ICEDIG.EU *Innovation and consolidation for large scale*

Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D6.2** / R / PU
Work package **6.3.2** / Lead: **CINES**
Delivery date Mars 2019

Digitisation infrastructure design for EUDAT / CINES

DELIVERABLE D6.2

CINES :

Nicolas Cazenave
Lorène Béchard
Olivier Rouchon

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

This deliverable specifies the requirements for adapting CINES & EUDAT services for long-term storage of large-scale digitised biodiversity data. The report describes the service(s) features, capacities, functions and costs, and its suitability. Recommendations of use and possible designs are also included. The report is structured in five parts:

1. The context section describes the setting of the pilot, including the institutions involved (sources and services providers) and the actors within those institutions, the processes covered, and services provided.
2. The infrastructure section depicts the facilities supporting the integration of service providers and consumers, particularly APIs, programming languages, protocols, and speeds.
3. The data section provides an overview of the data model used for storing digital specimen data requiring long-term preservation, including the minimal data required for making a deposit, as well as data and metadata, which are part of each digital specimen.
4. The design section describes the overall architectural view of the implemented solution.
5. Finally, the recommendation section provides practical recommendations to use the service(s) evaluated in the pilot, the benefits and trade-offs.

1. Context

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. The high resolution images of these specimens require substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using OCR, but the exploitation of this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts. Much of the information is only available using handwritten text and pattern recognition which is a less mature technology than OCR. Today our system provides the best OCR technology adapted to the requirements of herbarium specimen images and requires minimal installation in each institution. This is what we propose to make available to botanists with our portal.

The goal for a museum is to be able to submit a large number of scanned images easily in our long-term archiving system in order to automatically obtain OCR texts and retrieve them by a full text search on an open data portal.

Most of the images are provided through CC-BY licenses. In any case, the right associated to a data is specified in a specific metadata.

This pilot was an opportunity to test the long-term storage service provided by CINES. The services developed by EUDAT were used to facilitate the transfer of data to the storage repository and to provide indexing services for access to that repository.

The workflow deployed includes:

- a EUDAT B2SAFE node to transfer images;
- a EUDAT B2HANDLE service to get persistent Id.
- data quality function modules such as integrity and antivirus control, file format validation and metadata validation;

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- OCR analysis module performed on CINES's High Performance Computing (HPC) facilities ;
- indexation and access interface. The OCR results are indexed in full text in a search engine. This function provides means to monitor the process, and access to images, metadata and OCR results with a multi-criteria search capabilities.
- Long-term preservation performed on the CINES Trustworthy Digital Repository, which is compliant with the Core Trust Seal certification.

Table 1 Service Consumer Requirements

Customer(s)	Service(s) Required:	Data Set:
MNHN	Long term storage of digitised herbarium sheets (images) compatible with open access policies.	Herbadrop Dataset (4,628,651) unique herbarium sheet images.

Table 2 Main Services Consumed

Provider	Network	Services	Purpose
EUDAT	EOSC	B2SAFE + B2SHARE + B2HANDLE	Transfer Data to long term repository
CINES	NI	Archive	Long term data archiving
CINES	NI	HPC	OCR, QA/QC (Checksum), Indexing
EUDAT	EOSC	B2FIND	Data indexing (elasticsearch)
CINES	NI	CINES portal opendata	Data indexing (elasticsearch)
CINES	NI	User portal	Access repository
CINES	NI	Administration portal	Use the monitoring

NI: National Infrastructure
EOSC: European Open Science Cloud

Certifications:

Services described in this report are about to obtain the Core Trust Seal certification. The scope of the certification includes both CINES and EUDAT B2Safe infrastructure, and data supplied by the MNHN.
Once the Core Trust Seal will be delivered, the submission report will be freely accessible in their website: https://www.coretrustseal.org/why-certification/certified-repositories/
For example, at the moment you can find the certification of the main digital repository managed by the CINES: https://assessment.datasealofapproval.org/assessment_160/seal/pdf/

Business model:

