Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 6, No. 3, Maret 2026, Hal. 657-666
DOI: https://doi.org/10.52436/1.jpti.1425
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Akuisisi Data Menggunakan Teknik *Web Scraping* untuk Konstruksi Basis Data Riset

M. Miftakul Amin[*1], Sukmini[2], Nurhasanah[3], ST Amanda Almira[4], M. Indra Jaya Pratama Kusuma[5], Amielia Sinta Dewi[6]

[1,4,5,6]Jurusan Teknik Komputer, Politeknik Negeri Sriwijaya, Palembang, Indonesia
[2,3]Jurusan Akuntansi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
Email: [1]miftakul_a@polsri.ac.id, [2]sukmini_hartati@yahoo.co.id, [3]fatihfauziakbar@yahoo.co.id, [4]stamandaalmira@gmail.com, [5]indrajayapkk@gmail.com, [6]amielisintadewi02@gmail.com

Abstrak

Permasalahan utama dalam pengelolaan data riset di Indonesia adalah keterbatasan akses dan integrasi informasi dari berbagai sumber resmi, seperti SINTA, Scopus, Google Scholar, Garuda, buku, Hak Kekayaan Intelektual (HKI), penelitian, dan pengabdian masyarakat. Kondisi ini menyulitkan proses analisis kinerja riset, pemetaan kolaborasi, serta perencanaan strategis pengembangan riset. Penelitian ini bertujuan untuk membangun sistem akuisisi basis data riset yang mampu mengumpulkan data secara otomatis dari berbagai sumber tersebut menggunakan teknik *web scraping*. Metode penelitian mencakup perancangan arsitektur pengambilan data, pengembangan modul *scraping* untuk tiap sumber, proses ekstraksi dan transformasi data, serta penyimpanan hasil dalam basis data terintegrasi. Data hasil *scraping* kemudian disajikan melalui aplikasi berbasis web yang memungkinkan pencarian, visualisasi, dan analisis data riset secara terpadu. Hasil implementasi menunjukkan bahwa sistem mampu mengakuisisi dan menyimpan data penulis (*author*), publikasi *Scopus*, *Google Scholar*, Garuda, buku, HKI, penelitian, dan pengabdian secara konsisten dan terstruktur. Dampak dari penelitian ini adalah tersedianya basis data riset yang komprehensif dan terpusat, yang dapat mendukung pengambilan keputusan, meningkatkan transparansi kinerja riset, serta mempercepat proses kolaborasi dan inovasi di lingkungan akademik.

Kata kunci: *Akuisisi, Basis data riset, Integrasi informasi, Transformasi data, Web scraping.*

Data Acquisition Using Web Scraping Techniques for Research Database Construction

Abstract

The main problem in research data management in Indonesia is limited access to and integration of information from various official sources, such as SINTA, Scopus, Google Scholar, Garuda, books, Intellectual Property Rights (IPR), research, and community service. This condition complicates the process of analyzing research performance, mapping collaborations, and strategic planning for research development. This study aims to develop a research database acquisition system capable of automatically collecting data from various sources using web scraping techniques. The research methods include designing a data retrieval architecture, developing scraping modules for each source, extracting and transforming data, and storing the results in an integrated database. The scraped data is then presented through a web-based application that enables integrated research data search, visualization, and analysis. The implementation results show that the system is capable of acquiring and storing data on authors, Scopus publications, Google Scholar, Garuda, books, intellectual property rights, research, and community service in a consistent and structured manner. The impact of this research is the availability of a comprehensive and centralized research database, which can support decision making, increase the transparency of research performance, and accelerate the process of collaboration and innovation in the academic environment.

Keywords: *Acquisition, Research database, Information integration, Data transformation, Web scraping.*

1. PENDAHULUAN

Meningkatnya jumlah publikasi ilmiah dalam beberapa tahun terakhir menciptakan tantangan baru dalam pengelolaan data riset yang tersebar di berbagai *platform* seperti *Google Scholar*, *Scopus*, SINTA, dan Garuda. Data publikasi yang terfragmentasi tersebut menyulitkan institusi dalam melakukan analisis kinerja peneliti, pemetaan kolaborasi, maupun penyusunan strategi pengembangan riset [1]. Namun, proses pengumpulan data

p-ISSN: 2775-4227
e-ISSN: 2775-4219

secara manual terbukti lambat, tidak efisien, dan rawan kesalahan. Karena itu, dibutuhkan teknik otomatis seperti *web scraping* untuk mengakuisisi data riset secara cepat, akurat, dan terstruktur [2].

Fenomena lain yang muncul adalah belum tersedianya sistem informasi riset internal yang mampu menampilkan kinerja publikasi dosen dan peneliti secara *real-time* dan komprehensif. Banyak perguruan tinggi masih mengandalkan laporan manual atau input mandiri, padahal data publikasi merupakan indikator penting dalam penilaian akreditasi dan pemeringkatan institusi [3]. Hal ini membuktikan bahwa teknik *web scraping* mampu menghimpun data publikasi dalam skala besar secara efisien, yang sebelumnya sulit dicapai dengan cara konvensional. Gejala yang tampak akibat tidak adanya basis data riset terpusat adalah kesulitan memantau produktivitas peneliti, keterbatasan data untuk pengambilan keputusan strategis, serta lambatnya proses kolaborasi antar peneliti. Selain itu, institusi juga kesulitan menampilkan portofolio riset dosen secara aktual kepada publik dan mitra industri [4].

