D2.3 EcoTaxa data service

Work Package	WP2, Developing the Blue-Cloud discovery and access service and overall Blue-Cloud architecture
Lead Partner	MARIS
Lead Author (Org)	Sorbonne Université
Contributing Author(s)	Jean-Olivier Irisson (SU), Laurent Salinas (SU)
Reviewers	Dick M.A. Schaap (MARIS), Francesca Spagnoli (TRUST-IT)
Due Date	30-11-2020, M14
Submission Date	04-12-2020, M15
Version	0.6

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

DISCLAIMER

"Blue-Cloud, Piloting Innovative services for Marine Research & the Blue Economy" has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue-Cloud services, Grant Agreement n.862409.

This document contains information on Blue-Cloud core activities. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the Blue-Cloud Consortium, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

COPYRIGHT NOTICE

VERSIONING AND CONTRIBUTION HISTORY

Version	Date	Authors	Notes
0.1	17.11.2020	SU	First version
0.2	26.11.2020	SU	Review page layout
0.3	30.11.2020	MARIS	Reviewed and accepted as Technical Coordinator
0.4	30.11.2020	SU	Minor extensions based on review
0.5	01.12.2020	Trust-IT	Reviewed by the project coordinator
0.6	04.12.2020	SU	Further corrections based on reviews
1.0	04.12.2020	Trust-IT	Finalised for EU submission

Contents

1 Principle of EcoTaxa

1.1 General overview

EcoTaxa is a web application dedicated to the classification of large numbers of images of individual objects among a taxonomy of terms. It is currently mostly dedicated to the classification of images of planktonic organisms and marine snow, taken by quantitative imaging instruments, along the tree of life or within custom morphological categories. It produces ecological data in the form of concentration and biovolume of organisms in a given taxon, at a given station (lat, lon, depth, time). To speed up the classification, EcoTaxa assists the users through an ergonomic interface and the use of artificial intelligence. It was born from the experience developed at Laboratoire d'Océanographie de Villefranche (LOV) regarding the quantitative, high-throughput imaging of plankton and from the Oceanomics project, which covered the exploitation of data collected during the Tara Oceans cruise. It is now developed through the WWWPIC project funded by the Belmont Forum and, of course, as part of the Blue-Cloud project.

The main steps in the usage of EcoTaxa are:
1. Create a **project**, which is a collection of images.
2. **Import** images along with a metadata table containing latitude, longitude, date, time, depth but also descriptors computed from this image that describe the organism (such as its body length or is Equivalent Spherical Diameter).
3. **Classify** a few images in relevant classes by dragging and dropping them on the appropriate class name or by typing it.
4. **Train** an artificial intelligence classifier based on the manually classified images or on any other EcoTaxa project with similar images that is already appropriately sorted. The classifier proposes a name for each image.
5. **Validate or correct** the classifications proposed by the classifier.
6. Repeat steps 3 to 5 until all data is correctly classified.
7. **Export** a table of data that contains the same information as the imported one, plus the name (and lineage, and classification timestamp, etc.) of each image.

1.2 Data model

Data in EcoTaxa is organised along a hierarchical data model:
- **Projects**: a cruise or a leg of a cruise, a year in a time series; each project uses a single imaging instrument (so images are comparable).
- **Samples**: a location, a date at which water was sampled
- Acquisitions: an image acquisition event; several acquisitions of the same sample can be done with various lights, on various size fractions, etc.
- **Processes**: an image processing event; the same acquisition can be processed with various settings, leading to several processes per acquisition
- **Objects**: individual organisms/objects extracted from the image by the process.
- **Images**: an image of the object; there can be several images per object.

In addition, **Collections** allow aggregating several compatible projects into conceptual data units: aggregate several legs of a cruise into a cruise-wide collection, several years of a time series into a full series collection, etc. Collections allow keeping some granularity at project level for everyday management (access rights, monitoring of the completeness of sorting etc.) while showing which go together conceptually to form a dataset.

At several levels within a project, the user can store metadata such as:

- Samples: name of the operator, sea state, etc.
- Acquisition: settings of the sensor, size fraction targeted, etc.
- Process: software version, settings for the segmentation, etc.
- Objects: measurements of the object on the image: area, length, mean grey level, etc.

