ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO LAB

LAB 3

DỮ LIỆU XML

Giáo Viên HD:
Đỗ Văn Tiến

Thực hiện:

- Ngô Phúc Danh – 21521924

TP. Hồ Chí Minh, ngày 26 tháng 4 năm 2023

Lời Cảm Ơn

Bài lab “Dữ liệu xml” được dựa trên yêu cầu của môn học Tính Toán Đa Phương Tiện, thuộc lớp học CS232.N21, được thực hiện bởi sinh viên thuộc trường Đại học Công Nghệ Thông Tin - Đại học Quốc gia Thành phố Hồ Chí Minh. Bài lab “Dữ liệu xml” thực hành việc thu thập dữ liệu và lưu xuống một cơ sỡ dữ liệu quan hệ về các bài báo khoa học.

Qua bộ môn “Tính Toán Đa Phương Tiện” sinh viên có thể xây dựng chương trình có khả năng thu thập dữ liệu từ các trang web và lưu xuống nhằm phục vụ các nhu cầu về dữ liệu. Cụ thể với yêu cầu của bài lab 3: “Dữ liệu xml”, Sinh viên có thể thực hành được cách viết Chương trình đọc thông tin từ file xml của trang dblp sau đó lưu xuống một CSDL quan hệ về các bài báo khoa học.

I. Tổng hợp

1. Yêu cầu bài lab:

- Viết Chương trình đọc thông tin từ file xml của trang dblp sau đó lưu xuống một CSDL quan hệ về các bài báo khoa học.

2. Xác định yêu cầu:

- Dblp: DBLP (Digital Bibliography & Library Project) là một cơ sở dữ liệu trực tuyến miễn phí chứa các bài báo khoa học trong lĩnh vực khoa học máy tính và khoa học thông tin. DBLP được quản lý bởi một nhóm các nhà khoa học máy tính ở Đức và cung cấp thông tin về các bài báo, tạp chí, cuốn sách và các tài liệu liên quan khác được xuất bản trong lĩnh vực này. Các bài báo trong DBLP được tổ chức theo tên tác giả, tên tài liệu và các từ khóa liên quan. DBLP là một nguồn tài liệu quan trọng cho các nhà nghiên cứu, giảng viên và sinh viên trong lĩnh vực khoa học máy tính và khoa học thông tin.
- File xml: File có định dạng XML (eXtensible Markup Language) là một loại định dạng tệp tin được sử dụng để lưu trữ và truyền tải dữ liệu giữa các hệ thống khác nhau. XML được sử dụng rộng rãi trong các ứng dụng web, như là một cách để trao đổi dữ liệu giữa các ứng dụng khác nhau, hoặc trong các ứng dụng cơ sở dữ liệu, để lưu trữ và truy vấn dữ liệu. XML được thiết kế để dễ đọc và dễ hiểu cho con người và các máy tính. Nó sử dụng các thẻ để đánh dấu các phần khác nhau của dữ liệu, giúp cho dữ liệu được tổ chức và truy xuất một cách dễ dàng.
- CSDL quan hệ: CSDL quan hệ (Relational Database) là một loại cơ sở dữ liệu được tổ chức và lưu trữ dưới dạng bảng, trong đó mỗi bảng đại diện cho một thực thể và mỗi hàng trong bảng đại diện cho một bản ghi của thực thể đó. Mỗi cột trong bảng tương ứng với một thuộc tính của thực thể. Các bảng được liên kết với nhau thông qua các quan hệ (relationship), trong đó một quan hệ được xác định bởi một cặp khóa (key) giữa hai bảng.

➔ Sinh viên cần xây dựng crawler nhằm thu thập dữ liệu trong file xml thuộc trang web dblp về danh sách các bài báo khoa học sau đó lưu xuống database.

3. Công nghệ sử dụng:

- Ngôn ngữ lập trình(programming language): Python.
- Tool: selenium.
- Libraries: numpy, pypyodbc, xml.etree.ElementTree.

II. Nội dung

1. Thử thách (challenge):

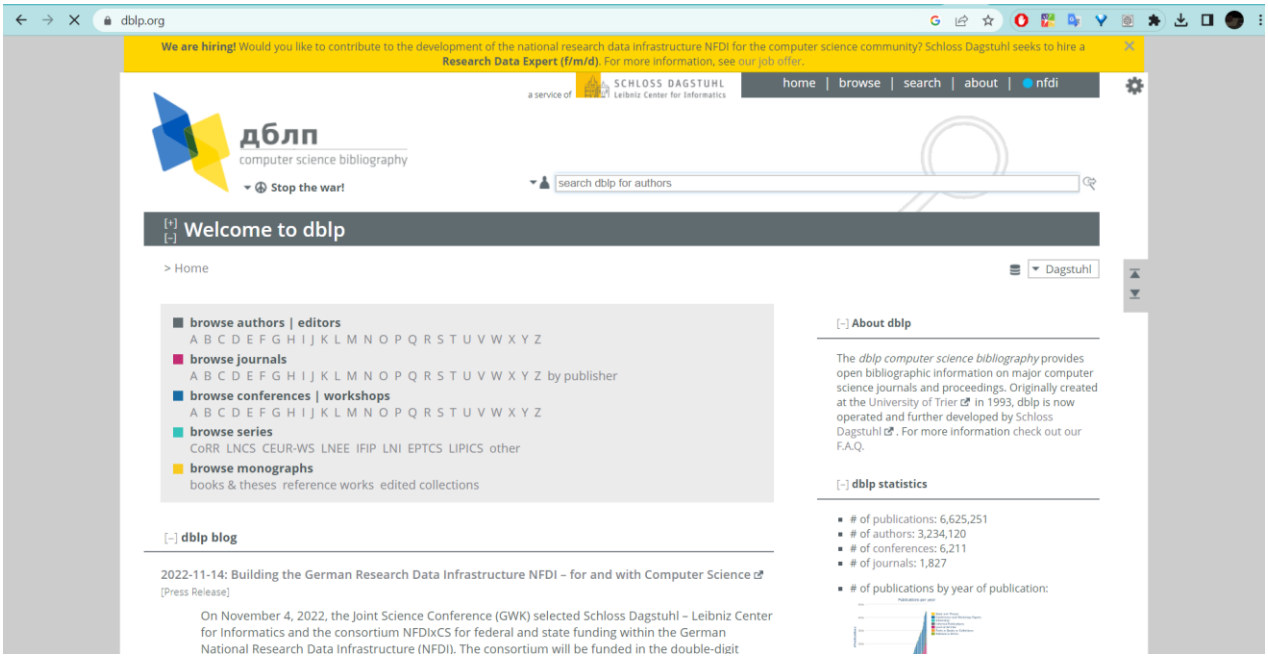
- Crawl dữ liệu trên dblp: Để crawl được dữ liệu về các bài báo trên dblp chúng ta cần phải truy cập vào trang web, chọn khung author và nhập vào tên tác giả sau đó nhấn vào tác giả vừa tìm được chúng ta sẽ được chuyển sang trang mới chứa danh sách các bài báo có sự tham gia của tên tác giả mà chúng ta vừa nhập, cuối cùng là crawl về máy tính cá nhân file xml.
- SQL server: Sau khi chúng ta giải quyết vấn đề crawl data thì chúng ta cần lưu xuống một cơ sở dữ liệu quan hệ.
- Vấn đề về chặn ip: trang web không có thao tác chặn bot như captcha

