



Université Mohamed Sedik Benyahia Jijel
Faculté de science de la nature et de la vie
Département de biologie moléculaire et cellulaire

TP :logiciel libre et open source

Etude théorique et pratique et exploration de zenodo

Préparé par : _ Youla Dounia

_ ladra riham

Réalisé par : A_Bensalem

Année universitaire :2025/2026

Plan de travail :

- **Partie 1** :Étude théorique
- Présentation générale de l'outil(EMBOSS)
- Fonctionnalités principales
- Aspects techniques
- Points forts Limites et points faibles
- Conclusion
- Référence
- **Partie 2** : Étude pratique Exploration de Zenodo :
 - Recherche,
 - sélection et téléchargement d'un dataset pertinent.
 - Analyse des métadonnées : Présentation structurée des métadonnées selon la norme choisie (Darwin Core ou Dublin Core). Qualité globale du rapport (2 points) Clarté, structuration, qualité de la forme et des références.
 - Bonus :Création d'un compte GitHub, dépôt

Partie 1 :étude théorique de l'outil EMBOSS

1 /présentation général de l'outil :

EMBOSS (European Molecular Biology Open Software Suite) est une suite de logiciels libres et open source dédiée à l'analyse des données biologiques, notamment les séquences d'ADN, d'ARN et de protéines. Elle a été développée pour fournir aux chercheurs et aux étudiants un ensemble d'outils fiables pour la bioinformatique et la biologie moléculaire. Développé à l'origine par l'EMBL-EBI, EMBOSS met à disposition plusieurs centaines d'outils en ligne de commande permettant d'effectuer des analyses courantes en biologie moléculaire, aussi bien en recherche qu'en enseignement ;EMBOSS est largement utilisée dans les laboratoires de recherche et les institutions académiques, et fonctionne principalement via une interface en ligne de commande.

2/Principales fonctionnalités :

EMBOSS comprend plus de 100 programmes en ligne de commande couvrant de nombreux aspects de la biologie moléculaire. Parmi ses fonctionnalités principales

- **Analyse de séquences** : traduction des séquences nucléotidiques en protéines, recherche de motifs, complément inverse de l'ADN, calcul de la composition.
- **recherche dans les bases de données** : alignements de paires de séquences, recherche de motifs et interrogation de bases de données locales ou distantes.
- **Analyse des protéines** : prédition de la structure secondaire, sites de clivage, hydrophobicité, calcul du poids moléculaire.
- Conversion de formats de séquences : conversion entre différents formats (FASTA, GenBank, EMBL...).
- **Outils de visualisation** : génération de graphiques, diagrammes et rapports pour représenter les caractéristiques des séquences.

3/Aspects techniques

- **Plateforme** : multiplateforme (Linux, Windows, macOS).
- **Langage de programmation** : principalement en C, avec certains scripts en Perl et Python.
- **Interface** : outils en ligne de commande, avec interfaces graphiques optionnelles comme JEMBOSS.
- **Intégration** : compatible avec les bases de données biologiques (NCBI, UniProt) et autres logiciels bioinformatiques.
- **Licence** : open-source sous GNU General Public License (GPL), permettant l'utilisation, la modification et la redistribution gratuites.

4/ Points forts :

Libre et open-source : gratuit et modifiable.

Large éventail d'outils : solutions pour de nombreuses analyses en biologie moléculaire.

Intégration facile : compatible avec d'autres logiciels et bases de données.

Multiplateforme : fonctionne sur différents systèmes d'exploitation.

Documentation **complète** : manuels, tutoriels et communauté active.

5. /Limites et faiblesses :

Interface en ligne de commande : peut être difficile pour les débutants sans connaissances en informatique.

GUI limitée : les interfaces graphiques sont moins développées que dans certains logiciels commerciaux.

Maintenance : certains outils anciens ne sont plus activement mis à jour.

Performance : pour des analyses à grande échelle, des scripts supplémentaires ou outils externes peuvent être nécessaires.

6/ Conclusion :

EMBOSS est une suite logicielle puissante, polyvalente et open-source pour la biologie moléculaire et la bioinformatique. Son large éventail d'outils et sa flexibilité en font un outil précieux pour la recherche et l'enseignement. Bien que l'interface en ligne de commande puisse représenter une difficulté pour les débutants, la disponibilité de tutoriels et le soutien de la communauté rendent EMBOSS accessible et fiable pour l'analyse de séquences et les flux de travail bioinformatique.