To allow for maximum flexibility, those metadata fields are not prescribed: the user can import anything and it will be stored, possibly used to sort/filter images in the interface, and re-exported. Those fields are therefore not the same from project to project. However, because many projects are prepared with the same processing toolbox, they often end up being homogeneous, at least per instrument. In addition, some key fields are standardized across the whole database, allowing to query data and extract ecologically usable datasets: latitude, longitude, depth, date and time.

1.3 User interface

The main page of the EcoTaxa application is the classification page. It has four sections:
1) A **header** (top) that gives status information and presents menus
2) A **filters** section (left) which has two modes:
   a) Taxonomy: allows to select a class or assign a class by drag and dropping images on it
   b) Other filters: filter images according to some criteria; see below.
3) A **tool bar** (top of the images section) that allows to sort images according to some criteria, display metadata under each image, define the density and scale of the images displayed etc.
4) An **image browser** (main content) that displays images and some extra data. Images displayed with a green background have been validated by users; those with a blue background are proposed by the artificial intelligence classifier.

Figure 1: The classification page of EcoTaxa

D2.3 EcoTaxa data service

Figure 2: Object filters available within EcoTaxa

The filter section allows to filter images according to the following criteria:

Sample name/code, with some facility to match several samples with a regular expression.

Depth of the objects.

Rectangle for the **location** of the objects.

with the possibility of drawing it on a map.

Date of collection of the object.

Time of collection of the object, with a facility to extract day, night, dawn and dusk samples (sun position computed from lat, lon and time).

Kind of **instrument** used to image the object. This is homogeneous per project but is useful across projects.

Person who classified the object

and time of this classification

As well as a slot to filter according to a field uploaded by the **user** (and therefore different from project to project)

D2.3 EcoTaxa data service

1.4 EcoTaxa today

Currently, EcoTaxa hosts 130 million objects, 55 millions of which have been classified by a human operator. Those objects are organised in ~2000 projects, have been uploaded from ~300 organisations and classified by ~900 users. Those objects are contained in samples that come from all over the world