2. Hướng tiếp cận (approach):

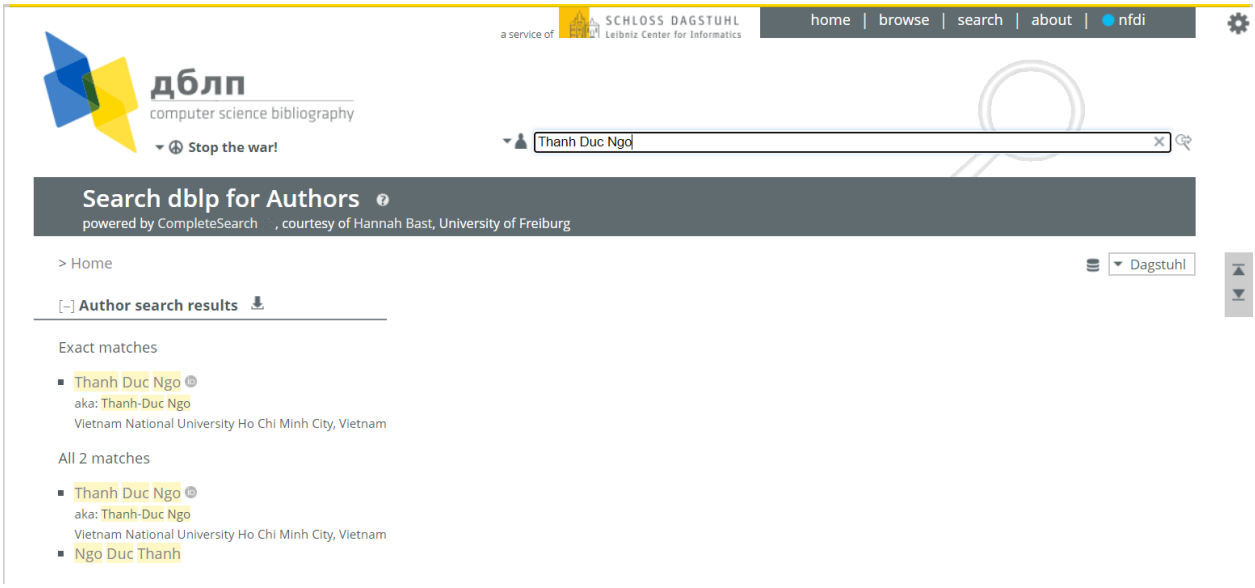
- Crawl dữ liệu: ta thấy rằng để đến được danh sách các bài báo cần thông qua nhiều quá trình tương tác, chính vì thế việc sử dụng selenium để tự động hóa quá trình crawl dữ liệu xml về máy tính cá nhân.
- SQL server: Mở SQL server để connect thông qua server name và tạo database. Để tiến hành phân chia các thẻ trong file xml thành các cột trong một bảng chúng ta dùng thư viện xml.etree.ElementTree.
- Sử dụng openvpn để đổi ip kết hợp với tùy chỉnh thời gian truy cập random.

3. Minh họa bằng hình ảnh (images illustration):

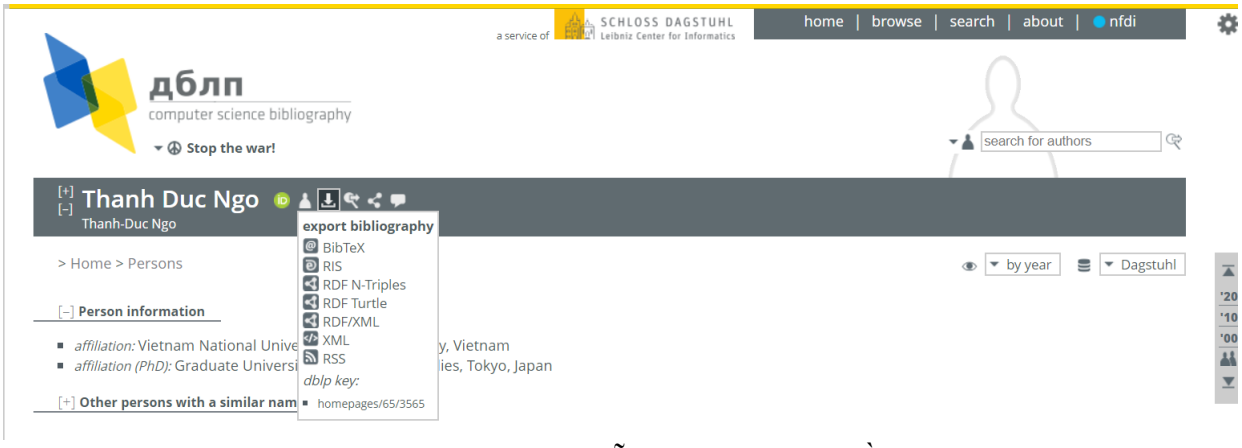
- Tự động hóa crawl dữ liệu bằng selenium:



Bước 1: Truy cập trang web.

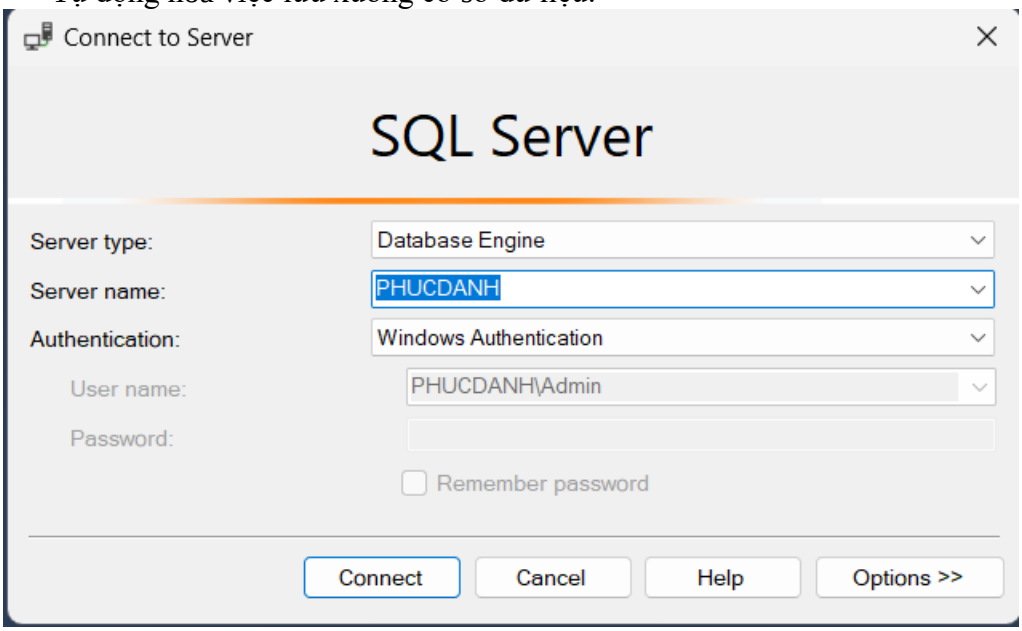


Bước 2: Nhập tên tác giả.

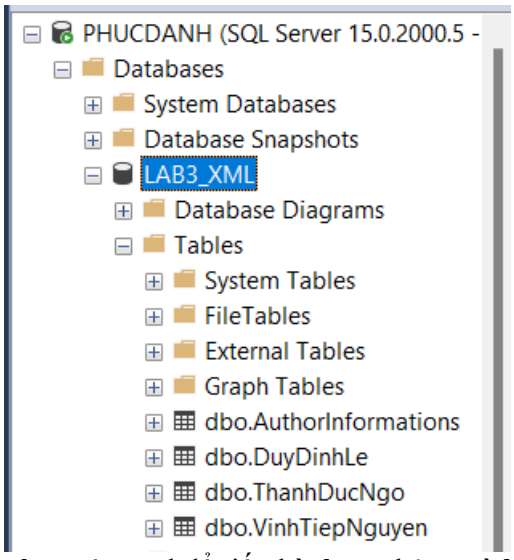


Bước 3: Truy cập vào đường dẫn XML và crawl về máy tính.

- Tự động hóa việc lưu xuống cơ sở dữ liệu:



Bước 1: Mở ứng dụng SQL server và tạo database



Bước 2: Sử dụng thư viện xml để tiến hành tạo bảng và lưu vào các bảng.

