contact@odatis-ocean.fr
www.odatis-ocean.fr

ODATIS
OCEAN DATA INFORMATION AND SERVICES

Atelier Technique Données Biologiques

Mise en accès de données de biodiversité de suivis à long-terme

*

Mark HOEBEKE
mark.hoebeke@sb-roscoff.fr

De la collecte à la mise en accès

- Acquisition / collecte
- Bancarisation
- Annotation / curation
- Mise en accès
- Bilan

Acquisition / Collecte

- Contexte institutionnel :
  - SBR : Station Marine ayant un statut d'OSU (depuis peu OSU STAMAR)
  - Dans ses missions d'observation :
    - Suivis à long terme de la biodiversité :
      - Du plancton : phytoplancton & zooplancton
      - Des communautés benthiques
    - Suivis à long terme de paramètres physico-chimiques
  - Intégrée à des réseaux nationaux d'observation

    RESOMAR/Pelagos, REBENT
  - Points de collecte faisant partie des SNO (PHYTOBS, *Benthobs,* SOMLIT)

Acquisition / Collecte

- Eléments opérationnels : suivis phytoplanctoniques
  - Objectif : caractériser la biodiversité des communautés planctoniques aux points de collecte :
    - 2 sites pour le nano et le picophytoplancton
    - 1 site pour le microphytoplancton

    → sites SOMLIT
  - Echantillonnages bimensuels (depuis 2000) :
    - Sorties terrain aux points de collecte, prélèvement d'échantillons d'eau de mer en surface (trait de filet à plancton ou bouteille Niskin) ou au fond (-60m)
    - Analyse après fixation (entre 15 jours et 1 an après échantillonnage) :
      - **Microphytoplancton : détermination taxonomique & abondances en microscopie optique**
      - Nano et picophytoplancton : quantification et classification en groupes fonctionnels basée sur la cytométrie en flux.
    - Bancarisation dans la base de données PELAGOS du RESOMAR.

Bancarisation des suivis (phyto + zoo)

- Base de données PELAGOS
  - En fonction depuis 2012, V2 en 2014, dernières évolutions fonctionnelles 2018.
  - Accès public aux « métadonnées » :
    - Descripteurs des suivis
  - Accès restreint aux données :
    - Données brutes sur les abondances, la taxonomie, les groupements fonctionnels.
    - Restreint aux utilisateurs ayant signé la charte et dont la demande a été approuvée par le comité de pilotage.

Bancarisation des suivis (phyto + zoo)

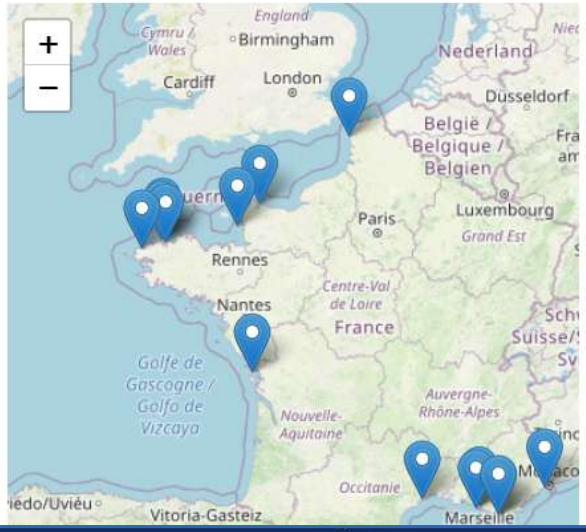
resomar

Pelagos | BDD Pelagos V2

Consultation

Statistiques de la base de données

Dernière mise à jour	04/03/2020
Jeux de données	18
Échantillons	4,777
Taxons	1,205



La base de données PELAGOS

La base de données Pelagos est le fruit d'un travail collaboratif du Réseau des Stations et Ob
l'écosystème pélagique côtier (dont des séries temporelles). L'un des objectifs est d'exploiter
concernant par exemple les facteurs qui contrôlent la distribution et l'abondance des orgar
temps.

Dans un premier temps, la base ne sera accessible qu'aux membres du réseau qui ont sign
d'utilisation des données).

