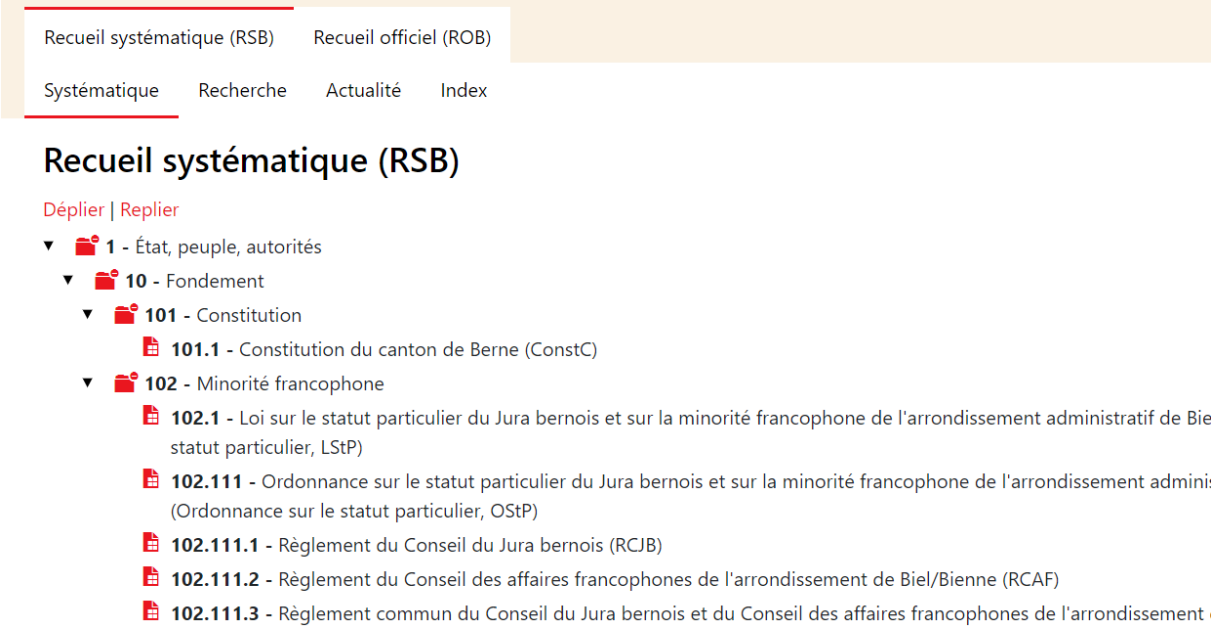
TP #7- Correction

Collecte de données (scraping) via Selenium

L'objectif de ce TP est de collecter l'intégralité des articles disponibles sur le site https://www.belex.sites.be.ch/app/fr/systematic/texts_of_law

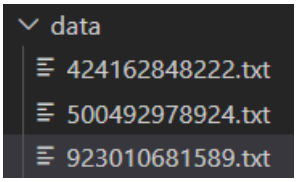


Nous souhaitons obtenir un JSON contenant les métadonnées suivantes :

- Nom
- Abréviation
- URL
- Numero_rs
- Provenance
- Fondement
- Date_entree_en_vigueur

Ces métadonnées sont trouvables sur le screenshot suivant :

Par ailleurs, nous souhaitons stocker le contenu de chaque article dans un fichier à l'intérieur d'un répertoire data/ :



923010681589.txt

```
data > 923010681589.txt
1   101.1
2   Constitution
3   du canton de Berne
4   (ConstC[1])
5   du 06.06.1993 (état au 15.05.2022)
6   Dans l'intention de protéger la liberté et le droit et d'aménager une collectivité dans laquelle tous
7   le peuple bernois se donne la Constitution suivante:
8   1 Principes généraux
9   Art. 1
10  Le canton de Berne
11  1
12  Le canton de Berne est un Etat de droit libéral, démocratique et social.
13  2
14  Le pouvoir de l'Etat appartient au peuple. Il est exercé par le corps électoral et les autorités.
15  Art. 2
16  Rapport avec la Confédération et les autres cantons
17  1
18  Le canton de Berne est l'un des Etats de la Confédération suisse.
19  2
20  Il coopère avec la Confédération et les autres cantons et se considère comme un lien entre la Suisse r
21  Art. 3
22  Territoire cantonal
23  1
24  Le canton de Berne comprend le territoire qui lui est garanti par la Confédération.
```

I. Création d'un nouveau projet

Ouvrez l'IDE de votre choix (VSCode ou PyCharm), et initiez un projet vierge (tp7).