Costs calculated based on two principles

- Level of service
  - SL1 - Copies on fast storage (e.g. disks)
  - SL2 - Copies on cold storage (eg. tapes)
- Size/volume
  - V1 - <10To
  - V2 - 10To-100To
  - V3 - >100To
- Model for 400To capacity
  - V1 → annual costs of HR (contractors) + H/W + S/W included in SL1, SL2

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- V2 → 50% of (annual costs of HR (contractors)) + H/W + S/W included in SL1, SL2
- V3 → H/W + S/W included in SL1, SL2

Annual fee based on storage space used

2. Infrastructure

The EUDAT services were used to speed up the implementation as they are mature "ready-to-use" tools which have been in production for quite a while, and provide data transferring functionalities. These services expose well-documented APIs, which were exploited in transferring and indexing datasets.

Table 3 Upload Services Details

Provider	Services	Function	Description
EUDAT	B2SAFE	Upload	See the "How to" section at the end of this document
EUDAT	B2HANDLE	Process	The B2HANDLE service is set up to retrieve a PID handle allowing the B2FIND service to come and retrieve metadata from CINES and post it on their portal.
CINES	SIPBUILDER	Process	This tool enables the user to create a package and automatically retrieve metadata from the GBIF database, create a tar.gz file to be sent in the Icedig workflow.

Table 4 Processing Services Details

Provider	Services	Function	Description
CINES	Ingest	Process	Internal process Java code
CINES	Archive	Store	Internal process Java code -Arcsys Software
CINES	HPC	Process	All the OCR checksum and validation processes of formats is performed on the OCCIGEN supercomputer hosted and operated at CINES. The Bull Occigen cluster is a parallel scalar supercomputer with a maximum theoretical power of 3.5 Pflops. With the use of the new HPC machine, the quality of the OCR results was optimized with an "improved" mode for Tesseract.
EUDAT	B2FIND	Upload	Use Webdav for the pictures

Table 5 Access Services

Provider	Services	Function	Description
EUDAT	B2FIND	Index	Use the Cines Elasticsearch index
CINES	User portal	Access/Search	Use the Cines Elasticsearch index
CINES	User portal	Download	Web site and API for downloading

3. Data model

The data is divided in archive metadata and object metadata (business data). Archive data refers to the data used to organise and identify the digital objects deposited (Institution, licensing, identifiers,

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

project, etc). Object data (business data) refers to the data and metadata, which the depositor considers relevant for the type of data objects being stored.

Field	Type	Parent	Description	Example
Record	**Element**	**Resource**		
Basis of record	Element	Record	Type of record	StillImage¦PRESERVED_SPECIMEN
Institution code	Element	Record	Institution code	MNHN
Occurrence Id	Element	Record	link to institutional record	http://coldb.mnhn.fr/catalognumber/mnhn/p/p00433086
License	Element	Record	Licensing	cc-by
Taxon	Element	Resource		
Family	Element	Taxon		Euphorbiaceae
Scientific name	Element	Taxon		Mallotus oppositifolius Müll.Arg.
Origin	Element	Resource		
Recorded by	Element	Origin	Collector name	Peltier, M.
RecordNumber	Element	Origin		
Event date	Element	Origin	Collection date	1964-12-29
FieldNumber	Element	Origin		
Field notes	**Element**	**Origin**		**unavailable**
Location	Element	Origin		Madagascar Pont de Kamoro (R.N.4)
Long term preservation	**Element**	**Resource**	**Details about transferring institution**	
Transferring agency	Element	Long term preservation	Institution code	MNHN
Archiving date	Element	Long term preservation	Archiving date	2018-10-01
Archiving Id	Element	Long term preservation	archive id	ark:/87895/1.90-366003
Licence	Element	Origin		« cc-by »
File information	Element	Resource		
File name	Element	File information	Name of image file	P00433086.jpg
File format	Element	File information	Image format	JPEG/1.01

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

4. Design

This section describes the overall architectural view of the implemented solution. The diagram in Figure 1 shows how the integration between EUDAT and CINES services as well as the main data flow paths, which are indicated with arrows.

