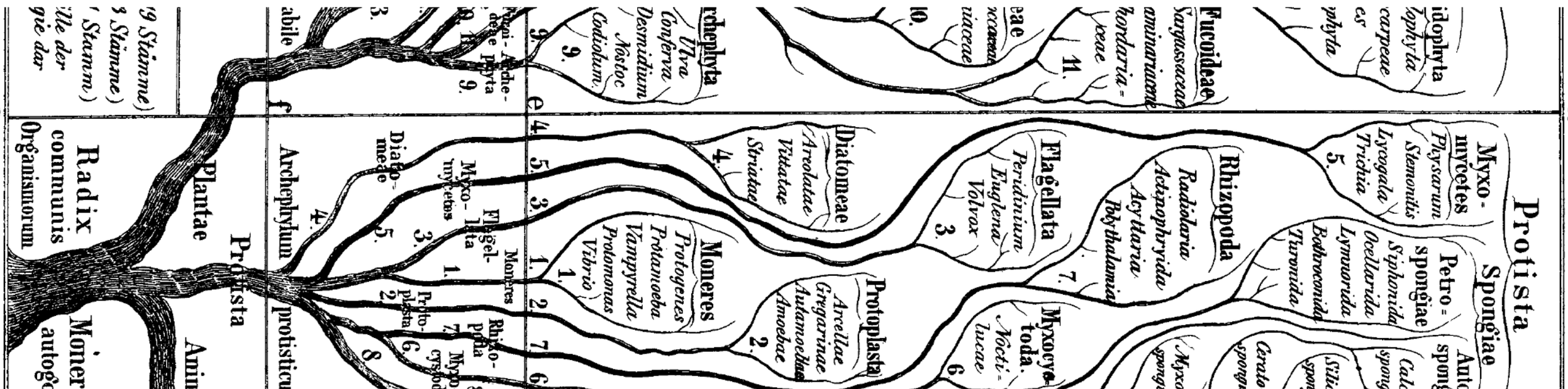
AT ODATIS, 2021-05-26
Jean-Olivier Irisson

