RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Le logiciel Collec-Science

Café numérique – juin 2022

Éric Quinton – eric.quinton@inrae.fr

20 juin 2022

EABX – *Écosystèmes aquatiques et changements globaux*

Collec-Science, un logiciel web pour gérer des échantillons

Origine du logiciel

- besoin identifié au sein d'EABX en 2015
  - pêches mensuelles depuis les années 80 en estuaire de la Gironde
  - 24 flacons d'1 litre avec des poissons conservés dans le formol, puis l'alcool, par campagne
  - mise en place d'une salle de lyophilisation (gain de place, sécurité de la manipulation)
- recherche d'un logiciel pré-existant :
  - opensource
  - sécurisé
  - plusieurs logiciels testés, mais aucun couvrant les besoins
- décision d'écrire un logiciel
  - collaboration au sein de l'OASU
    - EPOC : techniques de magasinage, tests des douchettes, imprimantes, étiquettes
    - LIENSs (Christine Plumejeaud) : animation au sein des zones ateliers
  - première version fin 2016

Objectifs

- Assurer la traçabilité des échantillons
  - où sont-ils stockés ?
  - d'où viennent-ils ?
  - que sont-ils ?
  - que deviennent-ils ?

Où sont-ils ?

- gestion des containers (ou contenants) sous forme de « poupées russes »
  - types de contenants configurables
    - ★ sites, bâtiments, pièces, armoires, caisses...
  - position géographique d'une pièce
    - ★ affichage sur une carte
- étiquetage
  - qrcode sur les étiquettes
  - lecture optique pour les entrées/sorties
- toutes les entrées/sorties sont conservées
- possibilité de faire des inventaires
  - identifier les objets manquants

Le logiciel ne gère que des mouvements. Les objets contenus dans une contenant sont calculés à partir de ceux-ci.
Pour retrouver un échantillon, on recherche son dernier mouvement.

D'où viennent-ils ?

- Saisie possible des coordonnées de prélèvement
  - coordonnées GPS
  - stations pré-déterminées, liées ou non à une collection
  - ajout du pays de collecte possible
- un échantillon peut être associé à une campagne de prélèvement
  - enregistrement des réglementations applicables (APA – protocole de Nagoya *p. e.*)
    - saisie d'un numéro d'autorisation et d'une date
    - possibilité de rechercher les échantillons associés à un numéro d'autorisation
  - ajout de pièces jointes (autorisations, etc.)
- gestion des prêts à d'autres laboratoires

Que sont-ils ?

- types d'échantillons totalement configurables
- un type d'échantillon peut être associé à un type de contenant
  - quand le contenu n'est pas dissociable du contenant (tube, sachet de lyophilisation, etc.) : l'échantillon n'a pas d'existence sans son contenant
    - ★ le contenant décrit les conditions de stockage (produit utilisé, risques associés)
  - l'étiquette est collée sur le contenant, pas sur l'échantillon lui-même
- échantillons dérivés :
  - échantillons créés à partir d'un échantillon
  - de nature différente ou non
  - exemple : écailles de poisson, organe, etc.
- suivi des prélèvements au sein d'un échantillon (sous-échantillonnage)
- ajout possible de métadonnées descriptives

Des identifiants multiples

- un numéro interne (UID), associé au code de l'instance Collec-Science
- un identifiant métier principal
  - si besoin, il est possible de le générer à partir des informations saisies
- des identifiants secondaires, de différents types (IGSN, numéro d'inventaire, etc.)
- un numéro universel : UUID (codé sur 32 caractères)
  - sa génération garantit son unicité au niveau mondial

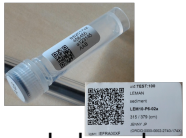
Aucune contrainte ne pèse sur les identifiants métier ou secondaires (saisie libre)

Description « métier » des échantillons – métadonnées

- il est possible d'associer des données « métier »
  - schémas de métadonnées associés aux types d'échantillons
  - possibilité de recherche dans les métadonnées
  - stockage au format JSON
- schémas simples
  - une liste de champs
    - ★ multi-valeurs possibles dans certains cas (listes)
  - ne remplace pas une application dédiée
- les métadonnées peuvent être imprimées sur l'étiquette

