

BHL Data Export

Updated December 7, 2011

Chris Freeland, BHL Technical Director

chris.freeland@mobot.org

Additional information on using these files is available at:

<http://biodiversitylibrary.blogspot.com/2008/09/export-of-titles-in-bhl-now-available.html>

Download

Individual tables

- Title: <http://www.biodiversitylibrary.org/data/title.txt> (16MB+)
- TitleIdentifier: <http://www.biodiversitylibrary.org/data/titleidentifier.txt> (4MB+)
- Subject: <http://www.biodiversitylibrary.org/data/subject.txt> (4MB+)
- Creator: <http://www.biodiversitylibrary.org/data/creator.txt> (5MB+)
- DOI: <http://www.biodiversitylibrary.org/data/doi.txt> (2MB+)
- Item: <http://www.biodiversitylibrary.org/data/item.txt> (19MB+)

To get Page and Name data, there is a ZIP archive containing the individual tables listed above, plus the Page and Name table data.

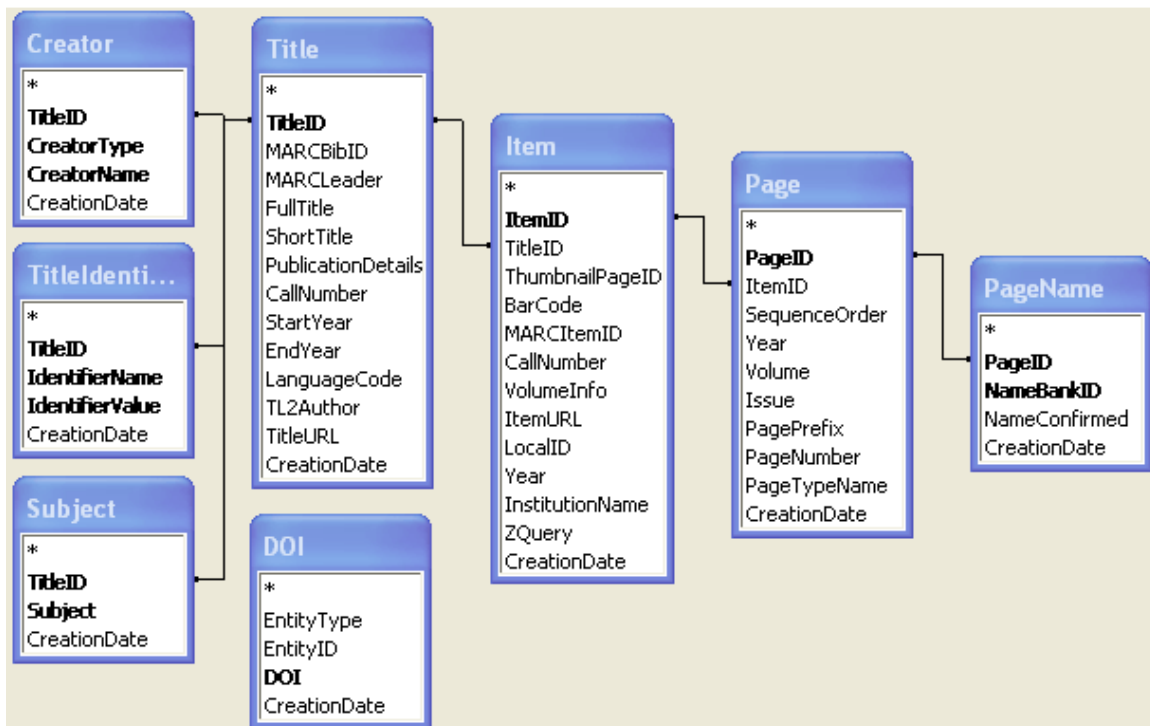
- <http://www.biodiversitylibrary.org/data/data.zip> (1+ GB)

Data Model

For a complete data model, including the tables described here, visit:

<http://biodiversitylibrary.blogspot.com/2008/08/revised-bhl-data-model.html>

The figure below shows how the tables available for download are related.