SCIENCES SORBONNE UNIVERSITÉ

Référentiels taxinomiques et formats de données

Retour d'expérience sur EcoTaxa



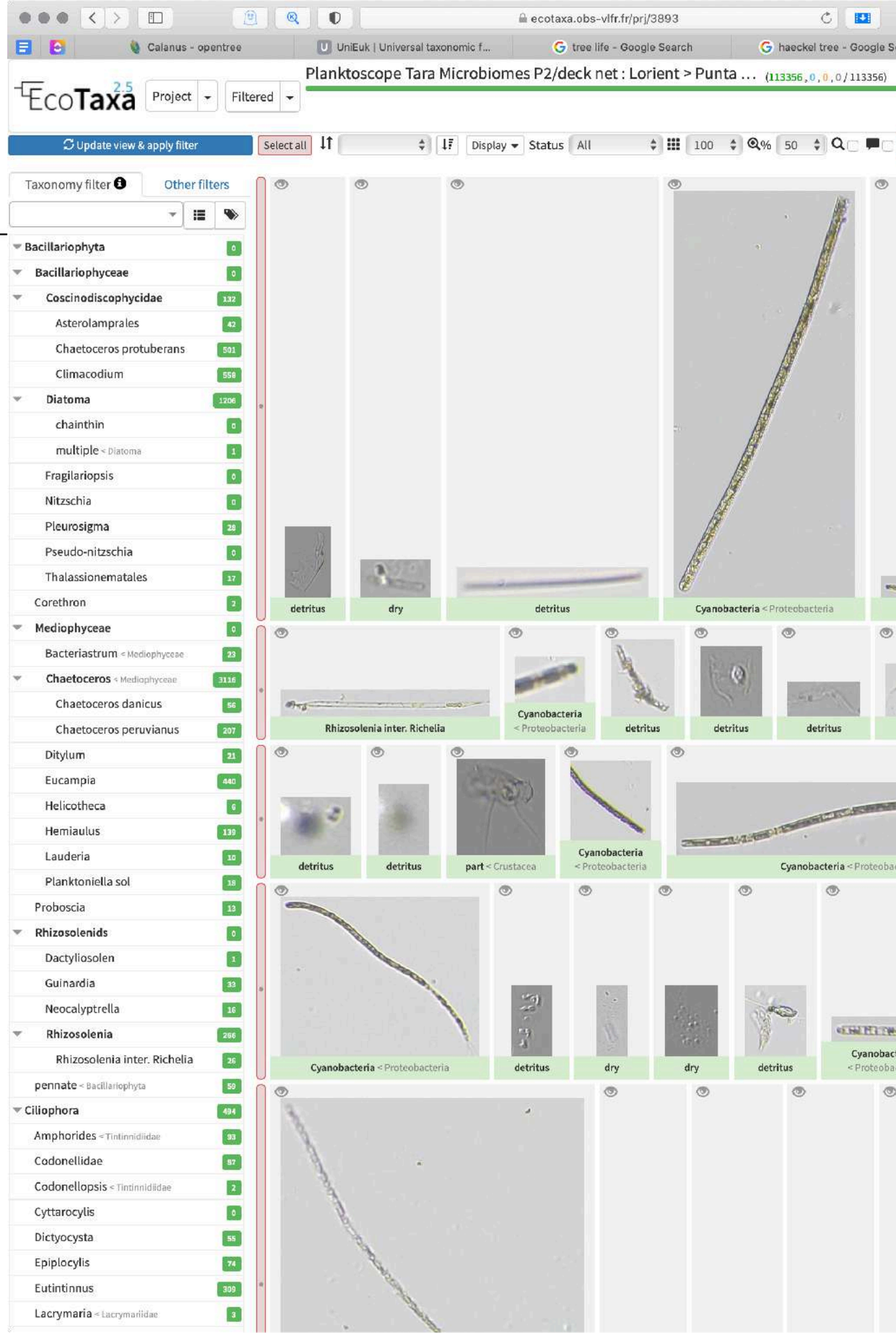
EcoTaxa

Une plateforme pour **stocker** et **classer** taxinomiquement des **images** d'organismes individuels (principalement du plancton)

Démarré comme un outil pour les données de Tara Océans

Contient actuellement ~**150M** d'objets, 63M nommés, en ~**45,000 points** de l'océan, collectés par ~**350 organisations** au niveau international

⟹ Nécessite un référentiel taxinomique universel et des formats d'import/export.



Les référentiels taxinomiques sont nombreux...

