BHL Tab-Separated Data Export

Updated November 15, 2022

Additional information on using these files is available at:
https://blog.biodiversitylibrary.org/2008/09/export-of-titles-in-bhl-now-available.html

Download

Two versions of these files are available. The first contains metadata for all items that are part of the BHL collection. The second contains metadata for only those items with content that is hosted by BHL (this includes the majority of the BHL collection). The data model and schema is the same for both sets of files. The only difference is the content.

The files containing metadata for the **entire BHL collection** can be downloaded from the following locations:

Individual tables

- Title: http://www.biodiversitylibrary.org/data/TSV/title.txt (55MB+)
- TitleIdentifier: http://www.biodiversitylibrary.org/data/TSV/titleidentifier.txt (17MB+)
- Subject: http://www.biodiversitylibrary.org/data/TSV/subject.txt (34MB+)
- Creator: http://www.biodiversitylibrary.org/data/TSV/creator.txt (30MB+)
- DOI: http://www.biodiversitylibrary.org/data/TSV/doi.txt (13MB+)
- Item: http://www.biodiversitylibrary.org/data/TSV/item.txt (106MB+)
- Part: http://www.biodiversitylibrary.org/data/TSV/part.txt (130MB+)
- PartCreator: http://www.biodiversitylibrary.org/data/TSV/partcreator.txt (24MB+)
- PartIdentifier: http://www.biodiversitylibrary.org/data/TSV/partidentifier.txt (11MB+)
- CreatorIdentifier: http://www.biodiversitylibrary.org/data/TSV/creatoridentifier.txt (7MB+)

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/TSV/data.zip (3.9+ GB)

The files containing metadata for **only items hosted by BHL** can be downloaded from the following locations:

Individual tables

- Title: http://www.biodiversitylibrary.org/data/TSV/hosted/title.txt (50MB+)
- TitleIdentifier: http://www.biodiversitylibrary.org/data/TSV/hosted/titleidentifier.txt (16MB+)
- Subject: http://www.biodiversitylibrary.org/data/TSV/hosted/subject.txt (33MB+)
- Creator: http://www.biodiversitylibrary.org/data/TSV/hosted/creator.txt (28MB+)
- DOI: http://www.biodiversitylibrary.org/data/TSV/hosted/doi.txt (13MB+)
- Item: http://www.biodiversitylibrary.org/data/TSV/hosted/item.txt (102MB+)
- Part: http://www.biodiversitylibrary.org/data/TSV/hosted/part.txt (115MB+)
- PartCreator: http://www.biodiversitylibrary.org/data/TSV/hosted/partcreator.txt (18MB+)
- PartIdentifier: http://www.biodiversitylibrary.org/data/TSV/hosted/partidentifier.txt (9MB+)
- CreatorIdentifier: http://www.biodiversitylibrary.org/data/TSV/hosted/creatoridentifier.txt (6MB+)