Créez un fichier de type python (tp7.py).

Installez le ChromeDriver, permettant de piloter le navigateur Chrome à travers Selenium.

II. Initialisation du json et du répertoire data

Exercice 1

Instanciez une variable json_data qui contiendra les métadonnées, et créez le répertoire data/ qui contiendra les fichiers de chaque article.

```
import os

isExist = os.path.exists('data/')
if not isExist:
    os.makedirs('data/')

json_data = []
```

III. Utilisation de Chrome depuis Python

Exercice 2

En utilisant Selenium, ouvrez chrome à l'adresse :

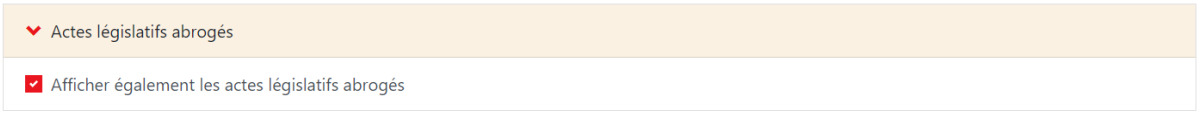
https://www.belex.sites.be.ch/app/fr/systematic/texts_of_law

```
url = 'https://www.belex.sites.be.ch/app/fr/systematic/texts_of_law'

chrome_options = Options()
# chrome_options.add_argument("--headless")
chrome_options.add_argument("--start-maximized");
driver = webdriver.Chrome(options=chrome_options)

driver.get(url)
time.sleep(3)
```

IV. Récupération de la liste de liens (articles à parcourir)

Exercice 3

Cochez la case « Afficher également les actes législatifs abrogés ».



Indice : Sélectionnez l'élément déroulant avec le XPATH et utilisez click() pour le déroulez. Ensuite, sélectionnez la checkbox et cliquez également dessus.

```
driver.find_element(By.XPATH, '//*[@id="page-content"]/ng-component/ng-
component/clex-texts-of-law-filters/div/div/div/span').click()
time.sleep(1)
driver.find_element(By.XPATH, '//*[@id="page-content"]/ng-component/ng-
component/clex-texts-of-law-filters/div/ul/a/div/label').click()
time.sleep(3)
```

Exercice 4

Récupérez la liste des liens.

Indice : Les liens sont compris dans l'attribut « href » des balises de classe « text-of-law-link ».

```
links = []
articles = driver.find_elements(By.CLASS_NAME, "text-of-law-link")

for article in articles:
    links.append(article.get_attribute('href'))
```

V. Récupération des métadonnées de chaque article

Exercice 5

Bouclez sur chaque article. Ouvrez chacun d'entre eux dans le navigateur piloté.

Attention à bien vérifier que la page est entièrement chargée avant de passer à la suite.

Indice : Vérifiez que la balise « h1 » est bien présente dans le code source. Vous pouvez tenter de récupérer l'élément dans un try / catch à l'intérieur d'une boucle while, et sortir de la boucle while uniquement lorsque le try s'exécute normalement (donc lorsque l'élément « h1 » est bien trouvable).

```
i = 0
for link in links:
    i = i + 1
```

```
    if i > 3:
        break;
    print()
    print("#", i)
    print(link)

    driver.get(link)
    time.sleep(1)

    ready = False
    while not ready :
        try:
            driver.find_element(By.TAG_NAME, "h1")
            ready = True
        except NoSuchElementException:
            print("NOT READY YET")
            time.sleep(1)
            ready = False
```

Exercice 6

Récupérez les métadonnées :