La gestion au quotidien

Des étiquettes configurables



- possibilité de créer autant de formats d'étiquettes que nécessaire
- depuis un format A4 (ou plus) pour poser sur une porte...
- jusqu'à des étiquettes d'1 cm de haut pour les tubes, avec QRcode sur le bouchon
  - la version 2.7 introduit la possibilité de générer des étiquettes au format EAN128 (tubes)
- texte libre
  - format défini en xml (exemples fournis)
- génération d'un QRCode :
  - il peut contenir des métadonnées (identifiant, coordonnées géographiques, etc.), stockées au format JSON
  - ou un texte fixe (identifiant de l'échantillon, adresse web)
  - le QRcode peut être lu avec un lecteur optique
    - ★ fonctions de magasinage (entrées, sorties du stock)
    - ★ les contenants sont dotés d'une étiquette
    - ★ utilisation de douchettes professionnelles (identiques à celles des supermarchés)
- les étiquettes peuvent être imprimées avec des imprimantes dédiées
- des supports résistants ont été testés
  - consultez le site web : https://www.collec-science.org

Les collections

- les échantillons sont regroupés au sein de collections
- une collection est un ensemble cohérent d'échantillons représentatifs d'une famille de recherche :
  - carothèque sédimentaire
  - poissons capturés selon un échantillonnage normalisé dans un bassin versant
  - peut comprendre des échantillons divers : poissons, otolithes, prélèvements génétiques, etc.
- un échantillon est rattaché à une et une seule collection
- seuls les personnes membres d'une collection peuvent :
  - créer un nouvel échantillon
  - modifier les informations
  - visualiser les métadonnées

Des fonctions d'importation et d'exportations

- Le logiciel propose plusieurs modules d'importation :
  - importation de masse
    - initialisation des données
    - ajout de données issues de systèmes tiers (saisie en tableur, par exemple)
  - exportation / importation d'échantillons vers une autre instance Collec-Science
    - permet d'échanger des informations entre instances
    - utilisable pour mettre à jour des échantillons en dehors du logiciel
  - exportation / importation d'un contenant avec son contenu
    - transmission d'un contenant avec les échantillons contenus à une autre instance (boite de tubes, par exemple)
- génération d'exportations « à façon »
  - compatibilité avec des formats tiers (GBIF, par exemple)

Accès direct aux échantillons

- Possibilité de déclarer une collection « publique »
  - un lien direct permet de récupérer la description d'un échantillon au format JSON
  - si un modèle d'export est fourni, les données sont formatées
  - les documents peuvent également être téléchargés directement (photos), sans identification préalable
- les liens peuvent être intégrés aux fichiers d'exports « à façon »

Installation

Plate-forme technique

- Serveur Linux, Apache, PHP 7.3, Postfix pour l'envoi des mails (mots de passe perdus)
- Base de données Postgresql 9.5 ou ultérieur (13 conseillé)
- navigateur web pour accéder à l'application
- imprimante à étiquettes, douchette ou smartphone équipé d'un lecteur laser
  - pour la lecture optique des échantillons
  - matériels testés, liste disponible (*cf.* https://www.collec-science.org)
  - modèles d'étiquettes proposés et adaptables par les utilisateurs
- possibilité de gérer plusieurs bases de données avec la même application
  - mécanisme utilisé par les plates-formes d'hébergement, comme celle d'INRAE Toulouse

Installation dans un serveur Linux

- plate-forme technique :
  - Ubuntu ou Debian préconisé (pas d'appui pour les autres OS)
  - script d'installation quasi-automatique fourni, pour installer le serveur web et la base PostgreSQL à partir d'une distribution « vierge »
- nécessite des composants systèmes :
  - une adresse DNS dédiée : *https ://instance.collec-science.inrae.fr* (*p. e.*)
  - un certificat adéquat
- une configuration à finaliser dans Apache (DNS, certificat)
- possibilité d'héberger la base de données dans un autre serveur

Script d'installation automatique :
https://github.com/Irstea/collec/raw/master/install/deploy_new_instance.sh

Plate-forme d'hébergement de Toulouse

- gérée par le CATI GEDEOP
- base de données indépendante pour chaque laboratoire, code partagé
- identification *via* un serveur CAS ou la fédération Renater
- demande d'instance par formulaire Ariane

Démarrer avec Collec-Science

Un projet d'unité

- Informatiser la gestion des échantillons, c'est :
  - identifier les processus actuels
  - rationaliser, harmoniser
  - changer les habitudes
- c'est un projet d'unité :
  - il faut plusieurs mois pour le préparer
  - un chef de projet doit être moteur
  - l'implication de la hiérarchie est essentielle

Un appui externe est possible

- Le concepteur du logiciel peut intervenir pour aider pendant les phases de démarrage
  - appui à la compréhension des concepts sous-jacents
  - conseils en matière de configuration
  - chaque implémentation présente ses propres particularités
- le projet Collec-Science est suivi par un comité de pilotage
  - une formation des administrateurs métiers devrait être proposée par Wilfried Heintz (DYNAFOR / GEDEOP) cet automne
    - pour ses collègues du CNRS intéressés
    - mais pas seulement...