Local checklists

WoRMS

AlgaeBase

UniEuk

Species 2000

Catalog of Life

Open Tree of Life

GBIF backbone

ITIS

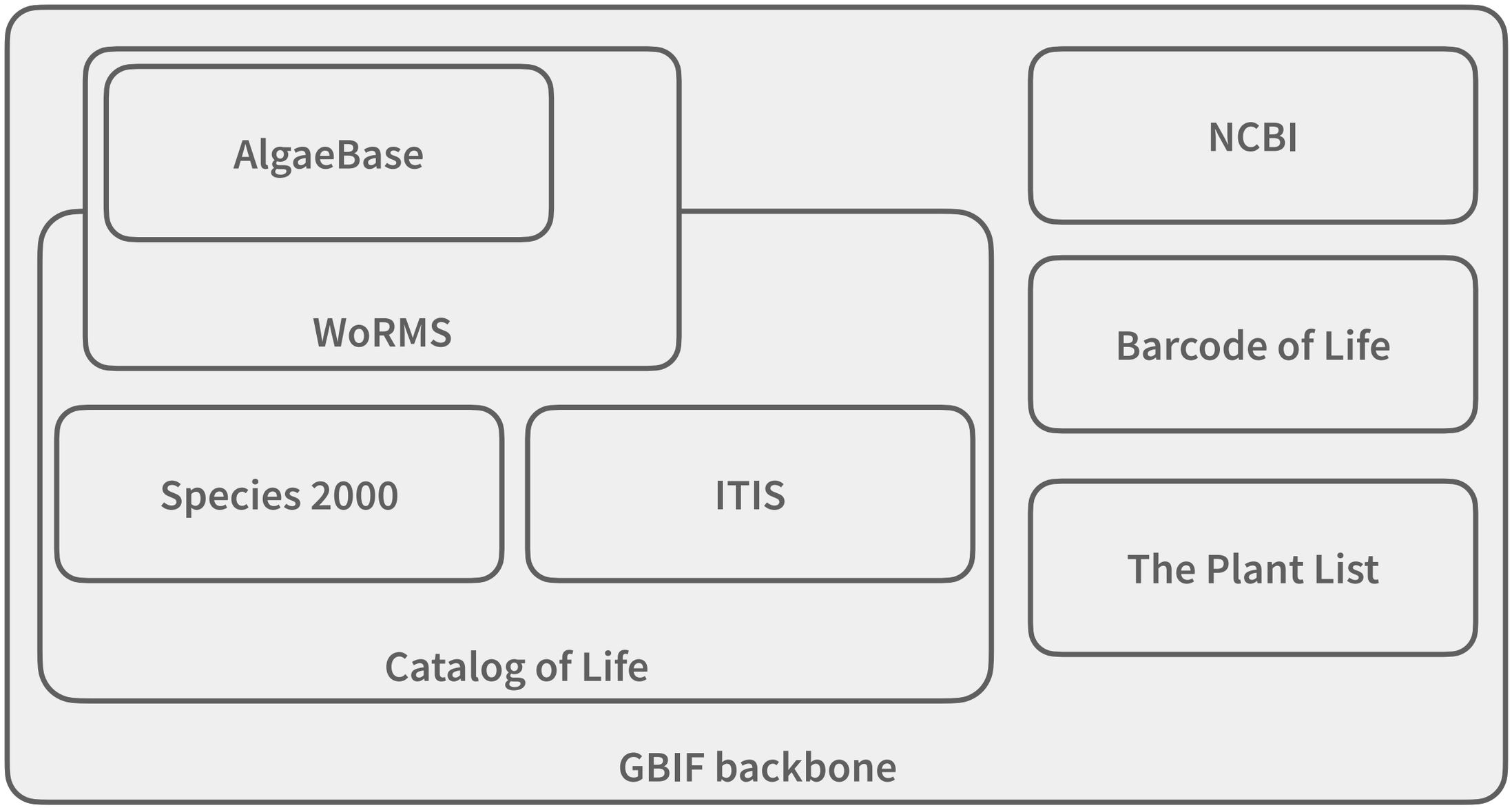
NCBI

TAXREF

Barcode of Life

The Plant List

...leurs relations sont complexes



Les “référentiels” taxonomiques sont en fait toutes des méta-référentiels

Grandes divergences à la racine de l'arbre…

Catalogue of Life

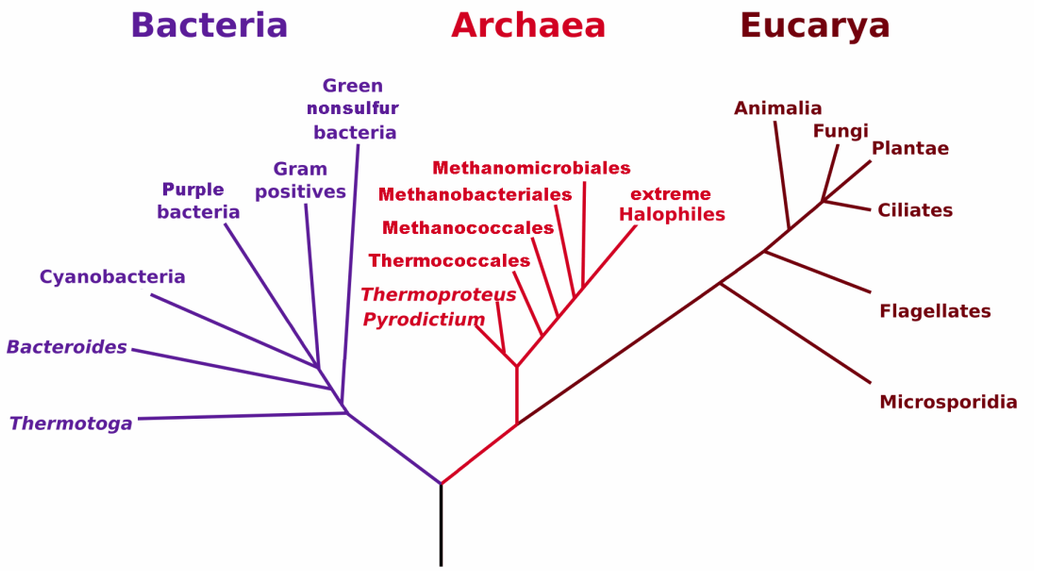
- unranked: *Biota*
  - kingdom: Animalia
  - kingdom: Archaea
  - kingdom: Bacteria Cavalier-Smith, 2002
  - kingdom: Chromista
  - kingdom: Fungi
  - kingdom: Plantae
  - kingdom: Protozoa
  - kingdom: Viruses

