BASE OAI Interface

Release 3.1

Bielefeld University Library

May 11, 2023

Contents

1 Introduction

This documentation describes the OAI-PMH interface of Bielefeld Academic Search Engine (BASE). BASE is an OAI search service that currently includes the contents of more than 3,000 document servers worldwide.

1.1 What is it for?

This API is especially suitable for clients that would like to get *subsets* of the BASE data. For instance, it can be used by *subject portals* to integrate subject-specific publication metadata from BASE into their indexes.

1.2 Alternatives

- If you would like to embed *search results* from BASE directly in your infrastructure, please consider using the BASE search API instead.
- If you need a *complete dump* of the BASE data for your non-commercial project, please contact us for an initial load.

1.3 How to get Access

Access to the BASE OAI-PMH interface is IP-restricted. Non-commercial projects may apply for access by contacting us via this form. Please specify your use case and an IP or IP range from which you need to access the API. You will get an email notification as soon as your IPs have been activated.

1.4 URL of the OAI Endpoint

The OAI endpoint of this API is located at http://oai.base-search.net/oai.

Note: If your IP is not registered yet (see above), you will face a custom OAI error with error code `restrictedInterface` when trying to access the base URL.

2 OAI-PMH Primer

The API implements the Open Archives Protocol for Metadata Harvesting (OAI-PMH). This section gives only a basic overview of OAI-PMH. For more information, please refer to the protocol specification.

2.1 Glossary of Important OAI-PMH Concepts

Repository A *repository* is a server-side application that exposes metadata via OAI-PMH. In the context of this API, the repository is the BASE search engine.

Harvester OAI-PMH client applications are called *harvesters*.

record A *record* is the XML-encoded container for the metadata of a single publication item. It consists of a *header* and a *metadata* section.

header The record *header* contains a unique identifier and a datestamp.

metadata The record *metadata* contains the publication metadata in a defined metadata format.

set A structure for grouping records for selective harvesting.

harvesting The process of requesting records from the repository by the harvester.

2.2 OAI Verbs

OAI-PMH features six main API methods (so-called "OAI verbs") that can be issued by harvesters. Some verbs can be combined with further arguments:

`Identify` Returns information about the repository. Arguments: None.

`GetRecord` Returns a single record. Arguments:

- `identifier` (the unique identifier of the record, *required*)
- `metadataPrefix` (the prefix identifying the metadata format, *required*)

`ListRecords` Returns the records in the repository in batches (possibly filtered by a timestamp or a `set`). Arguments:

- `metadataPrefix` (the prefix identifying the metadata format, *required*)
- `from` (the earliest timestamp of the records, *optional*)
- `until` (the latest timestamp of the records, *optional*)
- `set` (a set for selective harvesting, *optional*)
- `resumptionToken` (used for getting the next result batch if the number of records returned by the previous request exceeds the repository's maximum batch size, *exclusive*)

`ListIdentifiers` *Like `ListRecords` but returns only the record headers.*

`ListSets` Returns the list of sets supported by this repository. Arguments: None

`ListMetadataFormats` Returns the list of metadata formats supported by this repository. Arguments: None

2.3 Harvesting Records

In the OAI terminology, *harvesting* refers to the consecutive aggregation of metadata records from a repository. This is done by issueing an initial `ListRecords` request followed by potential resumption requests if the the number of records matching the inital request exceeds the maximum response batch size of the repository. In the latter case, the existence of further records is indicated by the repository through an XML element `resumptionToken` at the bottom of the response. The content of this element has to be provided in the subsequent request.

A valid example of an initial request would be:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=oai_dc

Note: The argument `metadataPrefix` specifying the metadata format for record dissemination is required.

Given the above request, the `resumptionToken` would be:

```
<resumptionToken completeListSize="41398100" cursor="0">
    fm9haV9kY34yMDB-fg==
</resumptionToken>
```

To fetch the next batch of records, the client would then need to issue the following request:

http://oai.base-search.net/oai?verb=ListRecords&resumptionToken=fm9haV9kY34yMDB-fg==

Note: The `resumptionToken` argument is *exclusive*. Additional arguments provided with the initial request like `metadataPrefix` or `set` therefore **must not** be included in resumption requests. Also note that the `resumptionToken` XML element carries two attributes `completeListSize` referring to the total number of records matching the request and `cursor` referring to the number of records returned so far. Clients are strongly encouraged to keep track of this information and include it in issue reports about the interface.

3 Metadata Formats

Currently, this API supports two metadata formats: OAI-DC (Dublin Core, metadata prefix `oai_dc`) and BASE-DC (OAI-DC extended with custom fields, metadata prefix `base_dc`). Further formats may follow in the future.