Schema Description

Title

The Title table contains bibliographic metadata about the journals and monographs represented in the BHL web portal, as extracted from the contributing library's catalogue at the time of scanning or applied post-scanning. **NOTE:** This export includes all of the Titles that have a status of "Published" in the BHL database.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
TitleID	PK, int	System-assigned identifier for the title within BHL database.
MARCBibID	nvarchar(50)	Identifier used to group titles during scanning. This field will be removed in future releases as it is now only used as an internal identifier.
MARCLLeader	nvarchar(24)	The Leader is the first field of a MARC record. It is fixed in length at 24 character positions (00-23). It consists of data elements that contain numbers or coded values and are identified by relative character position. ¹ http://www.loc.gov/marc/bibliographic/bdleader.html
FullTitle	ntext	The complete title for the work, as concatenated from MARC 245 subfields.
ShortTitle	nvarchar(255)	Same as FullTitle, but truncated to 255 characters.
PublicationDetails	nvarchar(255)	Publisher, Place of Publication, and Date, as concatenated from MARC 260 subfields.
CallNumber	nvarchar(100)	An alphanumeric code which identifies the shelf location of the title in the contributing library.
StartYear	smallint	Date extracted from MARC 008/07-10 http://www.loc.gov/marc/bibliographic/concise/bd008.html
EndYear	smallint	Date extracted from MARC 008/11-14 http://www.loc.gov/marc/bibliographic/concise/bd008.html
LanguageCode	nvarchar(10)	Language extracted from MARC 008/35-37 http://www.loc.gov/marc/bibliographic/concise/bd008.html
TL2Author	nvarchar(100)	TL-2 is the standard reference work for plant taxonomic literature from Linnean times to 1940. This field will be removed in future releases in favor of authors extracted from the MARC record. http://tl2.idcpublishers.info/
TitleURL	nvarchar(100)	Stable URL to the title in the BHL web portal.
CreationDate	Datetime	Date that the title was added to BHL.

¹ <http://www.itsmarc.com/crs/Gen0049.htm>

TitleIdentifier

The TitleIdentifier table relates digitized titles to standard identification schemes, as extracted from the contributing library's catalogue at the time of scanning or applied post-scanning.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
TitleID	FK, int	The TitleID to which this identifier is associated.
IdentifierName	nvarchar(40)	One of: <ul style="list-style-type: none">• Abbreviation Standard abbreviation by which this title is known.• BPH Identifier for title from Botanico-Periodicum-Huntianum• CODEN http://en.wikipedia.org/wiki/CODEN• DDC Dewey Decimal Classification• DLC• GPO Government Printing Office Identifier• ISBN International Standard Book Number• ISSN International Standard Serial Number• MARC001 Local control number assigned by the organization creating the record.• NAL Identifier assigned by the National Agricultural Library• NLM Identifier assigned by the National Library of Medicine• OCLC Identifier assigned by OCLC, available through www.worldcat.org• TL2 Identifier assigned by Taxonomic Literature-2.• WonderFetch Identifier assigned to the title by the contributing library at the time of digitization.
IdentifierValue	nvarchar(125)	The value extracted from the MARC record for the identifier
CreationDate	Datetime	Date that the identifier was added to BHL.

Creator

The Creator table contains the names of the authors of each journal and monograph.

NOTE: This export includes creator information for all Titles that have a status of “Published” in the BHL database.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
TitleID	FK, int	The TitleID to which this creator is associated.
CreatorType	nvarchar(50)	One of: <ul style="list-style-type: none">• Main – Personal Name• Main – Corporate Name• Main – Meeting Name• Added – Personal Name• Added – Corporate Name• Added – Meeting Name• Added – Uncontrolled Name “Main” and “Added” follow the MARC definitions for the same (see http://www.loc.gov/marc/bibliographic/ for more information).
CreatorName	nvarchar(450)	The creator name extracted from the MARC record for the title
CreationDate	Datetime	Date that the creator was added to BHL.

