nice example of this:

- sparse on meta-data
- multiple link hops to **image** of record (?!)

Importing covers and meta data

- DVD with scanned book front pages
  - various resolution (from stamp size to 300 dpi)
  - number_student/location/inventory_note.jpg
- Koha 3.8 has a tool to upload zip with cover images and idlink.txt
  - zip files big, and we don't have biblio records
- Create MARC21 records from file names (only metadata available to us)
- Write script which uses Koha API
  - create MARC21 using MARC::Record
  - AddItemFromMarc, PutImage
  - https://github.com/dpavlin/Koha/blob/koha_gimpoz/misc/gimpoz/import-images.pl

Scrape cataloging

- It's like copy cataloguing, but you don't have to use copy/paste in your browser to do it
- Instead, you use scraper to Z39.50 gateway: https://github.com/dpavlin/Biblio-Z3950
- Source formats:
  - Aleph - NSK, our national library
  - ~~COBISS~~ - they started serving images for records!
  - Google Books - another JSON API
  - vuFind - HathiTrust (MARC records export)
  - DPLA - JSON API (with broken UTF-8 encoding)
- Returns MARC21 records for Koha import

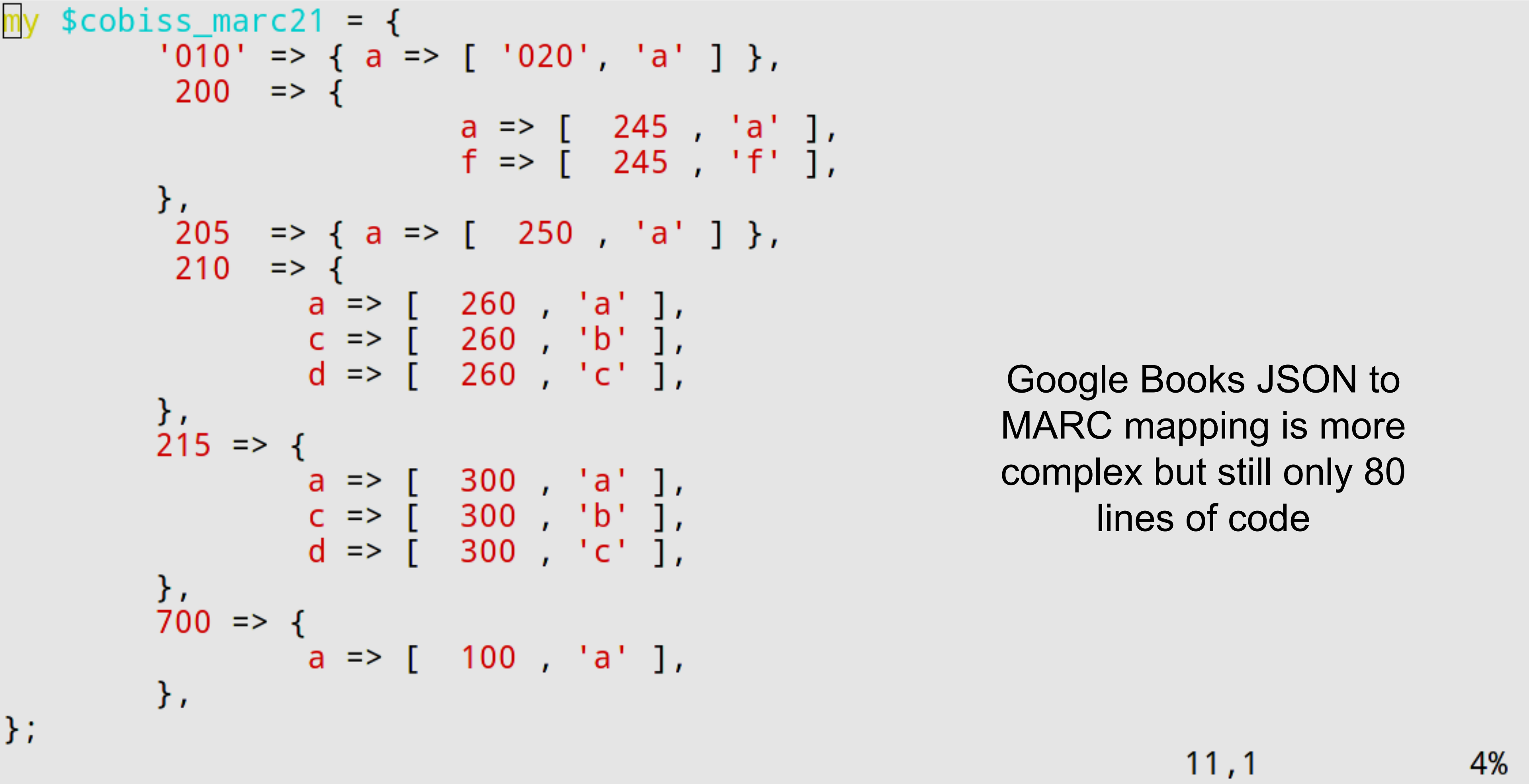
Scraping?!