3.1 oai_dc

The `oai_dc` format exposes the metadata encoded as Dublin Core. The following listing shows an example record encoded in `oai_dc`:

```
<record>
  <header>
    <identifier>ftubbiepub:oai:pub.uni-bielefeld.de:1680979</identifier>
    <datestamp>2016-02-21T23:44:21Z</datestamp>
  </header>
  <metadata>
    <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
               xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
               xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
               http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>Bielefeld Academic Search Engine (BASE) An end-user oriented
                institutional repository search service</dc:title>
      <dc:creator>Pieper, Dirk</dc:creator>
      <dc:creator>Summann, Friedrich</dc:creator>
      <dc:description>Purpose - The purpose of this paper is ...</dc:description>
      <dc:source>
        Pieper D, Summann F.: Bielefeld Academic Search Engine (BASE).
        An end-user oriented institutional repository search service.
```

(continues on next page)

(continued from previous page)

```
        Library Hi Tech. 2006; 24(4):614-619.
      </dc:source>
      <dc:source>ftubbiepub</dc:source>
      <dc:language>eng</dc:language>
      <dc:date>2006</dc:date>
      <dc:identifier>
        https://pub.uni-bielefeld.de/publication/1680979
      </dc:identifier>
      <dc:identifier>
        https://pub.uni-bielefeld.de/download/1680979/2535619
      </dc:identifier>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/doi/10.1108/07378830610715473
      </dc:relation>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/issn/0737-8831
      </dc:relation>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/wos/000242893300014
      </dc:relation>
      <dc:subject>Bielefeld Academic Search Engine</dc:subject>
      <dc:subject>ddc:020</dc:subject>
      <dc:type>info:eu-repo/semantics/article</dc:type>
      <dc:type>doc-type:article</dc:type>
      <dc:type>text</dc:type>
      <dc:type>121</dc:type>
      <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
    </oai_dc:dc>
  </metadata>
</record>
```

3.2 base_dc

The base_dc format extends the Dublin Core format with extra elements containing information that has been added or normalized by BASE. These elements are listed in the following table.

Namespace: http://oai.base-search.net/base_dc/

XML Schema: http://oai.base-search.net/base_dc/base_dc.xsd

Table 1: Additional XML elements of base_dc

Element	Value Format	Description
author_id	contains 2 XML elements:	<creator_name> and <creator_id>
autoclasscode	1-3 digit Dewey number	Automatically assigned Dewey number.
classcode	1-3 digit Dewey number	Manually assigned Dewey number.
collection	BASE collection name	Internal identifier of original repository.
collname	full collection name	Full name of the original repository.
continent	3 digit code (see below)	Continent of provenance (repository).
country	ISO 3166 country code	Country of provenance (repository).
creator_id	URI	ORCID iD written as a URL
creator_name	character string	repeats a name also given in <creator>
doi	URI	DOI (Digital Object Identifier) of this document
global_id	Free text (in UTF-8)	copy of *identifier* from the record header
lang	ISO 639-2/B language code	Three-letter document language code normalized by BASE, or unknown.
link	URI	Canonical link to the repository splash page.
oa	1 digit code	Open Access status (0 = Not Open Access, 1 = Open Access, 2 = unknown)
rightsnorm	controlled list see below	Licensing information normalized by BASE.
typenorm	alphanumeric code	Alphanumerically encoded normalized document type.
year	4 digit year	Normalized publication year.

Below is the example record from the previous section encoded in base_dc:

```
<record>
  <header>
    <identifier>ftubbiepub:oai:pub.uni-bielefeld.de:1680979</identifier>
    <datestamp>2016-02-21T23:44:21Z</datestamp>
  </header>
  <metadata>
    <base_dc:dc xmlns:base_dc="http://oai.base-search.net/base_dc/"
                xmlns:dc="http://purl.org/dc/elements/1.1/"
                xsi:schemaLocation="http://oai.base-search.net/base_dc/
                http://oai.base-search.net/base_dc/base_dc.xsd">
      <dc:title>Bielefeld Academic Search Engine (BASE) An end-user oriented
        institutional repository search service</dc:title>
      <dc:creator>Pieper, Dirk</dc:creator>
      <dc:creator>Summann, Friedrich</dc:creator>
      <base_dc:author_id>
        <base_dc:creator_name>Pieper, Dirk</base_dc:creator_name>
        <base_dc:creator_id>https://orcid.org/0000-0002-6083-9348</base_dc:creator_
 ↪id>
      </base_dc:author_id>
      <base_dc:author_id>
        <base_dc:creator_name>Summann, Friedrich</base_dc:creator_name>
        <base_dc:creator_id>https://orcid.org/0000-0002-6297-3348</base_dc:creator_
 ↪id>
      </base_dc:author_id>
      <dc:description>Purpose - The purpose of this paper is ...</dc:description>
      <dc:language>eng</dc:language>
      <dc:date>2006</dc:date>
```

