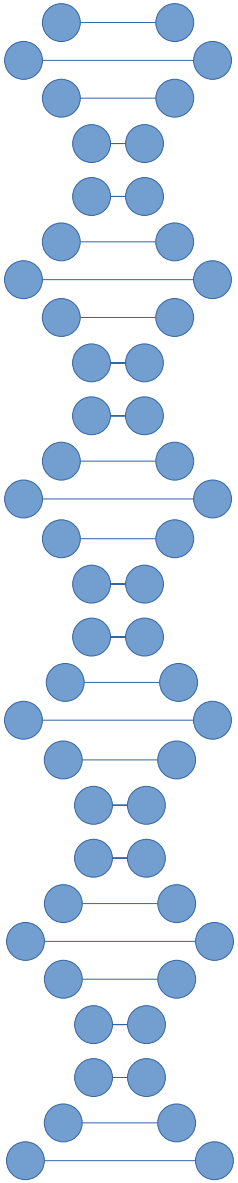
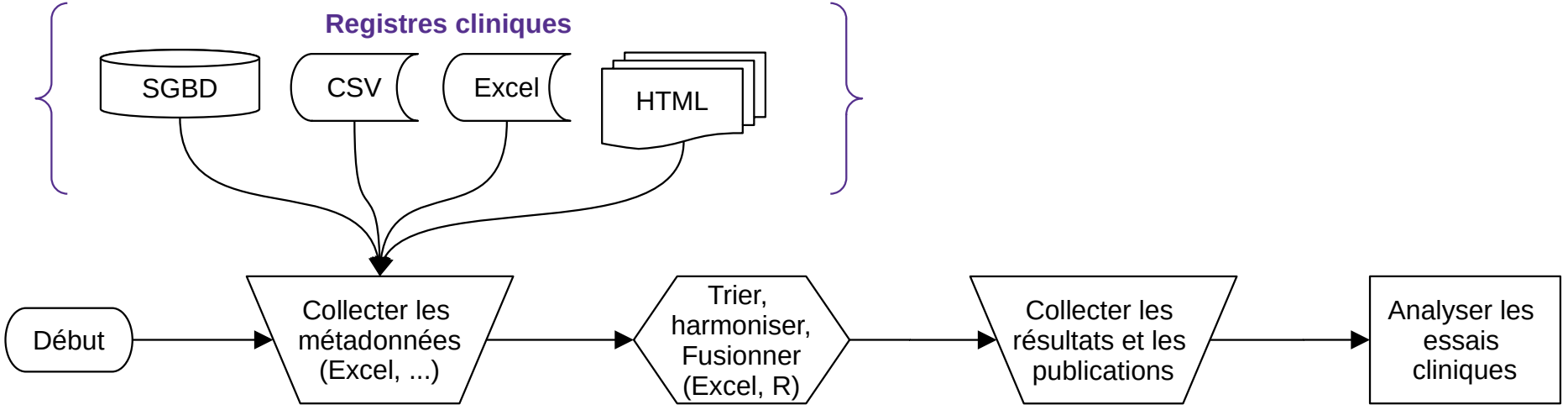


Histoire d’un code improbable pour le COVID

Françoise CONIL, meetup Python Lyon, 22 mars 2023

Le contexte : analyse d'essais cliniques

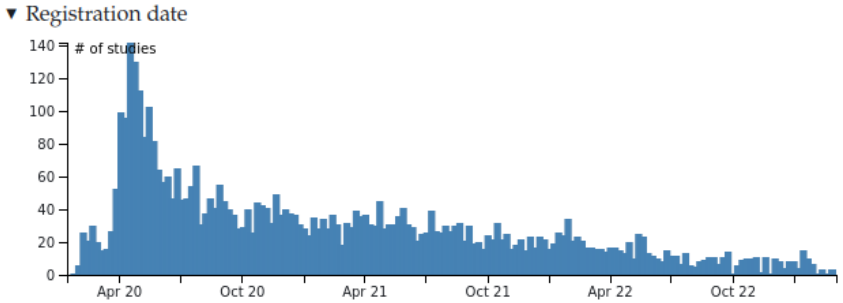
- Équipe METHODS [1] du CRESS (Centre of Research in Epidemiology and StatisticS)
- Méta-analyse des essais cliniques menées dans le monde
- Analyser la qualité des études et l'efficacité des traitements testés