Références:

1.Rice P., Longden I., Bleasby A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 16(6), 276–277.

2.EMBOSS official site: <https://emboss.open-bio.org>?

3.Wikipedia: EMBOSS – <https://en.wikipedia.org/wiki/EMBOSS>?

PARTIE2 -Étude pratique :

Exploration de zenodo :

1/présentation de plateforme zenodo :

Zenodo est une plateforme de dépôt numérique en libre accès destinée à la diffusion et à la conservation des productions scientifiques. Elle a été développée pour soutenir le mouvement de la science ouverte (Open Science) en offrant aux chercheurs un espace fiable pour partager leurs données et leurs travaux.

Objectifs de la plateforme :

- Assurer l'accès libre aux données scientifiques
- Permettre l'archivage à long terme des ressources de recherche
- Attribuer un identifiant permanent (DOI) à chaque dépôt
- Favoriser la réutilisation et la reproductibilité des recherches

Types de contenus hébergés :

Zenodo héberge plusieurs types de ressources scientifiques :

- Jeux de données (datasets)
- Logiciels et codes sources
- Publications scientifiques
- Présentations et posters
- Supports pédagogiques

Importance de Zenodo pour la science ouverte et la recherche en NLS :

Zenodo joue un rôle fondamental dans la diffusion des connaissances scientifiques. Il permet aux chercheurs en sciences naturelles et sciences du langage (NLS) de partager leurs données, d'améliorer la transparence des recherches et de renforcer la collaboration scientifique à l'échelle internationale.

2/ description des étapes réalisées :

2 .1 Accès a la plateforme zenodo :

La plateforme est accessible via l'adresse suivante :