Figure 3: Map of the samples currently hosted by EcoTaxa.
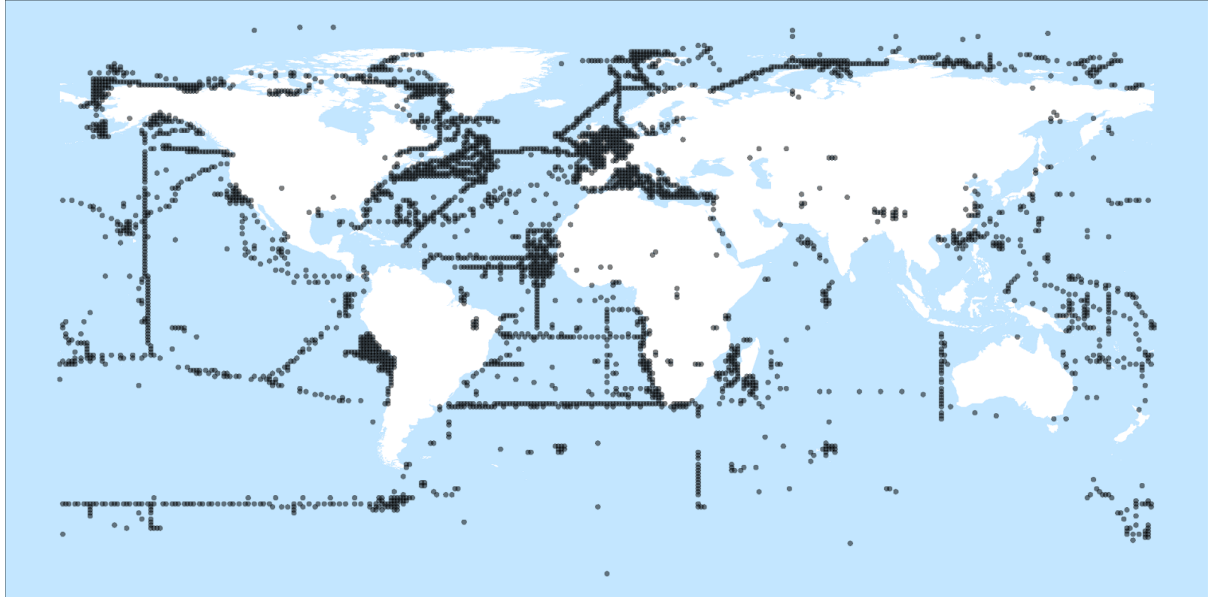
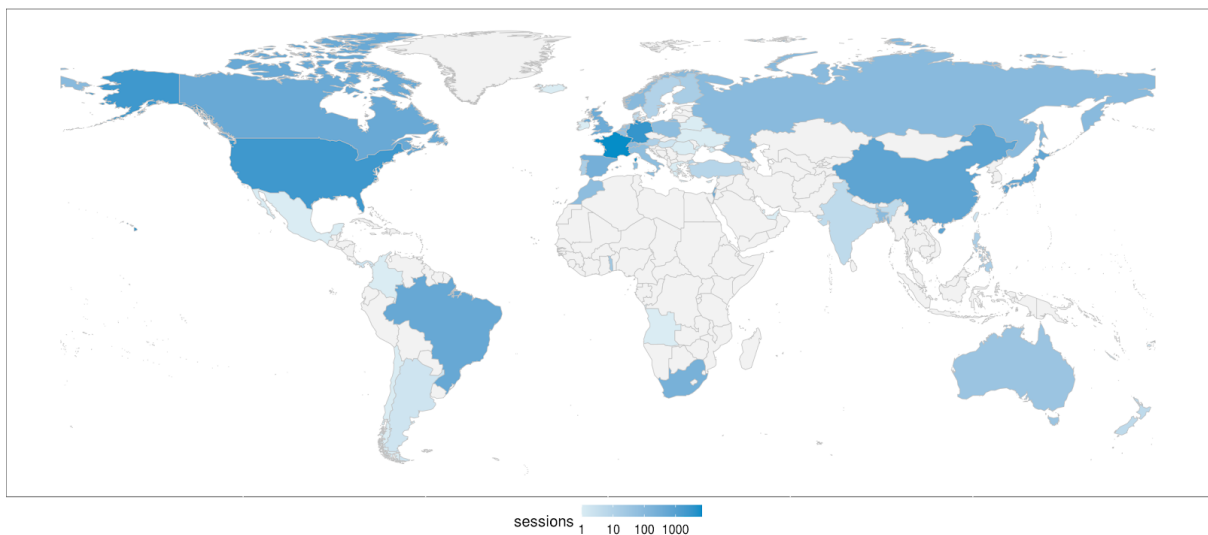And the classifications are also done by users from all over the world.

Figure 4: Choropleth of the number of user sessions per country since 2020-01-01.

NB: the colour scale is logarithmic.

Institutionally, EcoTaxa is operated as a service of EMBRC-France, the French branch of EMBRC-ERIC, a European Research Infrastructure Consortium.

D2.3 EcoTaxa data service

2 EcoTaxa in Blue-Cloud

Within Blue-Cloud, EcoTaxa will be used by, at least, the zooplankton part of Demonstrator 1: Zoo and phytoplankton essential ocean variables and Demonstrator 2: Plankton genomics. It will serve as a data *processing* facility and data *source*. Moreover, it will become part of the offer of the Blue-Cloud Data Discovery & Access service by which external users can also discover and use EcoTaxa data.

The initial plan was to integrate EcoTaxa data as a service of EuroBioImaging, for the Blue-Cloud access and discovery service to browse the relevant datasets from there. However, while EuroBioImaging is about images, it deals mostly with *biological* images (i.e. small number of large, high resolution images such as micrographs). EcoTaxa is about *ecological* images (i.e. large number of small, relatively low-resolution images). Furthermore, the data created by EcoTaxa are counts or concentrations of organisms per lat/lon/time point, which is typical of ecological/biodiversity studies.

Such biodiversity data is what EurOBIS collects and publishes. It was therefore decided to provide *summarised data* from EcoTaxa (concentrations per sample = lat/lon/time point) to EurOBIS, rather than all images to EuroBioImaging, and that the Blue-Cloud access and discovery service would fetch the metadata about those datasets from EurOBIS.