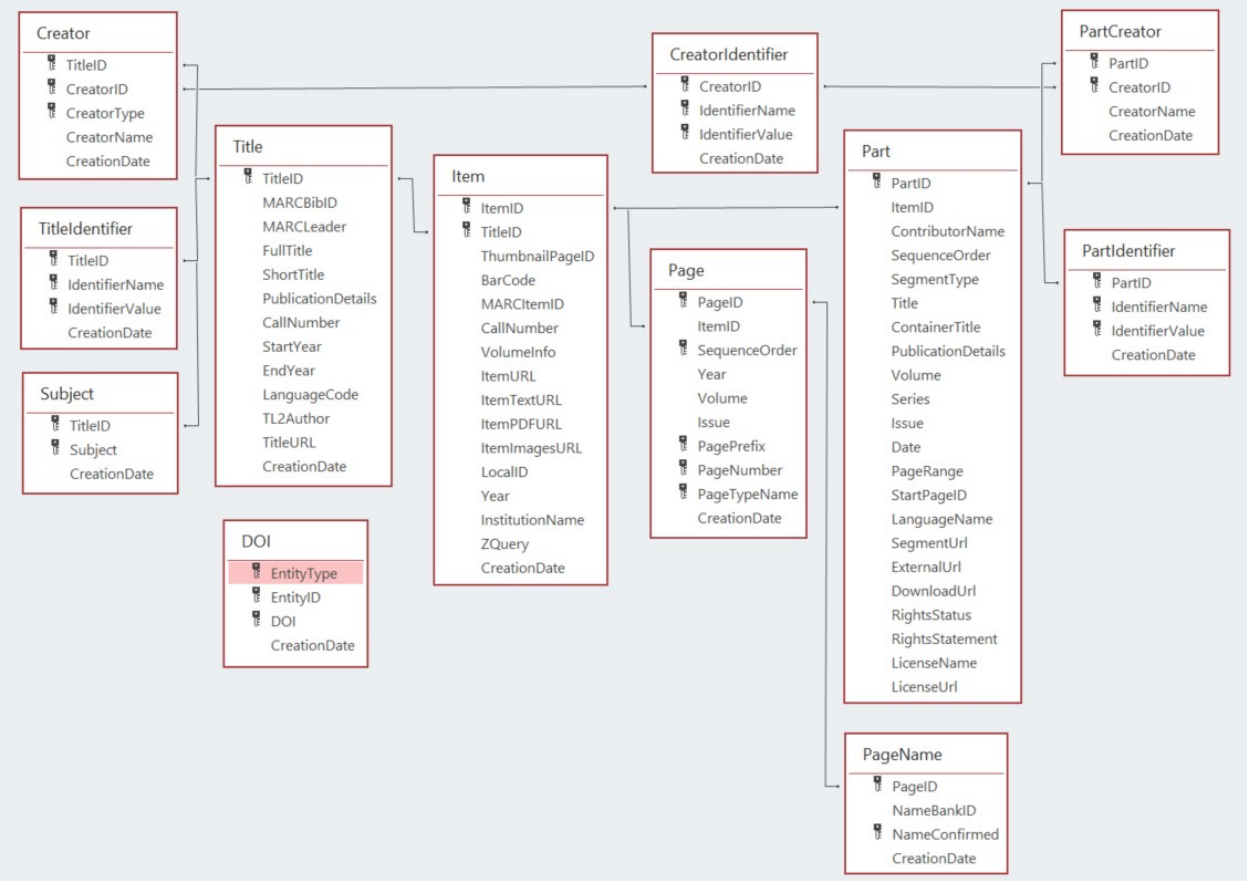
To get Page and Name data, there is a ZIP archive containing the individual tables listed above, plus the Page and Name table data.

- http://www.biodiversitylibrary.org/data/TSV/hosted/data.zip (3.9+ GB)

Data Model

For a complete data model and data dictionary, visit:
https://github.com/gbhl/bhl-us/tree/master/Documentation/DataModel

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



Users loading the data from these export files into a relational database are recommended to use surrogate primary keys for the tables. This avoids the possibility of duplicate primary key values caused by dirty data. BHL strives to produce clean data in these export files but cannot guarantee it. BHL uses surrogate keys internally within their own databases.

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.

Field	*Data Type*	*Description*
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.
MARCLeader	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. 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.

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.

Field	*Data Type*	*Description*
TitleID	PK, FK, int	The TitleID to which this identifier is associated.
IdentifierName	PK, nvarchar(40)	One of: • Abbreviation Standard abbreviation by which this title is known. • ARK Archival Resource Key • 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/eISSN 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 • Soulsby Soulsby number • TL2 Identifier assigned by Taxonomic Literature-2. • Wikidata "Q" identifier assigned by Wikidata • WonderFetch Identifier assigned to the title by the contributing library at the time of digitization.
IdentifierValue	PK, 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.

Field	*Data Type*	*Description*
TitleID	PK, FK, int	The TitleID to which this creator is associated.
CreatorID	PK, Int	System-assigned identifier for the creator within BHL database.
CreatorType	PK, nvarchar(50)	One of: • 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.

Field	*Data Type*	*Description*
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).

Field	*Data Type*	*Description*
EntityType	PK, nvarchar(50)	The type of Entity to which this DOI is associated. Valid values include "Title", "Item", "Page".
EntityID	PK, 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.

Field	*Data Type*	*Description*
ItemID	PK, int	System-assigned identifier for the item (volume or book) in the BHL database.
TitleID	PK, 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: http://biodiversitylibrary.org/pagethumb/<ID>
BarCode	nvarchar(200)	The identifier assigned to this book during scanning by the Internet Archive.
MARCItemID	nvarchar(200)	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.
ItemTextURL	nvarchar(100)	Stable URL to the text of the item.
ItemPDFURL	nvarchar(100)	Stable URL to the PDF for the item.
ItemImagesURL	nvarchar(100)	Stable URL to the page scans for the item.
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.

