POLITECHNIKA ŚLĄSKA

WYDZIAŁ INŻYNIERII MATERIAŁOWEJ

Kierunek: Informatyka Przemysłowa  
Profil praktyczny  
Rodzaj studiów: stacjonarne I stopnia

Projekt inżynierski  
Krzysztof MOCHOCKI

projekt i implementacja  
Systemu raportujący braki pomiędzy  
serwisami agregującymi publikacje  
naukowe napisany w języku C++

Design and implementation of  
reporting system for missing articles between science publications aggregation sites in C++

Kierujący pracą: Recenzent:

Dr inż. Adrian SMAGÓR dr hab. inż. Roman PRZYŁUCKI, prof. PŚ

Katowice, czerwiec 2021 r.

Druk jednostronny

Spis treści

1. Wstęp 4

1.1. Motywacja 4

1.2. Cel 4

2. Zakres Pracy 5

2.1. Przegląd Literaturowo - Technologiczny 5

2.1.1. Podobne Rozwiązania 5

2.1.2. Aktualny Model Synchronizacji 5

# Wstęp

## Inspiracja

Inspiracją dla niniejszego projektu inżynierskiego było dostrzeżenie problemu dotykającego pracowników naukowych Politechniki Śląskiej, którzy, aby móc dzielić się wynikami swoich prac z resztą świata nauki, zmuszeni są eksponować swój dorobek   
w różnych agregatach publikacji naukowych. Niestety w związku z różną datą pojawienia się agregatorów, oraz natłokiem innych zadań, ciężko jest dopilnować, aby wszystkie serwisy posiadały wszystkie publikacje.

## Cel

Celem projektu jest dostarczenie narzędzia pozwalającego zautomatyzować cały proces wyszukiwania brakujących publikacji wśród znanych i obsługiwanych serwisów.

Wśród obsługiwanych portali, powinny znaleźć się te, które posiadają najwięcej zbiorów, oraz te, które cieszą się największą popularnością. Dodatkowym kryterium jest możliwość dodania lub edycji danych w przeszukiwanym przez aplikację portalu.

Przy założonych kryteriach wyłaniają się 4 portale:

* DOROBEK, dostępny pod adresem *https://www.bg.polsl.pl/expertus/new/bib*
* ORCID, dostępny pod adresem *https://orcid.org*
* SCOPUS, dostępny pod adresem *https://www.scopus.com/*
* Web Of Science, dostępny pod adresem *https://webofknowledge.com*

Docelowo program, ma generować prosty w interpretacji raport w powszechnie znanym formacie, pozwalający na jednoznaczne wskazanie, zbiorów publikacji, które należy uzupełnić w poszczególnym portalu internetowym. Formatem spełniającym te warunki oraz dodatkowo zapewniający przenaszalność i wiele narzędzi w jego obsłudze, jest otwarty format opracowany przez firmę *Microsoft* dla arkuszy kalkulacyjnych: *xlsx.*

Biorąc pod uwagę możliwość pojawienia się nowych podobnych serwisów internetowych, projekt aplikacji powinien zapewniać łatwą rozszerzalność o nowe strony.

# Zakres Pracy

## Przegląd Literaturowo - Technologiczny

### Podobne Rozwiązania

W domenie publicznej niestety nie udało się znaleźć podobnych rozwiązań, co może być zrozumiałe z uwagi na charakter korzystania z tego typu serwisów. Wnioskując na podstawie dokumentacji, dostarczonej przez wyżej wymienione portale, programowalny interfejs jest udostępniany do integracji z wewnętrznymi zasobami ośrodków naukowych.

Zasoby tego typu, jeżeli będą otwarte, to nie będą skoncentrowane na twórczości pracowników Politechniki Śląskiej, lecz na swoich naukowcach. Fakt ten uniemożliwia korzystanie z potencjalnych portali, jako narzędzia do diagnozowania luk we wspieranych portalach, dla obcych badaczy.

### Aktualny Model Synchronizacji

W momencie rozpoczęcia prac brakowało narzędzi pozwalających na sprawną synchronizację publikacji naukowych pracowników Politechniki Śląskiej w internetowych bazach publikacji, takich jak: DOROBEK, SCOPUS, czy ORCID, ponieważ narzędzia, które miały im to umożliwić nie zapewniały wystarczających możliwości.