[1] https://cress-umr1153.fr/index.php/methods/

La pandémie

- Explosion des essais cliniques sur le COVID-19
- 20 annotateurs à temps plein complètent, analysent et synthétisent ces données
- Le CRESS fait part de son besoin d'automatisation de tâches



- Avril 2020 : le CNRS lance un appel à volontaires
- Une équipe hétérogène et dispersée se constitue

[1] CRESS : https://cress-umr1153.fr/
[2] LRI => LISN : https://www.lisn.upsaclay.fr/
[3] LIMOS : https://limos.fr/
[4] LIRIS : https://liris.cnrs.fr/

Un réel travail collaboratif

- On ne se connaît pas, on est dispersés, on est confinés
- On ne connaît pas le domaine

Il nous faut des outils de communication et de partage

- Chat : Rocket chat avec fédération d’identité
- Visio : Jitsi, Webex, BBB,
- GitLab, NextCloud

Définition de règles communes

- Formater : Black
- Linter : Flake8
- Documentation développeur sur le Wiki du GitLab
- Utilisation des issues et des merge request

Objectif : créer un entrepôt de données

Bases

- Framework Django et langage Python
- Base de données PostgreSQL
- Différentes librairies Python : Requests, Beautiful Soup, Pandas, Jupyter Notebooks, lxml, Selenium

Plus tard :

- Déploiement sur VM Proxmox avec Ansible
- Docker pour le développement

Registres d’essais cliniques : une obligation

- La déclaration de tous les essais « interventional » est considérée comme une obligation scientifique, éthique et morale [1]
- La déclaration d’Helsinki [2] stipule que chaque essai devrait être enregistré dans une base de données publiquement accessible avant le recrutement du premier patient
- Les essais sont obligatoirement déposés dans un registre
- Ils peuvent être déposés dans plusieurs registres

[1] https://www.who.int/news-room/questions-and-answers/item/clinical-trials
[2] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Rassembler les essais cliniques : intérêt

- Les essais sont dispersés dans plusieurs registres avec des doublons entre les registres
- L'OMS collecte les données de registres (data providers [1]) qui respectent un ensemble de critères [2]
- Les registres européens et américains ne sont pas considérés par l'OMS comme des registres nationaux
- Méta-analyses prospectives *(les chercheurs peuvent utiliser le projet « COVID-NMA » pour suivre les essais cliniques menés dans le monde)*
- Cette obligation devrait permettre d'éviter la duplication d'essais cliniques : plus de 150 essais sur l'hydroxychloroquine [3]

[1] https://www.who.int/clinical-trials-registry-platform/network/who-data-set/data-providers
[2] https://www.who.int/clinical-trials-registry-platform/network/registry-criteria
[3] https://covid-nma.com/treatments_tested

Registres d’essais cliniques : pb de qualité

- Standards sur l’échange de données provenant d’études cliniques [1] mais pas sur les métadonnées concernant ces études [2]
- Nombreuses publications sur les problèmes de qualité de données dans le registre OMS et dans les registres primaires et depuis longtemps
  - Obstacles to the reuse of study metadata in ClinicalTrials.gov [2]
  - The Quality of Registration of Clinical Trials: Still a Problem [3]
  - Exploring Data Quality Management within Clinical Trials [4]

[1] https://www.cdisc.org/standards/data-exchange/define-xml
[2] https://www.nature.com/articles/s41597-020-00780-z
[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084727
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5801732/

Répartition des sources 1/2

- **OMS** : https://www.who.int/clinical-trials-registry-platform
- **USA** : https://clinicaltrials.gov
- **Europe** : https://www.clinicaltrialsregister.eu/
- **Iranien** : https://en.irct.ir
- Indien : http://ctri.nic.in/Clinicaltrials/advancesearchmain.php
- Australien : https://www.anzctr.org.au/TrialSearch.aspx
- Anglais : https://www.isrctn.com/