Integrasi akuisisi data riset umumnya mengombinasikan pengambilan *metadata* terstruktur melalui OAI-PMH dengan *scraping* dinamis pada halaman web yang tidak menyediakan *endpoint* terstruktur, sehingga memanfaatkan kemampuan OAI-PMH untuk *harvesting* selektif dan format *Dublin Core* sekaligus menjangkau konten yang hanya dapat diekstrak melalui parsing DOM, dimana praktik ini telah dideskripsikan dalam dokumentasi dan paket implementasi OAI *harvester* modern [5]. Teknologi *scraping* modern memadukan *headless browser*, HTML *parsers*, dan teknik *rendering* untuk mengekstrak elemen yang dihasilkan secara dinamis, yang menjadi standar praktik dalam studi-studi tinjauan terbaru tentang *web scraping* untuk basis data riset [6].

Informasi yang bersumber dari halaman web seperti *Google Scholar* atau halaman jurnal tanpa *Application Programming Interface* (API) resmi, *pipeline crawling* yang andal menggabungkan strategi *crawling* terjadwal, penanganan paginasi, dan mekanisme *backoff* untuk menjaga keandalan dan kepatuhan terhadap kebijakan akses situs [7]. Setelah pengambilan, tahap ETL (*Extract-Transform-Load*) berfokus pada normalisasi metadata, pemetaan skema ke *canonical schema*, dan pembersihan nilai kosong/inkonsisten yang diperlukan sebelum integrasi ke basis data riset. Sedangkan pada aspek sumber data, menurut [6] menekankan bahwa *Entity resolution* (*record linkage*) menjadi komponen kunci untuk merekonsiliasi penulis dan publikasi yang muncul di beberapa sumber, dan pendekatan *graph-based* atau pembelajaran tak berlabel menunjukkan performa baik pada entitas kompleks. Sehingga, pedoman pembersihan dan penggabungan data bibliometrik modern merekomendasikan langkah-langkah standar (skema canonical, deduplikasi, verifikasi otoritas penulis, dan *metadata enrichment*) untuk menghasilkan basis data riset yang konsisten, dapat ditelusuri, dan siap pakai untuk visualisasi serta layanan rekomendasi [8].

Penelitian yang dilakukan oleh [9], telah membahas permasalahan kesulitan peneliti dalam menemukan jurnal yang relevan untuk publikasi ilmiah akibat kurangnya integrasi antara sistem manajemen jurnal seperti *Open Journal System* (OJS) dengan layanan pencarian yang efektif. Untuk mengatasinya, dalam penelitian yang dilakukan telah mengembangkan aplikasi *journal finder* berbasis web yang merekomendasikan jurnal berdasarkan tingkat kemiripan judul dan abstrak artikel. Metode yang digunakan meliputi penerapan algoritma *Jaccard Similarity* untuk menghitung kesamaan antara input pengguna dan metadata jurnal, serta pemanfaatan protokol *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) untuk mengumpulkan metadata artikel dari 59 repositori OJS yang menghasilkan 3.321 data artikel ilmiah. Hasil pengujian menunjukkan aplikasi mampu menampilkan daftar rekomendasi jurnal dengan tingkat kemiripan tertentu, mempermudah peneliti menemukan jurnal yang sesuai secara lebih cepat, serta meningkatkan efisiensi proses publikasi artikel ilmiah.

Lebih lanjut, penelitian yang dilakukan oleh [10] telah membahas permasalahan belum tersedianya layanan informasi produktivitas riset dosen dan peneliti di Politeknik Negeri Sriwijaya, yang menyulitkan pihak institusi dalam memantau dan mengevaluasi kinerja publikasi ilmiah. Untuk mengatasinya, penelitian ini menggunakan metode *web scraping* atau *crawling* pada *Google Scholar* dengan memanfaatkan bahasa pemrograman *Python* dan pustaka seperti *Selenium*, *BeautifulSoup*, dan Panda, serta menyimpan hasilnya ke dalam basis data MySQL. Proses *crawling* dilakukan terhadap 403 akun peneliti/dosen dan berhasil mengumpulkan total 9.511 publikasi ilmiah ke dalam dua tabel terhubung (*gs_author* dan *gs_paper*). Hasil penelitian menunjukkan bahwa basis data riset yang dibangun mampu menampilkan direktori peneliti dan berbagai parameter kinerja publikasi dalam aplikasi berbasis web, yang bermanfaat untuk mendukung pengambilan keputusan strategis dalam meningkatkan reputasi institusi.

Lingkungan pengembangan pada aplikasi web cukup banyak digunakan pada berbagai bidang, diantaranya implementasi pada sistem informasi lowongan pekerjaan [11], pada pelaporan keuangan [12], pendataan susu sapi [13], rekam medis elektronik pada pasien bersalin [14], dan manajemen stok [15]. Begitu besarnya spektrum pengembangan aplikasi web, maka dalam penelitian yang dilakukan juga menggunakan *web* sebagai lingkungan *deployment* aplikasi dengan fokus pada basis data riset. Berbeda dengan penelitian-penelitian sebelumnya, penelitian ini menggunakan teknik *web scraping* untuk menghimpun basis data riset yang berasal dari beberapa

p-ISSN: 2775-4227
e-ISSN: 2775-4219

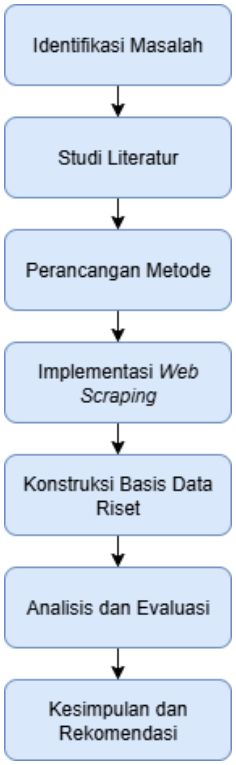
repository artikel ilmiah dengan menyimpan metadata artikel ke dalam sebuah basis data tunggal. Selanjutnya informasi yang berasal dari basis data riset tersebut dapat divisualkan dan ditampilkan dalam lingkungan aplikasi web, sehingga dapat dimanfaatkan lebih lanjut bagi para pihak yang membutuhkan.