Subject

The Subject table contains information about subject headings assigned to each journal and monograph represented in the BHL web portal. **NOTE:** This export includes subject headings for all Titles that have a status of “Published” in the BHL database.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
TitleID	PK, int	The TitleID to which this subject is associated.
Subject	PK, nvarchar(50)	A subject heading assigned to this title.
CreationDate	Datetime	Date that the subject was added to BHL.

DOI

The DOI table contains information about Digital Object Identifiers that have been assigned to BHL entities (Titles, Items, Pages, etc).

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
EntityType	nvarchar(50)	The type of Entity to which this DOI is associated. Valid values include “Title”, “Item”, “Page”.
EntityID	Int	The ID of the Entity to which this DOI is associated.
DOI	PK, nvarchar(50)	The DOI identifier assigned to the Entity.
CreationDate	Datetime	Date that the DOI was added to BHL.

Item

The Item table contains information about each bound object (or “book”) digitized from a contributing library. For a serial, journal, or multi-volume monograph, an item represents a volume or multiple volumes bound together. For a single-volume monograph an item represents the book. **NOTE:** This export includes all of the Items that have a status of “Published” in the BHL database.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
ItemID	PK, int	System-assigned identifier for the item (volume or book) in the BHL database.
TitleID	FK, int	The TitleID to which this item is associated.
ThumbnailPageID	int	The identifier of the page whose thumbnail can be used as a thumbnail image for the entire book. To construct the path to the thumbnail image, use this identifier in the following URL: <a href="http://biodiversitylibrary.org/pagethumb/<ID>">http://biodiversitylibrary.org/pagethumb/<ID>
BarCode	nvarchar(4)	The identifier assigned to this book during scanning by the Internet Archive.
MARCItemID	nvarchar(50)	Redundant with BarCode. Will be removed in future releases.
CallNumber	nvarchar(100)	The shelf location of the volume, as assigned by the contributing library.
VolumeInfo	nvarchar(100)	Free-text volume (and year, in some cases) for the volume scanned, as extracted from the contributing library’s catalogue or applied during scanning.
ItemURL	nvarchar(100)	Stable URL to the item in the BHL web portal.
LocalID	nvarchar(100)	Local library identifier assigned at the time of scanning.
Year	nvarchar(20)	Free-text year for the volume scanned, as extracted from the contributing library’s catalogue or applied during scanning.
InstitutionName	nvarchar(255)	The contributing library.
ZQuery	nvarchar(200)	String used to query the contributing library’s catalogue to return bibliographic information about the volume during scanning.
CreationDate	Datetime	Date that the item was added to BHL.

Page

The Page table contains the metadata about the scanned pages from an Item. **NOTE:** This export DOES NOT include all of the pages in the BHL database. It only contains pages on which taxonomic names have been identified.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
PageID	PK, int	System-assigned identifier for the page image in the BHL database.
ItemID	FK, int	The ItemID to which this page is associated.
SequenceOrder	Int	Position of the page within an Item. For example, a value of 1 indicates the first page. A value of 10 indicates the tenth page. This is NOT the assigned page number, but rather an indication of the order of the pages within an Item.
Year	nvarchar(20)	Year the page was published, assigned post-scanning.
Volume	nvarchar(20)	Volume in which the page was published, assigned post-scanning.
Issue	nvarchar(20)	Issue in which the page was published, assigned post-scanning.
PagePrefix	nvarchar(20)	Free-text descriptor for the page, assigned post-scanning.
PageNumber	nvarchar(20)	Page number as printed on the page, assigned post-scanning.
PageTypeName	nvarchar(40)	Kind of page, assigned post-scanning.
CreationDate	Datetime	Date that the page was added to BHL.

PageName

The PageName table lists all of the names that have been identified by TaxonFinder and the pages on which those names are found.

<i>Field</i>	<i>Data Type</i>	<i>Description</i>
NameBankID	PK, int	Taxon identifier from NameBank
NameConfirmed	nvarchar(255)	String extracted by TaxonFinder and confirmed by NameBank.
PageID	FK, int	The PageID on which this name is found.
CreationDate	Datetime	Date that the name was added to BHL.