En gras : les sources qui sont actuellement utilisée en production

Répartition des sources 2/2

- **OMS** : fichier CSV (Sarah, Marinette, Marie, Ruben)
- **USA** : accès à une base PostgreSQL (Sarah, Farouk, Brice, Emmanuel, Benoît, Ruben, Virginie, Nicolas)
- **Europe** : scraping (Sarah, Françoise, Bastien, Ruben)
- **Iranien** : XML et scraping (Emmanuel, Ruben)
- Indien : scraping (Merwan, Françoise)
- Australien : XML et Excel (Marinette, Marie, Françoise, Simon, Lamine)
- Anglais : API qui délivre du XML (Françoise)

En gras : les sources qui sont actuellement utilisée en production

Développements parallèles

- Démarrage : données OMS insuffisantes, arrivant avec beaucoup de délai au début (jusqu’à 3 semaines)
- Compléter les données des essais cliniques recensés par l’OMS avec les données des registres nationaux/internationaux
- Registres plus ou moins détaillés, données non homogènes selon les registres, accès variables et registres parfois inaccessibles (se produit encore)
- Répartition des registres entre les développeur·ses
- Travail en local, beaucoup d’A/R, de mise au point et d’échanges avec CRESS pour comprendre les données extraites et les transformations à appliquer qui évoluent au fil des échanges

Source OMS 1/3

- Les informations ne sont pas obtenues directement des registres
- L’OMS envoie un fichier aux registres chaque semaine et les registres répondent plus ou moins dans les temps, plus ou moins correctement
- C’est l’OMS qui décide du contenu et du format de ce fichier
- L’OMS ne fait aucun traitement
- Le registre américain a refusé de signer un accord avec l’OMS. Les américains envoient leurs informations à leur format, 2 fois par semaine. Les données du registre américain représenterait la majorité des lignes du fichier OMS (> 50%)

[1]

Source OMS 2/3

- Utilisation de Pandas [1] pour extraire et transformer les données
- Développement avec des notebooks Jupyter
- Agrégation des données des différentes sources, problème majeur de qualité de données
  - Types incohérents
  - Données manquantes
  - Utilisation de syntaxe multiples pour le même champ
  - Mélange de langues
  - Ne pas utiliser les ontologies du domaine pour certains champs (molécules pharmaceutiques)
- Pression forte lors du premier confinement, pas le temps d'étudier les outils en détail

[1] https://pandas.pydata.org/

Source OMS 3/3

Quelques exemples de valeurs, toutes les variations ne sont pas présentes !

champ « **Inclusion agemin** »		champ « **Inclusion gender** »		champ « **Date registration** »	
18 Years	5819	All	6813	20/04/2020	75
18	510	Both	2296	2020-03-15	33
18 years	487	 Female: yes Male: yes 	573	31-08-2020	15
no limit	93	Both, male and female	420		
18Y	93	Female	232		
20years-old	82	Male and Female	191		
>= 20age old	68	Both males and females	177		
no minimum age	57	-	147		
19 Year(s)	52	Male	98		
Not applicable	51	Male/Female	65		
1 Month	16	Females	7		
18?(Year)	8	 Female: yes Male: no 	4		
N/A (No limit)	8	BOTH	4		
0 Day(s)	2	F	3		
1months-old	1	B	2		
18 años	1	Not Specified	2		
2??	1	M	1		
0.1	1				

Registre USA

- Base PosgreSQL, grosse requête pour extraire et transformer les données
- Registre plus propre, source la plus accessible qui propose des dumps des données plus anciennes
- Pas réussi à utiliser les foreign data wrapper, trop de jointures peut-être
- Récemment des erreurs « too many connections »
- Stagiaire pour comparer la fréquence de changement des données, en comparant des dumps plus ou moins lointains [1]
- Stagiaire pour reprendre la requête et requêter uniquement les données brutes (historisation, séparation extraction / transformation)
- Tout est fait en SQL pour ce registre

[1] Étudie pour deux dumps, entre telles dates et telles dates, quelles sont les études pour lesquelles le nombre de « recruitment status » a changé. Pour valider que l'historisation du DWH fonctionne bien