2. METODE PENELITIAN

2.1. Tahapan Penelitian

Seperti yang disajikan pada Gambar 1, bahwa penelitian ini terdiri dari beberapa tahapan sebagai berikut:

1. Identifikasi masalah: Menemukan permasalahan terkait kesulitan memperoleh data riset secara terintegrasi.
2. Studi literatur: Mempelajari penelitian terkait *web scraping*, basis data, dan sistem informasi riset.
3. Perancangan metode: Menentukan sumber data (misalnya SINTA, *Scopus*, Garuda, *Google Scholar*), teknik scraping, serta rancangan basis data.
4. Implementasi *web scraping*: Membuat program *scraping* untuk mengumpulkan data publikasi, penulis, dan metadata.
5. Konstruksi basis data: Menyimpan hasil *scraping* ke dalam basis data terstruktur.
6. Analisis dan evaluasi: Mengevaluasi kualitas data (kelengkapan, akurasi, konsistensi) serta performa sistem.
7. Kesimpulan dan rekomendasi: Menyimpulkan hasil penelitian dan memberikan rekomendasi untuk pengembangan lanjutan.

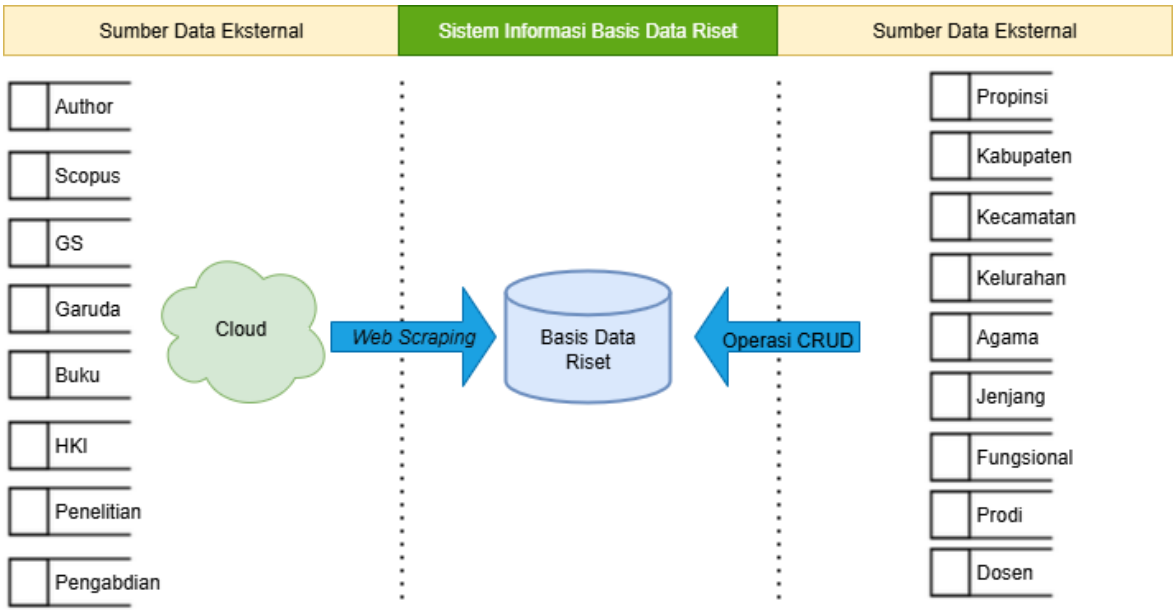


Gambar 1. Tahapan Penelitian

2.2. Arsitektur Sistem

Pada Gambar 2 memperlihatkan rancangan arsitektur sistem yang digunakan dalam penelitian ini. Basis data riset dikonstruksi menggunakan sumber data internal dan eksternal. Sumber data eksternal dilakukan dengan menggunakan *web scraping* untuk menghimpun basis data riset. Data yang dilakukan *scraping*, meliputi *author*, *scopus*, *google scholar*, garuda, buku, hak kekayaan intelektual (HKI), penelitian, dan pengabdian. Sedangkan sumber data internal, menggunakan operasi *Create*, *Read*, *Update*, dan *Delete* (CRUD) standar pada lingkungan aplikasi berbasis *web*. Selanjutnya basis data riset yang telah dihimpun, akan digunakan sebagai entitas tunggal dalam pengembangan aplikasi *web*.

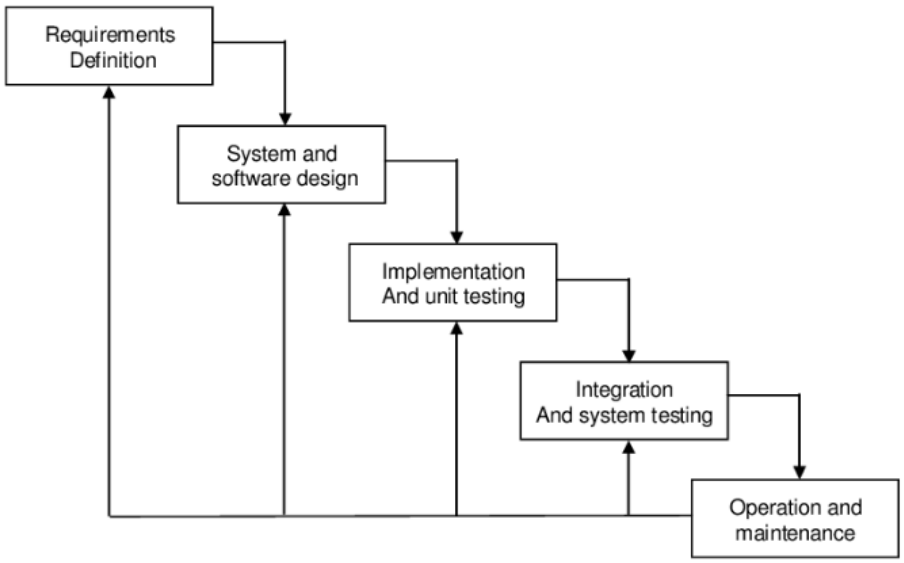
p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 2. Rancangan Arsitektur Sistem