Assurer la pérennité du logiciel : un projet en cours

- un comité de pilotage regroupe des personnes de divers organismes : INRAE, CEFE, CEREGE, OASU (Bordeaux), Université de Savoie-Mont-Blanc, INRAP, etc.
  - des réflexions en cours pour assumer le pilotage du projet
  - des développements hors « collec-science web » envisagés
  - recherche des sources de financement possibles
- parallèlement, des discussions entamées avec la DPTI et la DIPSO

Un souhait du comité de pilotage :

Quand on parle de gestion d'échantillons, on pense « Collec-Science » !

Évolutions

- logiciel OpenSource, disponible dans Github (licence AGPL)
- une nouvelle version par an (version 2.7 prévue courant juin 2022)
- dans les perspectives à moyen terme :
  - automatisation des échanges avec la cyber-carothèque sédimentaire (travaux avec l'OASU à Bordeaux)
  - génération de formulaires de saisie ODK Collect et récupération des données saisies sur le terrain, en partenariat avec le CATI GEDEOP
  - appli Android dédiée au magasinage (entrée/sortie d'échantillons) – collaboration envisagée avec le CEREGE à Montpellier
- modification de l'hébergement dans github : https://github.com/collec-science/collec-science
  - probablement à partir de la version 2.7

Questions ?

Annexes

Fonctions d’importation / exportation

Importation de masse

- Fonction d'importation de masse
  - création des contenants
  - création des échantillons
- utilisée pour :
  - initialiser le logiciel
  - importer les nouveaux échantillons saisis dans des outils tiers (logiciels spécialisés, tableurs, etc.)
- fichier au format CSV

Export-import externe

- utilisé pour échanger les informations sur des échantillons entre deux instances Collec-Science
  - exemple : saisie sur le terrain dans des instances dédiées (Raspberry/tablettes ou micro-ordinateurs portables)
- lors de l'importation, mécanisme d'appariement des types d'échantillons et autres tables de paramètres
- travaille en mode modification / insertion : peut être utilisé pour mettre à jour des échantillons en dehors du logiciel
- fichier au format CSV

Transfert des informations sur les échantillons prêtés

- lors du prêt des échantillons, possibilité de générer un fichier comprenant non seulement les échantillons, mais également les contenants qui les contiennent (boites de tubes *p. e*)
- transmission du fichier généré en même temps que les échantillons
- importation dans une autre instance Collec-Science
- fichier au format JSON

Génération de fichiers d'exports « à façon »

- Création préalable d'un lot d'échantillons
- format d'export totalement configurable :
  - peut contenir plusieurs fichiers, stockés dans un fichier zip
  - formats d'export en csv, json, xml
  - peut porter soit sur les échantillons, soit sur la définition de la collection, soit sur les documents associés
  - les libellés peuvent être transcodés :
    - ⋆ remplacement des noms de colonne par la valeur attendue
    - ⋆ remplacement de certains libellés par ceux attendus (exemple : remplacement de *TRF*, saisi dans Collec-Science, par *Salmo trutta*, nom latin de la truite fario)
- possibilité de relancer un export pour un lot d'échantillons
- possibilité de réaliser des exports dans des formats différents pour le même lot
- le module a été conçu pour permettre l'automatisation
  - échanges informatiques entre des services différents

Ce module permet de créer des fichiers compatibles GBIF (https://www.gbif.org)

Échantillons dérivés et sous-échantillonnage

Échantillons dérivés et sous-échantillonnage

À partir d'un échantillon :

- on extrait un élément
  - il est identifiable individuellement (étiquetage dédié, p. e.)
  - il peut être d'un type différent
- exemples : un otolithe de poisson, une section de carotte de sondage
- c'est un échantillon dérivé

Échantillons dérivés et sous-échantillonnage

À partir d'un échantillon :

- on extrait un élément
  - il est identifiable individuellement (étiquetage dédié, p. e.)
  - il peut être d'un type différent
- exemples : un otolithe de poisson, une section de carotte de sondage
- c'est un échantillon dérivé

- on dispose d'une quantité de matière ou de matériel
  - on peut prélever ou remettre une parcelle de celui-ci pour analyse (éqtitreuivalent à l'aliquote)
  - il est impossible d'identifier précisément ce qui est prélevé
- exemples : 5 écailles prélevées sur un poisson,10 cm^3 de matière
- c'est du sous-échantillonnage
- il est possible de suivre ce qui est prélevé

Un sous-échantillonnage peut devenir un échantillon dérivé.

