Harvard Library Bibliographic Dataset

Rev. April 23, 2012

Contents:

Dataset Overview

The Harvard Bibliographic Dataset contains over 12 million bibliographic records for materials held by the Harvard Library, including books, journals, manuscripts, archival materials, electronic resources, scores, audio, video and other materials.

The metadata has been created, acquired and modified over decades, and represents a range of cataloging rules and practices. The records have not been altered or quality-checked during the export process and are offered as is.

Updates

The download file will be replaced with updated data periodically (approximately weekly). Records in the dataset will change if they have been updated in Harvard's library processing system. Each week's export will contain a new full data set. As records are deleted in Harvard's system, they will be omitted from subsequent refreshed download files.

Use Terms

Pursuant to its Open Metadata Policy, the Harvard Library makes this set of bibliographic records and the metadata contained therein (together, the "Metadata") available for public use under the Creative Commons Zero (CC0) Public Domain Designation:

Although Harvard does not impose any legally binding conditions on access to the Metadata, Harvard requests that you act in accordance with the following Community Norms of the Harvard Library with respect to the Metadata:

First, Harvard requests that the Harvard Library, along with OCLC Online Computer Library Center, Inc. ("OCLC") and the Library of Congress be given attribution as a source of the Metadata, to the extent it is technologically feasible to do so.

Second, Harvard requests that you make the Metadata and any improvements thereto freely available on the same terms as Harvard has done, i.e., without claiming any legal right in, or imposing any legally binding conditions on access to, the Metadata or your improvements, and with a request to act in accordance with these Community Norms.

Third, with respect to Metadata consisting of or contained in records Harvard has obtained from the OCLC WorldCat database, Harvard requests that you respect and act in accordance with the community norms set forth in the *WorldCat Rights and Responsibilities for the OCLC Cooperative*. Use of metadata from the WorldCat database for study and research is consistent with those norms, but if you plan to use such Metadata for other purposes, whether or not you are an OCLC member, we ask that you review and comply with those norms.

Attribution

We suggest the following language to provide proper attribution when using this dataset:

This [title of report or article or dataset] contains information from the Harvard Library Bibliographic Dataset, which is provided by the Harvard Library under its Bibliographic Dataset Use Terms and includes data made available by, among others, OCLC Online Computer Library Center, Inc. and the Library of Congress.

Bulk Download

The file containing the MARC21 records can be downloaded from this link:
http://openmetadata.lib.harvard.edu/bibdata/data

This file has been tarred and gzipped. If you are in a Windows environment you'll need a tool that lets you untar and unzip files. A number of such utilities exist with 7-Zip (http://en.wikipedia.org/wiki/7-Zip) being one of the more popular options.

The dataset can be downloaded through your browser, but if you prefer more control and feel comfortable using the command line, here's the basic flow:

1. Make a directory to work in:
    - mkdir hlom
2. Retrieve the latest tarball from the HLOM site:
    - curl -LO http://openmetadata.lib.harvard.edu/bibdata/data
3. Unzip and unpack the download:
    - tar xvf harvard.tar.gz

API Access Using the Digital Public Library of America API

Please see DPLA documentation at
http://blogs.law.harvard.edu/dplatechdev/2012/04/24/going-live-with-harvards-catalog/

Download File Format

The data consists of records in the MARC21 bibliographic format.[1] MARC21 bibliographic data conforms to the *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media.*[2]

Character Set

The records in this dataset are encoded in UTF-8.[3] See also the Character Sets and Encoding Options section of *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media.*

Local Characteristics

The records were exported from Harvard's library processing system and have some local characteristics:

1. ***Harvard typically catalogs with the so-called single record approach***, where multiple formats of the same content are cataloged using a single bibliographic record. Information specific to alternate formats of publication is moved out of the bibliographic record and into holdings records. This means that characteristics specific to microform and digital copies, in particular, will be absent from the bibliographic records in this dataset.

    a. The following fields are managed in holding records at Harvard and therefore are not included in the bibliographic records in this dataset:

MARC Bibliographic Field	MARC Holdings Equivalent	Element Name
007	007	Physical Description Fixed Field
506	506	Restrictions on Access Note
533	843	Reproduction note
540	845	Terms Governing Use and Reproduction note

[1] *MARC21 Format for Bibliographic Data*. Washington, D.C.: Library of Congress, 1999-2011. <http://www.loc.gov/marc/bibliographic/ecbdhome.html>
[2] *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media.* Washington, D.C.: Library of Congress, 2000-2007. < http://www.loc.gov/marc/specifications/spechome.html>
[3] UTF-8, UTF-16, UTF-32 & BOM Frequently Asked Questions <http://unicode.org/faq/utf_bom.html>

541	541	Immediate source of Acquisition Note
561	561	Ownership and Custodial History
562	562	Copy and Version Identification Note
563	563	Binding information
583	583	Action Note
85X-87X	85X-87X	Holdings, location, item information

b. Because of the single record approach, identifiers (e.g. OCLC numbers) for separate records for each format may all be present in the single bibliographic record.

c. Because the MARC field 856 is managed at the holdings level, links to electronic resources are not included in this dataset.