Field	*Data Type*	*Description*
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	PK, 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	PK, nvarchar(40)	Free-text descriptor for the page, assigned post-scanning.
PageNumber	PK, nvarchar(20)	Page number as printed on the page, assigned post-scanning.
PageTypeName	PK, 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.

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

PartCreator
The PartCreator table contains the names of the authors of each part.

Field	*Data Type*	*Description*
PartID	PK, FK, int	The PartID to which this creator is associated.
CreatorID	PK, Int	System-assigned identifier for the creator within BHL database.
CreatorName	nvarchar(450)	The creator name
CreationDate	Datetime	Date that the creator was added to BHL.

Part

The Part table contains information about articles/chapters/treatments/etc. These parts may or may not be contained in material scanned by BHL. **NOTE:** This export includes all of the Parts that have a status of "Published" in the BHL database.

Field	*Data Type*	*Description*
PartID	PK, int	System-assigned identifier for the part within BHL database.
ItemID	Int	The ItemID to which this part is associated. May be blank.
ContributorName	nvarchar(255)	Name of the institution or person that contributed the part.
SequenceOrder	Int	Sequence order of the part within its container work. (Example: First article in a journal is assigned a value of 1, second is assigned 2, and so on).
SegmentType	nvarchar(50)	The type of part. Possible values include Article, Chapter, and Treatment.
Title	nvarchar(2000)	The complete title for the part.
ContainerTitle	nvarchar(2000)	The title of the work in which this part appears.
PublicationDetails	nvarchar(400)	Publisher, Place of Publication, and Date
Volume	nvarchar(100)	Volume of the container work
Series	nvarchar(100)	Series of the container work
Issue	nvarchar(100)	Issue of the container work
Date	nvarchar(20)	Date of publication of the part
PageRange	nvarchar(50)	The range of pages in the container work on which the part appears.
StartPageID	Int	PageID of the first page of the part. May be blank.
LanguageName	nvarchar(100)	Language in which the part is written.
SegmentUrl	nvarchar(200)	Stable URL to the part in the BHL web portal.
ExternalUrl	nvarchar(200)	URL of a location external to BHL at which the part is located
DownloadUrl	nvarchar(200)	URL for downloading the part
RightsStatus	nvarchar(500)	Rights, usage, and copyright information.
RightsStatement	nvarchar(500)	Rights, usage, and copyright information.
LicenseName	nvarchar(200)	Rights, usage, and copyright information.
LicenseUrl	nvarchar(200)	Rights, usage, and copyright information.

PartIdentifier

The PartIdentifier table relates parts to standard identification schemes.

Field	*Data Type*	*Description*
PartID	PK, FK, int	The PartID to which this identifier is associated.
IdentifierName	PK, nvarchar(40)	One of: • Abbreviation Standard abbreviation by which this part is known. • ARK Archival Resource Key • BioStor Identifier of the part within BioStor (https://biostor.org/) • DLC • ISBN International Standard Book Number • ISSN/eISSN International Standard Serial Number • JSTOR Identifier assigned by JSTOR • OAI • OCLC Identifier assigned by OCLC, available through www.worldcat.org • Soulsby Soulsby number • TL2 Identifier assigned by Taxonomic Literature-2. • Wikidata "Q" identifier assigned by Wikidata
IdentifierValue	PK, nvarchar(125)	The value for the identifier
CreationDate	Datetime	Date that the identifier was added to BHL.

CreatorIdentifier

The CreatorIdentifier table relates creators to standard identification schemes.

Field	*Data Type*	*Description*
CreatorID	PK, FK, int	The CreatorID to which this identifier is associated.
IdentifierName	PK, nvarchar(40)	One of: • Abbreviation Standard abbreviation for this creator • ARK ARK identifier for this creator • BioStor Identifier of the creator within BioStor (https://biostor.org/) • DLC • OCLC Identifier assigned by OCLC, available through www.worldcat.org • ORCID Identifier assigned by the ORCID organization • ResearchGate https://www.researchgate.net • SNAK ARK https://snaccooperative.org • Soulsby Soulsby number • Tropicos Identifier of the creator within Tropicos (https://www.tropicos.org) • VIAF Identifier assigned by VIAF • Wikidata Identifier assigned at Wikidata
IdentifierValue	PK, nvarchar(125)	The value for the identifier
CreationDate	Datetime	Date that the identifier was added to BHL.