- It's 2012, where is my semantic web?!
- Various reasons why scraping is easier
  - no public Z39.50 server
    - or there is one but has wrong encoding
  - data source isn't MARC21
    - older national MARC standards, UNIMARC or JSON for Google Books
- This is open source projects
  - all parts, but some assembly required
    - URLS to resources, mapping to MARC
  - modify existing scrapers to create new ones
- Let the data flow!

Biblio::Z3950

- based on Net::Z3950::SimpleServer
- convert Z39.50 RPN query to URL params
  - API support for and/or/not operators
  - enter just one field in Koha
- use WWW::Mechanize to issue search
  - advanced search syntax is best choice if available
  - scrape web page for results
    - web page with MARC-like structure
    - export formats
- use MARC::Record to create MARC21
  - web pages have utf-8 encoding
  - mapping to MARC specified in code

Mappings easy to define (in code :-)



```
my $cobiss_marc21 = {
        '010' => { a => [ '020', 'a' ] },
         200  => {
                        a => [  245 , 'a' ],
                        f => [  245 , 'f' ],
        },
         205  => { a => [  250 , 'a' ] },
         210  => {
                a => [  260 , 'a' ],
                c => [  260 , 'b' ],
                d => [  260 , 'c' ],
        },
        215 => {
                a => [  300 , 'a' ],
                c => [  300 , 'b' ],
                d => [  300 , 'c' ],
        },
        700 => {
                a => [  100 , 'a' ],
        },
};
                                                              11,1          4%
```

Google Books JSON to MARC mapping is more complex but still only 80 lines of code

Questions?

- Do you have nicely formatted web pages which need conversion to MARC21 for Koha?
- Is storing cover images in database the right way? (4.9Gb gziped SQL dump)

- This presentation: http://bit.ly/gimpoz
- Koha instance: http://gimpoz.koha.rot13.org
- Blog: http://blog.rot13.org

Abstract

We live in a world of data. However, data doesn't always come in a format that is as easy to share as we would expect.

We had approximately **6000 scans of book front pages** coming from the Teachers' library stock of the Gymnasium in Požega, which was proclaimed the movable monument of culture that carries national significance. Our **goal was to make the library stock visible to public** and we needed to add metadata to those images. Fortunately, some of that data was already available on the web: in National Library's Aleph system, several Croatian libraries using Koha, Hathi Trust digital library (VuFind), Open Library, Google Books, Europeana etc.

Importing local images is now standard part of Koha, so we decided to import all those images and to create the initial biblio records using the only kind of metadata that we had: structured directories and filenames which represent some kind of identifier number. After that, we started cataloguing our items. There is a convenient method for adding bibliographic data to a catalogue: using Z39.50 search. Unfortunately, not all of our metadata sources provided Z39.50 interface.

Our solution to the problem was to use **scrape-cataloguing**, which provided us with a way to avoid infinite copy & paste cycles or manual data entry. Instead, the job was done by our script that provides Z39.50 interface for Koha.