<https://zenodo.org>

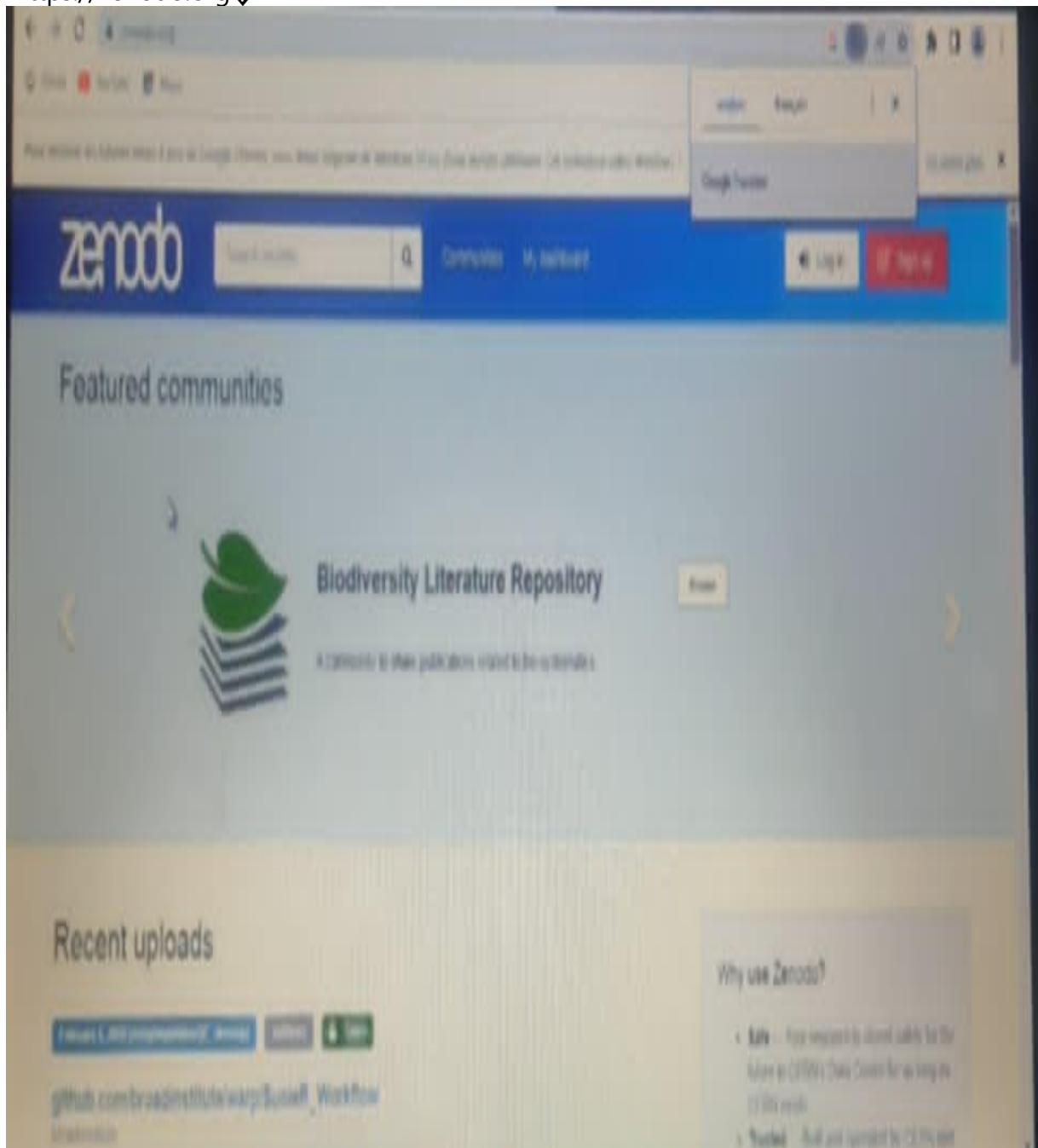


Figure 1 :page d'accueil de la plateforme zenodo

2.2 recherche effectuée d'un dataset

The screenshot shows the Zenodo search interface with the following details:

- Search results**: zenodo.org
- zenodo** logo
- 29,873 result(s) found**
- Sort by**: Best match
- June 19, 2024 (2.0.0)**, **Software**, **Open**
- github.com/genome/qc-analysis-pipeline**
genome
- qc-analysis-pipeline Workflows used for QC of WGS or
WES data Single Sample QC This WDL pipeline...
- Part of Dockstore
- Uploaded on June 11, 2025
- Views: 57, Downloads: 31
- March 7, 2023 (v1.0.1.beta.zenodo)**, **Software**, **Open**
- shohei-kojima/MEGANE:**
v1.0.1.beta.zenodo
- ShoheiKojima; Genome Immuno
- To obtain DOI.
- Uploaded on March 7, 2023
- Views: 266, Downloads: 40
- February 27, 2019 (022719)**, **Dataset**, **Open**
- proportion expressed across transcripts**
(pext)

Figure 2 :search résulte for mot-clé :génome

2.3 sélection du dataset :

Parmi les résultats obtenus, un dataset pertinent a été sélectionné selon les critères suivants :

Type de document : Dataset

Domaine scientifique : Génomique

Métadonnées complètes et bien structurées

Accès libre (Open Access)

Published February 27, 2019
| Version 022719

Dataset

Open

proportion expressed across transcripts (pext)

Genome Aggregation Database Production Team  :
Genome Aggregation Database Consortium 

Contributors

Producers:

Genome Aggregation Database Production Team  :
Genome Aggregation Database Consortium 

Original file that was previously found here:

https://storage.googleapis.com/gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvts.annotated.022719.tsv.bgz

Now only a newer file is available:

https://storage.googleapis.com/gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvts.annotated.GTFv_v7_021520_tsv.hg38



Figure 3 :page de dataset proportion expressed across transcripts

Le dataset sélectionné est intitulé Proportion expressed across transcripts. Il a été choisi car il est lié à l'analyse de l'expression génétique dans le domaine du génome

Now only a newer file is available.

https://storage.googleapis.com/gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvs.tx_annotated.GTEx.v7.021520.tsv.bgz

which is differs from the original.

