WorldCat knowledge base
Content provider data feeds

Overview

This document is intended to help content providers provide up-to-date metadata for subscribed online content (electronic journals, ebooks, eaudiobooks, video, and still images) for use in the WorldCat knowledge base. The WorldCat knowledge base enables discovery and delivery of library resources for information seekers throughout both OCLC services and non-OCLC systems. At its core, the WorldCat knowledge base is driven by a comprehensive set of electronic resource collections and titles offered by a variety of content providers.

The WorldCat knowledge base is built to support a wide range of uses throughout different applications. Libraries and end users across the globe may use these different applications to determine their access to electronic resources. As a result, it is critically important that content providers work with OCLC to accurately capture information about their collections and titles in any region that these materials are available.

This document describes the OCLC suggested format for providing these feeds. This format is based on a set of NISO recommendations for transmission of this type of data called KBART. The format described in the KBART recommendations is endorsed and supported by all major link resolver vendors. OCLC's suggested format differs from KBART only by adding information above and beyond KBART to allow content providers to share data that is not currently covered by the recommendations.

What is a collection?

A collection is a grouping of titles that should be recognizable to librarians working with knowledge base data. These may be sold to libraries as subject or year based groups of titles that are covered by one subscription. If a la carte or custom purchases are offered, the WorldCat knowledge base supports a single, large collection that allows a subset of titles to be selected by the library.

Content providers should work with OCLC to describe all possible collections that may be offered to libraries. Some examples of different collection types are:

- A collection containing all titles on offer
- A subject based collection (e.g. a Medicine Collection)
- A year based collection (e.g. All titles published in 2013)
- A consortial collection (e.g. All titles available for purchase only to the XYZ consortium)
- A regional version of a collection (e.g. Books available for purchase in the UK)
- An open access collection (e.g. All open access journals)
- A collection of titles available for Patron Driven Acquisition
- Some combination of the above

In all cases, the collections should match as closely as possible to the different subscription and access models for content being sold to libraries.

Library specific holdings

Providing library specific holdings to OCLC is strongly recommended. The WorldCat knowledge base uses global level collections and titles lists as its base, but having holdings for each library is what makes the system meaningful to users. Traditionally, each subscribing library manages this data manually. By providing these holdings directly to OCLC, a content provider can significantly ease the burden of maintenance and increase the satisfaction of libraries with their products.

Detailed specifications about the data format are provided below. We look forward to working with you to update your metadata.

Data Format Overview

This document describes the desired format. If it is not possible to provide data in exactly this format, OCLC must evaluate the format that will be used to ensure all data will map to the WorldCat knowledge base correctly. If a different format is used, this document should still be used to determine at the field level the kinds of information that should be sent.

Files should be created in tab-delimited format and saved with UTF-8 encoding. Field values can be quoted if necessary.

If there is data available beyond the fields outlined in this document, those fields can be added to the format and the name of the field should be prefixed with the 'provider_id' value as defined below. *If using this format all column/field headings should always be present 'as is' and in the specified order, but the value can be left blank*.

Frequency

Ideally, OCLC would like to support weekly feeds for this metadata. This ensures that the WorldCat knowledge base has an up to date representation of the contents of any platform. For many content providers, this data may not change as often as weekly and other schedules can be discussed. We would like at least a refreshed copy of all data on a monthly basis even if very little has changed.

OCLC can receive file updates for automated update processing in a number of ways. We can provide an FTP account, use your onsite FTP account or retrieve files from static URLs on your web pages. It is important to communicate the update schedule and date for your metadata for example, 'monthly on the first of the month' or 'weekly on Mondays'. This will ensure that we can load your data as quickly as possible alleviating lags in data coverage for users.

OCLC does not guarantee any data provided will be accepted for ingest into the WorldCat knowledge base. Certain common errors may sometimes prevent accurately interpreting these data feeds. If errors are detected which prevent the loading of data, we will follow up with our contacts at your organization to correct the data.

Collection File Naming

There will be a per organization identifier or 'provider_id' assigned which should be used in the naming of any files contributed to OCLC for use in this processing. The <provider_id> should be contiguous text with no special characters. The second element in the collection filename is <collection_id> and this ID must be a static identifier as it is used for subsequent machine updates and match processing. Please omit any punctuation or special characters in <collection_id>. The final element in the collection filename is an optional date stamp in ISO format.

<provider_id>_<collection_id>_YYYY-MM-DD.txt

For Region or Consortia include the additional information as in this example

<provider_id>_<region/consortium>_<collection_id>_<YYYY-MM-DD>.txt

<ProviderName>_<Region/Consortium>_<PackageName>_<YYYY-MM-DD>.txt

This format identifies the source of the files. An example might be “oclc” as the identifier for OCLC and is illustrated below:

- oclc_alltitles_2011-05-05.txt
- oclc_history01_2011-07-04.txt
- oclc_soh12345_2011-11-11.txt

In this example using the ‘oclc’ ‘provider_id’, all collection files are prefixed with ‘oclc’ followed by the ‘collection_id’. * Note that ‘collection_id’ must be persistent as it is used for harvest-load matching.