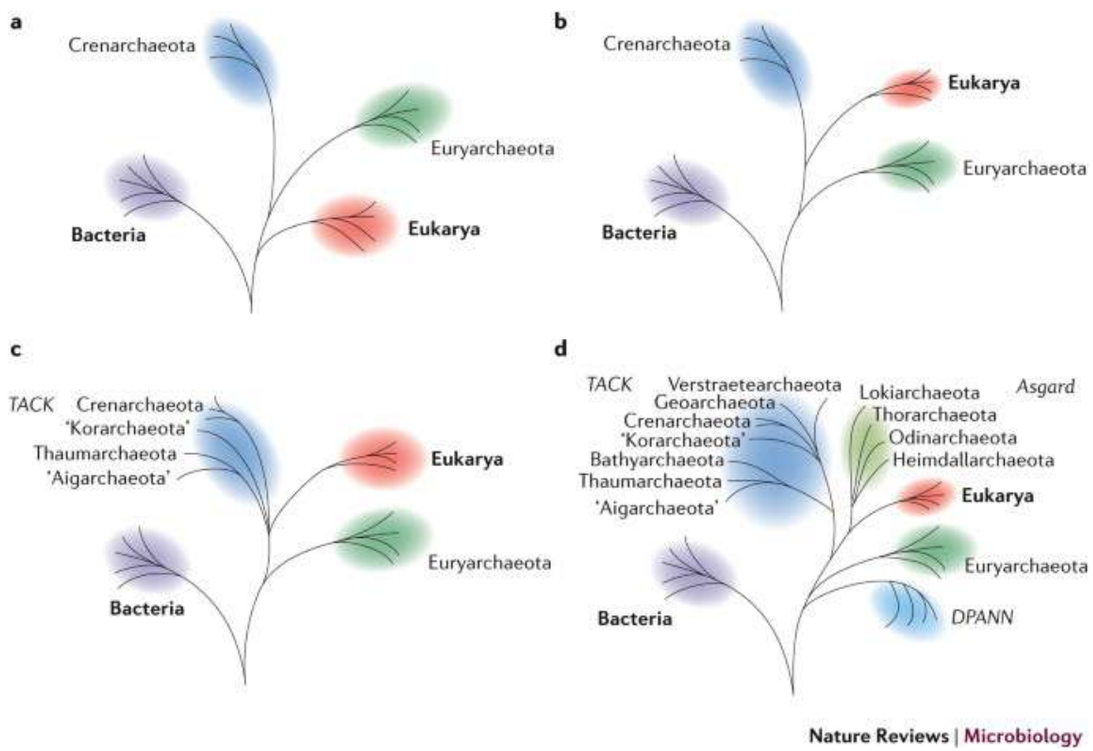
WoRMS World Register of Marine Species

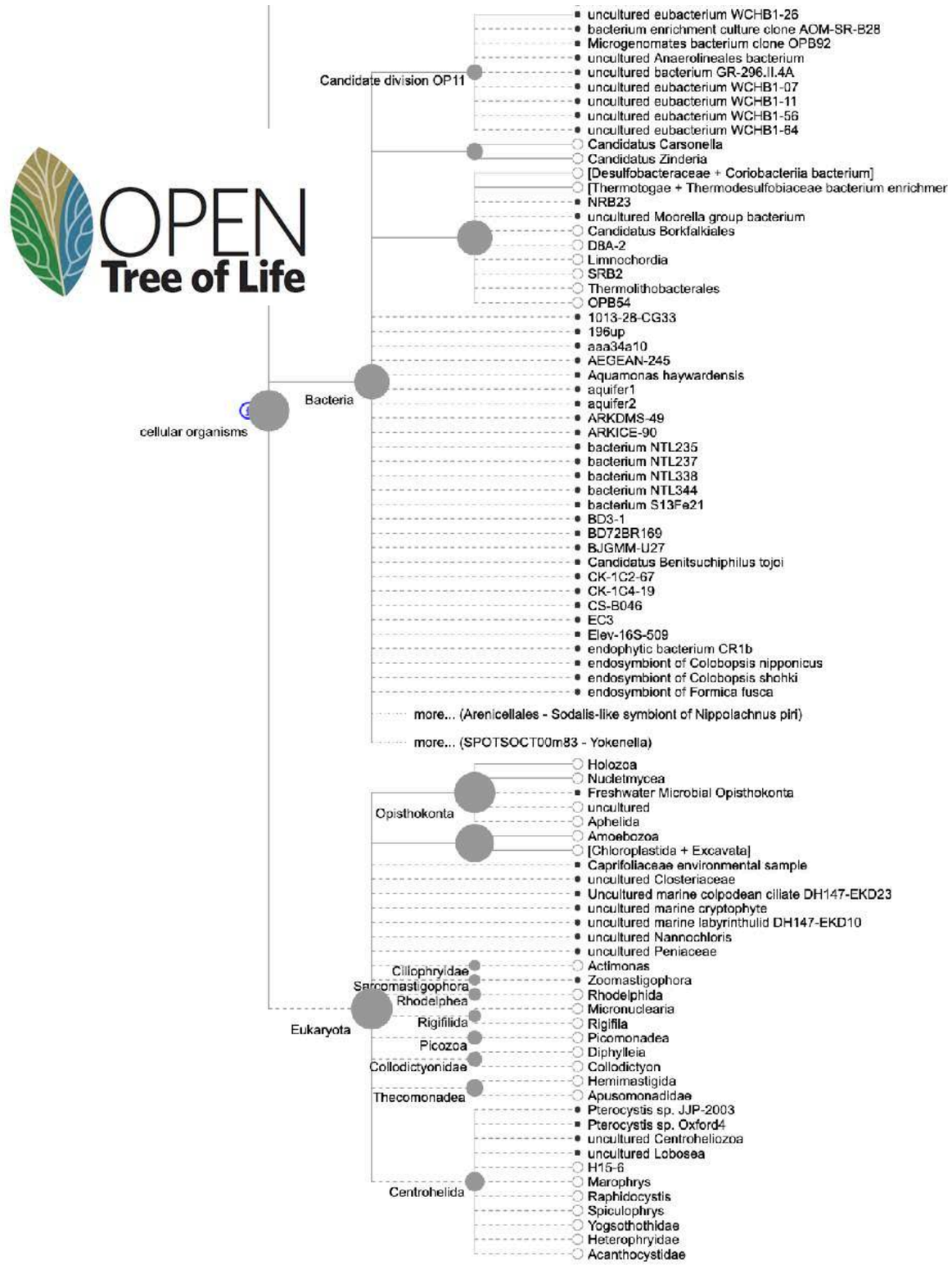
- Biota (238 252)
  - + Kingdom Animalia (203 311)
  - + Kingdom Archaea (116)
  - + Kingdom Bacteria (2 077)
  - + Kingdom Chromista (20 237)
  - + Kingdom Fungi (1 382)
  - + Kingdom Plantae (10 378)
  - + Kingdom Protozoa (633)
  - + Kingdom Viruses (115)