2.3. Metode Pengembangan Sistem

Penelitian ini menggunakan metode pengembangan sistem *Waterfall* seperti tersaji pada Gambar 3, karena sistem yang digambarkan memiliki kebutuhan yang cukup jelas dan terstruktur [16], yaitu pengelolaan berbagai data akademik, penelitian, dan publikasi yang dapat diakses oleh *Admin* dan *User*. Dengan pendekatan *Waterfall*, tahapan analisis kebutuhan dapat dilakukan terlebih dahulu untuk mendefinisikan semua entitas data (seperti dosen, prodi, publikasi, dan wilayah administratif), kemudian dilanjutkan dengan tahap perancangan sistem yang dituangkan dalam diagram UML seperti *use case diagram*. Setelah itu, proses implementasi dapat dilakukan berdasarkan rancangan yang sudah matang, lalu diuji melalui tahap pengujian agar setiap fungsi berjalan sesuai kebutuhan. Pendekatan ini cocok karena sistem lebih menekankan pada kejelasan proses, konsistensi pengolahan data, serta integrasi informasi antar-entitas yang membutuhkan dokumentasi detail dan pengendalian ketat pada setiap tahap pengembangan.

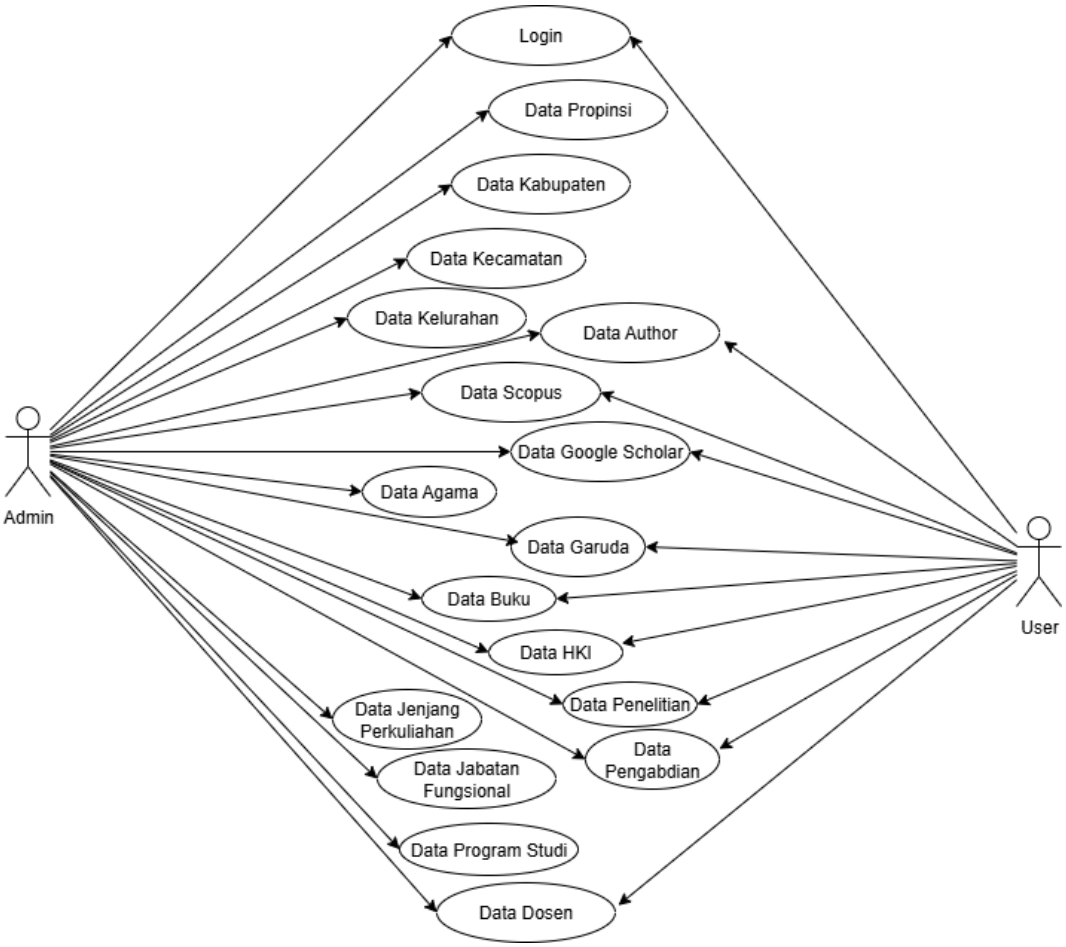


Gambar 3. Model *Waterfall* [16]

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.4. Rancangan *Use Case Diagram*

Use Case Diagram pada Gambar 4 menggambarkan interaksi dua aktor utama, yaitu ***Admin*** dan ***User***, terhadap sistem pengelolaan basis data riset. *Admin* memiliki hak akses penuh untuk mengolah berbagai entitas data, mulai dari data wilayah administratif (propinsi, kabupaten, kecamatan, kelurahan), data akademik (agama, jenjang perkuliahan, jabatan fungsional, program studi, dosen, *author*), hingga data ilmiah dan publikasi (dokumen *Scopus*, *Google Scholar*, Garuda, buku, hak kekayaan intelektual, penelitian, dan pengabdian). *User* juga berinteraksi dengan sistem untuk mengakses informasi dari entitas yang sama, meskipun pada umumnya perannya lebih terbatas dibandingkan *Admin* yang dapat melakukan pengolahan data. Selain itu, terdapat *use case* ***Login*** yang menjadi pintu masuk bagi kedua aktor sebelum dapat mengakses fungsionalitas lainnya. Diagram ini menekankan bahwa sistem dirancang untuk mendukung kebutuhan manajemen data terintegrasi yang relevan baik bagi administrator maupun pengguna umum.