The Icedig architecture is split into a sequence of functions that process one-step of the workflow. The image replication uses the EUDAT B2SAFE service. B2HANDLE is required for PID (Persistent Identifier) generation and then to guarantee data access through the B2FIND portal. The B2FIND portal and API provide user with advanced search functionalities and allow access to the data resources associated to the metadata found in the catalogue.

The infrastructure has been developed specifically for the Herbadrop pilot using the existing CINES Trustworthy Digital Repository, which is OAIS (ISO 14721) compliant. Since the end of the pilot, he functionalities of the platform have been continuously improved, particularly for the validation of archives and the processing times. Once transferred, the images have to pass several quality controls. The main ones are detailed below:

- The first step is the file integrity check. This is performed during the transfer phase by forcing all files to be controlled by the iRODS protocol. The checksum is calculated before and after the upload operation and the values are compared. If they are different, the operation returns an error to the transfer initiator.
- The second is the antivirus check, which is done for each file.
- Then, the packaging of the deposit (including metadata) is checked to detect any potential non-compliance, in particular with the scheme of the deposit form. The structure of the package must comply with a given specification with a deposit form in XML to declare metadata.
- Finally, the files format has to be validated against a list of accepted formats.

If all the checks above are successful, an automatic OCR generator processes the image. Once completed, the image is stored in the CINES archiving platform, and the long-term preservation process kicks-off.

Regarding the security, many solutions have been put in place to provide a sufficient level of protection at different levels (physical/logical access, power backup, etc.):

- A decree published in 2015 officially established the CINES as a ZRR (Zone à Régime Restrictif), meaning that, besides enhancing safety of the information infrastructure, the physical access to the site is restricted. The access to the machine rooms inside the building is also limited to a list of identified persons. Every access is logged.
- The power supply relies on two distinct electric circuits. In case of an outage, two UPS provide short-term power (10-15 minutes). If the problem is not fixed during that period, two power-generators start automatically to provide the electricity required by the infrastructure.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