Pour accéder aux données (et éventuellement insérer de nouvelles données) vous devez :

- signer la charte RESOMAR PELAGOS et l'envoyer à contact.pelagos@sb-roscoff.fr.
- demander l'ouverture d'un compte via le formulaire disponible à http://abims.sb-roscoff

Un retour sera fait sur la validation de la demande de création de compte. Pour les comptes q
sur l'affiliation du demandeur pourront être demandées.

Un environnement intégré pour l'analyse en ligne de la base de données est proposé à la cor
de scripts R interfacés sous la plateforme Galaxy.

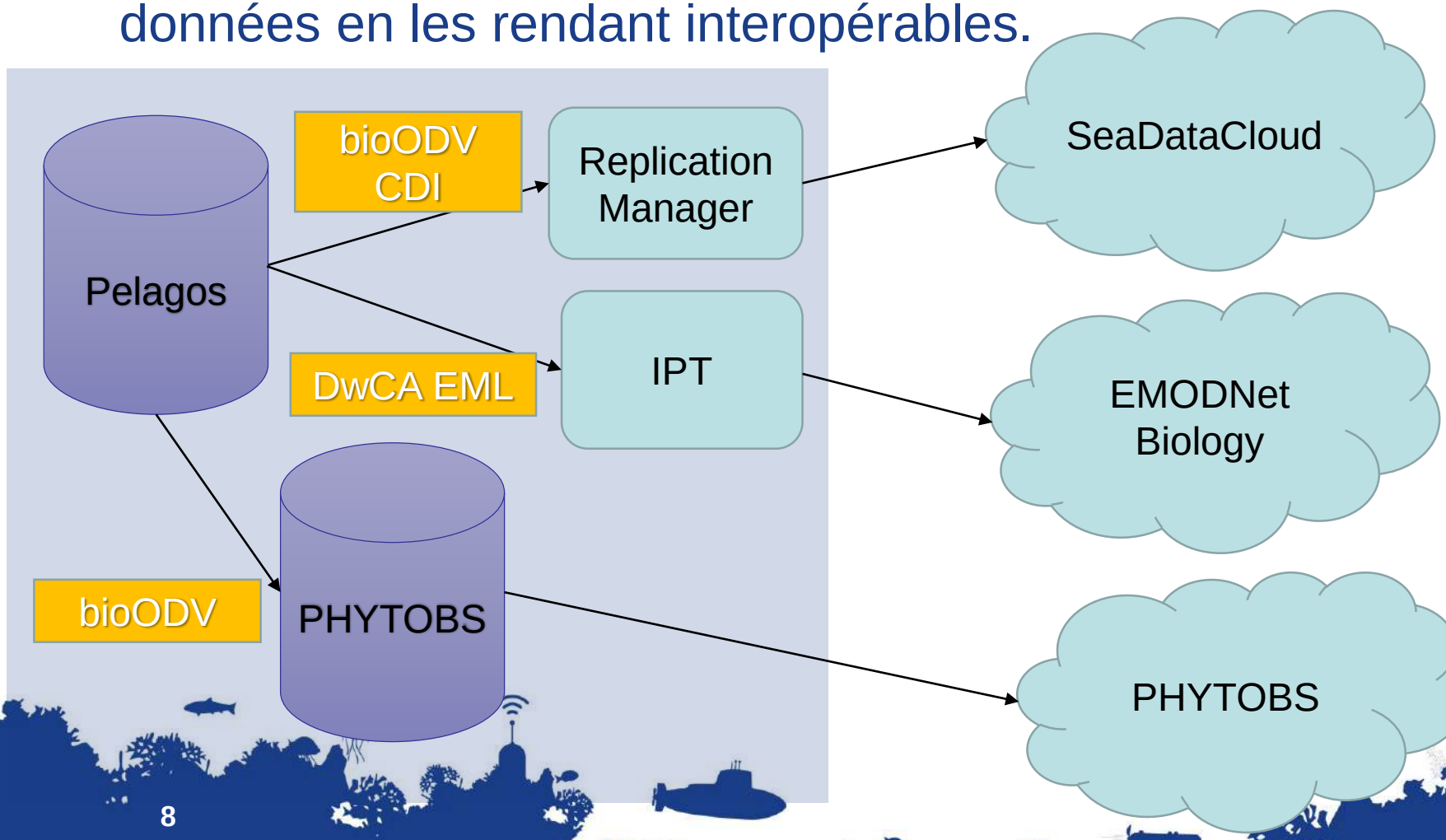
Pour de plus amples informations : contact.pelagos@sb-roscoff.fr.

