Herbadrop project

A Quick User Guide
Searching data using HTTP REST API Version 3 (2017-12)

1. Introduction

As part of the Herbadrop project, CINES exposes an HTTP REST API in order to search the data associated to a deposit according to criteria on OCR results optionally combined with criteria on metadata and to retrieve an image searching by its identifier.

In this quick user guide, we will:

1. describe all the available parameters of this API.

2. give an example of how to use this web service.

2. REST API Interface description

The web service is available at the following URLs:

Function	URL	Protocol
Search in OCR result and in metadata	https://herbarium.cines.fr/access/rest/data/search	HTTP POST*
Retrieve the OCR result of an image	https://herbarium.cines.fr/access/rest/data/	HTTP GET
Get the image	https://herbarium.cines.fr/access/rest/image/	HTTP GET

* This REST API uses HTTP POST protocol with JSON format for the data.

2.1. Authentication

To use the REST API, you need to be authenticated using the HTTP Basic method. In

consequence, in your HTTP request you must specify the HTTP header "Authorization".

It must contain the « Basic » method and the Base64 representation of your user name and password separated by the character ":".

For security reason, you will use a connection encrypted by SSL to call the REST API.

2.2. REST API parameters

a) Request for a Full-Text search in OCR results

All the request parameters are transmitted using the JSON Format.

The following table contains the detailed description of all available parameters.

Category	JSON Parent	Parameter name	Description	Type	Required
Search		text	Full-text fragment you want to search in the OCR results.	String	Yes
		page	Page number of the result.	Integer	No (default 1)
		size	Number of results per page.	Integer	No (default 20)
		language	Dictionary in which the keyword will be search and which will appear in the response	String	No (default all)
		transferringAgency	The short name of the transferring agency.	String	No
		strictCharacterSearch	A flag used to specify that Full-Text search on OCR results must match exact words.	Boolean	Yes (default false)
		searchTextInAdditionalData	A flag used to specify that the scope of the search includes the	Boolean	Yes (default

Category	JSON Parent	Parameter name	Description	Type	Required
			OCR results.		true)
		searchTextInMetadata	A flag used to specify that the scope of the search includes the metadata.	Boolean	Yes (default true)
Personalization	highlight	preTag	Start tag in HTML used to wrap highlighted text. Example : “”	String	No
		postTag	End tag used to wrap highlighted text . Example : “”	String	No
		fragmentSize	The size (in characters) of the highlighted fragment.	Integer	No
		fragmentsCount	The maximum number of fragments in the result.	Integer	No
Metadata criteria	metadataCriteria		List of ‘criterion’ objects	List	No
		field	The full path of the field. See the list below. Example : “aip.meta.note”	String	Yes
		operator	The name of the operator. See the list below. Example : “CONTAINS”	String	Yes
		not	A flag used to specify that the condition must be inverted or not.	Boolean	Yes

Category	JSON Parent	Parameter name	Description	Type	Required
		values	A text describing one or more values separated by the '|' character. If value is a date, the ISO8601 format must be used. Examples : - "StillImage" - "2017-01-01|2017-12-31"	String	Yes

b) Request for a search in metadata

One ore more criteria can be added to search for values in the metadata.

Each criterion describes:

- a path to the field from one of the list described below,
- an operator from one of the list described below,
- a flag used to specify if the condition must be reverted or not,
- one or more values to search for.

Request parameters

The available paths for fields of the metadata are described in this table :

Path of the field	Description	Type
aip.dc.contributor	Contributor (part of the Dublin Core metadata)	String
aip.dc.coverage	Location of the collect. Format: 'Country|StateProvince|County|Municipality|Locality|VerbatimLocality'	String
aip.dc.creator	Collector name	String
aip.dc.description	Description of the specimen	String
aip.dc.endDate	Collect date (same as start date, specified using ISO8601	Date

	format)	
aip.dc.format	Image format or mime type Example: 'image/jpeg'	String
aip.dc.identifier	The unique identifier computed during the archiving process (aka ARK)	String
aip.dc.language	Language of the specimen or "und" if undefined	String
aip.dc.publisher	Name (or acronym) of the institution in charge of the specimen	String
aip.dc.rights	License associated to the specimen, Example: 'cc-by'	String
aip.dc.source	Event number and collect number separated by '|'.	String
aip.dc.startDate	Collect date (same as end date, specified using ISO8601 format)	Date
aip.dc.subject	Family of the specimen	String
aip.dc.title	Name of the specimen	String
aip.dc.type	Specimen type. Example: 'PreservedSpecimen' or 'StillImage.'	String
aip.files.format	The format of the image file (filled by the SipBuilder tool)	String
aip.files.name	The name of the image file (filled by the SipBuilder tool)	String
aip.files.originalChecksum	The checksum of the image file (computed and filled by the SipBuilder tool)	String
aip.files.originalChecksumType	The type of the checksum associated to the file (filled by the SipBuilder tool)	String
aip.meta.archivingDate	The arching date (specified using ISO8601 format)	Date
aip.meta.filePlan	The file plan of the institution Example: 'Herbarium'	String
aip.meta.producerIdentifier	Identifier of the specimen according to the repository of the institution	String
aip.meta.transferringAgency	The full name of the institution having sent the data	String
aip.meta.version	The version of the deposit	String

Note: The contents of these fields have been discussed during one of the workshop sessions.