Sécurité

Gestion des droits

- Cinq droits globaux gérés :
  - consult : visualiser les échantillons, les containers, les mouvements
  - gestion : créer des mouvements, des containers, des échantillons
  - import : import de masse
  - collection : import de masse, gestion des collections
  - param : paramétrage global
  - admin : gestion des droits et des utilisateurs

Identification des utilisateurs

- plusieurs mécanismes utilisables :
  - gestion des comptes dans la base de données
  - utilisation d'un annuaire d'entreprise à la norme LDAP pour identifier les utilisateurs
    - si l'annuaire possède des groupes, possibilité de les utiliser pour donner des droits
  - identification d'abord auprès de l'annuaire LDAP, et si en échec, recherche dans la base des comptes locaux
  - identification sous-traitée auprès d'un serveur d'identification (serveur CAS – *Common Access Service*)
  - identification sous-traitée à une fédération d'identités, comme Renater
    - permet d'identifier des utilisateurs de plusieurs organismes en même temps
    - impose une configuration particulière du serveur web
    - est actuellement incompatible avec l'accès sans identification aux échantillons pour les collections publiques

Sécurité

- traces applicatives dans la base de données
  - qui fait quoi ?
  - purge au bout d'un an
  - consultables directement par les administrateurs de l'application
- messages d'erreur stockés dans *syslog* (système de gestion de traces de Linux)
  - possibilité de moissonnage par des systèmes de gestion de traces
- en cas de mot de passe erroné (identification par la base de données) :
  - blocage temporaire du mot de passe
  - envoi d'un mail aux administrateurs
- sécurité du code : conformité vis à vis de la nomenclature ASVS https://www.owasp.org/index.php/Category:OWASP_Application_Security_Verification_Standard_Project
  - résistance aux attaques dites opportunistes (niveau 1)

Démarrer un projet

Identifiez les personnes impliquées

Rôle	Description
directeur	Ancrage des pratiques dans le laboratoire, attribution des moyens humains et financiers
chef de projet	pilotage du projet
administrateur des données	responsable de la qualité des données, animation du projet
administrateur système	maintien en condition opérationnelle des plates-formes techniques
responsables applicatifs	utilisateurs avancés du logiciel, responsables de collections, etc.
utilisateurs	utilisateurs du logiciel

Une personne peut assumer plusieurs rôles.

(Source : https://hal.inrae.fr/hal-02615072)

Comprenez les principes mis en œuvre dans le logiciel

- logiciel hautement configurable :
  - il peut s'adapter à des contextes différents
- points principaux à aborder :
  - fonctionnement général, règles d'ergonomie
  - qu'est-ce qu'un contenant ?
  - qu'est-ce qu'un échantillon ?
  - la gestion des mouvements
  - qu'est-ce qu'une collection ?
  - comment sont gérés les droits ?
  - les étiquettes, les QRcodes

Apprenez le fonctionnement du logiciel

- support de formation disponible dans le site web
- les formations organisées, dans le cadre des zones-ateliers ou autres, seront publiées dans le site web
- utilisez la liste de diffusion pour poser des questions :
  - collec-users@groupes.renater.fr
- sollicitez l'intervention d'un référent pour expliciter certains points
  - mis en place dans le cadre des zones-ateliers ?
  - appui possible du développeur en troisième niveau
- gestion des droits :
  - des vidéos disponibles dans le site

Pourquoi tester ?

- les premières configurations sont rarement les meilleures
  - difficultés de compréhension de certains mécanismes internes au logiciel
  - mauvaise définition des types d'échantillons
    - recherche de normalisation inter-laboratoires parfois souhaitable (carothèques, par exemple)
- l'apprentissage est plus facile en testant et en recommençant
- il est possible de revenir à une base de données propre
  - création d'un autre schéma dans la base de données
  - récupération des paramètres
    - soit par export/import pour certaines informations
    - soit par recopie des données depuis le schéma de test

Choisissez une collection pilote

- critères de choix :
  - pas trop complexe
  - des utilisateurs motivés et acceptant « d'essuyer les plâtres »
  - limitez le nombre d'échantillons à intégrer dans une première phase
- définissez les personnes ou les groupes de personnes qui vont la manipuler
- définissez le référent de la collection
  - peut être un nom générique (équipe de recherche)

Définissez les contenants

- décrivez les lieux de stockage
  - bâtiments, salles, armoires, frigos...
- décrivez les boites, caisses, etc. utilisés pour ranger les échantillons
  - limitez-vous à ceux qui seront effectivement utiles pour la phase de démarrage
- vous pouvez utiliser l'importation de masse pour faciliter la création des contenants