(continues on next page)

(continued from previous page)

```
      <dc:identifier>
        http://pub.uni-bielefeld.de/publication/1680979
      </dc:identifier>
      <dc:identifier>
        http://pub.uni-bielefeld.de/download/1680979/2535619
      </dc:identifier>
      <base_dc:doi>https://doi.org/10.1108/07378830610715473</base_dc:doi>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/issn/0737-8831
      </dc:relation>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/doi/10.1108/07378830610715473
      </dc:relation>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/urn/urn:nbn:de:0070-pub-16809798
      </dc:relation>
      <dc:relation>
        info:eu-repo/semantics/altIdentifier/wos/000242893300014
      </dc:relation>
      <dc:subject>Bielefeld Academic Search Engine</dc:subject>
      <dc:subject>DDC:020</dc:subject>
      <dc:type>info:eu-repo/semantics/article</dc:type>
      <dc:type>doc-type:article</dc:type>
      <dc:type>text</dc:type>
      <dc:source>
        Pieper D, Summann F.: Bielefeld Academic Search Engine (BASE).
        An end-user oriented institutional repository search service.
        Library Hi Tech. 2006; 24(4):614-619.
      </dc:source>
      <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
      <base_dc:collection>ftubbiepub</base_dc:collection>
      <base_dc:collname>
        PUB - Publications at Bielefeld University
      </base_dc:collname>
      <base_dc:continent>ceu</base_dc:continent>
      <base_dc:country>de</base_dc:country>
      <base_dc:lang>eng</base_dc:lang>
      <base_dc:link>https://pub.uni-bielefeld.de/publication/1680979</base_dc:link>
      <base_dc:oa>1</base_dc:oa>
      <base_dc:typenorm>121</base_dc:typenorm>
      <base_dc:year>2006</base_dc:year>
    </base_dc:dc>
  </metadata>
</record>
```

4 Record Headers

4.1 identifier

The unique identifier in the record headers consists of the OAI identifier assigned by the original repository, prefixed with the internal BASE repository identifier. For instance, in the following example identifier

`ftubbiepub:oai:pub.uni-bielefeld.de:2083906`

the original identifier was `oai:pub.uni-bielefeld.de:2083906` and the prefix `ftubbiepub:` is BASE's internal name for the repository of Bielefeld University, "PUB".

4.2 `datestamp`

The `datestamp` element in the record header contains the date of harvesting into BASE.

4.3 Deleted Records

This OAI-PMH interface does **not** keep track of deleted records.

5 Date Ranges

OAI-PMH allows selective harvesting by date via the `from` and `until` parameters. In the BASE OAI-PMH API, the semantics of these dates is the *date of harvesting*, i.e., when the content was fetched from the original repositories into the BASE infrastructure. For instance, to get all contents that have been included into BASE from May to June 2012, you could use:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=oai_dc&from=2012-05-01&until=2012-06-30

If you would like to filter for the *date of publication* instead, please specify a dynamic set using the `date` field (see below).

6 Dynamic Sets

Traditionally, OAI-PMH structures content into collections by using sets. However, the protocol does not support combinations of multiple sets. To overcome this static nature of OAI sets, this API uses *dynamic sets* (inspired by DataCite's OAI interface).

This means that you can specify sets using the Solr query syntax `field:value`.

For instance, if you would like to filter for manually classified records from the field of economics from Germany, you could use the following set:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=oai_dc&set=classcode:33*+country:de

The supported fields are documented in the next section.

7 Indexed and Normalized Fields

BASE puts considerable efforts into the normalization of the (often heterogenously used) Dublin Core fields harvested from the original repositories. This section gives an overview of the queryable fields, their contents, and (if applicable) their normalization. The fields can be queried by using dynamic sets, as described in the previous section.

7.1 Overview of Indexed Fields

Table 2: Queryable fields.

Field	Value Format	Description
autoclasscode	1-3 digit Dewey number	Automatically assigned Dewey number.
classcode	1-3 digit Dewey number	Manually assigned Dewey number.
collection	BASE collection name	Original repository.
continent	3 digit code (see below)	Continent of provenance (repository).
contributor	Free text (in UTF-8)	Contributor to the publication
country	ISO 3166 country code	Country of provenance (repository).
creator	Free text (in UTF-8)	Author of the publication.
date	Free text (in UTF-8)	Date of publication.
deweyfull	1-3 digit Dewey number	Manually + automatically assigned Dewey numbers.
description	Free text (in UTF-8)	Abstract.
format	Free text (in UTF-8)	Document format (e.g., MIME).
identifier	Free text (in UTF-8)	Document identifier (e.g., URI).
lang	ISO 639-2/B language code	Document language as 3 letter code normalized by BASE, or "unknown".
language	Free text (in UTF-8)	Document language as in the original repository.
link	URI	Canonical link to the repository splash page.
oa	1 digit code	Open Access status (1 = "Open Access", 2 = "unknown")
person	Free text (in UTF-8)	Authors + contributors.
rightsnorm	controlled list see below	License information normalized by BASE.
subject	Free text (in UTF-8)	Subject headings.
title	Free text (in UTF-8)	Document title.
typenorm	alphanumeric code	Alphanumerically encoded normalized document type.
year	4 digit year	Normalized publication year.