Hereunder is the list of supported operators :

Name	Supported type(s)	Expected number of values
EQUALS	String, Date, Number	One value
CONTAINS	String	One value
MATCHES	String	One value
MATCHES_REGEX	String	One value
STARTS_WITH	String	One value
AFTER	Date, Number	One value
BEFORE	Date, Number	One value
BETWEEN	Date, Number	Two values

Note: The 'MATCHES_REGEX' operator uses an expression supported by the Lucene solution and is not fully compatible with the Perl expressions. To get more details, please refer to the documentation of our current indexing engine at 'https://www.elastic.co/guide/en/elasticsearch/reference/5.6/query-dsl-regexp-query.html#regexp-syntax'.

c) Examples of search requests

Example of a JSON query for a Full-Text search only on OCR results without specifying a language:

```
{
   "text":"Herbarium",
   "strictCharacterSearch":false,
   "searchTextInAdditionalData":true,
   "searchTextInMetadata":false,
   "page":1,
   "size":20
}
```

Example of a JSON query for a Full-Text search on OCR results and metadata using a particular language:

```
{
   "text":"Herbarium",
   "strictCharacterSearch":false,
   "searchTextInAdditionalData":true,
   "searchTextInMetadata":true,
   "page":1,
   "size":20,
   "language":"eng"
```

```
}
```

Example of a JSON query for a search only on metadata without specifying a language:

```
{
   "strictCharacterSearch":false,
   "searchTextInAdditionalData":true,
   "searchTextInMetadata":true,
   "page":1,
   "size":20,
   "language":"",
   "metadataCriteria":[
      {
         "field":"aip.dc.type",
         "operator":"CONTAINS",
         "not":"false",
         "values":[
            "StillImage"
         ]
      }
   ]
}
```

Example of a JSON query for a search on metadata combined with Full-Text search on OCR results, still without specifying a language:

```
{
   "text":"Herbarium",
   "strictCharacterSearch":false,
   "searchTextInAdditionalData":true,
   "searchTextInMetadata":true,
   "page":1,
   "size":20,
   "metadataCriteria":[
      {
         "field":"aip.dc.type",
         "operator":"CONTAINS",
         "not":"false",
         "values":[
            "StillImage"
         ]
      }
   ]
}
```

d) Response structure

The REST API returns a response in JSON format with the following structure:

JSON Parent	Parameter name	Description	Type
	maxScore	Indicates the max score of all the results :	Float
	total	Total number of results.	Integer
Result	Score *[1]	Score of the result.	Float
	depositId	Identifier of the submitted image corresponding to the OCR content.	String
	transferringAgency	Name of transferring agency who has transferred the image corresponding to the OCR content	String
	transferringAgencyId	Identifier of transferring agency who has transferred the image	String
Result/image	fileName	File name of the image corresponding to the OCR content.	String
	fileFormat	File format of the image corresponding to the OCR content.	String
Result/contentOcr	und	The OCR content processed with the **selected** (und, French, Spanish, German, Latin, English) dictionary. «und» means the result of the OCR processing with the 5 dictionaries used together. Note: Values are the ISO3 codes associated to the language.	String
	fra		
	spa		
	deu		
	lat		
	eng		
Result/highlight	contentOcr.fra	List of fragments containing the matching	String

JSON Parent	Parameter name	Description	Type
	contentOcr.all	term(s) according to the fields used for the Full-Text search. Fields can be one (or more) of the OCR results or one (or more) of the metadata according to the search parameters.	(HTML)
	contentOcr.deu		
	contentOcr.spa		
	contentOcr.eng		
	contentOcr.lat		
Result/metadata		The list of fields of metadata described as a pair, having a path and at least one value. Refer to the list of fields detailed below and the following examples.	Object

*[1] This score indicate the relevance of each result. The higher the score, the more relevant document.

The max_score value is the highest score of any document that matches the query.