4. Source code:

- Ý tưởng:
  - Crawler: sử dụng selenium để tự động hóa việc lấy dữ liệu.
    - Bước 1: Truy cập vào đường link: https://dblp.org/.
    - Bước 2: Chọn chế độ tìm kiếm bằng tên tác giả.
    - Bước 3: Thực hiện việc nhập tên tác giả vào ô tìm kiếm.
    - Bước 4: Truy cập vào mục xml và tiến hành crawl file.xml.
  - Sql: sử dụng pypyodbc kết hợp với xml.etree.ElementTree.
    - Bước 1: Tạo bảng (table) cho từng file xml.
    - Bước 2: Tạo từng cột với tên ứng với từng thẻ tag.
    - Bước 3: Lưu dữ liệu xuống cơ sở dữ liệu.
  - Run code:
    - Bước 1: Import 2 module crawler và sql đã thực hiện ở trên.
    - Bước 2: Nhập tên Server và database.
    - Bước 3: Nhập tên các tác giả umốn crawl bài báo.
    - Bước 5: Tiến hành chạy chương trình.

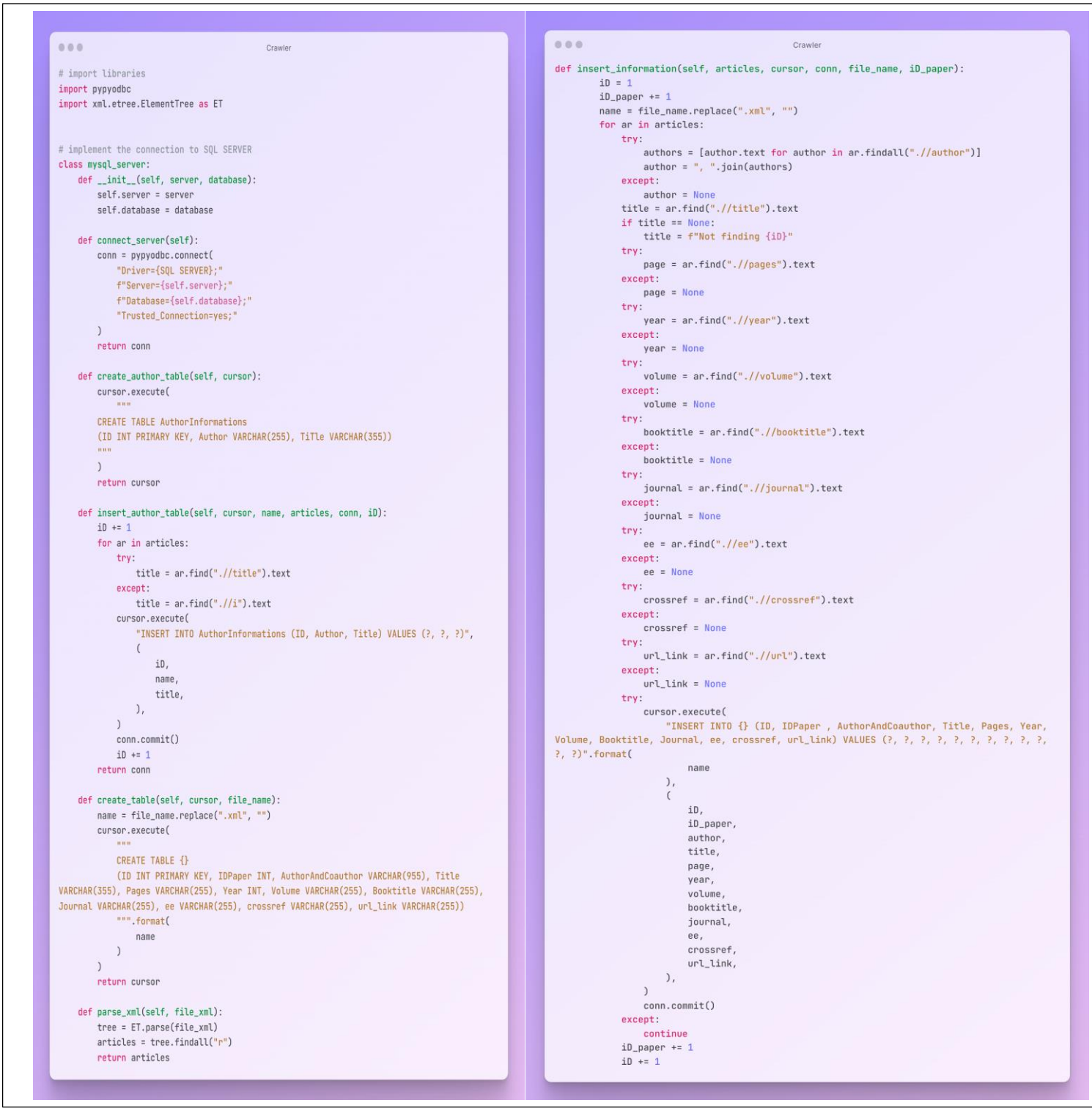
- Code:
  - Crawler:

Crawler

```python
# import libraries
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from time import sleep
import numpy as np
import requests


# implement crawler for crawling xml file
class crawling:
    def __init__(self, browser):
        self.browser = browser

    def navigate_to_dblp(self):
        sleep(np.random.randint(2, 6))
        dblp_url = "https://dblp.org/"
        self.browser.get(dblp_url)
        sleep(np.random.randint(2, 6))
        return self.browser

    def author_searching_function(self, authorname):
        searching_options_selector = "#search-mode-selector > div.head > img"
        searching_options = self.browser.find_element(
            By.CSS_SELECTOR,
            searching_options_selector,
        )
        searching_options.click()
        searching_author_selector = "#search-mode-author"
        searching_authors = self.browser.find_element(
            By.CSS_SELECTOR,
            searching_author_selector,
        )
        searching_authors.click()
        sleep(np.random.randint(2, 6))
        author_name_xpath = "/html/body/div[2]/div[2]/form/input"
        author_name_searching = self.browser.find_element(
            By.XPATH,
            author_name_xpath,
        )
        author_name_searching.send_keys(authorname)
        author_name_searching.send_keys(Keys.ENTER)
        sleep(np.random.randint(2, 6))
        author_chosen_selector = (
            "#completesearch-authors > div > ul:nth-child(2) > li > a"
        )
        author_chosen = self.browser.find_element(
            By.CSS_SELECTOR,
            author_chosen_selector,
        )
        author_chosen.click()
        return self.browser

    def access_author_link(self):
        sleep(np.random.randint(2, 6))
        xml_selector = "#headline > nav > ul > li.export.drop-down > div.body > ul:nth-
child(2) > li:nth-child(6) > a"

        xml_option = self.browser.find_element(By.CSS_SELECTOR, xml_selector)
        xml_link = xml_option.get_attribute("href")
        return xml_link

    def download_xml(self, author_name, xml_link):
        author_name = author_name.replace(" ", "")
        filename = f"{author_name}.xml"
        response = requests.get(xml_link)
        if response.status_code == 200:
            with open(filename, "wb") as f:
                f.write(response.content)
            print(f"Downloaded {filename} successfully.")
        else:
            print(f"Failed to download {filename}. Status code: {response.status_code}")
        sleep(5)
        self.browser.close()
        return filename
```

Hình 2.4.1: Source code crawler.