Grandes divergences à la racine de l'arbre…

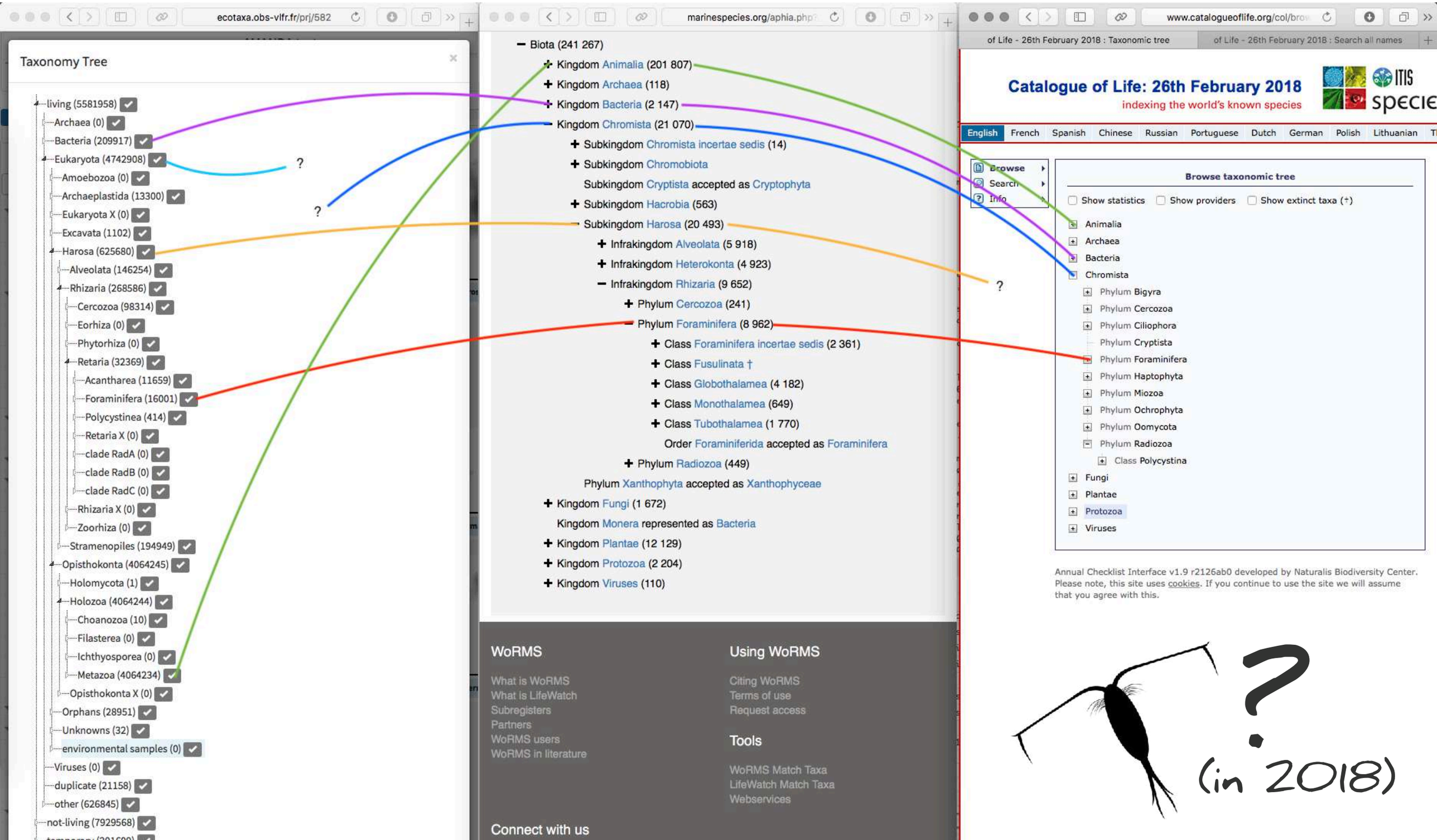
Woese & Fox, 1977







… et sur la topologie de l'arbre



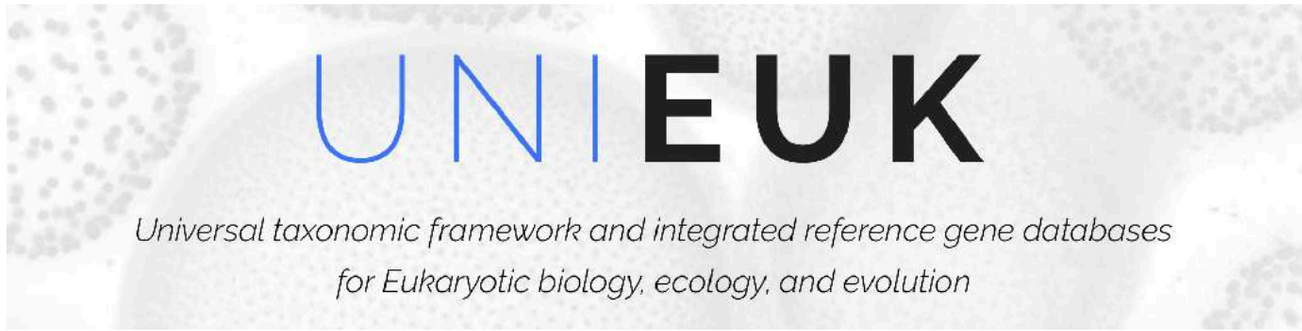
Et dans EcoTaxa, finalement?

Taxonomie initiale par **UniEuk**, pour des raisons historiques

Ajouts, par les utilisateurs, de taxa manquants (et de “taxa” morphologiques)

Transition vers **WoRMS**, pour compatibilité avec OBIS

À terme, espérons l’émergence d’un **méta**référentiel taxinomique assurant la correspondance avec l’existant, universel, phylogénétiquement correct…



Le format d'EcoTaxa est peu contraignant

Dossier d'images + 1 fichier .tsv avec une ligne par image (donc bcp de répétitions)