Files



Additional details

Related works

Is part of

Journal article: [10.1038/s41586-020-2329-2 \(DOI\)](https://doi.org/10.1038/s41586-020-2329-2)

References



Figure 4 :téléchargement du dataset depuis la plateforme zenodo

Les fichiers du dataset ont été téléchargés à partir de la plateforme Zenodo afin de permettre leur analyse ultérieure

3/métadonnées du dataset

3_1 norme utilisée :

Les métadonnées ont été récupérées selon la norme Dublin Core, une norme internationale largement utilisée pour la description des ressources numériques

2/29

Log of changes

Date	Mvm	Who	Description
27/11/2023	0V1	Troels Rasmussen, NGC	First Draft
13/02/2024	0V2	Mercedes Rothschild Steiner (ELIXIR Hub)	Copy circulated to the GDI-MB for review
20/02/2024	0V3	Troels Rasmussen, NGC	Comments addressed
29/02/2024	1V0	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to the EC portal

Table of contents

Contents

1. Executive Summary	4
Context	4
2. Contribution towards project outcomes	5
3. Methods and framework	7
3.1 Cost estimation framework	7
4. Cost framework for 1+MG EDIC	11
4.1 Central Hub cost units	12
Organisational costs - Hub Coordination and Management	13
Operational costs - Central data and user services	14
Development of Services - Central infrastructure	14
4.2 Cost units with Local/national nodes	15
Organisational costs - National coordination and management	15
Organisational costs related to data inclusion (including storage) at national level	16
Organisational costs related to User Services	17
Operational costs - SME Serving user requests (computing)	17

2

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

18

Development of services of the national infrastructure

5. Initial discussion on cost elements	10
5.1 Central Hub Staffing	10
5.2 Considerations on hardware investments and cost recovery models	21
5.2.1 Calculating compute costs	21
5.2.2 Free-at-the-point-of-use vs. user fee	23
5.4 Data Inclusion	24
5.5 Data Storage	26
5.6 Information security management	27
References	29

1

Deliverable D2.1

First analysis of cost elements for the setup of 1+MG infrastructure

Project Title	Grant agreement no
Genomic Data Infrastructure	Grant agreement 101081813
Project Acronym (EC Call)	GDI
WP No & Title	WP2: Long term sustainability
WP Leaders	Regina Becker (S. UNILU), Troels Rasmussen (13. NGC)
Deliverable Lead Beneficiary	13. NGC
Contractual delivery date	31/01/2024
Actual delivery date	29/02/2024
Delayed	[Yes]
Partner(s) contributing to deliverable	NGC, ISCIII, HRI
Authors	Troels Rasmussen, NGC, DK
Contributors	Rob Hooft, Health-RI, NL
Acknowledgements	N/A
Reviewers	Regina Becker, PNED, LU, Tommi Nyrönen FI

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Log of changes

Date	Mvm	Who	Description
27/11/2023	0V1	Troels Rasmussen, NGC	First Draft
13/02/2024	0V2	Mercedes Rothschild Steiner	Copy circulated to the GDI-MB for review

Figure 5 :dataset téléchargé

Tableau des métadonnées(dublin core) :

Elément (dubline core)	Information
Titre	Proportion expressed across transcripts (pext)
Créateur	Genome Aggregation Database Production Team Genome Aggregation Database Consortium
Subject	Genome gene expression transcripts pext
Description	Ce dataset fournit des informations sur la proportion d'expression des variants génétiques à travers différents transcrits, permettant une meilleure interprétation fonctionnelle des variantes génomiques
Publisher	Zenodo
Contributor	Genome Aggregation Database Consortium
Date	27 février 2019
Type	dataset
Format	TSV,BGZ,ZIP
Identifiant	DOI:10.5281/ Zenodo .4447230
Source	Genome aggregation Database (gnomAD)
Rights	Open access
Langue	English

Les métadonnées ont été extraites à partir de la page du dataset sur la plateforme Zenodo selon la norme Dublin Core. Ces informations permettent de décrire la ressource numérique et de faciliter sa citation et sa réutilisation scientifique

Conclusion :

Cette étude pratique a permis d'explorer la plateforme Zenodo, d'effectuer une recherche basée sur le mot-clé genome, de sélectionner un dataset pertinent (pext) et d'extraire ses métadonnées selon le standard Dublin Core. Zenodo constitue un outil fondamental pour la diffusion et la valorisation des données scientifiques en accès libre

Partie 3 :BONUS :GitHub

Etape 1 :créer un compte GitHub :

Étapes suivies :

- Accès au site officiel de GitHub via le lien suivant : <https://github.com>
- Cliquer sur le bouton "Sign up".
- Renseigner les informations demandées : Nom d'utilisateur (Username) Adresse e-mail Mot de passe Valider la vérification de sécurité (captcha).
- Confirmer l'adresse e-mail via le message reçu.