The available fields of the metadata returned by this API are described in this table :

Path of the field	Description	Type
aip.dc.contributor	Contributor (part of the Dublin Core metadata)	String
aip.dc.coverage	The container describing the location of the collect per language (as ISO3) Example: { 'und' : 'Country|StateProvince|County|Municipality|Locality|VerbatimLocality' }	Object
aip.dc.creator	Collector name	String
aip.dc.description	The container describing the descriptions of the specimen by language (as ISO3)	Object
aip.dc.endDate	Collect date (same as start date)	Date
aip.dc.format	The container describing the formats (or mime types) of the image per language (as ISO3) Example: { 'eng' : 'image/tiff' }	Object
aip.dc.identifier	The unique identifier computed during the archiving process (aka ARK) Example: 'ark:/87895/1.herbadrop_test=1'	String

aip.dc.language	Language of the specimen or 'und' if undefined	String
aip.dc.publisher	Name (or acronym) of the institution in charge of the specimen	String
aip.dc.relation	*Not used in Herbadrop at this stage*	String
aip.dc.rights	The container describing the licenses of the specimen per language (as ISO3) Example: { 'und' : 'cc-by' }	Object
aip.dc.source	Event number and collect number separated by '|'. The values are described in a container per language (as ISO3) where the default language code is 'und'	Object
aip.dc.startDate	Collect date (same as end date)	Date
aip.dc.subject	The container describing the family of the specimen per language (as ISO3) Example: { 'lat': 'Amaranthaceae' }	Object
aip.dc.title	The container specifying the names of the specimen per language (as ISO3)	Object
aip.dc.type	The container describing the types of the specimen per language (as ISO3) Example: { 'eng': 'PreservedSpecimen|StillImage' }	Object
aip.files	The container describing a list of files objects. See lines below	Object
aip.files.compression	The information of compression	String
aip.files.format	The format of the image file (filled by the SipBuilder tool)	String
aip.files.formatVersion	The version of the format for the file (filled by the SipBuilder tool)	String
aip.files.name	The name of the image file (filled by the SipBuilder tool)	String
aip.files.note	The note associated to the image file (filled by the SipBuilder tool)	String
aip.files.checksum	The checksum of the image file (computed and filled during the archiving process)	String
aip.files.checksumType	The type of the checksum associated to the file (computed and filled during the archiving process)	String
aip.files.encoding	The encoding of the file Example: 'UTF-8'	String
aip.files.id	The unique identifier of the file computed during the archiving process Example: 'ark:/87895/1.herbadrop_test=2/2'	String
aip.files.originalChecksum	The checksum of the image file (computed and filled by the SipBuilder tool)	String
aip.files.originalChecksumType	The type of the checksum associated to the file	String

	(filled by the SipBuilder tool)	
aip.files.sizeInBytes	The size of the image file (in bytes) (filled during archiving process)	Long
aip.files.structure	The structure information associated to the image file (filled by the SipBuilder tool)	String
aip.meta.archivingDate	The arching date (specified using ISO8601 format)	Date
aip.meta.depositIdentifier	The deposit identifier	String
aip.meta.filePlan	The container with the file plan of the institution per language (as ISO3) Example: { 'eng' : 'Herbarium' }	Object
aip.meta.finalAction	The container describing the final action per language (as ISO3), reserved for archiving purpose.	Object
aip.meta.note	The container detailing the notes per language (as ISO3)	Object
aip.meta.pacIdentifier	Internal identifier of the deposit in the archiving solution	Long
aip.meta.previousVersion	The reference on the previous version of the deposit	String
aip.meta.producerIdentifier	Identifier of the specimen according to the repository of the institution	String
aip.meta.project	The project associated to the deposit	String
aip.meta.structure	The structure information associated tot the deposit	String
aip.meta.transferringAgency	The full name of the institution having sent the data	String
aip.meta.version	The version of the deposit	String