7.2 Document Types

As the categorization of document types is very heterogenous across repositories, BASE normalizes them by mapping types into consistent categories which are identified by a numerical code. The table below lists normalized document types, which can be queried by using the field `typenorm`.

Table 3: Numeric codes for normalized document types.

document type	numeric code
text	1
book	11
book part	111
journal/newspaper	12
article in journal/newspaper	121
other non-article part of journal/newspaper	122
conference object	13
report	14
review	15
course material	16
lecture	17
thesis	18
bachelor thesis	181
master thesis	182
doctoral or postdoctoral thesis	183
manuscript	19
patent	1A
musical notation	2
map	3
audio	4
image or video	5
still image	51
moving image (video)	52
software	6
dataset	7
other/unknown material	F

Example Query:

Filter for books:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=typenorm:11

Encoding in the `base_dc` format:

```
<base_dc:typenorm>11</base_dc:typenorm>
```

7.3 Continents and Countries

BASE keeps track of the origins of its contents by storing the continent and country of the original repositories. Countries are encoded by using ISO 3166 country codes. Continents are encoded as shown in the following table:

Table 4: Codes for continents.

Continent	Code
Africa	caf
Australia	cas
Australia/Oceania	cau
Europe	ceu
North America	cna
South America	csa
Web server without geographic relation (org)	cww

Country Example Query:

Filter for documents from Germany:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=country:de

Encoding in the base_dc format:

```
<base_dc:country>de</base_dc:country>
```

Continent Example Query:

Filter for documents from North America:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=continent:cna

Encoding in the base_dc format:

```
<base_dc:continent>cna</base_dc:continent>
```

7.4 Subject Classification

The BASE index supports the Dewey Decimal Classification (DDC) for subject categorization. The assignment of Dewey classes to documents is established in two ways:

1. **manually** for contents from repositories that use the DDC.
2. **automatically** through machine learning-based document categorization.

Depending on their origin, the Dewey numbers are either stored in the field `classcode` (for manually assigned numbers) or `autoclasscode` (for automatically assigned numbers). The field `deweyfull` can be used for querying both manually and automatically classified documents.

Example Queries:

Filter for mathematical documents (manually and automatically classified):

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=deweyfull:51*

Filter for manually classified mathematical documents:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=classcode:51*

Filter for automatically classified mathematical documents:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=autoclasscode:51*

Encoding in the base_dc format:

```
<base_dc:classcode type="ddc">510</base_dc:classcode>
```

Conversely, if the class was assigned automatically, it would look like this:

```
<base_dc:autoclasscode type="ddc">510</base_dc:autoclasscode>
```

7.5 Open Access Status

BASE indexes the Open Access status of full text documents where this information is available. The status is stored numerically encoded in the field oa.

Table 5: Encoding of Open Access

Status Code	Description
0	Not Open Access
1	Open Access
2	Unknown

Example Query:

Filter for Open Access documents:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=oa:1

Encoding in the base_dc format:

```
<base_dc:oa>1</base_dc:oa>
```

7.6 Licensing Information

A great variety of values for the dc:rights field can be encountered in the wild. BASE maps those values it can recognise onto the following list of license codes. No attempt is made to recognise version numbers.

Table 6: Encoding of licensing information

Rightsnorm Code	Description
CC-BY-NC-ND	Creative Commons Attribution-NonCommercial-NoDerivatives
CC-BY-NC-SA	Creative Commons Attribution-NonCommercial-ShareAlike
CC-BY-SA	Creative Commons Attribution-ShareAlike
CC-BY-ND	Creative Commons Attribution-NoDerivatives
CC-BY-NC	Creative Commons Attribution-NonCommercial
CC-BY	Creative Commons Attribution
CC0	Public Domain Dedication
PDM	Public Domain Mark

Example Query:

Filter for Creative Commons Attribution (CC-BY) licensed documents:

http://oai.base-search.net/oai?verb=ListRecords&metadataPrefix=base_dc&set=rightsnorm:CC-BY

Encoding in the `base_dc` format:

```
<base_dc:rightsnorm>CC-BY</base_dc:rightsnorm>
```