Here are additional examples including consortium and other special cases:

- JSTOR_Global_AllTitles_2013-01-14.txt
- Taylor&Francis_Global_AllTitles_2012-08-30.txt

Title list is consortium-specific, for a specific package:

- IOP_NESLi2_Option1-2011_2012-05-31.txt (includes a year as part of the package name)
- OxfordJournals_SCELC_AllTitles_2013-01-09.txt (contains all titles that the consortium has subscribed to)

Title list is region-specific, for a specific package:

- Springer_Asia-Pacific_Medicine_2013-01-28.txt

Global (Bibliographic level) data

The basis for the knowledge base is a simple listing of all content available from a particular content provider. This global data describes the complete set of title and collections of titles that are available from a particular content provider.

Global data should always be sent as complete replacement files. All data, regardless of whether there have been changes should be sent with each set of files. If records are removed or deleted, those should not be present at all. One complete data file per 'collection' should be sent. See the collection section for a description of collections.

Global data format

This format is based on the KBART Phase 2 NISO RP-9-201x recommendations.

Column number	*Column heading*	*Description*
01	publication_title	Publication title for serial or monograph. (Conference proceedings series titles are treated as serials while volume titles are treated as monographs)
02	print_identifier	Print-format identifier (i.e. ISSN, ISBN, etc.)
03	online_identifier	Online-format identifier (i.e. eISSN, eISBN, etc.)
04	date_first_issue_online	Date of first issue available online in ISO format: YYYY-MM-DD or YYYY-MM or YYYY; *applies only to serials*
05	num_first_vol_online	Number of first volume available online; applies only to serials
06	num_first_issue_online	Number of first issue available online; applies only to serials
07	date_last_issue_online	Date of last issue available online (or blank, if coverage is to present) in ISO format: YYYY-MM-DD or YYYY-MM or YYYY; applies only to serials
08	num_last_vol_online	Number of last volume available online (or blank, if coverage is to present) applies only to serials
09	num_last_issue_online	Number of last issue available online (or blank, if coverage is to present) applies only to serials
10	title_url	Title-level URL; Applicable to both serials and monograph For conference proceedings, the title url for the proceedings series and the title_url for each volume should be different.
11	first_author	First author (for monographs only)
12	title_id	Title ID - A unique identifier for this title in the data provider's system (this item must be persistent to facilitate update matching) Applicable to both serials and monographs. For conference proceedings, the title_id for

		the conference proceedings series and the title_id for each proceeding volume should be different.
13	embargo_info	Embargo information Describes any limitations on when resources become available online
14	coverage_depth	Coverage depth (e.g., abstracts or full text)
15	notes	Free-text field to describe the specifics of the coverage policy.
16	publisher_name	Publisher name (Not to be confused with platform hosting name)
17	publication_type	Serial or monograph. Use serial for journals and conference proceeding series. Use monograph for books, e-books, and conference proceeding volumes.
18	date_monograph_published_print	Date the monograph is first published in print (in ISO format: YYYY-MM-DD or YYYY-MM or YYYY)
19	date_monograph_published_online	Date the monograph is first published online (in ISO format: YYYY-MM-DD or YYYY-MM or YYYY)
20	monograph_volume	Number of volume for monograph. Applicable to e-books and conference proceedings. For proceedings, the volume within the conference proceedings series.
21	monograph_edition	Edition of the monograph
22	first_editor	First editor. Applicable to monographs, i.e., e-books or conference proceedings volumes.
23	parent_publication_title_id	Title identifier of the parent publication. For a conference proceeding volume, the parent_publication_title_id is the title_id of the conference proceedings series.
24	preceding_publication_title_id	Title identifier of any preceding publication title. Applicable to serials and conference proceedings series.
25	access_type	Access type may be fee-based (P) or Open Access (F)

26	oclc_number	The OCLC number (if available)

Collections

One file (in the above format) should be sent for each collection or package offered by the provider. A collection is a grouping of titles that will be shown together and, typically, can be purchased as a single unit by a customer. Possible reasons for grouping titles in collections are listed in the overview above.

Note that there may not be any different collections. If there is only one group of titles that subscribers can pick from, there may be only one collection of titles.

There should not be one collection per site/customer.

Collections Description File

In addition to the per collection files, one file should be provided which lists some information about the collections that are being transferred. This is the Collections Description File.