Bancarisation (Micro)Phytoplankton

- Nature de l'information bancarisée :
  - Caractérisation de l'échantillonnage (nom du suivi, géolocalisation, horodatage, intervenants, instruments)
  - Caractérisation de l'analyse (intervenants, date, méthode)
  - Caractérisation du spécimen :
    - **Dénomination originelle** donnée par l'expérimentateur
    - **Identifiant** dans le référentiel taxonomique **WoRMS**
    - **Dénomination** normalisée dans la base **PELAGOS** :
      - Dénomination officielle extraite du référentiel taxonomique WoRMS
      - Dénomination interne à la base (*taxons ou groupes de taxons orphelins)*
- Processus de bancarisation :
  - Utilisation de gabarits (Excel)
  - Chaque utilisateur authentifié dispose de son espace de travail pour téléverser / réaliser des corrections assistées / insérer dans la base.

Annotation / Curation

- Motivation : assurer une diffusion plus large des jeux de données en les rendant interopérables.



Annotation/Curation

- Eléments communs à l'ensemble des exports :
  - Utilisation du référentiel taxonomique WoRMS
  - Utilisation des vocabulaires contrôlés SeaDataCloud (P01/P06) *à l'exclusion des termes taxonomiques désignant des taxons, remplacés par un terme générique permettant l'utilisation des identifiants WoRMS => au lieu d'une colonne par taxon avec des abondances ; une colonne avec un identifiant WoRMS et l'abondance associée.*
  - Processus de génération des fichiers de données :
    - Génération de fichiers tabulés à l'aide de routines (Python/Notebook ou Java) : extractions des données de la base et mise en forme des fichiers (bioODV ou DwCA)
    - *Pas de modifications aux schémas de la base de données.*

Annotation/Curation

- Format (bio)ODV : structure du fichier

métadonnées

Annotation/Curation

- Format (bio)ODV : métadonnées

```
//SDN_parameter_mapping
//<subject>SDN:LOCAL:MinimumObservationDepth</subject><object>SDN:P01::MINWDIST</object><units>SDN:P06::ULAA</units>
//<subject>SDN:LOCAL:MaximumObservationDepth</subject><object>SDN:P01::MAXWDIST</object><units>SDN:P06::ULAA</units>
//<subject>SDN:LOCAL:SampleID</subject><object>SDN:P01::SAMPID01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:SamplingEffort</subject><object>SDN:P01::VOLWBSMP</object><units>SDN:P06::ULIT</units>
//<subject>SDN:LOCAL:SubsampleID</subject><object>SDN:P01::SSAMID01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:SubSamplingCoefficient</subject><object>SDN:P01::SSAMPC01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:ScientificName</subject><object>SDN:P01::SCNAME01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:ScientificNameID</subject><object>SDN:P01::SNANID01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:Sex</subject><object>SDN:P01::ENTSEX01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:LifeStage</subject><object>SDN:P01::LSTAGE01</object><units>SDN:P06::UUUU</units>
//<subject>SDN:LOCAL:DensityPerUnitEffort</subject><object>SDN:P01::SDBIOL01</object><units>SDN:P06::UCPL</units>
//<subject>SDN:LOCAL:PresenceOrAbsence</subject><object>SDN:P01::PRESABS1</object><units>SDN:P06::UUUU</units>
//
```