Gambar 4. Rancangan Use Case Diagram

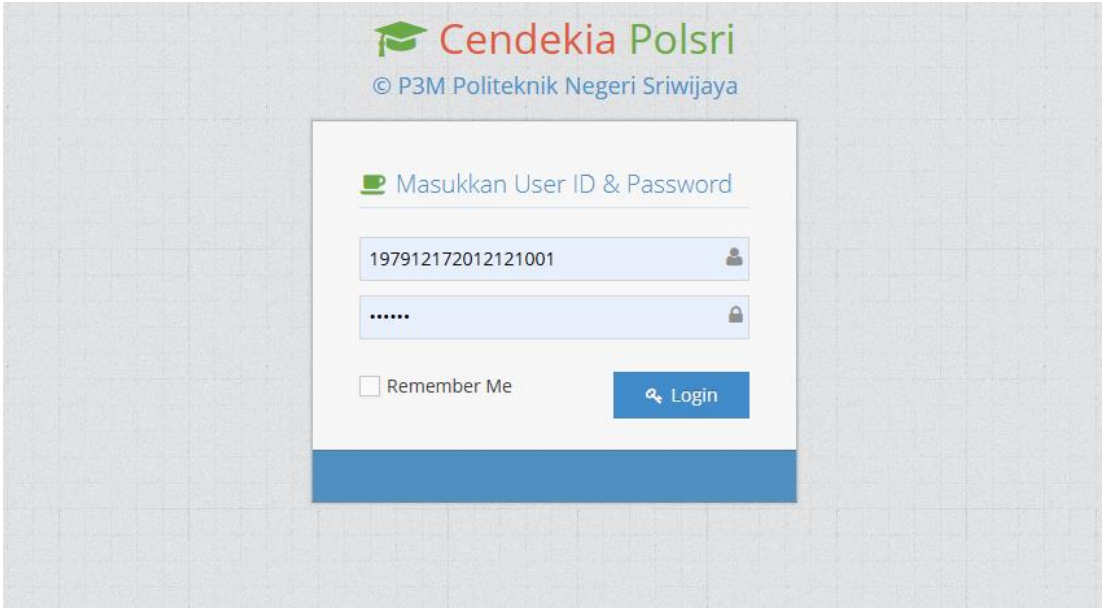
3. HASIL DAN PEMBAHASAN

3.1. Hasil Pengembangan Aplikasi

3.1.1. Halaman Login

Pada pengembangan aplikasi yang dijalankan pada lingkungan berbasis *web*, diperoleh sebuah aplikasi yang diberi nama ”Cendekia Polsri”. Pertama kali dijalankan, maka akan diperoleh tampilan seperti diperlihatkan pada Gambar 5. Terdapat 2 buah informasi penting yang dapat dimasukkan oleh user, yang pertama adalah *username* dan yang kedua adalah *password*. Selanjutnya, jika user terlah berhasil melakukan login, maka akan dibawa ke menu *dashboard* atau menu utama. Namun, jika proses otentikasi user gagal, maka akan diminta memasukkan *username* dan *password* sampai terverifikasi dengan benar.

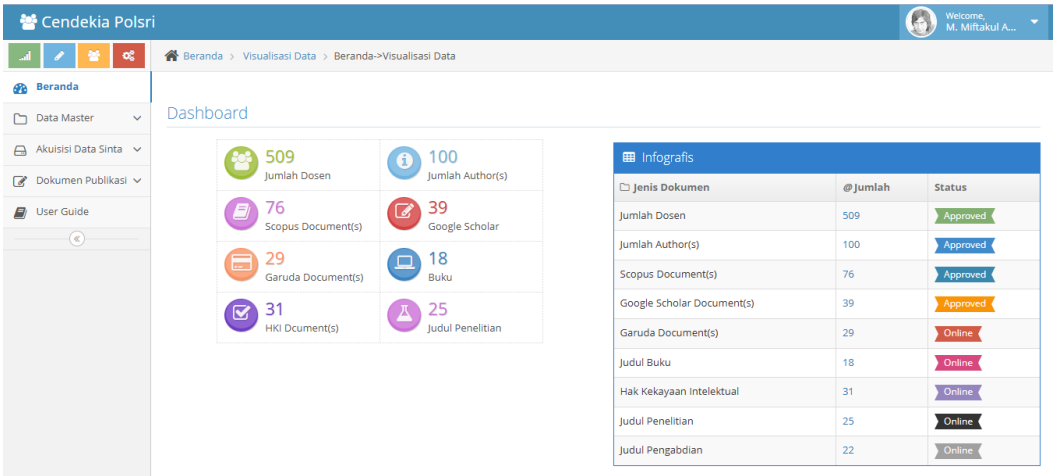
p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 5. Halaman Login

3.1.2. Menu Utama

Aplikasi yang dihasilkan menyediakan sebuah menu utama atau *dashboard* aplikasi seperti tersaji pada Gambar 6. Tampilan ini akan disajikan, ketika user telah berhasil melakukan login terlebih dahulu. Pada Gambar 6 ini terdapat beberapa segmen yang dapat dipilih user, seperti pada bagian atas terdapat informasi tentang user yang berhasil login ke dalam sistem. Kemudian pada bagian kiri, terdapat menu untuk bekerja dengan fitur yang disediakan oleh sistem seperti data master, akuisisi data SINTA, dan dokumen yang telah berhasil dikumpulkan. Sedangkan pada bagian tengah, terdapat tampilan jumlah dokumen yang telah berhasil dikumpulkan dan disimpan dalam basis data riset.