➢ Sql:

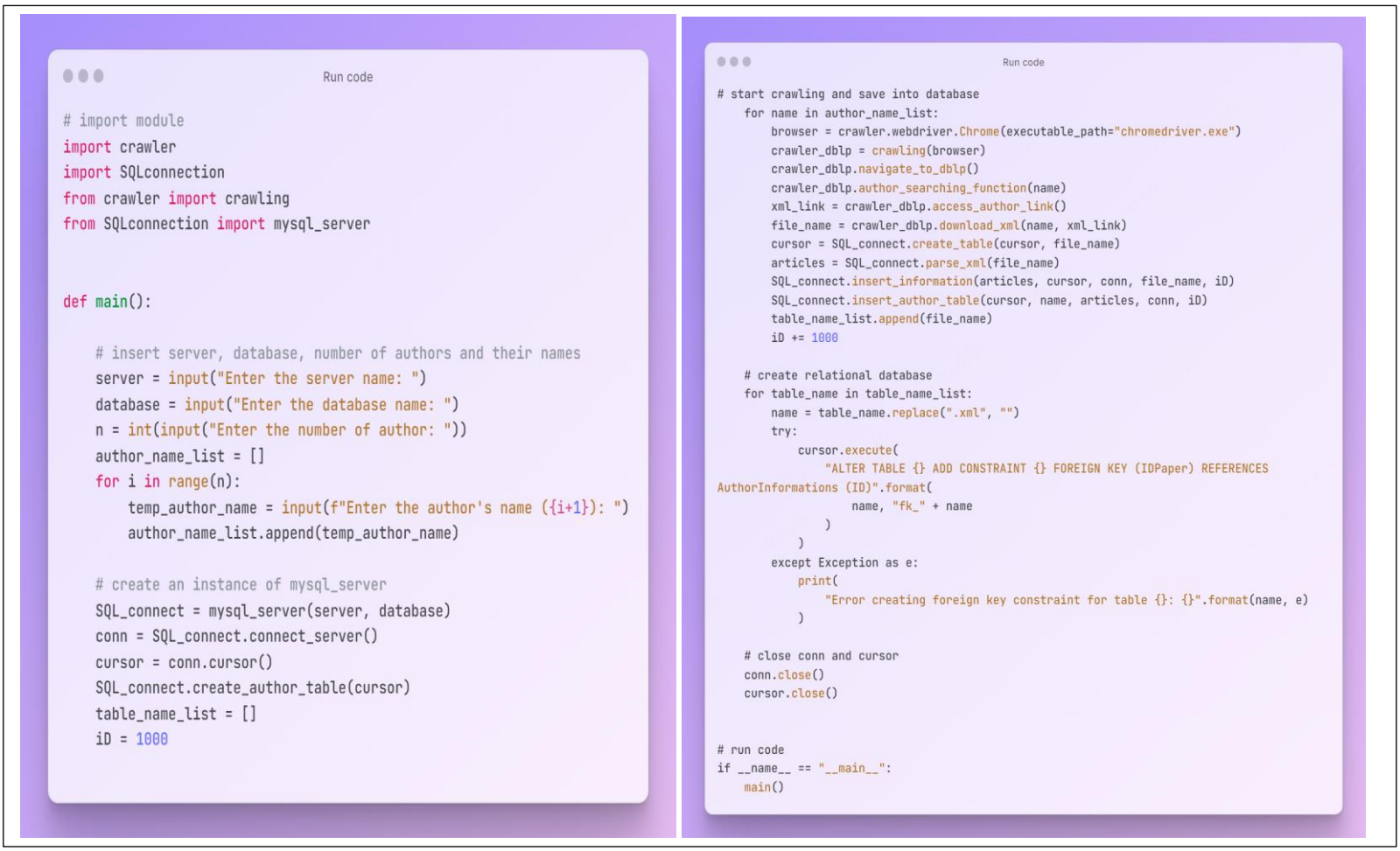


```
# import libraries
import pypyodbc
import xml.etree.ElementTree as ET


# implement the connection to SQL SERVER
class mysql_server:
    def __init__(self, server, database):
        self.server = server
        self.database = database

    def connect_server(self):
        conn = pypyodbc.connect(
            "Driver={SQL SERVER};"
            f"Server={self.server};"
            f"Database={self.database};"
            "Trusted_Connection=yes;"
        )
        return conn

    def create_author_table(self, cursor):
        cursor.execute(
            """
        CREATE TABLE AuthorInformations
        (ID INT PRIMARY KEY, Author VARCHAR(255), TiTle VARCHAR(355))
        """
        )
        return cursor

    def insert_author_table(self, cursor, name, articles, conn, iD):
        iD += 1
        for ar in articles:
            try:
                title = ar.find(".//title").text
            except:
                title = ar.find(".//i").text
            cursor.execute(
                "INSERT INTO AuthorInformations (ID, Author, Title) VALUES (?, ?, ?)",
                (
                    iD,
                    name,
                    title,
                ),
            )
            conn.commit()
            iD += 1
        return conn

    def create_table(self, cursor, file_name):
        name = file_name.replace(".xml", "")
        cursor.execute(
            """
            CREATE TABLE {}
            (ID INT PRIMARY KEY, IDPaper INT, AuthorAndCoauthor VARCHAR(955), Title
VARCHAR(355), Pages VARCHAR(255), Year INT, Volume VARCHAR(255), Booktitle VARCHAR(255),
Journal VARCHAR(255), ee VARCHAR(255), crossref VARCHAR(255), url_link VARCHAR(255))
            """.format(
                name
            )
        )
        return cursor

    def parse_xml(self, file_xml):
        tree = ET.parse(file_xml)
        articles = tree.findall("r")
        return articles
```

```
def insert_information(self, articles, cursor, conn, file_name, iD_paper):
        iD = 1
        iD_paper += 1
        name = file_name.replace(".xml", "")
        for ar in articles:
            try:
                authors = [author.text for author in ar.findall(".//author")]
                author = ", ".join(authors)
            except:
                author = None
            title = ar.find(".//title").text
            if title == None:
                title = f"Not finding {iD}"
            try:
                page = ar.find(".//pages").text
            except:
                page = None
            try:
                year = ar.find(".//year").text
            except:
                year = None
            try:
                volume = ar.find(".//volume").text
            except:
                volume = None
            try:
                booktitle = ar.find(".//booktitle").text
            except:
                booktitle = None
            try:
                journal = ar.find(".//journal").text
            except:
                journal = None
            try:
                ee = ar.find(".//ee").text
            except:
                ee = None
            try:
                crossref = ar.find(".//crossref").text
            except:
                crossref = None
            try:
                url_link = ar.find(".//url").text
            except:
                url_link = None
            try:
                cursor.execute(
                    "INSERT INTO {} (ID, IDPaper , AuthorAndCoauthor, Title, Pages, Year,
Volume, Booktitle, Journal, ee, crossref, url_link) VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?,
?, ?)".format(
                        name
                    ),
                    (
                        iD,
                        iD_paper,
                        author,
                        title,
                        page,
                        year,
                        volume,
                        booktitle,
                        journal,
                        ee,
                        crossref,
                        url_link,
                    ),
                )
                conn.commit()
            except:
                continue
            iD_paper += 1
            iD += 1
```

Hình 2.4.2: Source code Sql connecting.

➢ Run code:



Run code

```
# import module
import crawler
import SQLconnection
from crawler import crawling
from SQLconnection import mysql_server


def main():

    # insert server, database, number of authors and their names
    server = input("Enter the server name: ")
    database = input("Enter the database name: ")
    n = int(input("Enter the number of author: "))
    author_name_list = []
    for i in range(n):
        temp_author_name = input(f"Enter the author's name ({i+1}): ")
        author_name_list.append(temp_author_name)

    # create an instance of mysql_server
    SQL_connect = mysql_server(server, database)
    conn = SQL_connect.connect_server()
    cursor = conn.cursor()
    SQL_connect.create_author_table(cursor)
    table_name_list = []
    iD = 1000
```

Run code

```
# start crawling and save into database
    for name in author_name_list:
        browser = crawler.webdriver.Chrome(executable_path="chromedriver.exe")
        crawler_dblp = crawling(browser)
        crawler_dblp.navigate_to_dblp()
        crawler_dblp.author_searching_function(name)
        xml_link = crawler_dblp.access_author_link()
        file_name = crawler_dblp.download_xml(name, xml_link)
        cursor = SQL_connect.create_table(cursor, file_name)
        articles = SQL_connect.parse_xml(file_name)
        SQL_connect.insert_information(articles, cursor, conn, file_name, iD)
        SQL_connect.insert_author_table(cursor, name, articles, conn, iD)
        table_name_list.append(file_name)
        iD += 1000

    # create relational database
    for table_name in table_name_list:
        name = table_name.replace(".xml", "")
        try:
            cursor.execute(
                "ALTER TABLE {} ADD CONSTRAINT {} FOREIGN KEY (IDPaper) REFERENCES
AuthorInformations (ID)".format(
                    name, "fk_" + name
                )
            )
        except Exception as e:
            print(
                "Error creating foreign key constraint for table {}: {}".format(name, e)
            )

    # close conn and cursor
    conn.close()
    cursor.close()


# run code
if __name__ == "__main__":
    main()
```

Hình 2.4.3: Source code main.py.