Définissez les types d'échantillons

- identifiez la généalogie des échantillons :
  - comment les échantillons évoluent-ils ?
    - un otolithe, des écailles sont prélevés depuis un poisson
    - une carotte est découpée en sections, puis en demi-sections. Des prélèvements de différents types sont réalisés ensuite
- caractérisez chaque type d'échantillon :
  - existe-t-il en plusieurs exemplaires dans le même sachet, peut-on lui attribuer un volume ?
    - gestion du sous-échantillonnage
  - présente-t-il des caractéristiques qu'il convient d'identifier ?
    - *top* et *bottom* d'une section de carotte
    - *taxon* d'un poisson
    - notion de métadonnées
  - décrivez les types de métadonnées dont vous aurez besoin
    - normalisez les libellés !

Importez les échantillons

- utilisez la fonction d'importation de masse :
  - commencez par créer les contenants
  - puis ajoutez les échantillons « parents »
  - et finissez par les échantillons « enfants »
- après chaque importation, notez les UID générés : en cas de problème, ce sera plus simple de les retrouver

Identifiez les étiquettes nécessaires

- pour chaque type d'échantillons :
  - identifiez les étiquettes dont vous aurez besoin
    - taille
    - informations textuelles à afficher
    - informations à intégrer dans le QRcode
- créez vos modèles d'étiquettes
  - une liste de modèles est disponible dans le site web
- identifiez les imprimantes, douchettes ou smartphones dotés d'un lecteur laser dont vous aurez besoin
  - décrivez les modes opératoires envisagés
- des recommandations de matériel sont proposées dans le site web
  - travaux d'expérimentation réalisés dans le cadre des zones-ateliers

Décrivez le mode opératoire

- décrivez comment vont être gérés les échantillons : mode de création, rangement, etc.
- identifiez les ressources dont vous aurez besoin :
  - imprimantes
  - douchettes
  - smartphones dotés d'un lecteur laser pour les opérations de magasinage

Réalisez un bilan

- réalisez un bilan de l'expérimentation
  - avantages escomptés
    - structuration
    - facilité de gestion
    - ...
  - contraintes rencontrées
    - besoin en structuration
    - temps passé à la phase d'initialisation
    - ...
- présentez les résultats au laboratoire
  - facilitation de l'appropriation
  - identification des collections à intégrer
  - calendrier

C’est parti !

- achetez le matériel adéquat
  - imprimantes, lecteurs laser, smartphones avec lecteurs laser, etc.
  - rouleaux d’étiquettes
- si nécessaire, réinitialisez la base de données
- formez les utilisateurs
- faites des bilans réguliers de mise en œuvre
  - au moins pendant la phase de démarrage
  - vérification de l’adéquation des moyens (humains, financiers)
  - accompagnement des utilisateurs
  - s’assurer que les anciennes habitudes ne reprennent pas le dessus
- remontez vos besoins ou les bogues rencontrés :
  https://github.com/Irstea/collec/issues/new

API

Les API

- plusieurs API définies
  - une API par fonction (pas d'API générique)
- liste des API :
  - consultation d'un échantillon
  - création/modification d'un échantillon
  - récupération d'un document
- à venir (version 2.8) :
  - création d'un mouvement d'entrée/sortie
  - API de recherche
- une réflexion en cours au sein du comité de pilotage pour (re)développer les API dans une appli tierce

Gestion des droits dans les API

- un compte dédié, géré comme les autres comptes de l'application
  - le mot de passe est remplacé par un *token* cryptographique
  - le compte doit être associé à une collection pour pouvoir récupérer les informations
- pour les collections déclarées « publiques »
  - création d'un *template* permettant de définir les informations affichées et de transcoder les libellés (nom des champs, voire contenu)
  - accès sans compte ni *token*

Installation Docker

Installation Docker

- Pour quels usages ?
  - instance « terrain » pour faire de la saisie hors labo
  - pour tester sur son poste de travail
- compatible Raspberry, Windows 10...
  - Raspberry : le terminal de saisie se connecte en wifi
  - Windows : on utilise le navigateur du PC
    - ⋆ configuration particulière du fichier *hosts*
- deux containers créés :
  - un pour la base de données
  - un pour l'application
- lors des mises à jour, seul le container applicatif est régénéré
- des scripts Docker et un mode d'emploi fournis

https://github.com/jancelin/docker-collec

Mise au point réalisée par Julien Ancelin et Christine Plumejeaud