Registre Européen 1/4

- Un export texte plein [1], sans séparateur, difficilement exploitable
- Pages web plus structurées : scraping pour extraire les données
- Scraping en Python avec BeautifulSoup [2] et Requests [3]
- Exécution d'une recherche simple, extraction des informations des pages de résultats et des pages de chaque essai
- Enregistrement de toutes les pages sur disque, au départ pour pouvoir reprendre l'extraction même en cas d'indisponibilité du site. Sert maintenant à l'historisation
- Les champs ne sont affichés que s'ils ont des données. On découvre de nouveaux champs suivant les essais affichés. Impossible d'obtenir la structure complète (réunion avec dev registre européen le 16/7/2020)

[1] https://www.clinicaltrialsregister.eu/ctr-search/trial/2021-004016-26/FI => cliquer sur le lien « download »
[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] https://requests.readthedocs.io/en/latest/

Registre Européen – texte plein – 2/4

```
Summary^M
EudraCT Number: 2021-004016-26^M
Sponsor's Protocol Code Number: Clin_COVID-19_Corok^M
National Competent Authority: Finland - Fimea ^M
Clinical Trial Type: EEA CTA^M
Trial Status: Ongoing^M
Date on which this record was first entered in the EudraCT database: 2021-09-14^M
Link: https://www.clinicaltrialsregister.eu/ctr-search/trial/2021-004016-26/FI/^M
^M
F.3 Group of trial subjects^M
F.3.1 Healthy volunteers: Yes^M
F.3.2 Patients: No^M
F.3.3 Specific vulnerable populations: No^M
F.3.3.1 Women of childbearing potential not using contraception  ^M
(For clinical trials recorded in the database before the 10th March 2011 this question read: "Women
of childbearing potential" and^M
did not include the words "not using contraception". An answer of yes could have included women of
child bearing potential whether^M
or not they would be using contraception. The answer should therefore be understood in that
context. This trial was recorded in the^M
database on 2021-09-14) : Yes^M
F.3.3.2 Women of child-bearing potential using contraception: Yes^M
```

[1] https://www.clinicaltrialsregister.eu/ctr-search/rest/download/trial/2021-004016-26/FI

Registre Européen – page web - 3/4

Summary	
EudraCT Number:	2021-004016-26
Sponsor's Protocol Code Number:	Clin_COVID-19_Corok
National Competent Authority:	Finland - Fimea
Clinical Trial Type:	EEA CTA
Trial Status:	Ongoing
Date on which this record was first entered in the EudraCT database:	2021-09-14

A. Protocol Information		
A.1	Member State Concerned	Finland - Fimea
A.2	EudraCT number	2021-004016-26
A.3	Full title of the trial	Substudy "Responses to Covid-19 vaccines" in research "Clinical picture, immunology, genetics and pathogenesis of COVID-19 infection" "COVID-19 infektion taudinkuva, immuunivaste, genetiikka ja patogeneesi"- tutkimuksen alatutkimus "Koronarokotteen aikaan saama vaste"

[1] https://www.clinicaltrialsregister.eu/ctr-search/trial/2021-004016-26/FI

Registre Européen 4/4

- Structure hiérarchique avec des sections qui peuvent apparaître plusieurs fois
- Extraction des données dans un dictionnaire, enregistrement des données brutes dans un fichier JSON avant la transformation
- Permet de faire des contrôles avec l'outil jq [1]

```
$ ls -1 EudraCT*.json | while read A; do echo "$A"; cat "$A" | jq '[.sections[] | select (.id == "section-b") | .subsections[] | .data[] | select (.id == "B.4.1")] | length' ; done
```

- Technologie fragile qui sera cassée lors des modifications de site et on perdra tout sur un changement de site [2] mais cela fonctionne depuis 2 ans
- Les changements doivent être approuvés par les 27 états membres
- Peu de champs obligatoires pour ne pas décourager la saisie
- Transformation des données avec Pandas

[1] https://stedolan.github.io/jq/
[2] Nouveau site : https://euclinicaltrials.eu/search-for-clinical-trials ?