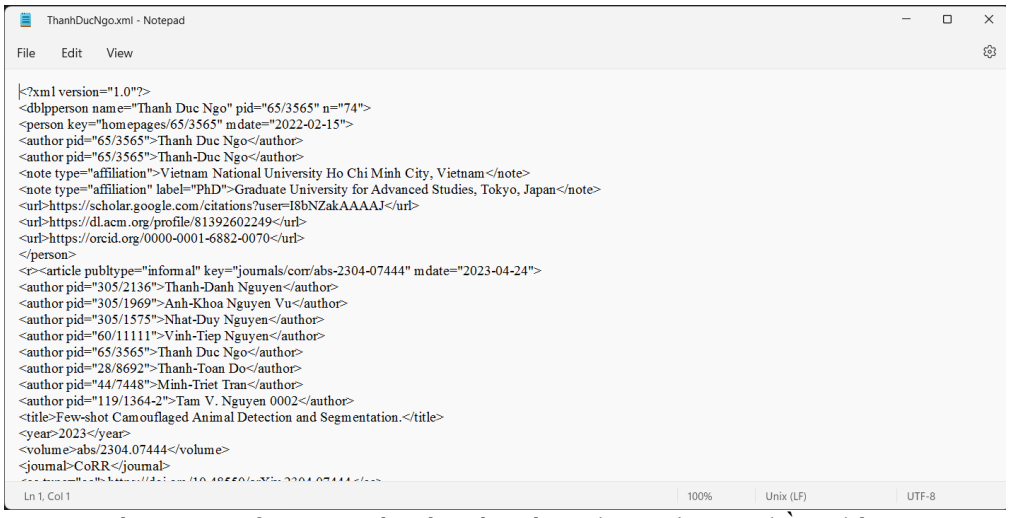
III. Cài đặt và kiểm thử kết quả

1. Cài đặt:

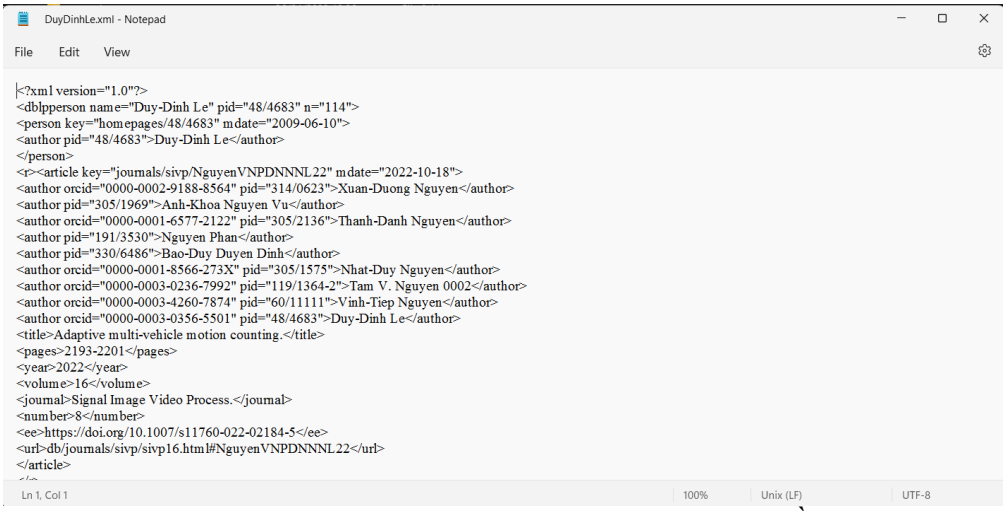
- Folder chứa source code gồm 3 file .py bao gồm (crawler.py, SQLconnection, main.py).
- Người dùng thực hiện việc download chromedriver trước khi thực hiện run code.
- Người dùng mở ứng dụng sql server management chọn server sau đó kết nối và tạo databse. Ở bài báo cáo này, sinh viên thực hiện sử dụng server: **PHUCDANH** và database: **LAB3_XML.**
- Sau khi chuẩn bị xong, người dùng chạy file main.py.
- Nhập vào tên server và database để lưu trữ.
- Nhập vào số lượng tác giả và tên của từng tác giả, ở đây sinh viên thực hiện kiểm thử với 3 tác giả.
- Chờ và ghi lại kết quả.

2. Kiểm thử kết quả:

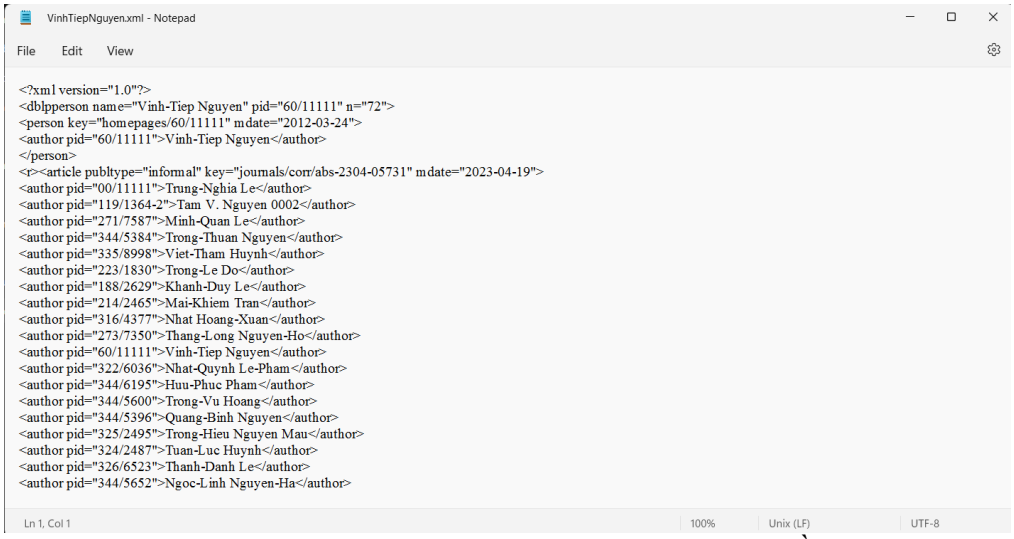
- Kết quả sau khi crawl tất cả bài báo của 3 tác giả là: thầy Ngô Đức Thành, thầy Lê Đình Duy và thầy Nguyễn Vinh Tiệp.



Hình 3.2.1: file xml các bài báo có sự tham gia của thầy Thành.

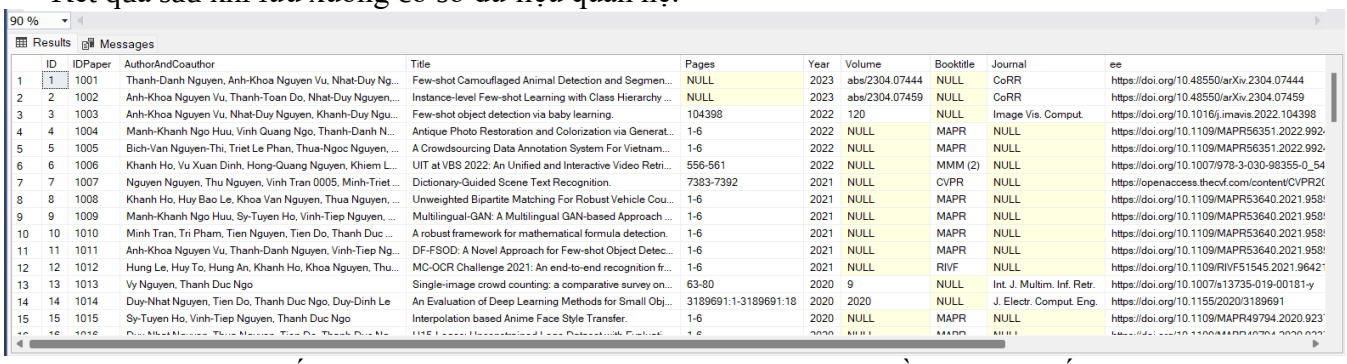


Hình 3.2.2: file xml các bài báo có sự tham gia của thầy Duy.