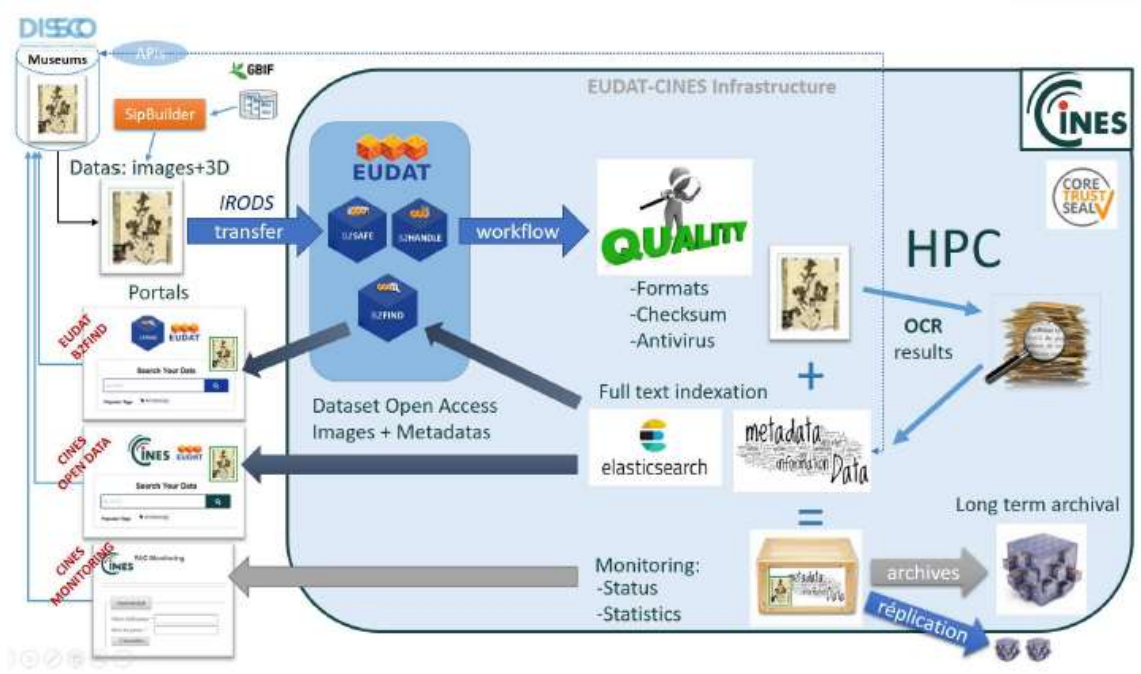

Figure 1 EUDAT/Cines Workflow

5. Recommendation – How to.

Digitized herbariums images are transferred to the CINES repository using a data synchronisation and exchange service, B2SAFE. A dedicated storage space on the CINES infrastructure is made available to each museum for uploads. It consists of a specific iRODS collection with full read write permissions which can be accessed using a login and a password provided by CINES. They IP addresses of the servers that will interact with the platform have to be provided prior to the submission, asthey will be checked during the login phase.

The transfer consists of the ingest of packages (SIP, Submission Information Package) containing one single image, with a limited number of metadata elements.

B2SAFE/IRODS: How to transfer data

Install the iRODS client with provided packages

Go to https://irods.org/download

Choose the appropriate version under the section « iCommands CLI »

Compile the irods client (only if you cannot find your distrib under iCommands CLI)

Go to : https://github.com/irods/irods_client_icommands

```
git clone git://github.com/irods/irods_client_icommands
```

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

```
cd irods_client_icommands
git submodule init
git submodule update
./packaging/build.sh icommands # for a standard installation
./packaging/build.sh icommands --run-in-place # for testing purposes
```

User accounts

iRODS user accounts have been created (one per institution) :

```
mnhnftp, rbgeftp, bgbmftp, naturalisftp, digitariumftp and cines
```

In the following section, choose on of the user name above instead of myaccount.

Inializing a connection

Once you have properly done the installation of the iRODS client, you will be prompted, the first time, to answer these questions after running the command iinit.

```
$ iinit
One or more fields in your iRODS environment file (irods_environment.json)
are missing; please enter them.
Enter the host name (DNS) of the server to connect to: ariane.cines.fr
Enter the port number: 1247
Enter your irods user name: myaccount
Enter your irods zone: CINES
Those values will be added to your environment file (for use by other
iCommands) if the login succeeds.

Enter your current irods password: myaccount(withoutftp)pwd (eg. mnhnpwd,
rbgepwd, bgbmpwd, naturalispwd, digitariumpwd)
```

Several files under a directory called .irods should be created. One is a token containing your credentials (.irodsA), and another one contains the connection parameters with the values you entered (irods_environment.json).

Thus, the next time you connect, you wont be promted to enter the connection information.

Change password

You may want to choose your own password.

```
$ ipasswd
Enter your current iRODS password:
```

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

File naming convention

The images must be transferred in the root directory of the user home (eg. /CINES/home/myaccount/.)

The images files are named in the following way.

ABC	1	2	3	4	5	6	7	8	_	s	.	jpg
Catalog number ABC (some letters) +8 digits										suffixe		extension

The prefixed letters (eg. ABC) are different from an institute to another and the number of letter is indefinite. Are they always the same for a given institute ?

The name of the files explicitely include the catalog number and optionally a suffixe.

The recognized extensions are (jpg anf tiff).

The naming is case insensitive, however the archive process will convert all names into caps for the name and extenxion in lower case.

It is recommended to use uppercase for prefix letters ABC and lowercase for extensions.

No dots (.) should be used in the name except the one separating the ID and the extension.

Another character than _ may be used (except dot .) but in the future this may be more complicated to identify suffixes.

Exemples of recommended naming formats

```
Good file formats
ABC12345678_a.jpg
ABC12345678.tiff
Not recommended or bad names
ABC12345678-a.jpg
ABC.12345678.jpg
ABC12345678.a.jpg
```

Transferring files

The transfer of files can be performed by one of the command iput or irsync.

The option -K (integrity check, it calculates and verifies the checksum) must be use.