EurOBIS will only be publishing "Collections" (i.e. datasets) to the Blue-Cloud access and discovery service; those would map to EcoTaxa "Collections". Therefore, it was also decided that an Application Programming Interface (API) on EcoTaxa's side should be developed to allow browsing *lower-level* information (projects to objects). Blue-Cloud users should be able to discover the EcoTaxa datasets through EurOBIS and then, if more details are needed, turn to EcoTaxa to get them at granule (i.e. sample or object) level.

Not all datasets in EcoTaxa will be uploaded to EurOBIS; the choice to do so remains at the hands of the data *owners*, not at Sorbonne Université or EMBRC who *operate* EcoTaxa but do *not own* all the data. However, to demonstrate the data flow and capabilities, the datasets corresponding to the Tara Oceans cruise as well as the long-term plankton monitoring time series of Villefranche-sur-Mer will be uploaded. This amounts to several thousand data points containing approximately 5 million images. The goal is that, once such a data flow has been facilitated by the developments made in Blue-Cloud, it will generalise to the wider community. So far, two Tara datasets have been uploaded to EurOBIS.

3 EcoTaxa's Application Programming Interface

The API of EcoTaxa is an OpenAPI/Swagger API. Its specification is therefore open and it benefits from many functionalities from this widely used toolbox.

One such facility is automatic documentation, browsable at https://ecotaxa.obs-vlfr.fr/api/docs. The interested user will browse this URL to find the latest API endpoints and the parameters they require. Only the functionalities relevant to Blue-Cloud are presented below.

3.1 Authentication

login endpoint: Provide a username and password to get an authentication token.

Example: Log in with a default user

```
$ curl -s -X POST 'https://ecotaxa.obs-vlfr.fr/api/login' \
> -d '{"username":"ecotaxa.api.user@gmail.com","password":"test!"}'|jq .
"eyJ1c2VyX2lkIjo5OTN9.X7BjGQ.iI_3kjlfKRuMwwW9CqcfK81RR8s
```

NB: jq is just a tool to prettify the JSON output of the command.

users/me endpoint: Shows information about the current authenticated user.

Example: Show the information of this default user

```
$ curl -s -X GET 'https://ecotaxa.obs-vlfr.fr/api/users/me' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
{
  "id": 993,
  "email": "ecotaxa.api.user@gmail.com",
  "name": "Test API user",
  "organisation": "Oceanographic Laboratory of Villefranche sur Mer - LOV",
  "active": true,
  "country": "France",
  "usercreationdate": "2020-11-05T12:31:48.299713",
  "usercreationreason": "Test API"
}
```

3.2 Collections

collections/search endpoint: search collections by title.

Example: Search all collections with "tara" in their name.

```
$ curl -s -X GET 'https://ecotaxa.obs-vlfr.fr/api/collections/search?title=%tara%' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
[
  {
    "project_ids": [
      395,
      397
    ],
```

= list of the ids of projects contained in the collection

```
    "provider_user": {
      "id": 4,
      "email": "irisson@obs-vlfr.fr",
      "name": "Jean-Olivier Irisson",
[...]
    },
    "contact_user": {
```

```
[...]
    },
    "creator_users": [
      {
[...]
      },
      {
[...]
      },
[...]
```

= specification of the users (same slots as above) users that provided the dataset, will be a contact for it (one each), create the dataset (several)

```
    "id": 43,
    "title": "Plankton community in Bongo net (300µm), Tara Oceans cruise",
```

= information about the collection

```
    "citation": "Tara Oceans Consortium, Tara GOSEE, Laboratoire d'Océanographie de Villefranche
(LOV), Sorbonne Université/CNRS; 2020; Plankton community in Bongo net (300µm), Tara Oceans
cruise",
    "license": "CC BY 4.0",
```

= permissions and citation for the collection

```
    "abstract": " Tara expeditions sampled the world's oceans with standardized protocols,
putting an exceptional effort [...]
    "description": "Bongo net (300µm mesh) hauled vertically from 500m depth to the surface,
once during the day, and once during the night. [...]
  },
```

= long form description of the content.

```
  {
    "project_ids": [
      492,
      497,
      896
    ],
 [...]
```

= another collection, etc.