Annotation/Curation

- Format (bio)ODV : données

Dénomination Expérimentateur

Identifiant WoRMS

Abondance

ScientificName:INDEXED_TEXT	QV:SEADATANET	ScientificNameID:INDEXED_TEXT	QV:SEADATANET	DensityPerUnitEffort [#/l]	QV:SEADATANET
Coscinodiscus	2	urn:lsid:marinespecies.org:taxname:148917	2	80	1
Coscinodiscus	2	urn:lsid:marinespecies.org:taxname:148917	2	20	1
Dictyocha speculum	2	urn:lsid:marinespecies.org:taxname:157260	2	20	1
Dinophysis	2	urn:lsid:marinespecies.org:taxname:109462	2	140	1
Dinophysis	2	urn:lsid:marinespecies.org:taxname:109462	2	100	1
Diploneis	2	urn:lsid:marinespecies.org:taxname:149018	2	40	1
Diploneis	2	urn:lsid:marinespecies.org:taxname:149018	2	20	1
Ditylum brightwellii	2	urn:lsid:marinespecies.org:taxname:149023	2	20	1
Guinardia delicatula	2	urn:lsid:marinespecies.org:taxname:149112	2	1700	1
Guinardia delicatula	2	urn:lsid:marinespecies.org:taxname:149112	2	2360	1
Leptocylindrus danicus	2	urn:lsid:marinespecies.org:taxname:149106	2	120	1
Leptocylindrus danicus	2	urn:lsid:marinespecies.org:taxname:149106	2	40	1
Navicula	2	urn:lsid:marinespecies.org:taxname:149142	2	40	1
Navicula	2	urn:lsid:marinespecies.org:taxname:149142	2	20	1
Navicula transitans	2	urn:lsid:marinespecies.org:taxname:149320	2	20	1

Code qualité

Annotation/Curation

- bioODV :
  - Génération simple : un seul fichier
  - Panoplie d'outils GUI pour la préparation (NEMO, Mikado), vérification (Octopus), et l'exploration (ODV)
  - Peu adapté aux jeux de données « creux » : une colonne par paramètre.

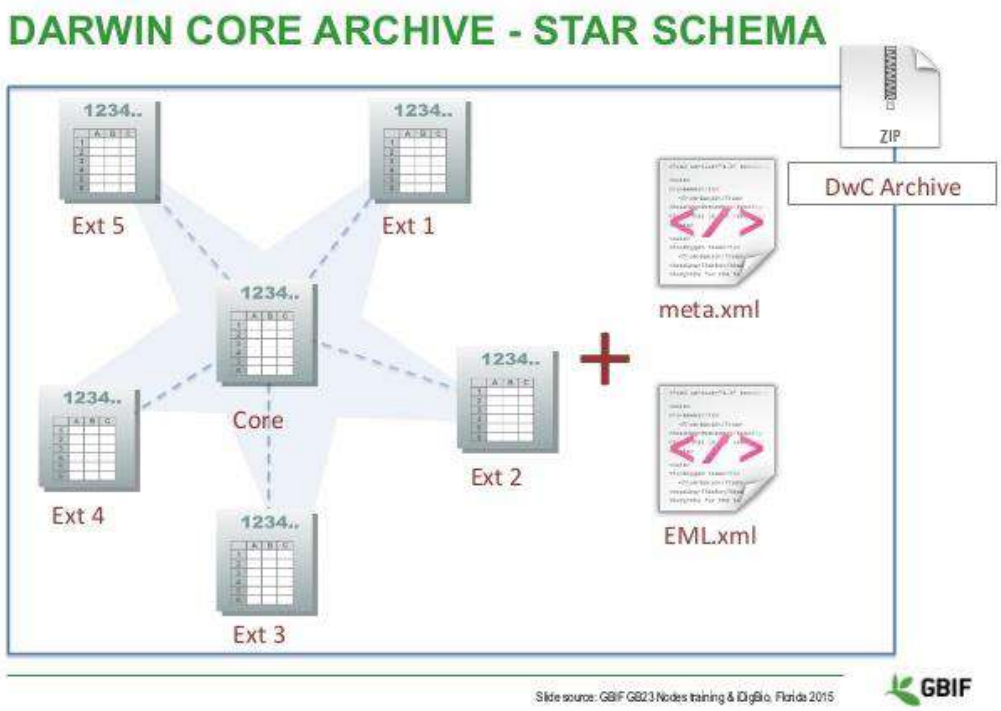
Annotation/Curation

- Format Darwin Core Archive

The conceptual data model of the Darwin Core Archive is a "**star schema**" (Robertson et al. 2014):