Organisation hiérarchique des métadonnées

Seuls **deux** champs sont **obligatoires**

Certains champs ont un **format** imposé

La **plupart** sont **libres** et juste stockés par dataset

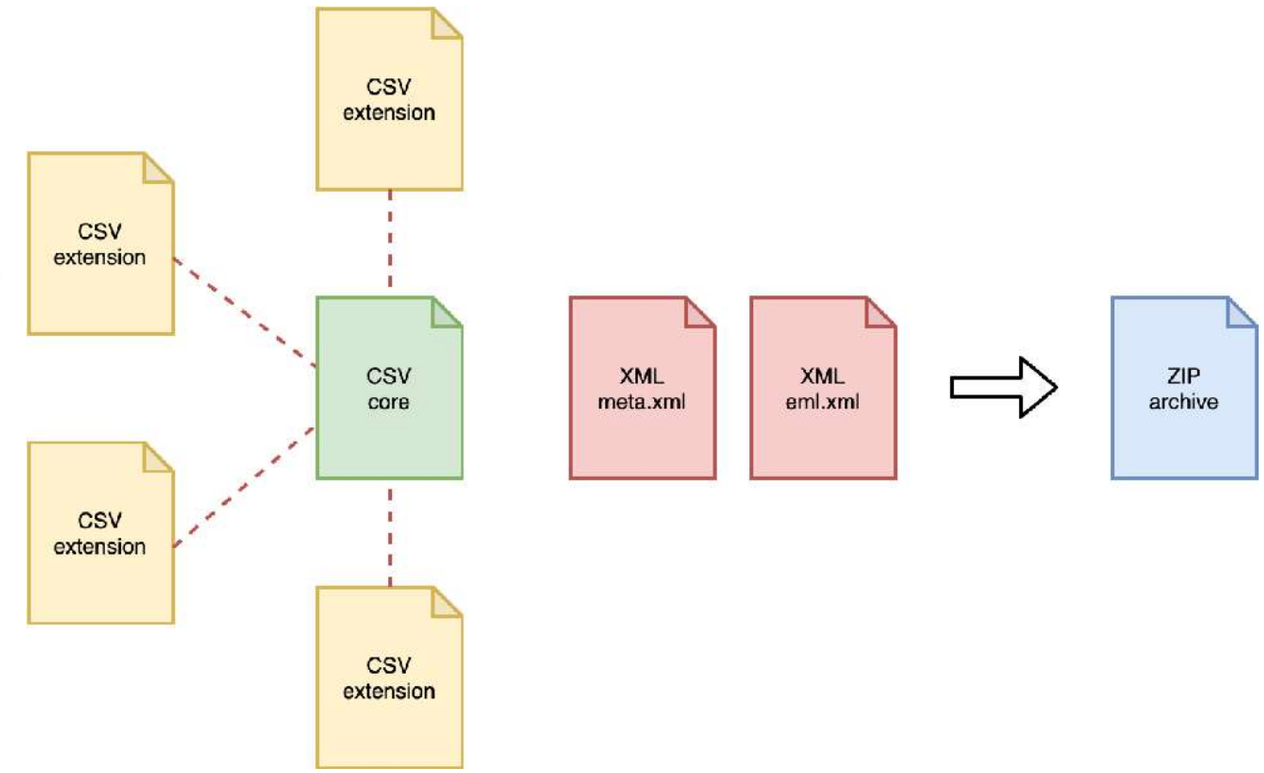
⟹ **Flexible** pour les utilisateurs mais **peu standard**

- IMAGE
  - `img_file_name` [t]: name of the image file in the folder (including extension)
  - `img_rank` [f] : rank of image to be displayed, in case of existence of multiple (<
    Starts at 1.
- OBJECT: one object to be classified, usually one organism. One object can be represente
  `tsv` file, there is one line per image which means the object data gets repeated on seve
  - `object_id` [t] : identifier of the object, must be unique in the project. It will be
  - `object_link` [f] : URL of an associated website
  - `object_lat` [f] : latitude, decimal degrees
  - `object_lon` [f] : longitude, decimal degrees
  - `object_date` [f] : ISO8601 YYYYMMJJ UTC
  - `object_time` [f] : ISO8601 HHMMSS UTC
  - `object_depth_min` [f] : minimum depth of object, meters
  - `object_depth_max` [f] : maximum depth of object, meters

  And, for already classified objects
  - `object_annotation_date` [t] : ISO8601 YYYYMMJJ UTC
  - `object_annotation_time` [t] : ISO8601 YYYYMMJJ UTC
  - `object_annotation_category` [t] : class of the object with optionally its dir
    by left angle bracket without whitespace "Cnidaria<Hydrozoa" or old style betwe
    (Hydrozoa)"
  - `object_annotation_category_id` [f] : Ecotaxa ID of the class of the object,
    export
  - `object_annotation_person_name` [t] : name of the person who identified t
  - `object_annotation_person_email` [t] : email of the person who identified
  - `object_annotation_status` [t] : predicted, dubious, or validated