These are all fields requested by the export to EurOBIS.

3.3 Projects

projects/search endpoint: search projects by title, instrument. By default, returns only projects accessible by the current user.

Example: Search for projects with "demo" in their title. This returns a list with only one project inside (this one: https://ecotaxa.obs-vlfr.fr/prj/185) and the following information:

```
$ curl -s -X GET 'https://ecotaxa.obs-vlfr.fr/api/projects/search?title_filter=demo' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
[
  {
    "obj_free_cols": {
[...]
      "area": "n03",
      "mean": "n04",
      "stddev": "n05",
[...]
    },
    "sample_free_cols": {
      "scan_operator": "t01",
      "ship": "t02",
[...]
```

= definition of the columns of data imported by the user: area, mean, stddev (of grey levels) at object level, name of the scan operator and of the ship at sample level, etc.

NB: some content is omitted as marked by [...]

```
    },
    "init_classif_list": [
        72398,
        61996,
[...]
    ],
```

= list of the taxa used for sorting images in this project (as numeric identifiers).

```
    "managers": [
        {
            "id": 5,
            "email": "someone@obs-vlfr.fr",
            "name": "Foo Bar",
[...]
        },
[...]
    ],
    "annotators": [
 [...]
        {
            "id": 993,
            "email": "ecotaxa.api.user@gmail.com",
            "name": "Test API user",
 [...]
    ],
```

= list of different kinds of users in this project: managers (can do everything), annotators (can classify images; this is the case of the test user), viewers (can only view images)

```
    "can_administrate": false,
    "license": "Copyright",
```

= permissions on this project

```
    "projid": 185,
    "title": "Demo Zooscan for API tests (point B WP2 200)",
```

= project id (useful for other queries) and title.

```
[...]
    "classiffieldlist": "depth_min=depth_min\r\ndepth_max=depth_max\r\narea=area
[pixel]\r\nmean=mean [0-255]\r\nfractal=fractal\r\nmajor=major
[pixel]\r\nsymetrieh=symetrieh\r\ncirc.=circ\r\nferet = Feret [pixel]",
[...]
```

= list of user-supplied fields that the images can be sorted by in the classification page.

```
    }
]
```

3.4 Taxonomy

taxa/search endpoint: Search the taxonomy list by name

Example: Search for taxa named "Calanidae"; only one answer here which gives its unique identifier and its name (pr is useless for now).

```
$ curl -s -X GET 'https://ecotaxa.obs-vlfr.fr/api/taxa/search?query=Calanidae&project_id=185' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
[
  {
    "id": 61993,
    "text": "Calanidae",
    "pr": 1
  }
]
```

3.5 Objects

object_set/{project_id}/query endpoint: Return a set of objects from a given project that match some conditions. The conditions that can be used correspond to all the filters described above in the user interface section: sample, lat/lon, depth, time, etc.

Example: Filter objects that have a taxonomic id of 61993 (i.e. Calanidae) and that are of status "V" (i.e. validated by a human operator). Returns 4 such objects as a set of their various ids.

```
$ curl -s -X POST 'https://ecotaxa.obs-vlfr.fr/api/object_set/185/query' \
> -d '{"taxo":"61993","statusfilter":"V"}' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
{
  "object_ids": [
    24475011,
    24474962,
    24474943,
    24474930
  ],
  "process_ids": [
    186,
    186,
    186,
    186
  ],
  "acquisition_ids": [
    204,
    204,
    204,
    204
  ],
  "sample_ids": [
    15743,
    15706,
    15719,
    15724
  ],
  "project_ids": [
    185,
    185,
    185,
    185
  ]
}
```

Example: Filter validated objects that were collected between September 1st and September 30th.
```
$ curl -s -X POST 'https://ecotaxa.obs-vlfr.fr/api/object_set/185/query' \
> -d '{"fromdate":"2016-09-01","todate":"2016-09-30","statusfilter":"V"}' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
{
  "object_ids": [
    24478668,
    24480301,
    24480034,
    24478471,
    24476866,
    24480777,
    24480646,
    24480338,
    24478832,
    24475709,
    24475165,
    24475156,
```

D2.3 EcoTaxa data service

```
        24475144,
        24475143,
        24475116,
        24479503,
        24474385,
        24479945,
        24479865,
        24479720,
        24480035,
        24479933,
        24479790,
        24477740
    ],
[...]
```

[object/{object_id}](#) endpoint: Extract all information about an object identified by its id.

