CD-ROM Redigitization Project (CDRiP)

Electronic Resources and Libraries Conference

February 22, 2007

Harrison Dekker
Coordinator of Data Services

Doe and Moffitt Libraries, UC Berkeley
Background: CD-ROMs and Library Data Services
Why do it?

**Pros:**
- Convenience
- Better access
- Preservation
- Hardware availability

**Cons:**
- Copyright issues
- Low use (?)
CDRiP: Production

- Student workers
- "Cheap" hardware
- Free software when possible
  - e.g. daemon tools, lcsisocreator
- Some locally developed applications (Perl, XML, XSLT)
- Use existing standards (e.g. iso, mods, xml/xslt)
• good for hierarchical data
• existing tools to save time
• easier coding for new tools
• platform independence
Description

MARC/Perl is a Perl 5 library for reading, manipulating, outputting and converting bibliographic records in the MARC format. The MARC format was designed by the Library of Congress in the late 1960s in order to allow libraries to convert their card catalogs into a digital format. The MARC format was standardized as ANSI/ISO 239.2 in 1971 and is now being used by libraries around the world to provide computerized access to their collections.

It is common for corporate, public and academic libraries to purchase Integrated Library Systems to organize and manage their collections. However, it is often advantageous for libraries to be able to create or massage pre-existing MARC data on the way into or out of such systems. This is the niche that MARC/Perl was designed to fill. If you can think of any other niches (e.g., Open Source Software for libraries) please let us know.

Development of this software library began with MARC.pm in the summer of 1999. After much experimentation, enhancement and discussion the MARC::Record framework was adopted which provides a more scalable approach to processing MARC data with Perl. While you can still download and get the documentation for MARC.pm, it is recommended that new users download MARC::Record. Future development of the MARC namespace on CPAN will fit into the framework provided by MARC::Record. If you have ideas for new modules please join the development team, or consider posting to the perl4lib list.

If you have any feature requests, or find any bugs (gasp) please report them via RT, the CPAN request tracking system.

Downloads

The latest stable version of the MARC::Record package is available from Comprehensive Perl Archive Network (CPAN). Experimental versions can often be found in the download area here at Sourceforge. Although older versions of MARC.pm are available for download, we recommend you use the MARC::Record package since future development will use it as a foundation.

Documentation

- MARC::Record::Tutorial: pdf
- MARC::Record: html
- MARC::Lint: html
REST – “representational state transfer”

- Facilitates “mash-ups” other forms of re-use.

**Example:**


**vs**

http://librrc-rh.berkeley.edu/crax/find/gl/8433509
  <mods version="3.1">
    <titleInfo>
      <title>Governor's budget summary</title>
    </titleInfo>
    <titleInfo type="alternative">
      <title>State of California, Governor's budget summary</title>
    </titleInfo>
    <name type="corporate">
      <namePart>California</namePart>
      <namePart>Governor</namePart>
    </name>
    <typeOfResource/>
    <originInfo>
      <place>
        <placeTerm type="street">Sacramento, Calif.</placeTerm>
      </place>
      <publisher>California Office of State Planning</publisher>
      <issuance>
        <frequency>Annual</frequency>
      </issuance>
      <originInfo/>
    </physicalDescription>
    <extent>4 ill. 28 cm.</extent>
    <physicalDescription/>
    <note type="statement of responsibility">
      submitted by ... Governor, to the California Legislature
    </note>
    <note>Some issues accompanied by CD-ROM.</note>
    <note>Report year ends June 30.</note>
    <note>Also available via the World Wide Web.</note>
    <subject>
      <geographicCode authority="marc46">m-usc-a</geographicCode>
    </subject>
    <subject authority="loc">
      <topic>Budget</topic>
      <geographic>California</geographic>
      <topic>Statistics</topic>
      <topic>Periodicals</topic>
    </subject>
  </mods>
</modsCollection>
CDRiP: User tools

- Wiki pages
- Virtual machine technology
  - allows users to safely install applications from otherwise “locked-down” workstations
- Search (to be determined)
  - how to integrate with other databases
  - how to leverage additional metadata
DREF

The following is a partial inventory of the statistical titles in the Doe Reference Collection. Please consult the Data Lab staff or a reference librarian if you wish to use one of the following or are having difficulty locating the data you need.

Statisticka rocenka Slovenskej republiky = Statistical yearbook of the Slovak Republic
Publisher: Statisticky urad Slovenskej republiky : Bratislava
Access: see PathFinder®

100 percent summary file: Census 2000 dress rehearsal
Access: see PathFinder®

108th Congressional district summary files: Census of Population and housing, 2000
Access: see PathFinder®

1987 census of agriculture
Access: see PathFinder®

1987 economic censuses
Access: see PathFinder®

1989 USSR population census
Publisher: East View Publications: Minneapolis, MN (1996)
Access: see PathFinder®

1990 Census of population and housing: P.L. 94-171 redistricting data
Access: see PathFinder®

1990 census of population and housing
CDRiP Future

- Preservation pipeline
- Publish to web
- Data mining
- File sharing
- Virtualization projects
SUDOC Virtualization Project

This site provides access to an online database containing government documents originally published on CDROM. There are two interfaces, a Database interface that enables browsing the entire collection and a Search interface that allows searching for specific documents using DQL queries.

Through either interface it is possible to access the original CDROM contents exported as WebDAV. These WebDAV directories can be directly “mounted” on some operating systems (OS X, Linux). In Windows they can be accessed through the network neighborhood or mounted using Novell Netdrive. WebDAV access does require a password. Please contact Geoffrey Brown for access.

On the Indiana University Bloomington campus, the CDROMs can be imported using windows file sharing. Interested users may contact Geoffrey Brown for details.
Summary

CDRiP rationale:
- Better long-term access

CDRiP technologies:
- ISO files and tools
- Virtual machine software
- REST
- Perl, XML, XSLT

Questions/comments: hdekker@library.berkeley.edu