  And additional object-related fields
  - `object_***` [f] or [t] : other fields relative to the object. Up to 500 [f] fields and
- PROCESS: metadata relative to the processing of the raw images
  - `process_id` [t] : identifier. The processing information is associated with the a
    missing, a dummy processing identifier will be created.
  - `process_***` [t] : other fields relative to the process. Up to 30 of them.
- ACQUISITION: metadata relative to the image acquisition
  - `acq_id` [t] : identifier of the image acquisition, must be unique in the project. If
    identifier will be created.
  - `acq_instrument` [t] : name of the instrument (UVP, ZOOSCAN, FLOWCAM, etc.)
  - `acq_***` [t] : other fields relative to the acquisition. Up to 30 of them.
- SAMPLE: a collection event
  - `sample_id` [t] : identifier of the sample, must be unique in the project. If missir
    will be created.
  - `sample_***` [t] : other fields relative to the sample. Up to 30 of them.

Ajout d’un export au format DwCA



Format beaucoup plus **compliqué**

Format beaucoup plus **verbeux**

Format beaucoup plus **standardisé** (termes DarwinCore, vocabulaire du BODC, etc.)

⟹ Adapté pour l’échange de **machine à machine** mais peu agréable pour travailler

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

Considération du format BioODV

Format “**tableur**” relativement classique = lisible par ODV mais aussi “Excel” (et R, et Python, etc.)

Format relativement **compact** (car les champs répétés peuvent être omis)

Format pouvant être **standardisé** (par des références au vocabulaire BODC dans l’**en-tête**)

⟹ **Bon compromis** entre flexibilité et standardisation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	//														
2	//<sdn_reference xlink:href="http://seadata.bsh.de/cgi-csr/XML/xmlDownload_V2.pl?edmo=269&identifier=GN36199704604" xlink:role="isObservedBy" xlink:type="SDN:L23::CSR"/>														
3	//<sdn_reference xlink:href="http://vocab.nerc.ac.uk/collection/C17/current/36AE" xlink:role="isObservedBy" xlink:type="SDN:L23::NVS2CON"/>														
4	//<sdn_reference xlink:href="http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp?edmo=269&identifier=GN36199764604001_0MNB4_269_G04" xlink:role="isDescribedBy" xlink:type="SDN:L23::CDI"														
5	//<sdn_reference xlink:href="http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp?edmo=269&identifier=GN36199764604001_0MSB5_269_G04" xlink:role="isDescribedBy" xlink:type="SDN:L23::CDI"														
6	//<sdn_reference xlink:href="http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp?edmo=269&identifier=GN36199764604001_0MSB6_269_G04" xlink:role="isDescribedBy" xlink:type="SDN:L23::CDI"														
7	//SDN_parameter_mapping														
8	//<subject>SDN:LOCAL:COREDIST</subject><object>SDN:P01::COREDIST</object><units>SDN:P06::ULAA</units>														
9	//<subject>SDN:LOCAL:Lithium Li</subject><object>SDN:P01::LISEDAD1</object><units>SDN:P06::UMKG</units>														
10	//<subject>SDN:LOCAL:Scandium Sc</subject><object>SDN:P01::HCMR0011</object><units>SDN:P06::UMKG</units>														
11	//<subject>SDN:LOCAL:Titanium Ti</subject><object>SDN:P01::RWSSED04</object><units>SDN:P06::UGKG</units>														
12	//<subject>SDN:LOCAL:Vanadium V</subject><object>SDN:P01::RWSSED02</object><units>SDN:P06::UMKG</units>														
13	//														
14	Cruise	Station	Type	YYYY-MM-DI	Longitude [d	Latitude [deg	LOCAL_CDI_I	EDMO_code	Bot. Depth [r	COREDIST [m	QV:SEADATA	Lithium Li [		c	QV:SEADATA
15	MTPII-MATE	MNB4	*	1997-09-01T	25.511	39.778	GN36199764	269	95	0.005	1	0		0	1
16										0.105	1	0		0	1
17										0.205	1	0		0	1
18	MTPII-MATE	MSB5	*	1997-09-01T	24.683	36.193	GN36199764	269	1255	0.005	1	0	1	0	1
19										0.105	1	0	1	0	1
20										0.205	1	0	1	0	1
21										0.26	1	0	1	0	1
22										0.315	1	0	1	0	1
23	MTPII-MATE	MSB6	*	1997-09-01T	25.705	36	GN36199764	269	1300	0.005	1	0	1	0	1

EcoTaxa 2.5

SCIENCES SORBONNE UNIVERSITÉ

puf Partner University Fund

BELMONT FORUM

EMBRC FRANCE
CENTRE NATIONAL DE RESSOURCES BIOLOGIQUES MARINES

Merci