Registre Iranien

- Export au format XML mais incomplet, extraction des données en pur Python au lieu d'utiliser la librairie lxml, notamment lxml.objectify [1] qui permet de traiter le XML comme une hiérarchie d'objets Python
- Scraping [2] des pages du registre pour compléter les données
- Transformation des données avec Pandas

[1] https://lxml.de/objectify.html
[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Registre Australien

- Export au format XML et Excel
- Pas d’accès direct aux données par des URL, nécessité d’utiliser un outil d’automatisation du navigateur pour déclencher le téléchargement : Selenium [1]
- Peu d’essais (politique zéro COVID), source écartée tant que les sources actuelles suffisent au CRESS pour le COVID

[1] https://selenium-python.readthedocs.io/index.html

Registre Indien

- Scraping engagé par un étudiant
- Pas d'accès direct au XML par des URL, nécessité d'utiliser un outil d'automatisation du navigateur pour déclencher le téléchargement :
  Selenium [1]
- Protection de l'accès au site par un captcha [2]
- Filtrage des adresse IPs
- Échec de l'enregistrement de certains essais en UTF-8 (pb d'encodage)
- Trop lourd, source écartée tant que les sources actuelles suffisent au CRESS pour le COVID

[1] https://selenium-python.readthedocs.io/index.html
[2] OCR Tesseract https://tesseract-ocr.github.io/tessdoc/ avec wrapper Python https://pypi.org/project/pytesseract/
[3] https://requests.readthedocs.io/en/latest/

Registre Anglais

- Une API d’accès qui permet de récupérer les données au format XML
- Limitation du nombre d’essais récupérés (1000 max), mais pas de pagination disponible
- Ce n’est pas vraiment un registre, c’est un site éditeur (Springer)
- Suite au BREXIT, il était probable que les anglais mettraient leurs essais clinique sur le registre américain, plus visible ce qui est le cas pour de nombreux essais
- Source écartée tant que les sources actuelles suffisent au CRESS pour le COVID

[1] https://lxml.de/objectify.html
[2] https://requests.readthedocs.io/en/latest/

Première livraison au bout de 4 mois

- Déclenchement des collectes par CRON
- Données brutes enregistrées sur disque jusqu’à la prochaine collecte
- Stockage des données collectées dans la base PostgreSQL
- Export CSV / Excel pour chaque source via une interface Web
- On doit reprendre nos projets / enseignements
- Recrutement d’un développeur qui va améliorer ce prototype pendant 2 ans

Aujourd’hui

- Toutes les données brutes sont conservées en évitant de stocker des données redondantes
- Les données ont été harmonisées grâce à des tables que le CRESS peut renseigner
- Un historique détaillé des modifications des essais existe
- Le déploiement est automatisé (Ansible, Docker)
- Toutes les annotations sont faites dans l’entrepôt directement
- Plus personne pour le développement : ouvrir le code en open source pour éviter qu’il disparaisse

Projet récompensé par un cristal collectif CNRS

Tous ceux qui ont participé à ce projet

- **Développeur·se·s**
  - Marinette BOUET (LIMOS – Clermont-Ferrand)
  - Nicolas CHAMAND (stagiaire LIRIS – Lyon)
  - Sarah COHEN-BOULAKIA (LISN – Paris)
  - Françoise CONIL (LIRIS – Lyon)
  - Emmanuel COQUERY (LIRIS – Lyon)
  - Bastien DOREAU (LIMOS - Clermont-Ferrand)
  - Merwan EL ASRI (stagiaire LRI - Paris)
  - Benoît GROZ (LRI - Paris)
  - Simon KLOPFENSTEIN (stagiaire LIRIS - Lyon)
  - Ruben MARTINEZ (LIMOS - Clermont-Ferrand)
  - Lamine MBOUP (stagiaire LIRIS - Lyon)
  - Brice MEYER
  - (stagiaire LIMOS - Clermont Ferrand)
  - Virginie NOËL (LISN - Paris)
  - Marie SAUVANT (stagiaire LIMOS - Clermont Ferrand)
  - Farouk TOUMANI (LIMOS - Clermont-Ferrand)
- **CRESS**
  - Hillary BONNET (CRESS – Paris)
  - Isabelle BOUTRON (CRESS – Paris)
  - Gabriel FERRAND
  - Thu Van NGUYEN
  - Carolina RIVEROS (CRESS - Paris)

Utilisation des données des essais

Contact de 224 auteurs d’articles liés COVID-19 en leur demandant le partage des données (avec association à l'article final + soutien logistique et anonymisation)

- 50% de réponses
- 22% ont envoyés leurs données
- Différence importante entre ce qui est dit (ouverture) et objectivement ce qui est fait