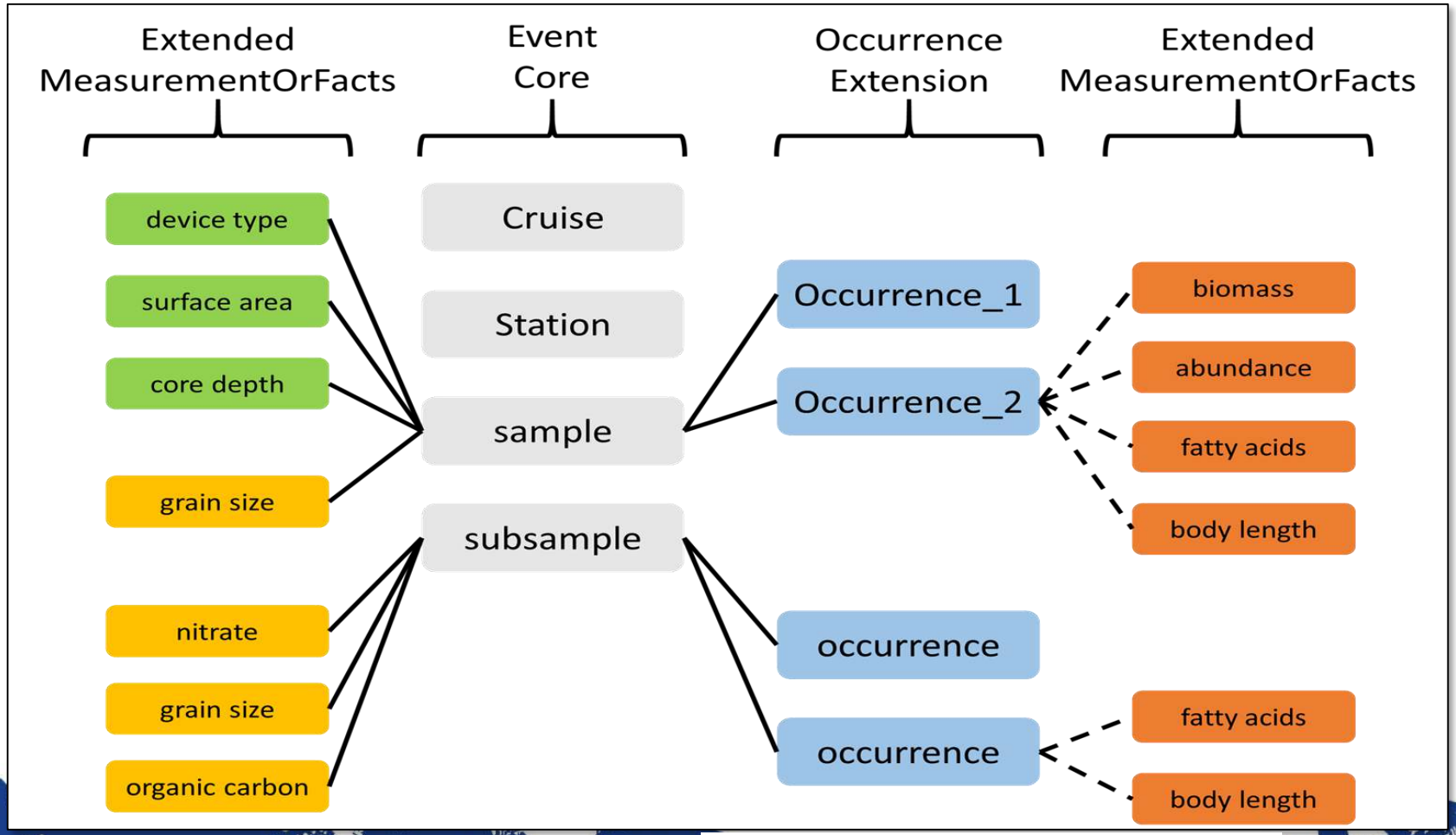
•**Core record**, such as an occurrence or an event, as the center of the star.

•**Extension records**, radiating out of the star, can optionally be associated with the core, linked by database keys such as an ID column.