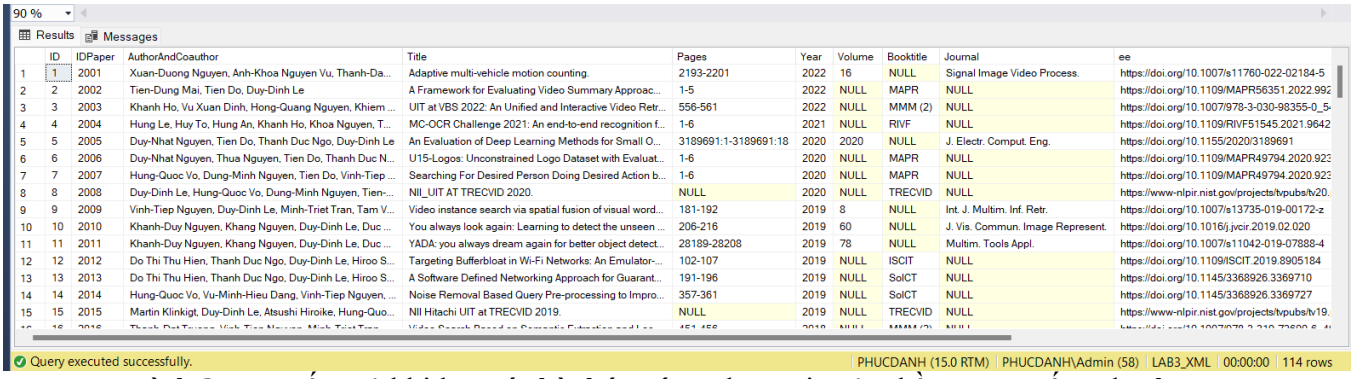
Hình 3.2.3: file xml các bài báo có sự tham gia của thầy Tiệp.

- Kết quả sau khi lưu xuống cơ sở dữ liệu quan hệ:



	ID	IDPaper	AuthorAndCoauthor	Title	Pages	Year	Volume	Booktitle	Journal	ee
1	1	1001	Thanh-Danh Nguyen, Anh-Khoa Nguyen Vu, Nhat-Duy Ng...	Few-shot Camouflaged Animal Detection and Segmen...	NULL	2023	abs/2304.07444	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.07444
2	2	1002	Anh-Khoa Nguyen Vu, Thanh-Toan Do, Nhat-Duy Nguyen,...	Instance-level Few-shot Learning with Class Hierarchy ...	NULL	2023	abs/2304.07459	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.07459
3	3	1003	Anh-Khoa Nguyen Vu, Nhat-Duy Nguyen, Khanh-Duy Ngu...	Few-shot object detection via baby learning.	104398	2022	120	NULL	Image Vis. Comput.	https://doi.org/10.1016/j.imavis.2022.104398
4	4	1004	Manh-Khanh Ngo Huu, Vinh Quang Ngo, Thanh-Danh N...	Antique Photo Restoration and Colorization via Generat...	1-6	2022	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR56351.2022.992
5	5	1005	Bich-Van Nguyen-Thi, Triet Le Phan, Thua-Ngoc Nguyen, ...	A Crowdsourcing Data Annotation System For Vietnam...	1-6	2022	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR56351.2022.992
6	6	1006	Khanh Ho, Vu Xuan Dinh, Hong-Quang Nguyen, Khiem L...	UIT at VBS 2022: An Unified and Interactive Video Retri...	556-561	2022	NULL	MMM (2)	NULL	https://doi.org/10.1007/978-3-030-98355-0_54
7	7	1007	Nguyen Nguyen, Thu Nguyen, Vinh Tran 0005, Minh-Triet ...	Dictionary-Guided Scene Text Recognition.	7383-7392	2021	NULL	CVPR	NULL	https://openaccess.thecvf.com/content/CVPR2(
8	8	1008	Khanh Ho, Huy Bao Le, Khoa Van Nguyen, Thua Nguyen, ...	Unweighted Bipartite Matching For Robust Vehicle Cou...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.958
9	9	1009	Manh-Khanh Ngo Huu, Sy-Tuyen Ho, Vinh-Tiep Nguyen, ...	Multilingual-GAN: A Multilingual GAN-based Approach ...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.958
10	10	1010	Minh Tran, Tri Pham, Tien Nguyen, Tien Do, Thanh Duc ...	A robust framework for mathematical formula detection.	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.958
11	11	1011	Anh-Khoa Nguyen Vu, Thanh-Danh Nguyen, Vinh-Tiep Ng...	DF-FSOD: A Novel Approach for Few-shot Object Detec...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.958
12	12	1012	Hung Le, Huy To, Hung An, Khanh Ho, Khoa Nguyen, Thu...	MC-OCR Challenge 2021: An end-to-end recognition fr...	1-6	2021	NULL	RIVF	NULL	https://doi.org/10.1109/RIVF51545.2021.9642
13	13	1013	Vy Nguyen, Thanh Duc Ngo	Single-image crowd counting: a comparative survey on...	63-80	2020	9	NULL	Int. J. Multim. Inf. Retr.	https://doi.org/10.1007/s13735-019-00181-y
14	14	1014	Duy-Nhat Nguyen, Tien Do, Thanh Duc Ngo, Duy-Dinh Le	An Evaluation of Deep Learning Methods for Small Obj...	3189691:1-3189691:18	2020	2020	NULL	J. Electr. Comput. Eng.	https://doi.org/10.1155/2020/3189691
15	15	1015	Sy-Tuyen Ho, Vinh-Tiep Nguyen, Thanh Duc Ngo	Interpolation based Anime Face Style Transfer.	1-6	2020	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR49794.2020.923

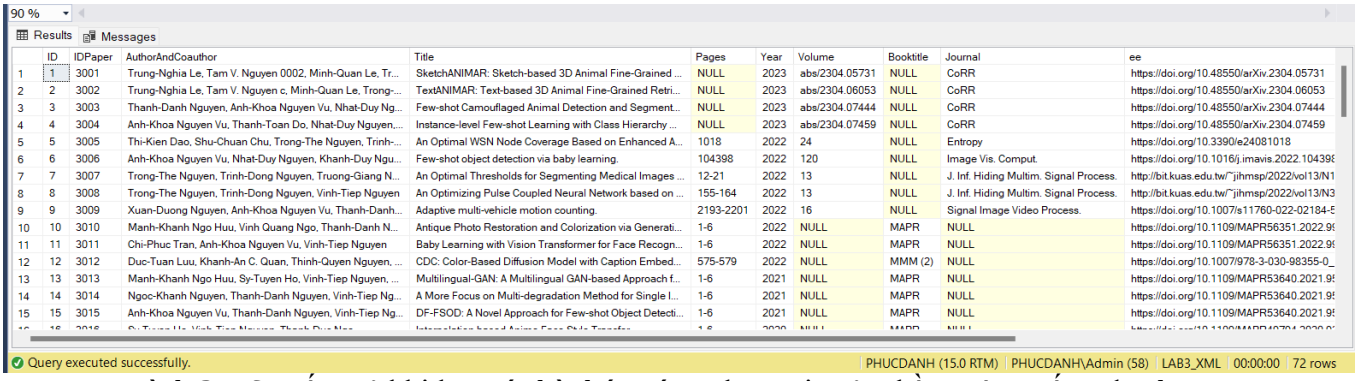
Hình 3.2.4: Kết quả khi lưu các bài báo có sự tham gia của thầy Thành xuống database.