Example: extract information about an object in the sets above (NB: output reorganised for clarity).

```
$ curl -s -X GET 'https://ecotaxa.obs-vlfr.fr/api/object/24475011' \
> --oauth2-bearer 'eyJ1c2VyX2lkIjo5OTN9.X7BcGA.J_Nk-FDbu9fDAC9_DHWxAHoZLMA'|jq .
{
  "orig_id": "wp220160817_d1_1_164",
  "objid": 24475011,
  "processid": 186,
  "acquisid": 204,
  "sampleid": 15743,
  "projid": 185,
```
= identifiers for this object (as imported, internal, etc.)
```
  "objdate": "2016-08-17",
  "objtime": "12:00:00",
  "latitude": 43.685,
  "longitude": 7.31566666666667,
  "depth_min": 0,
  "depth_max": 75,
  "sunpos": "D",
```
= set of objects properties that can be queried.
```
  "classif_id": 61993,
  "classif_qual": "V",
  "classif_who": 4,
  "classif_when": "2020-11-13T21:09:49.470878",
  "classif_auto_id": 61993,
  "classif_auto_score": 0.949359155170351,
  "classif_auto_when": "2019-02-26T11:36:50.748984",
  "classif_crossvalidation_id": null,
```
= information about the classification of this object: taxon, status, when, by whom, etc.
```
  "img0id": 24469465,
  "imgcount": 1,
[...]
  "images": [
    {
      "imgid": 24469465,
      "objid": 24475011,
      "imgrank": 0,
      "file_name": "2446/9465.jpg",
      "orig_file_name": "wp220160817_d1_1_164.jpg",
      "width": 357,
      "height": 605,
      "thumb_file_name": "2446/9465_mini.jpg",
      "thumb_width": 236,
      "thumb_height": 400
    }
```
= link to the images for this object (only one here)

D2.3 EcoTaxa data service

```
  ],
  "free_columns": {
[...]
    "area": 22813,
    "mean": 152.79,
    "stddev": 52.109,
[...]
```
= list (and values) of descriptors measured on the object and imported together with it (their storage is defined at project level; Cf above).
```
  }
}
```

4 Upcoming developments and conclusions

While the EcoTaxa application and API are available for use right now, development will continue, as part of Blue-Cloud and then other projects (and Blue-Cloud will benefit from these improvements). Below we mention the developments of interest for Blue-Cloud.

The database levels "acquisitions" and "processes" will be *merged* in a single level called "subsample" ([#367](#)). Currently, no acquisition contains more than one process, so offering this possibility was overkill and there is an opportunity for simplification. In addition, when several acquisitions exist per sample, they correspond to subsamples so changing the name will have a clearer meaning for users. Finally, subsample will match better with the [OBIS data model](#) which will make exporting data to EurOBIS easier.

Then we will propagate standard metadata fields (lat, lon, depth, date and time) to every level in the hierarchy (object but also subsample and sample), notably to ease queries from Blue-Cloud's data discovery service.

The taxonomic database that EcoTaxa uses is old and not a universal standard. All taxa are being matched to the World Register of Marine Species (WoRMS) which is the standard EurOBIS uses ([#454](#)).

To allow export of ecologically-relevant data, additional metadata fields will be standardised based on BODC vocabularies, which are currently being expanded to accommodate plankton imaging specificities. In particular, subsampling rate and volume sampled will be added, hence allowing the computation of concentrations in a generic manner.

In addition to the DarwinCore Archive export, which is a standard archive and data exchange format (used by EurOBIS), a simple .tsv table export is available in the application, which is an easier to parse format, more suitable for everyday work. An endpoint to export this format from a project or an object set (result of a search according to some criteria) will be added to the API.