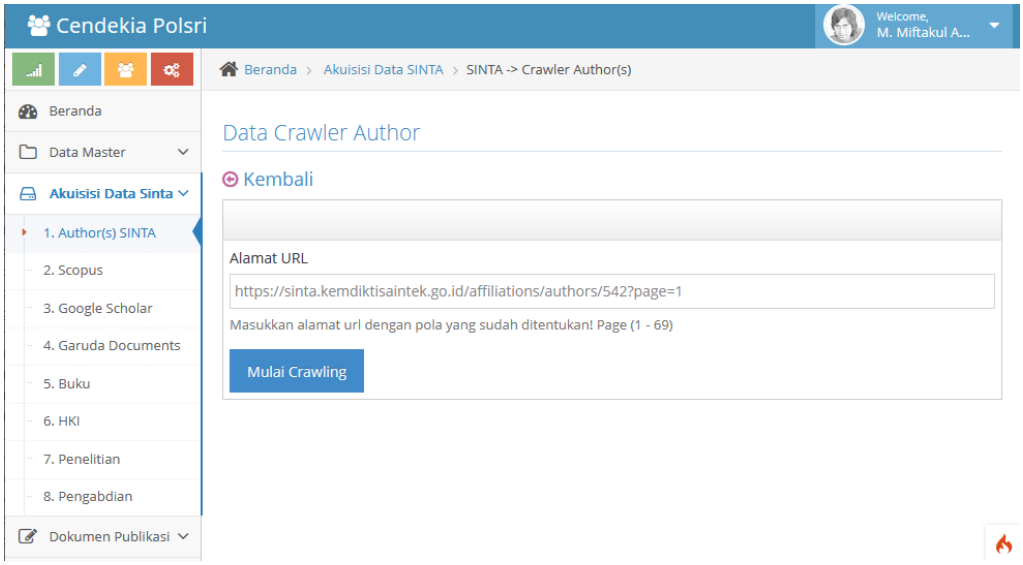


Gambar 6. Halaman Menu Utama

3.1.3. Menu Akuisisi Data SINTA

Basis data riset dihimpun dari laman SINTA, yang terdiri dari data *author*, *scopus*, *google scholar*, garuda, buku, hki, penelitian, dan pengabgian. Tampilan dari halaman akuisisi data SINTA dapat dilihat pada Gambar 7. Proses menghimpun data *author* dijadikan sebagai acuan awal sebelum melakukan akuisisi data lainnya. Sedangkan pengambilan data publikasi didasarkan pada informasi ID SINTA, setiap *author* yang telah berhasil dihimpun.

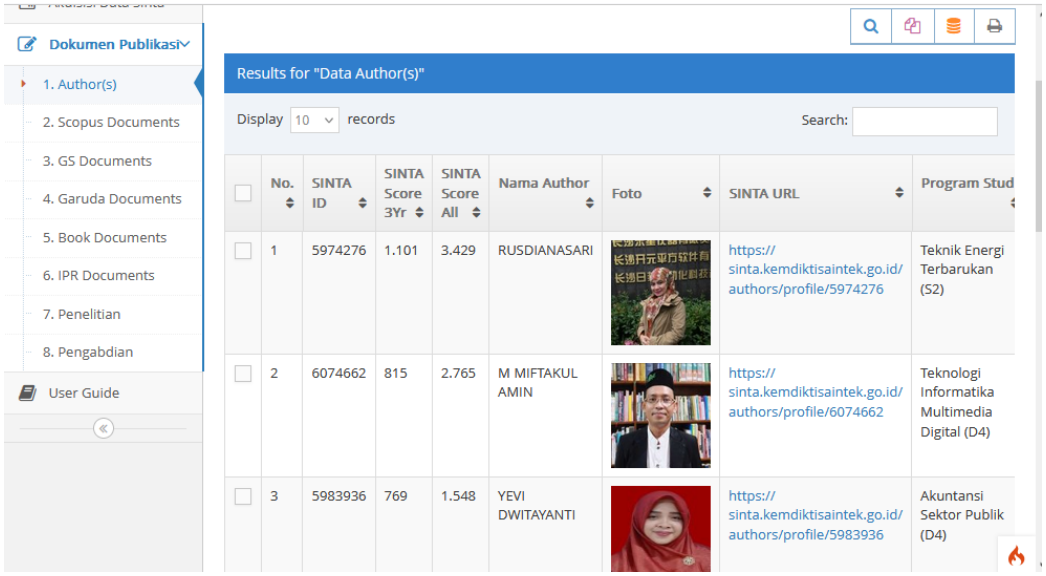
p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 7. Menu Akuisisi Data SINTA

3.1.3.1. Menu Dokumen Publikasi

Metadata yang telah berhasil dihimpun, selanjutnya dapat disajikan berupa tampilan data seperti tersaji pada Gambar 8. Pada tampilan data disajikan informasi Sinta ID, *Sinta Score 3 year*, *Sinta Score All*, Nama, Foto, Sinta *URL*, dan program studi yang menjadi *homebase* Dosen. Saat ini telah terhimpun sebanyak 690 Dosen dan Peneliti di Politeknik Negeri Sriwijaya.