2. Some fields in the metadata contain subfield 5 (Institution to which field applies), which is used to flag information about a particular copy. The values in subfield 5 in this dataset are Harvard internal values for individual Harvard collections rather than MARC21 organization codes. Not all codes represent currently active libraries.

Code	**Display text**
AFR	Afro-American Studies
AJP	Botany Arboretum
ANT	Andover Newton Theol
ARG	Botany Gray/Arnold
ARN	Botany Arnold (Cambr.)
ART	Harvard University Art Museums
BAK	Baker Business
BER	Biblioteca Berenson
BIO	Biological Labs
BIR	Birkhoff Math
BLH	Blue Hill
BRM	Busch-Reisinger Mus
BUT	BU School of Theol
CAB	Cabot Science
CAR	Carpenter Center
CEA	East Asian Res Ctr
CEL	Robinson Celtic
CFI	Ctr Intl Affairs
CHE	Chemistry
CHI	Child Memorial
CRL	CRL (Ctr for Research Libs)
DAN	Derek Bok Center

DCJ	Doc Ctr Japan
DDO	Dumbarton Oaks
DES	Loeb Design
DEV	Development Office
DIV	Andover-Harv. Theol
DOC	Documents (Lamont)
ECB	Botany Econ. Botany
EDS	EDS/Weston
ENV	Environmental Information Ctr
EUR	Ctr Eur Studies
FAL	Fine Arts
FAR	Botany Farlow Library
FIG	Networked Resource
FOG	Fogg Museum
FOR	Harvard Forest
FUN	Fung Library
GCT	Gordon-Conwell
GDC	Harvard Data Center
GEO	Kummel Geological Sci
GIB	Gibb Islamic
GRA	Botany Gray Herbarium
GRO	Grossman
GUT	Gutman Education
HCR	Holy Cross Orthodox
HEL	Ctr Hellen Studies
HFA	Harvard Film Archive
HIL	Hilles
HIS	History Dept
HOU	Houghton
HPO	Harvard Planning & Real Estate
HSI	Sci Instruments
HSL	History of Science
HUA	Harvard Archives
HYL	Harvard-Yenching
KIR	Kirkland House
KSG	Kennedy Sch of Gov
LAM	Lamont
LAW	Law School
LIN	Linguistics
LIT	Littauer

MAP	Map Coll (Pusey)
MCK	McKay Applied Sci
MCZ	Museum Comp Zoology
MED	Countway Medicine
MIC	Microforms (Lamont)
MMF	Master Microforms
MUR	Murray Research Ctr
MUS	Loeb Music
NEL	Near Eastern Lib
NET	Networked Resource
NMM	National master micro
OPH	Ophthalmology
ORC	Botany Ames Orchid
PAL	Medieval Studies Lib
PEA	Peabody Museum
PHI	Robbins Philosophy
PHY	Physics Research
POE	Poetry Room (Lamont)
PRI	Primate Center Lib
PSY	Psychology Research
PUS	Pusey
QUA	Quad Library
RCA	Radcliffe Archives
RRC	Russian Res Ctr
RUB	Rubel (Fine Arts)
SAN	Sanskrit Library
SBC	Solidarity Bibl Center
SCC	Straus Conservation
SCH	Schlesinger
SFL	Schering Health Care
SIA	Sci & Intl Affairs
SMY	Smyth Classical
SOC	Social Rel-Sociol
STA	Statistics
STJ	St John's Seminary
TBC	Boston College
THE	Theatre Collection
TOZ	Tozzer
URI	Ukrainian Res Inst
WAM	Warren Anatomical
WAR	Charles Warren Ctr Lib
WEI	Weissman Preservation Ctr

WES	EDS/Weston
WID	Widener
WOL	Wolbach Library

3. The records contain a local field, 906, which contains a code for the "governing source" of records eligible for inclusion in the dataset. The field will include one of the codes below in a subfield 0. This field will be created during export and does not exist in Harvard's library processing system. Each record receives a single code. Once a record receives a code it is removed from further evaluation. The code is set using a combination of the MARC 040 $a (Original Cataloging Agency) and the presence of various identifiers. It is a determination based on those factors but cannot be treated as definitive.

Source Code	**Source Name**
DLC	Library of Congress
MH	Harvard
RLIN	Research Libraries Group
VEN	Vendor
OCLC	OCLC

4. The records contain a local field, 988, with a date formatted as YYYYMMDD. This date reflects a reasonable determination of when the record was added to Harvard's library processing system. Records added prior to June 2002 are all dated June 2002, when they were converted to a new library processing system.