The organization identifier or ‘provider_id’ assigned (mentioned above) should also be used in this file name and in the following format which follows conventions for matching and automated processing:

COLLECTIONS_<provider_id>.txt

Note: date stamp is optional as OCLC’s automated processing automatically date stamps submitted files

COLLECTIONS_<provider_id>_YYYY-MM-DD.txt

Information and Format of Collections Description File

Column number	***Column heading***	***Description***
01	collection_name	Human readable name for this collection
02	collection_id	Identifier used for this collection by the data provider's system (sometimes called a database ID). *This string must remain static and free of punctuation* and it also must be contained in the collection file name described above.
03	selectable	Yes/No indication of whether individual records can be purchased/selected from this collection by a customer, essentially à la carte subscriptions within the greater collection.

04	custom_coverage	Yes/No indication of whether the coverage dates (for individual journals) can be changed on a per customer basis.
05	open_access	Yes/No indication of whether this collection contains only open access titles.
06	pda	Yes/No indication of whether this collection contains only titles available for Patron Driven Acquisitions (also called Demand Driven Acquisitions)
07	collection_url	URL pointing to a collection level page (can be a search form or a listing of titles in the collection) on the provider's site.
08	record_count	The number of records in this collection. This is used to help verify the contents of the data files.

Library specific holdings data format

After the global level data is synchronized between systems, it is also desirable to have customer specific holdings files that describe the particular subscriptions of each customer.

Holdings data can be contributed in full replacement files each time or by delivering updates to the previous files. If delivering updates, be sure to use the 'status' field to indicate the type of change from previous.

Format of holdings data

Column number	***Column heading***	***Description***
01	site_id	A site/customer identifier *See note below
02	collection_id	A collection_id that matches to one of the collection_id values in the collections information
03	title_id	A title_id that matches to one of the title_ids in the data file for this collection. Should be left blank if the entire collection is subscribed by this customer and there are no overrides in columns 4-10.
04	date_first_issue_online	Date of first issue available online in ISO format: YYYY-MM-DD or YYYY-MM or YYYY; *for ebooks, the date of publication*

05	num_first_vol_online	Number of first volume available online
06	num_first_issue_online	Number of first issue available online
07	date_last_issue_online	Date of last issue available online (or blank, if coverage is to present) in ISO format: YYYY-MM-DD or YYYY-MM or YYYY
08	num_last_vol_online	Number of last volume available online (or blank, if coverage is to present)
09	num_last_issue_online	Number of last issue available online (or blank, if coverage is to present)
10	title_url	Title-level URL (if different from the global URL)
11	status	values: ADD/DELETE/UPDATE. Optional field used when sending changes only (not full files) to indicate whether this is a new holding (ADD) or an old holding to be removed (DELETE) or just a changed holding (UPDATE).

- Note: Columns 4-10 should only contain values if this customer's subscription is different from the data in the global (bib) data files.

Site IDs

These two types of customer identifiers are supported by default:

- OCLC Symbol (http://www.oclc.org/contacts/libraries/)
- WorldCat Registry Identifier (http://worldcat.org/registry)

If it is not possible to use one of those identifiers or if the identifier is not guaranteed to be available for all customers, a provider specific customer identifier can be used. If a provider specific site_id is used, the customer data file (below) is required.

Library (about) data

This file is optional. It can be used to convey customer specific information if necessary. This file should always be included when the 'site_id' values specified in the holdings information are not one of the two officially supported identifiers or when per customer configuration (account id or custom location id) is required by the provider's platform in order to link to this content.

Column number	Column heading	Description
01	site_id	A site/customer identifier, matching the value in the holdings file.
02	site_name	The name of this site/customer
03	oclc_symbol	The OCLC symbol(s) for this library. If more than one, they should be comma separated (i.e. "TRN,SER,OZY")
04	registry_id	The WorldCat Registry id for this library.
05	local_site_id	An identifier for this site/customer used by the data provider's site.
06	country	The country this library is located in. Used to help match to OCLC customer record.
07	notes	Any other pertinent information

File names to use

There are potentially many different files which will be transferred through this process. Specific file names are required to distinguish between each type of data. The naming scheme to use is:

- Global (title list) metadata – many files: DATA_<provider_id>_<collection_id>.txt
- Collection information file – one file: COLLECTIONS_<provider_id>.txt
- Library specific holdings file – one file: HOLDINGS_<provider_id>.txt
- Library (about) data file – one file: CUSTOMERS_<provider_id>.txt

Notes:

- "<provider_id>" should be replaced with the agreed upon identifier for the organization providing the data
- "<collection_id>" should be replaced with the collection_id used in the collections description file
- The extension ".txt" can be replaced with the appropriate extension, such as ".xml" or ".csv" if providing the data in an alternate format
- If the data is compressed as a zip file the name should be: DATA_<provider_id>.zip

Support

For support issues, questions concerning data feed formatting or to request an FTP account send inquiries to the Knowledge Base Data Team and please include ‘provider_id’ in the subject line.