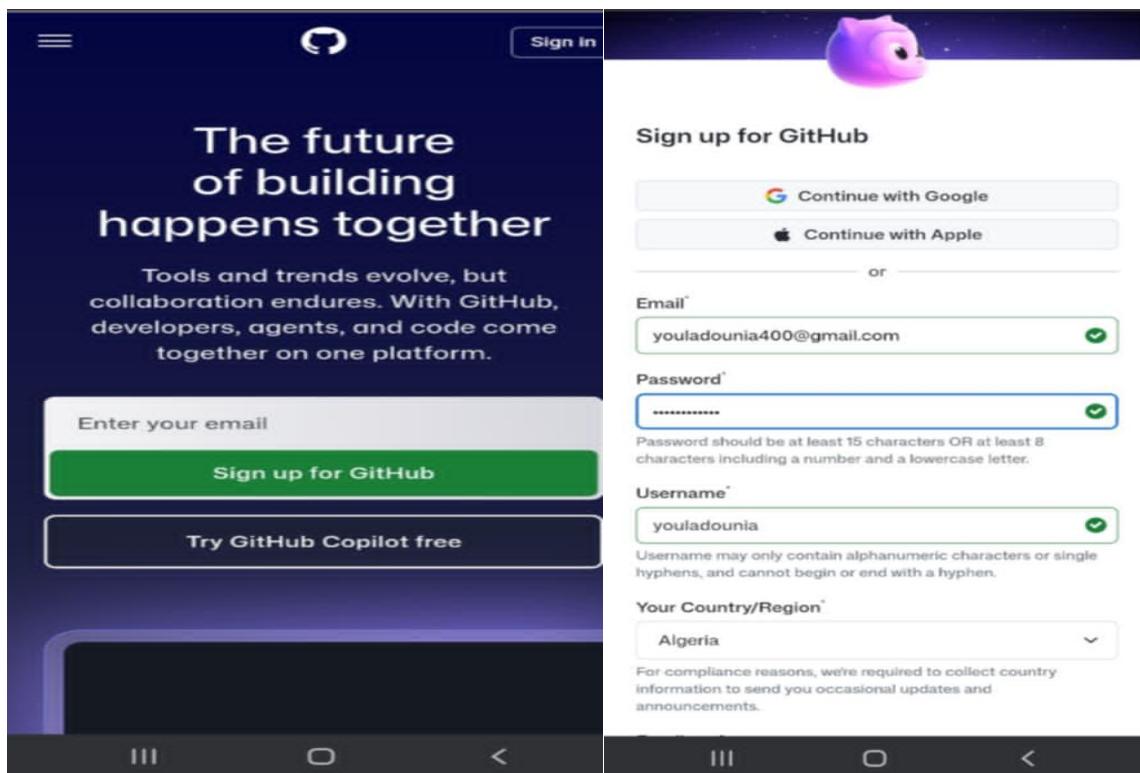
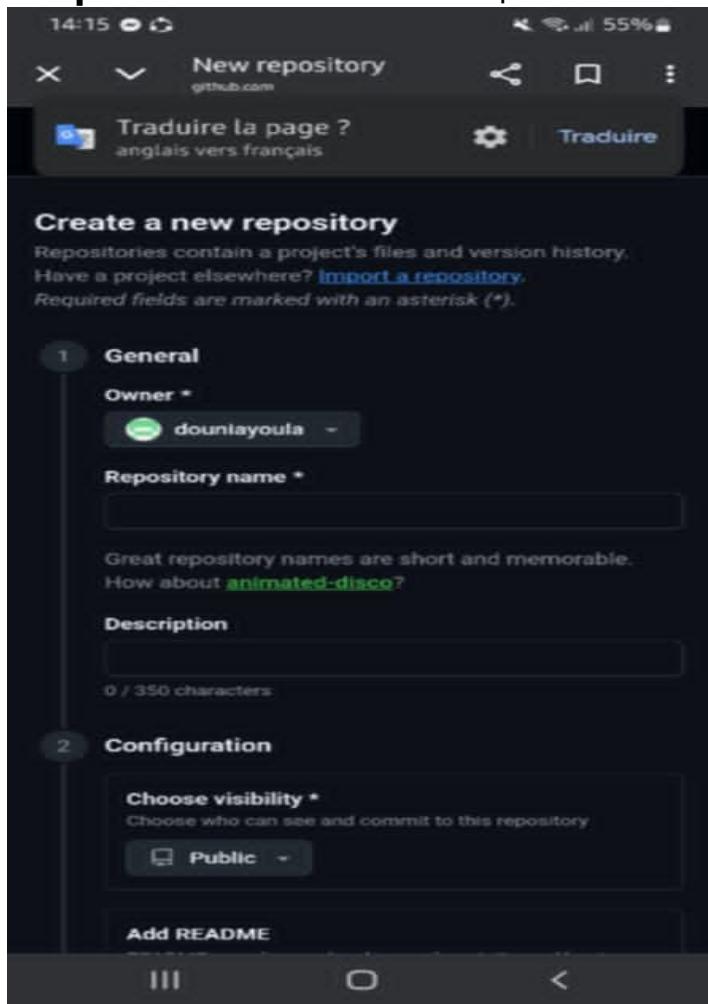