Example of JSON response:


```
{
   "maxScore":0.012074512,
   "total":2,
   "result":[
      {
        "depositIdentifier": "B100380787BIS",
        "metadata": {
          "aip.dc.contributor": "Some people",
          "aip.dc.coverage": {
            "und": "Yemen |||| Gov. Hadhramout. ...adi E of Alkadi al Beida. |"
          },
          "aip.dc.creator": "P. Hein",
          "aip.dc.description": {
            "und": "unavailable"
          },
          "aip.dc.endDate": "2002-09-08T00:00:00+0200",
          "aip.dc.format": {
           "eng": "Image/tiff"
```

```
},
"aip.dc.identifier": "ark:/87895/1.herbadrop_test=1",
"aip.dc.language": "und",
"aip.dc.publisher": "B",
"aip.dc.relation": {
  "eng": "relation"
},
"aip.dc.rights": {
  "und": "cc-by"
},
"aip.dc.source": {
  "und": "unavailable"
},
"aip.dc.startDate": "2002-09-08T00:00:00+0200",
"aip.dc.subject": {
  "lat": "Amaranthaceae"
},
"aip.dc.title": {
  "lat": "Chenopodiaceae"
},
"aip.dc.type": {
  "eng": "PreservedSpecimen|StillImage"
},
"aip.files": [
{
  "checksum": "ca1748e459d7102...78dd62d044c293292",
  "checksumType": "SHA-256",
  "encoding": "UTF-8",
  "format": "TIFF",
  "formatVersion": "NA",
  "identifier": "ark:/87895/1.herbadrop_test=1/1",
  "name": "B_10_0380787_bis.tiff",
  "originalChecksum": "d1d7760ef920ae98273bf9038000a4a7",
  "originalChecksumType": "MD5",
  "sizeInBytes": 24189012
}
],
"aip.meta.archivingDate": "2017-11-13T11:08:40+0100",
"aip.meta.depositIdentifier": "B100380787BIS",
"aip.meta.filePlan": {
  "eng": "/"
},
"aip.meta.finalAction": {
  "fra": "Conservation définitive"
},
"aip.meta.note": {
  "eng": "unaivalable"
},
"aip.meta.pacIdentifier": 1,
"aip.meta.previousVersion": "1",
```

```
            "aip.meta.producerIdentifier": "http://herbarium....object/B100380787",
            "aip.meta.project": "herbadrop_test",
            "aip.meta.transferringAgency": "BGBM",
            "aip.meta.version": "2"
          },
          "score": 0,
          "transferringAgencyIdentifier": "B100380787BIS"
        },
        {
           ...
        }
    ]
}
```


Note: The fields, metadata and languages are sorted in alphabetical order when possible.

e) Retrieve the indexed data associated to an image

You can retrieve the indexed data of an image by using the image identifier (depositIdentifier) from a HTTP GET request.

URL to use:

https://herbarium.cines.fr/access/rest/data/***<depositIdentifier>***

where <depositIdentifier> must be replaced by one of the deposit identifiers

Response structure

The REST API return a response in JSON format similar to the one returned y the search query excepted that only one result is returned.

Refer to the ‘Result’ details of the response section of the search request feature.

f) Get the image

This API provided a way to get the image as a binary content by using the image identifier(depositIdentifier) from a HTTP GET request.

URL to use:

https://herbarium.cines.fr/access/rest/image/***<depositIdentifier>***

where <depositIdentifier> must be replaced by one of the deposit identifiers

3. Examples of calls on the REST API

The examples described below use the 'curl' command available on Unixes (sometimes as an additional package). For windows operative systems, a binary can be downloaded at the URL: https://curl.haxx.se/download.html. Install and use of this third party at your own risk.

3.1. Full-text search in OCR results only

```
curl --user <yourusername>:<yourpassword> -k -H "Content-Type: application/json" -X POST
-d '{ "text" : "Paris", "strictCharacterSearch" : false, "searchTextInAdditionalData" :
true, "searchTextInMetadata" : false, "page" : 1, "size" : 3, "language" : "fra",
"highlight" : { "preTag" : "<em>", "postTag" : "</em>", "fragmentSize" : 10,
"fragmentsCount" : 2 }}' https://herbarium.cines.fr/access/rest/data/search
```

3.2. Full-text search in metadata only with criteria on metadata

```
curl --user <yourusername>:<yourpassword> -k -H "Content-Type: application/json" -X POST
-d '{ "searchTextInAdditionalData" : false, "searchTextInMetadata" : true, "page" : 1,
"size" : 3, "language" : "fra", "highlight" : { "preTag" : "<em>", "postTag" : "</em>",
"fragmentSize" : 10, "fragmentsCount" : 2 },"metadataCriteria":[{ "field":"aip.dc.type",
"operator":"CONTAINS", "not":"false", "values":[ "StillImage" ]}]}'
https://herbarium.cines.fr/access/rest/data/search
```

3.3. Full-text search in metadata and OCR results

```
curl --user <yourusername>:<yourpassword> -k -H "Content-Type: application/json" -X POST
-d '{ "text" : "Paris", "searchTextInAdditionalData" : true, "searchTextInMetadata" :
truee, "page" : 1, "size" : 3, "language" : "fra", "highlight" : { "preTag" : "<em>",
"postTag" : "</em>", "fragmentSize" : 10, "fragmentsCount" : 2 },"metadataCriteria":
[{ "field":"aip.dc.type", "operator":"CONTAINS", "not":"false", "values":
[ "StillImage" ]}]}' https://herbarium.cines.fr/access/rest/data/search
```

3.4. Retrieve the indexed data associated to an image

```
curl --user <yourusername>:<yourpassword> -k -H "Content-Type: application/json" -X GET
-v  https://herbarium.cines.fr/access/rest/data/P01742198
```

3.5. Get the image

```
curl --user <yourusername>:<yourpassword> -k -H "Content-Type: application/json" -X GET
-v  https://herbarium.cines.fr/access/rest/image/P01742198
```