Dla pracowników Politechniki Śląskiej, punktem wyjścia ręcznego porównania spisów swoich publikacji w internetowych bazach, jest baza DOROBEK. Baza ta powinna zawierać wszystkie publikacje, pracowników Politechniki Śląskiej. Aby możliwe było w prosty i systematyczny sposób porównywać dane pomiędzy portalami potrzebna jest funkcjonalność prostego zbierania danych, czego baza DOROBEK nie zapewnia.

Naturalnym podejściem osoby niepotrafiącej programować, byłoby ręczne kopiowanie tytułów i wklejane ich do innych wyszukiwarek, celem weryfikacji ich istnienia. Temu podejściu nie sprzyja jednak mocno dynamiczny wygląd strony, który uniemożliwia sprawne zaznaczanie, interesującego ciągu znaków kursorem.

Alternatywą dla pierwszego sposobu mogłoby być skorzystanie z generacji pliku CSV, który jednak jest mocno niekompletny, między innymi przez brakujące numery ORCID, znacznie ułatwiające wyszukiwanie, a dane zarówno w nim jak i na stronie są nierzadko zapisane w nieprzystępnej formie.

Często widywaną praktyką jest wpisywanie tytułu w języku polskim i nowożytnym po przecinku, średniku lub kropce, co generuje kolejny problem. Problem powodujący niejednoznaczność w dalszych poszukiwaniach, ponieważ pojawiają się dwie różne wartości dla jednego nie tablicowego pola. Strona, oprócz pliku w formacie CSV oferuje również generowanie pliku w formacie RTF, który względem poprzedniego pliku ma przewagę w postaci kompletnych danych.

Największe trudności związane z automatycznym poborem danych, z bazy DOROBEK jest brak możliwości zdalnego generowania plików CSV i RTF, ponieważ są one generowane przez skrypty klienckie, co wymusza przetwarzanie gołego kodu strony w języku HTML wraz ze wszystkimi stylami i definicjami skryptów klienckich.

Domyślną postawą podczas komunikacji strony internetowej wraz z bazą danych, jest korzystanie z programowalnego interfejsu aplikacji (ang. API – Application Programmable Interface), za pomocą protokołu *REST*, czy *GRPC*. Można przypuszczać, że w przypadku tego serwisu, również zachodzi taka wymiana, co potwierdza szybka analiza ruchu sieciowego strony. Skorzystanie jednak z wyłonionego zapytania, poprzez wyszukanie największego typu metody POST (Rysunek 1), również nie napawa optymizmem, ponieważ zapytanie to, zwraca kawałek strony internetowej wygenerowanej po stronie serwera.

Ponadto zapytanie zbudowane przez stronę, wygląda na pozór na zaszyfrowane (Rysunek 1), ponieważ w rzeczywistości, są to trudno identyfikowalne nazwy wprowadzonych pól. Stosowanie takiej praktyki sprawia, że wykorzystanie tego zapytania przez zewnętrzny serwis wymaga zarówno wcześniejszego zgadywania przez programistę za pomocą prób i błędów znaczenia poszczególnych pól, a także żmudnego i wymagającego sporych zasobów przetwarzania dostarczonych danych w formacie HTML.

Wielokrotne próby komunikacji z administratorem strony zakończyły się kategoryczną odmową dostępu do jakiegokolwiek punktu dostępowego dla osób trzecich, co wymusza na osobie zainteresowanej automatyzacją procesu weryfikacji spójności danych w innych serwisach na skorzystanie z przetwarzanie danych uzyskanych ze źródła przeznaczonego dla użytkownika przeglądarki internetowej.



Zrzut Ekranu 1 Przedstawia zapytanie na jednym polu - autorze, wygenerowane przez serwis (skala szarości)

Kolejne serwisy, między innymi ORCID, oraz SCOPUS zapewniają bardzo obszerne dokumentacje, z przykładami oraz dokładnymi objaśnieniami, w przeciwieństwie do bazy DOROBEK są przyjaźnie nastawione do użytkownika i oferują zwięzły panel użytkownika