Annotation/Curation

- Format Darwin Core Archive



(Courtesy EMODNet Biology)

Annotation/Curation

- Format Darwin Core Archive

	C	D	E	F	G	H
	eventID	parentEventID	Type	eventDate	LocationID	DecimalLatitud
	BIOFUN1		Cruise			
	BIOFUN1_BF1M1	BIOFUN1	Sample	2009-05-30	WM1200	38.3881
	BIOFUN1_BF1M2	BIOFUN1	Sample	2009-05-30	WM1200	38.3915
	BIOFUN1_BF1M3	BIOFUN1	Sample	2009-06-01	WM2000	38.038
	BIOFUN1_BF1M4	BIOFUN1	Sample	2009-06-01	WM2000	38.0482

	A	B	D	F	H	I
1	eventID	CollectionCc	occurrenceID	ScientificName	scientificNameID	occurrer
2	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_1	Alepocephalus rostratus	urn:lsid:marinespecies.org:taxname:12668	present
3	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_2	Bathypterois mediterraneus	urn:lsid:marinespecies.org:taxname:29994	present
4	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_3	Coelorinchus mediterraneus	urn:lsid:marinespecies.org:taxname:28031	present
5	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_4	Galeus melastomus	urn:lsid:marinespecies.org:taxname:10581	present
6	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_5	Lepidion lepidion	urn:lsid:marinespecies.org:taxname:12649	present

	A	B	C	D
1	eventID	occurrenceID	MeasurmentType	MeasurmentTypeID
2	BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Density	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02/
409	BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Abundance	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01
810	BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Wet Weight Biomass	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL05
843	BIOFUN1_BF1M1		Sampling instrument nam	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/
870	BIOFUN1_BF1M1		net cod-end mesh size	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/

(Courtesy EMODNet Biology)

Annotation/Curation

- DwCA :
  - Expressivité :
    - Hiérarchisation possible des événements
    - Description plus générique de la géolocalisation (*footprintWKT, coordinateUncertainty)*
    - Plus adapté à la composition de jeux de données hétérogènes (nombreux paramètres différents présents pour quelques occurrences seulement) : un paramètre par ligne.
  - Pas d'outils *user-friendly* pour en visualiser/exploiter les contenus.
  - Génération plus coûteuse qu'un tableau simple (gestion des identifiants répétés dans les fichiers).

Mise en accès

- SeaDataCloud :
  - Création du fichier CDI (XML) avec les métadonnées relatives à chaque suivi (empreinte spatio-temporelle, licence, participants & rôles…) :
    - Les routines d'export génèrent un fichier SQLite avec toutes les métadonnées nécessaires et propres au jeu de données.
    - Utilisation *one-shot* de **Mikado** pour générer une configuration incluant les métadonnées constantes (licence, code EDMO…) ainsi que les requêtes extrayant les métadonnées spécifiques du fichier SQLite
    - Utilisation *batch* de **Mikado** pour générer à partir de la configuration et d'un fichier SQLite le CDI pour chaque jeu de données et compléter la *coupling table.*
  - Déploiement dans le **Replication Manager** (webapp Java)