Single transfer with iput :

```
$ iput -K ABC12345678_a.jpg
$ ils
```

Multiple transfer with irsync :

eg. you want to transfer all files in your local dir mydir/

```
mydir/
      ABC12345678.jpg
```

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

```
ABC12345678_a.jpg
```

```
$ irsync -rK mydir i:/CINES/home/username
```

Viewing all metadata associated with an image

```
$ imeta ls -d ABC12345678_a.jpg

attribute: ID                          # eg. ABC12345678_a
attribute: EXTENSION  # eg. jpg, tiff
attribute: OWNER                # eg. mnhnftp
attribute: STATUS     # eg. TRANSFER_OK, OCR_OK, ERROR
attribute: PID        # persitent identifier provided by EUDAT EPIC service
                                                                [note
that the PID will be added in the future]
```

Monitoring the progress of the workflow

```
$ imeta ls -d ABC12345678_a.jpg STATUS
attribute: STATUS
value: TRANSFER_OK            # possible values are currently TRANSFER_OK
and OCR_OK
```

View raw ocr results

```
$ ils -l /CINES/home/herbadrop

/CINES/home/herbadrop/ABC12345678

When the OCR will be performed :

$ iget /CINES/home/herbadrop/ABC12345678/ABC12345678.ocr
```

SipBuilder tool: Create submission information package (SIP)

The SipBuilder is a tool developed by the CINES to enable the museum information system to create the final submission information package (SIP) which contains images and metadata.

The application detects the new file as soon as it is deposited and then initializes the main processing.

The processing contains the following stages:

- retrieval of some metadata associated to the deposit by requesting an external rest API (GBIF),
- build of the archive structure according to the configuration and provided mapping,
- export of the final package on the local file-system.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

All the partners have agreed on a mapping between DarwinCore metadata and CINES metadata (based on DublinCore). The tool queries the GBIF API to collect the DarwinCore metadata using an unique identifier corresponding to a single resource (i.e. one herbarium image). First, a file that contains all the DarwinCore metadata of the resource is produced. Integrated to the data package as "community metadata", this TXT file complies with the JSON format. An XML file is generated from the approved mapping and is used as the deposit form for the digital repository. Its content is indexed in the indexing engine so that a cross-reference is possible with the OCR results. Then, the tool formats the data (the image + the 2 metadata files) in a package of which structure follows the digital repository requirements. The folder created is packed into a .tar.gz file; that file is then made available to the user to be sent in the Icedig workflow.

This freely licensed tool is made available to the partners to help them associate the metadata with the data to be preserved. We used the GBIF portal to unify the interfacing with the platform. This allowed all museums displaying their data on the GBIF to automatically have a connector for archiving at CINES. In addition, GBIF also performs a small qualitative work on metadatas. But the SIPbuilder being outside our channel, it is quite possible for the services that wish to provide us the data from their own systems or the future unified interface of DissCo. The SIPbuilder can then be tailored by museums showing interest, as the code is in open access.

How to Use CINES API for download

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://opendata.cines.fr/herbadrop-api/rest/data/search	HTTP POST*

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Retrieve the OCR result of an image	https://opendata.cines.fr/herbadrop-api/rest/data/	HTTP GET
Get the image	https://opendata.cines.fr/herbadrop-api/rest/image/	HTTP GET

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

2.1.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	Require
Search		text	Full-text fragment you want to search in the OCR results.	String	Yes
		page	Page number of the result.	Integer	No (default
		size	Number of results per page.	Integer	No (default
		language	Dictionary in which the keyword will be search and which will appear in the response	String	No (default
		transferringAgency	The short name of the transferring agency.	String	No

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

		strictCharacterSearch	A flag used to specify that Full-Text search on OCR results must match exact words.	Boolea	Yes (default
		searchTextInAdditionalDa	A flag used to specify that the scope of the search includes the OCR results.	Boolea	Yes (default
		searchTextInMetadata	A flag used to specify that the scope of the search includes the metadata.	Boolea	Yes (default
Personalizatio	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 ia	metadataCrite		List of ‘criterion’ objects	List	No