	ID	IDPaper	AuthorAndCoauthor	Title	Pages	Year	Volume	Booktitle	Journal	ee
1	1	2001	Xuan-Duong Nguyen, Anh-Khoa Nguyen Vu, Thanh-Da...	Adaptive multi-vehicle motion counting.	2193-2201	2022	16	NULL	Signal Image Video Process.	https://doi.org/10.1007/s11760-022-02184-5
2	2	2002	Tien-Dung Mai, Tien Do, Duy-Dinh Le	A Framework for Evaluating Video Summary Approac...	1-5	2022	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR56351.2022.992
3	3	2003	Khanh Ho, Vu Xuan Dinh, Hong-Quang Nguyen, Khiem ...	UIT at VBS 2022: An Unified and Interactive Video Retr...	556-561	2022	NULL	MMM (2)	NULL	https://doi.org/10.1007/978-3-030-98355-0_5
4	4	2004	Hung Le, Huy To, Hung An, Khanh Ho, Khoa Nguyen, T...	MC-OCR Challenge 2021: An end-to-end recognition f...	1-6	2021	NULL	RIVF	NULL	https://doi.org/10.1109/RIVF51545.2021.9642
5	5	2005	Duy-Nhat Nguyen, Tien Do, Thanh Duc Ngo, Duy-Dinh Le	An Evaluation of Deep Learning Methods for Small O...	3189691:1-3189691:18	2020	2020	NULL	J. Electr. Comput. Eng.	https://doi.org/10.1155/2020/3189691
6	6	2006	Duy-Nhat Nguyen, Thua Nguyen, Tien Do, Thanh Duc N...	U15-Logos: Unconstrained Logo Dataset with Evaluat...	1-6	2020	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR49794.2020.923
7	7	2007	Hung-Quoc Vo, Dung-Minh Nguyen, Tien Do, Vinh-Tiep ...	Searching For Desired Person Doing Desired Action b...	1-6	2020	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR49794.2020.923
8	8	2008	Duy-Dinh Le, Hung-Quoc Vo, Dung-Minh Nguyen, Tien-...	NII_UIT AT TRECVID 2020.	NULL	2020	NULL	TRECVID	NULL	https://www-nlpir.nist.gov/projects/tvpubs/tv20.
9	9	2009	Vinh-Tiep Nguyen, Duy-Dinh Le, Minh-Triet Tran, Tam V...	Video instance search via spatial fusion of visual word...	181-192	2019	8	NULL	Int. J. Multim. Inf. Retr.	https://doi.org/10.1007/s13735-019-00172-z
10	10	2010	Khanh-Duy Nguyen, Khang Nguyen, Duy-Dinh Le, Duc ...	You always look again: Learning to detect the unseen ...	206-216	2019	60	NULL	J. Vis. Commun. Image Represent.	https://doi.org/10.1016/j.jvcir.2019.02.020
11	11	2011	Khanh-Duy Nguyen, Khang Nguyen, Duy-Dinh Le, Duc ...	YADA: you always dream again for better object detect...	28189-28208	2019	78	NULL	Multim. Tools Appl.	https://doi.org/10.1007/s11042-019-07888-4
12	12	2012	Do Thi Thu Hien, Thanh Duc Ngo, Duy-Dinh Le, Hiroo S...	Targeting Bufferbloat in Wi-Fi Networks: An Emulator-...	102-107	2019	NULL	ISCIT	NULL	https://doi.org/10.1109/ISCIT.2019.8905184
13	13	2013	Do Thi Thu Hien, Thanh Duc Ngo, Duy-Dinh Le, Hiroo S...	A Software Defined Networking Approach for Guarant...	191-196	2019	NULL	SoICT	NULL	https://doi.org/10.1145/3368926.3369710
14	14	2014	Hung-Quoc Vo, Vu-Minh-Hieu Dang, Vinh-Tiep Nguyen, ...	Noise Removal Based Query Pre-processing to Impro...	357-361	2019	NULL	SoICT	NULL	https://doi.org/10.1145/3368926.3369727
15	15	2015	Martin Klinkigt, Duy-Dinh Le, Atsushi Hiroike, Hung-Quo...	NII Hitachi UIT at TRECVID 2019.	NULL	2019	NULL	TRECVID	NULL	https://www-nlpir.nist.gov/projects/tvpubs/tv19.

Query executed successfully. | PHUCDANH (15.0 RTM) | PHUCDANH\Admin (58) | LAB3_XML | 00:00:00 | 114 rows

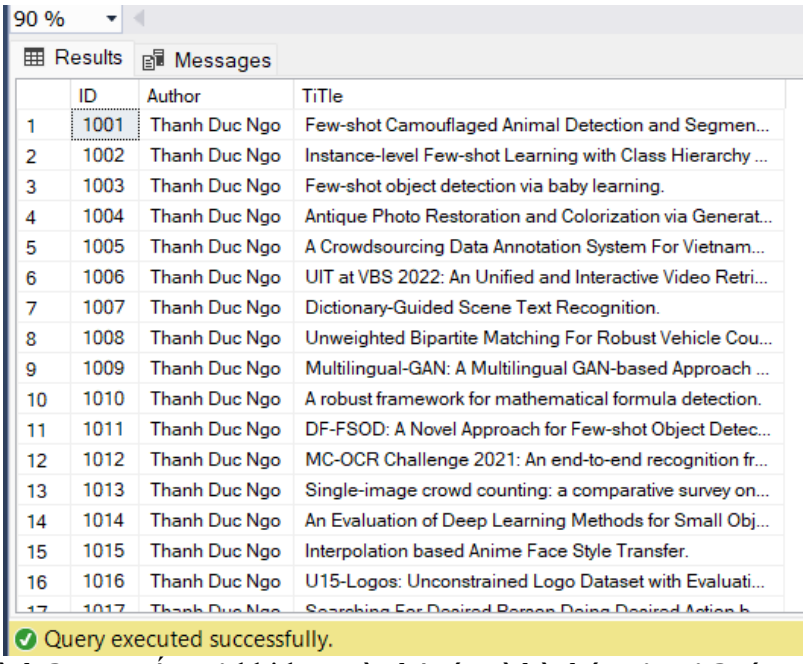
Hình 3.2.5: Kết quả khi lưu các bài báo có sự tham gia của thầy Duy xuống database.



	ID	IDPaper	AuthorAndCoauthor	Title	Pages	Year	Volume	Booktitle	Journal	ee
1	1	3001	Trung-Nghia Le, Tam V. Nguyen 0002, Minh-Quan Le, Tr...	SketchANIMAR: Sketch-based 3D Animal Fine-Grained ...	NULL	2023	abs/2304.05731	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.05731
2	2	3002	Trung-Nghia Le, Tam V. Nguyen c, Minh-Quan Le, Trong-...	TextANIMAR: Text-based 3D Animal Fine-Grained Retri...	NULL	2023	abs/2304.06053	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.06053
3	3	3003	Thanh-Danh Nguyen, Anh-Khoa Nguyen Vu, Nhat-Duy Ng...	Few-shot Camouflaged Animal Detection and Segment...	NULL	2023	abs/2304.07444	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.07444
4	4	3004	Anh-Khoa Nguyen Vu, Thanh-Toan Do, Nhat-Duy Nguyen,...	Instance-level Few-shot Learning with Class Hierarchy ...	NULL	2023	abs/2304.07459	NULL	CoRR	https://doi.org/10.48550/arXiv.2304.07459
5	5	3005	Thi-Kien Dao, Shu-Chuan Chu, Trong-The Nguyen, Trinh-...	An Optimal WSN Node Coverage Based on Enhanced A...	1018	2022	24	NULL	Entropy	https://doi.org/10.3390/e24081018
6	6	3006	Anh-Khoa Nguyen Vu, Nhat-Duy Nguyen, Khanh-Duy Ngu...	Few-shot object detection via baby learning.	104398	2022	120	NULL	Image Vis. Comput.	https://doi.org/10.1016/j.imavis.2022.104398
7	7	3007	Trong-The Nguyen, Trinh-Dong Nguyen, Truong-Giang N...	An Optimal Thresholds for Segmenting Medical Images ...	12-21	2022	13	NULL	J. Inf. Hiding Multim. Signal Process.	http://bit.kuas.edu.tw/~jihmsp/2022/vol13/N1
8	8	3008	Trong-The Nguyen, Trinh-Dong Nguyen, Vinh-Tiep Nguyen	An Optimizing Pulse Coupled Neural Network based on ...	155-164	2022	13	NULL	J. Inf. Hiding Multim. Signal Process.	http://bit.kuas.edu.tw/~jihmsp/2022/vol13/N3
9	9	3009	Xuan-Duong Nguyen, Anh-Khoa Nguyen Vu, Thanh-Danh...	Adaptive multi-vehicle motion counting.	2193-2201	2022	16	NULL	Signal Image Video Process.	https://doi.org/10.1007/s11760-022-02184-5
10	10	3010	Manh-Khanh Ngo Huu, Vinh Quang Ngo, Thanh-Danh N...	Antique Photo Restoration and Colorization via Generati...	1-6	2022	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR56351.2022.9
11	11	3011	Chi-Phuc Tran, Anh-Khoa Nguyen Vu, Vinh-Tiep Nguyen	Baby Learning with Vision Transformer for Face Recogn...	1-6	2022	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR56351.2022.9
12	12	3012	Duc-Tuan Luu, Khanh-An C. Quan, Thinh-Quyen Nguyen, ...	CDC: Color-Based Diffusion Model with Caption Embed...	575-579	2022	NULL	MMM (2)	NULL	https://doi.org/10.1007/978-3-030-98355-0_
13	13	3013	Manh-Khanh Ngo Huu, Sy-Tuyen Ho, Vinh-Tiep Nguyen, ...	Multilingual-GAN: A Multilingual GAN-based Approach f...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.9
14	14	3014	Ngoc-Khanh Nguyen, Thanh-Danh Nguyen, Vinh-Tiep Ng...	A More Focus on Multi-degradation Method for Single I...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.9
15	15	3015	Anh-Khoa Nguyen Vu, Thanh-Danh Nguyen, Vinh-Tiep Ng...	DF-FSOD: A Novel Approach for Few-shot Object Detecti...	1-6	2021	NULL	MAPR	NULL	https://doi.org/10.1109/MAPR53640.2021.9