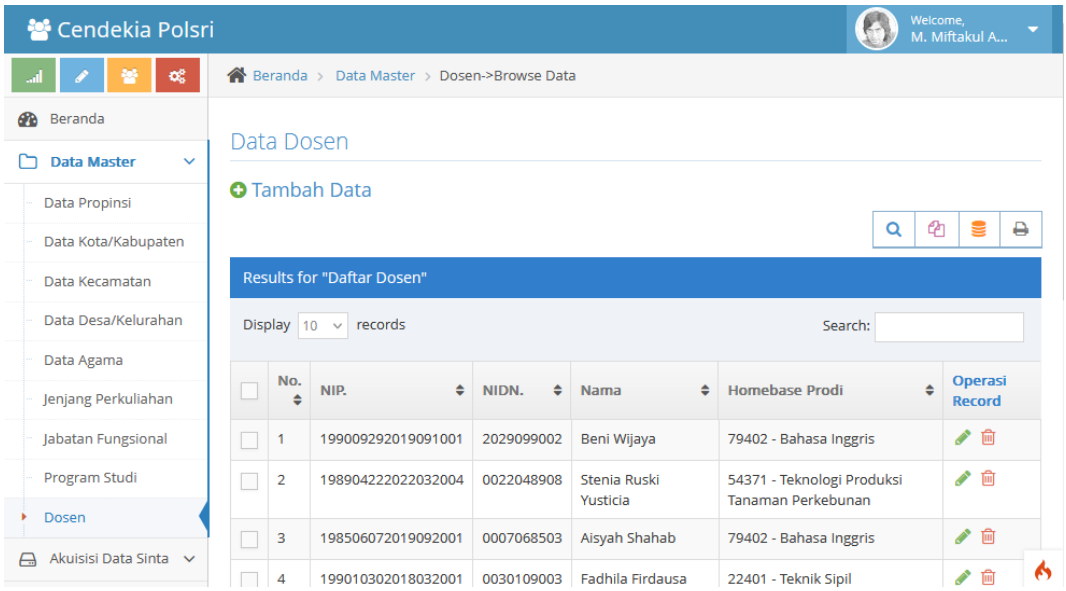


Gambar 8. Menu Akuisisi Data SINTA

3.1.3.2. Menu Data Master

Dalam aplikasi juga dikembangkan pemrosesan data yang berasal dari internal perguruan tinggi, seperti data Dosen, jenjang perkuliahan, data wilayah (propinsi, kabupaten, kecamatan, kelurahan), agama, jenjang akademik, dan program studi. Tampilan aplikasi data internal perguruan tinggi yang dikelompokkan pada Data Master dapat dilihat pada Gambar 9. Setiap pengolahan data, dilengkapi dengan berbagai fungsionalitas seperti Tambah Data, Edit Data, Update Data, Delete Data, Pencarian Data, Ekspor Data dan Pengecekan Data.

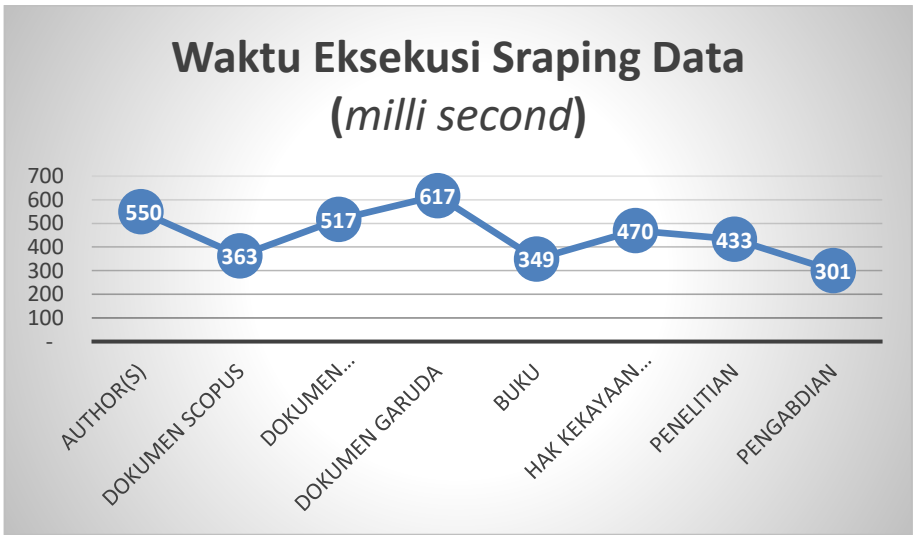
p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 9. Menu Akuisisi Data SINTA

3.2. Pembahasan

Penelitian ini menekankan pentingnya integrasi data riset dari berbagai sumber resmi yang selama ini terfragmentasi, seperti SINTA, Scopus, Google Scholar, Garuda, dan repositori lainnya. Fragmentasi data tersebut mengakibatkan kesulitan dalam melakukan analisis kinerja riset maupun dalam mengidentifikasi potensi kolaborasi antar peneliti. Dengan mengimplementasikan metode *web scraping* yang terstruktur, penelitian ini berhasil mengotomatisasi proses akuisisi data sehingga mempercepat pengumpulan informasi yang sebelumnya dilakukan secara manual. Selain itu, arsitektur sistem yang dibangun mendukung proses ekstraksi, transformasi, dan penyimpanan data ke dalam basis data yang terintegrasi, sehingga menghasilkan repositori riset yang lebih konsisten dan mudah diakses.