Figure 6 :création d'un compte sur la plateforme GitHub

Etape 2 : Créer un nouvel entrepôt



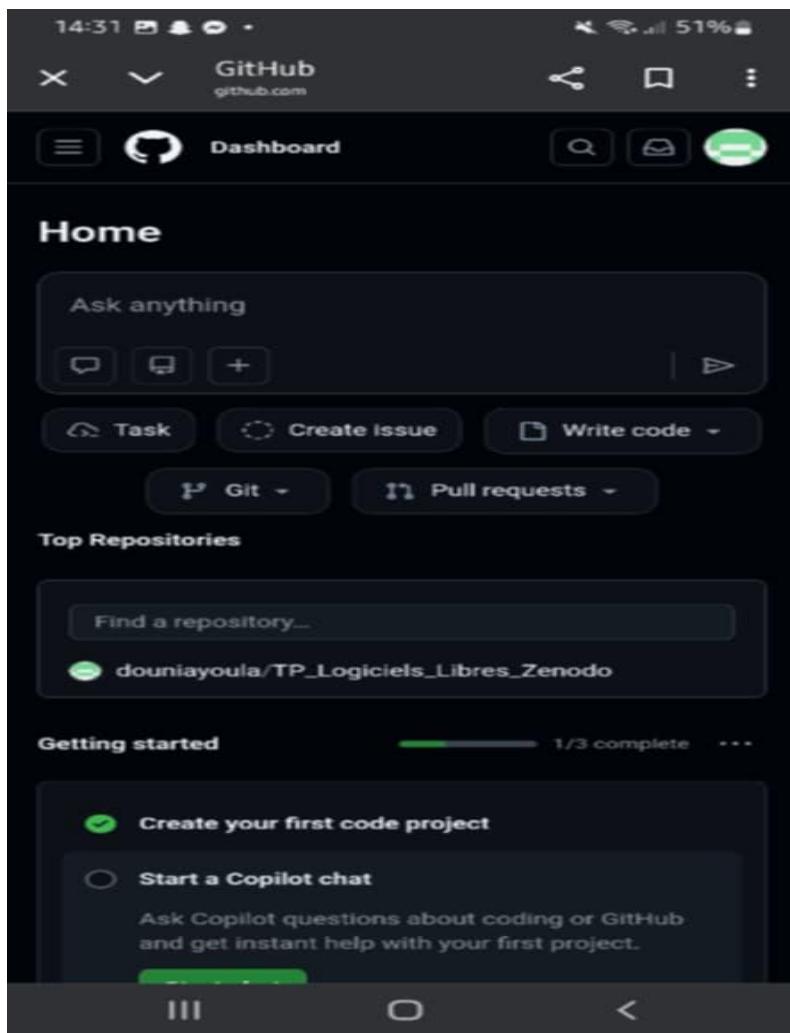


Figure 7 :création un dépôt (repository)