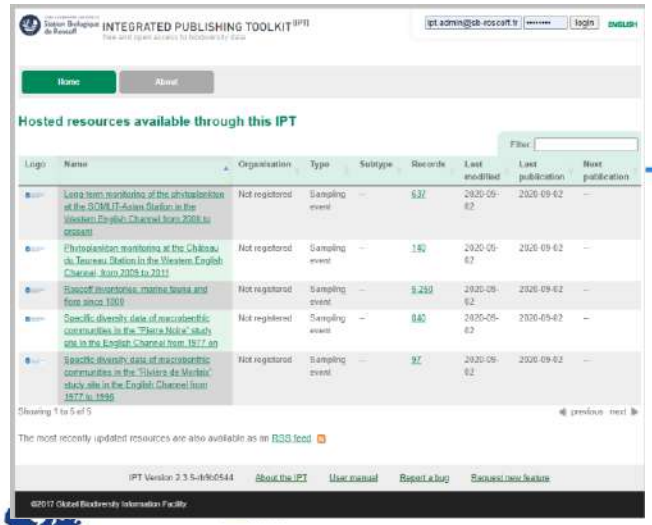
Mise en accès

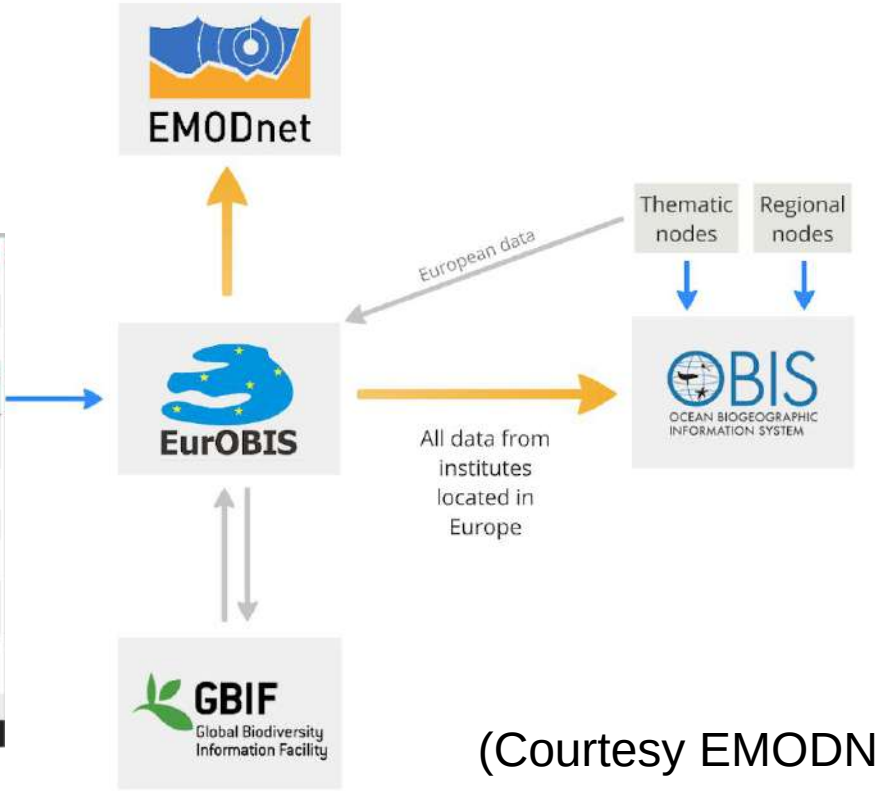
- EMODNet Biology:
  - Génération des 3 fichiers DwCA par suivi :
    - Events, Occurrences, ExtendedMeasurementsOrFacts.
  - Contrôle Qualité à l'aide de l'outil en ligne (http://rshiny.lifewatch.be/BioCheck/
  - Déploiement dans l'**Integrated Publishing Toolkit :**

    (https://ipt.sb-roscoff.fr)
    - Téléversement des trois fichiers
    - Si nécessaire, mise en correspondance de colonnes non-standardisées
    - Saisie des métadonnées dans un formulaire (stockées dans un fichier au format EML)
    - Confirmation explicite de la publication (avec gestion de version intégrée).