Query executed successfully. | PHUCDANH (15.0 RTM) | PHUCDANH\Admin (58) | LAB3_XML | 00:00:00 | 72 rows

Hình 3.2.6: Kết quả khi lưu các bài báo có sự tham gia của thầy Tiệp xuống database.



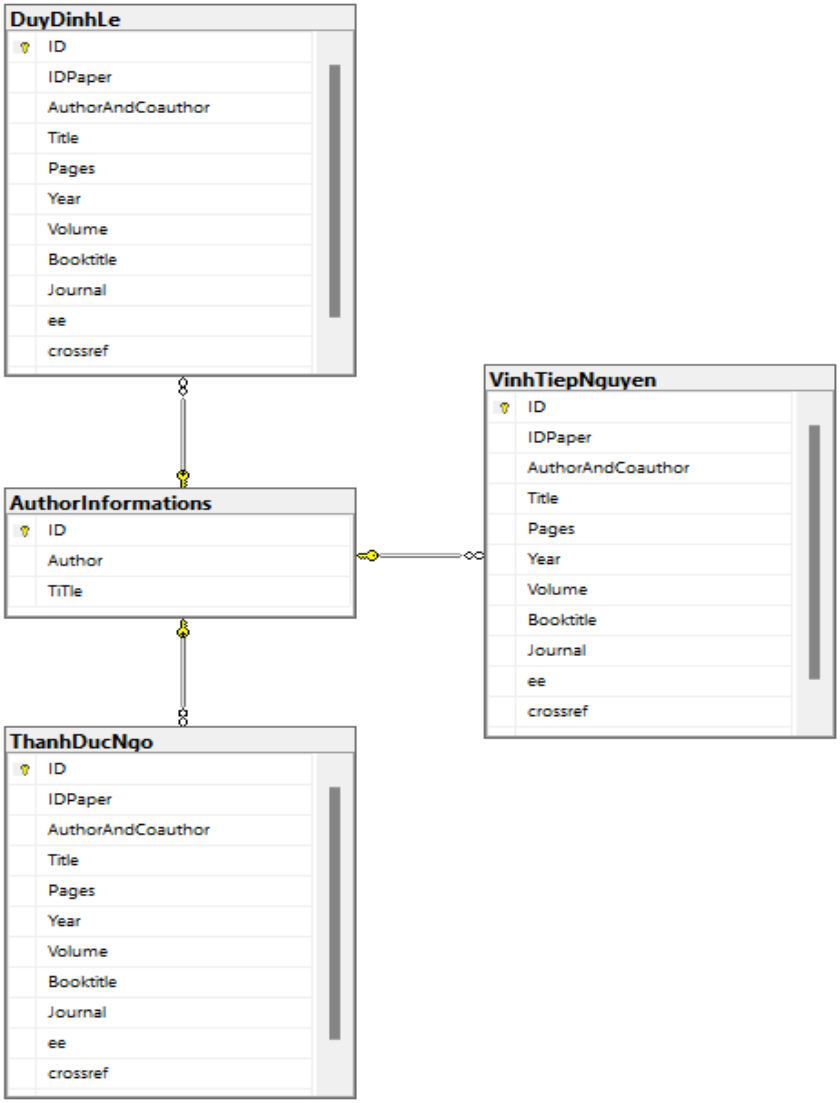
90 %

Results | Messages

	ID	Author	TiTle
1	1001	Thanh Duc Ngo	Few-shot Camouflaged Animal Detection and Segmen...
2	1002	Thanh Duc Ngo	Instance-level Few-shot Learning with Class Hierarchy ...
3	1003	Thanh Duc Ngo	Few-shot object detection via baby learning.
4	1004	Thanh Duc Ngo	Antique Photo Restoration and Colorization via Generat...
5	1005	Thanh Duc Ngo	A Crowdsourcing Data Annotation System For Vietnam...
6	1006	Thanh Duc Ngo	UIT at VBS 2022: An Unified and Interactive Video Retri...
7	1007	Thanh Duc Ngo	Dictionary-Guided Scene Text Recognition.
8	1008	Thanh Duc Ngo	Unweighted Bipartite Matching For Robust Vehicle Cou...
9	1009	Thanh Duc Ngo	Multilingual-GAN: A Multilingual GAN-based Approach ...
10	1010	Thanh Duc Ngo	A robust framework for mathematical formula detection.
11	1011	Thanh Duc Ngo	DF-FSOD: A Novel Approach for Few-shot Object Detec...
12	1012	Thanh Duc Ngo	MC-OCR Challenge 2021: An end-to-end recognition fr...
13	1013	Thanh Duc Ngo	Single-image crowd counting: a comparative survey on...
14	1014	Thanh Duc Ngo	An Evaluation of Deep Learning Methods for Small Obj...
15	1015	Thanh Duc Ngo	Interpolation based Anime Face Style Transfer.
16	1016	Thanh Duc Ngo	U15-Logos: Unconstrained Logo Dataset with Evaluati...

Query executed successfully.

Hình 3.2.7: Kết quả khi lưu toàn bộ tác và bài báo của cả 3 tác giả.



Hình 3.2.8: Cơ sở dữ liệu quan hệ.

IV. Đánh giá và kết luận

- **Đánh giá**:
  - Kết quả thu thập được sau một quá trình tự động hóa khá tốt.
  - Mỗi bài báo lưu trên file xml có cấu trúc thẻ không cố định chính vì thế các thẻ mang thông tin không có trong một bài báo sẽ được gán giá trị None hay NULL.

- **Kết Luận**:
  - Ưu điểm:
    - Dễ hiểu: việc source code được chia thành các class theo hướng đối tượng giúp code được sắp xếp rõ ràng, dễ nhìn.
    - Dễ thực thi: để chạy code cần chạy file main.py nên không cần phải chạy nhiều file một lúc.
    - Kết nối nhanh: Việc lưu một lượng lớn dữ liệu xuống database diễn ra khá nhanh.

  - Nhược điểm:
    - Tính toán thời gian crawl: Việc sử dụng selenium có tính tương tác cao đòi hỏi người dùng cần canh chỉnh thời gian hợp lý.
    - Yêu cầu kết nối internet ổn định: Việc loading website đòi hỏi đường truyền ổn định tuy nhiên không đòi hỏi đường truyền quá tốt.