- Numéro
- Nom
- Abréviation
- Provenance
- Fondement

```
    if numero_rs is not None:
        numero_rs = numero_rs.text
        print("NUMERO :", numero_rs)

    nom = driver.find_element(By.CLASS_NAME, "title")
    if nom is not None:
        nom = nom.text.replace("\n", " ")
        print("NOM :", nom)

    abbreviation = driver.find_element(By.CLASS_NAME, "abbreviation")
    if abbreviation is not None:
        abbreviation = abbreviation.text
        print("ABREVIATION :", abbreviation)
    else:
        abbreviation = ""

    provenance = driver.find_element(By.CLASS_NAME, "ingress_author")
```

```
    if provenance is not None:
        provenance = provenance.text
        print("PROVENANCE :", provenance)
    else:
        provenance = ""

    fondement = driver.find_element(By.CLASS_NAME, "ingress_foundation")
    if fondement is not None:
        fondement = fondement.text
        print("FONDEMENT :", fondement)
    else:
        fondement = ""
```

Exercice 7

Récupérez la date en vigueur avec une regex.

```
    date_entree_en_vigueur = ""
    sous_titre = driver.find_element(By.XPATH, '//*[@id="page-content"]/ng-
component/div[1]/div/clex-meta-info/p')
    if sous_titre is not None:
        sous_titre = sous_titre.text
        print("SOUS TITRE :", sous_titre)
        regex = 'du (\d+[.]\d+[.]\d+)'

        date_entree_en_vigueur = re.search(regex, sous_titre)
        if date_entree_en_vigueur :
            date_entree_en_vigueur =  date_entree_en_vigueur[1]
            print("DATE ENTREE EN VIGUEUR : " + date_entree_en_vigueur)
        else :
            date_entree_en_vigueur = ""
    else:
        sous_titre = ""
```

Exercice 8

Pushez les métadonnées de l'article dans le json.

```
json_article = {
                    "id" : id,
                    "nom" : nom,
                    "abreviation" : abbreviation,
                    "url" : link,
                    "numero_rs" : numero_rs,
                    "provenance" : provenance,
                    "fondement" : fondement,
                    "date_entree_en_vigueur" : date_entree_en_vigueur,
                    "file_path" : filepath
                }
    json_data.append(json_article)
```

VI. Récupération du contenu de chaque article

Exercice 9

Récupérez le contenu de l'article dans la balise de classe « document ».

```
content = driver.find_element(By.CLASS_NAME, "document")
    if content is not None:
        content = content.text
        # print("CONTENT :", content)
```

Exercice 10

Ecrivez le contenu dans le répertoire /data

```
open(filepath, 'w', encoding='utf-8', newline="").write(content)
```

VII. Enregistrement du fichier JSON

Exercice 11

Enregistrez le fichier sur le disque.

```
open("data.json", "w", encoding='utf-8').write(json.dumps(json_data,
                                                          indent = 4,
                                                          sort_keys = False,
                                                          ensure_ascii=False))
```


```
[
    {
        "id": "923010681589",
        "nom": "Constitution du canton de Berne",
        "abreviation": "(ConstC[1])",
        "url": "https://www.belex.sites.be.ch/app/fr/texts_of_law/101.1",
        "numero_rs": "101.1",
        "provenance": "",
        "fondement": "Dans l'intention de protéger la liberté et le droit et d'aménager une collectivité dans
        "date_entree_en_vigueur": "",
        "file_path": "data/923010681589.txt"
    },
    {
        "id": "500492978924",
        "nom": "Ordonnance sur les mesures urgentes destinées à maîtriser la crise du coronavirus *",
        "abreviation": "(OCCV)",
        "url": "https://www.belex.sites.be.ch/app/fr/texts_of_law/101.2",
        "numero_rs": "101.2",
        "provenance": "Le Conseil-exécutif du canton de Berne,",
        "fondement": "vu l'article 91, alinéa 1 de la Constitution cantonale (ConstC)[1],\nsur proposition de
        "date_entree_en_vigueur": "21.03.2020",
        "file_path": "data/500492978924.txt"
```