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

		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.	Boolea	Yes
		values	A text describing r more values ated by the ‘|’ cter. If value is a the ISO8601 format be used. Examples : - “StillImage” - “2017-01-01|2017-12-	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.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

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|VerbatimLocali	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 at)	Date
aip.dc.format	Image format or mime type Example: 'image/jpeg'	String
aip.dc.identifier	The unique identifier computed during the archiving ess (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 men	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 at)	Date
aip.dc.subject	Family of the specimen	String

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

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 nstitution	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

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

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

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

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
	additionalIdentifiers	Collection of additional identifiers	Object
	additionalIdentifiers,type	Type of the identifier	String
	AdditionalIdentifiers.identifier	Value associated to the identifier	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	String
	fra		
	spa		
	deu		
	lat		

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

	eng	associated to the language.	
Result/highlight	contentOcr.fra	List of fragments containing the matching 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.	String (HTML)
	contentOcr.all		
	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

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

aip.dc.coverage	The container describing the location of the collect per uage (as ISO3) Example: { 'und' : try|StateProvince|County|Municipality|Locality|VerbatimLocality' }	Object
aip.dc.creator	Collector name	String
aip.dc.description	The container describing the descriptions of the specimen by uage (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 e per language (as ISO3) Example: { 'eng' : 'image/tiff' }	Object
aip.dc.identifier	The unique identifier computed during the archiving process 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 uage (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) e 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 uage (as ISO3) Example: { 'lat': 'Amaranthaceae' }	Object

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

aip.dc.title	The container specifying the names of the specimen per ıage (as ISO3)	Object
aip.dc.type	The container describing the types of the specimen per ıage (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 ess Example: 'ark:/87895/1.herbadrop_test=2/2'	String
aip.files.originalChecksu	The checksum of the image file (computed and filled by the SipBuilder tool)	String

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

aip.files.originalChecksu)e	The type of the checksum associated to the file (filled by the SipBuilder tool)	String
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.depositIdentifi	The deposit identifier	String
aip.meta.filePlan	The container with the file plan of the institution per language O3) Example: { 'eng' : 'Herbarium' }	Object
aip.meta.finalAction	The container describing the final action per language (as), 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.previousVersio	The reference on the previous version of the deposit	String
aip.meta.producerIdenti	Identifier of the specimen according to the repository of the ution	String
aip.meta.project	The project associated to the deposit	String
aip.meta.structure	The structure information associated tot the deposit	String
aip.meta.transferringAg	The full name of the institution having sent the data	String
aip.meta.version	The version of the deposit	String

Example of JSON response:

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

```
{
   "maxScore":0.012074512,
   "total":2,
   "result":[
       {
       "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": "/"
       },
```


Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

```
        "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": "Agency1",
            "aip.meta.version": "2"
        },
          "score": 0,
          "depositIdentifier": "B100380787BIS",
          "transferringAgencyIdentifier": "agency1ftp",
          "additonalIdentifiers": [
        {
            "type": "HANDLE",
            "identifier": "APIDHANDLE1"
        }
        ]
        },
        {
        ...
        }
    ]
}
```


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 (depositId) and the transferring agency identifier from a HTTP GET request.

URL to use:

https://opendata.cines.fr/herbadrop-api/rest/data/**<transferringAgencyId>**/***<depositId>***

where <transferringAgencyId> must be replaced by the transferring agency associated to the archive,

<depositId> 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.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

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 (depositId) and the transferring agency identifier from a HTTP GET request.

URL to use:

https://opendata.cines.fr/herbadrop-api/rest/image/**<transferringAgencyId>**/***<depositId>***

where <transferringAgencyId> must be replaced by the transferring agency associated to the archive,

<depositId> 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://opendata.cines.fr/herbadrop-api/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://opendata.cines.fr/herbadrop-
api/rest/data/search
```

3.3. Full-text search in metadata and OCR results

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

```
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://opendata.cines.fr/herbadrop-api/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://opendata.cines.fr/herbadrop-api/rest/data/agencyftp/P01742198
```

3.5.Get the image

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

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483