Mise en Accès

- EMODNet Biology

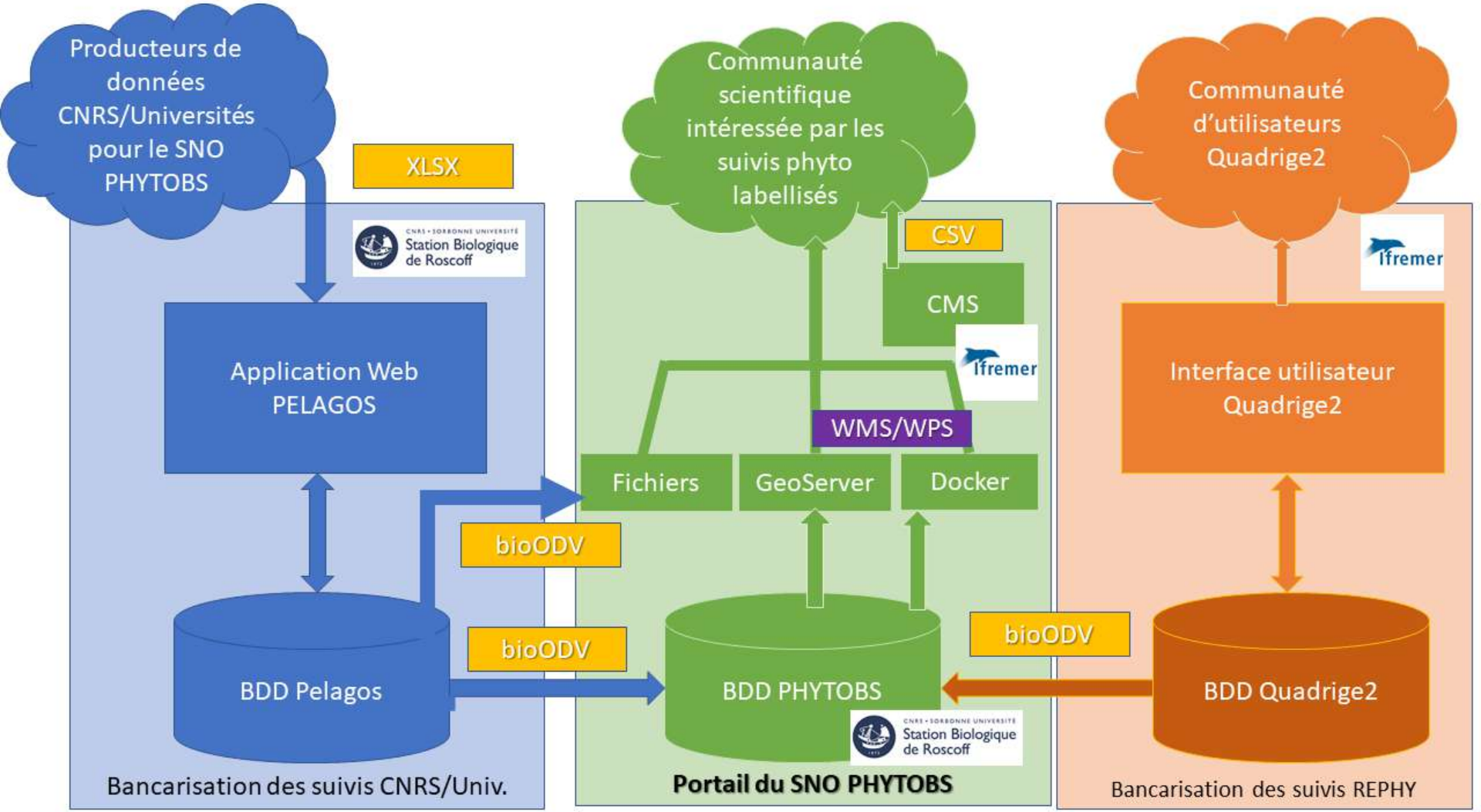




(Courtesy EMODNet Biology)

Mise en accès :

- SNO PHYTOBS :



Retour d’expérience

- Sensibilisation indispensable des experts du domaine :
  - « Investissement » humain non-négligeable dans la préparation des données :
    - Familiarisation avec les référentiels
    - Harmonisation des données avant leur mise en accès (préférablement avant leur bancarisation)

Retour d’expérience

- Bon support de la part des deux interlocuteurs (SDC, EMODNet Bio) :
  - Réactifs par la messagerie
  - Organisation de sessions de formation
  - Documentation abondante et globalement de bonne qualité pour toutes les étapes.
- Faible coût en ressources matérielles :
  - Vms (ou conteneurs) faisant tourner Tomcat.

Retour d'expérience

- SDC :
  - Interface d'accès aux données puissante, bien adaptée à la fabrication d'agrégations de jeux de données (« panier »).
  - Tableau de bord pour le suivi des requêtes concernant « nos » jeux de données.
  - Degré d'automatisation élevé pour les mises à jour.
  - Gestion des DOI ?

Retour d'expérience

- EMODNet Biology :
  - L'IPT offre :
    - Un point d'entrée « générique » et indépendant pour l'accès à l'ensemble des jeux de données mis en accès par une structure (équipe/laboratoire/réseau métier…)
    - Une gestion intégrée des DOI
  - Interface de recherche extraction moins ergonomique que SDC (pas de « panier »)
  - Etapes manuelle encore indispensables lor des mises à jour.
  - Passerelle vers GBIF ?

Bilan

- Capitalisation de l'expérience acquise :
  - Publication de nouveaux jeux de données dans SDC et/ou EMODNet Biology ?
  - Transposition du « modèle » PHYTOBS au SNO Benthobs *en incubation* :
    - Annotation & Curation : démarche d'harmonisation des données / réutilisation des référentiels
    - Mise en accès : même type d'architecture & partage des responsabilités entre partenaires.

contact@odatis-ocean.fr
www.odatis-ocean.fr

Questions ?