Gambar 10. Waktu Eksekusi *Web Scraping*

Lebih lanjut, hasil implementasi membuktikan bahwa sistem mampu menangani berbagai jenis data, mulai dari publikasi ilmiah, buku, hingga HKI dan aktivitas pengabdian masyarakat. Data yang diperoleh kemudian disajikan melalui aplikasi berbasis *web* yang memudahkan pengguna dalam melakukan pencarian, analisis, serta visualisasi. Keberadaan sistem ini memiliki dampak strategis karena tidak hanya meningkatkan transparansi dan akurasi dalam pelaporan kinerja riset, tetapi juga mendukung pengambilan keputusan berbasis data. Dengan demikian, penelitian ini memberikan kontribusi signifikan terhadap pengembangan infrastruktur riset nasional, terutama dalam mendorong kolaborasi lintas institusi serta mempercepat proses inovasi di bidang akademik.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Mengacu pada Gambar 10, terdapat beberapa kali pengujian dari proses *web scraping* dari 8 (delapan) sumber *dataset*. Waktu eksekusi rata-rata diperoleh data 450 ms (*milli second*). Waktu eksekusi paling lama 617 ms, dan paling kecil 301 ms. Perbedaan waktu eksekusi dipengaruhi oleh beberapa faktor, diantaranya adalah besarnya payload, koneksi jaringan internet, dan spesifikasi perangkat keras yang digunakan.

4. KESIMPULAN

Penelitian yang telah dilakukan telah menghasilkan sebuah aplikasi berbasis *web* yang diberi nama ”Cendekia Polsri”, sebagai sebuah repositori basis data riset pada tingkat perguruan tinggi. Aplikasi telah berhasil menyediakan sebuah infrastruktur akuisisi data untuk melakukan konstruksi data menggunakan teknik *web scraping*, untuk menghimpun beragam metadata *author*, dokumen *scopus*, dokumen *google scholar*, dokumen garuda (garba rujukan digital), buku, hak kekayaan intelektual, penelitian, dan pengabdian kepada masyarakat. Teknik *web scraping* dapat dilakukan secara periodik, sesuai dengan kebutuhan pengumpulan data yang diperlukan. Data dapat dihimpun dari eksternal, dan juga internal perguruan tinggi, dengan menambahkan fitur input data yang berasal dari berbagai kegiatan tridarma internal perguruan tinggi.

Penelitian lanjutan dapat dilakukan dengan membangun *Application Programming Interface* (API) maupun *web services* yang dapat menghubungkan beragam repositori ilmiah untuk menghimpun basis data riset dari berbagai sumber data. Pengembangan pengukuran kinerja berdasarkan capaian publikasi ilmiah dapat dilakukan, sebagai upaya untuk mengukur tingkat produktifitas para Dosen dan peneliti. Dari sisi *UI/UX* juga dapat dilakukan penyempurnaan untuk memudahkan pengakses data dan informasi, serta penelusuran dan navigasi fungsionalitas dari sistem yang dikembangkan. Ekspansi dari sisi lingkungan pengembangan seperti aplikasi *mobile* juga menjadi salah satu alternatif perluasan akesibilitas dari pengembangan perangkat lunak. Penambahan fitur-fitur dalam aplikasi juga dapat diperluas, sehingga memberikan nilai tambah terhadap sistem yang telah dikembangkan.

DAFTAR PUSTAKA

[1] J. Mingers, J. R. O. Hanley, and M. Okunola, “Using Google Scholar institutional level data to evaluate the quality of university research,” *Scientometrics*, vol. 113, no. 3, pp. 1649–1665, 2017, doi: 10.1007/s11192-017-2532-6.

[2] J. Tjaden, “Web Scraping for Migration, Mobility, and Migrant Integration Studies: Introduction, Application, and Potential Use Cases,” *Int. Migr. Rev.*, vol. 15, no. 3, 2023, doi: https://doi.org/10.1177/01979183231208428.

[3] R. Ulloa, F. Mangold, F. Schmidt, and J. Gilsbach, “Beyond time delays : how web scraping distorts measures of online news consumption,” *Commun. Methods Meas.*, vol. 00, no. 00, pp. 1–22, 2025, doi: 10.1080/19312458.2025.2482538.

[4] I. Finocchi, A. Martino, F. Ranjbar, and B. Sinaimeri, “Data cleaning and enrichment through data integration : networking the Italian academia,” *Sci. Data*, vol. 12, pp. 1–16, 2025, doi: 10.1038/s41597-025-04608-6.

[5] K. Hornik, “Metadata Harvesting with R and OAI-PMH,” in *OAIHarvester vignette*, 2024, pp. 1–7.

[6] M. A. Brown, A. Gruen, G. Maldoff, S. Messing, and M. Zimmer, “Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations,” in *arXiv*, 2024, pp. 1–43.

[7] C. Lotfi, S. Srinivasan, M. Ertz, and I. Latrous, “Web Scraping Techniques and Applications : A Literature Review,” in *SCRS Conference Proceedings on Intelligent Systems*, 2021, pp. 381–394. doi: https://doi.org/10.52458/978-93-91842-08-6-38.

[8] M. Nowakowska, “A comprehensive approach to preprocessing data for bibliometric analysis,” *Scientometrics*, no. 0123456789, 2025, doi: 10.1007/s11192-025-05415-x.

[9] M. M. Amin, A. Firdaus, and Y. Dwitayanti, “Model Rekomendasi Jurnal dengan Algoritma Jaccard Similarity dan Protokol OAI- PMH Journal Recommendation Model with Jaccard Similarity Algorithm and OAI-PMH Protocol,” *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. 10, pp. 489–499, 2024, doi: https://doi.org/10.52436/1.jpti.725.

[10] M. M. Amin, A. Sutrisman, and Y. Dwitayanti, “Google Scholar Crawling for Constructing Research Database,” in *7th FIRST 2023 International Conference on Global Innovations (FIRST-ESCSI 2023)*, Atlantis Press International BV, 2024, pp. 331–337. doi: 10.2991/978-94-6463-386-3.

[11] A. Hamid *et al.*, “Pengembangan Aplikasi Lamar Bagawi dengan Metode SDLC Waterfall untuk Pengelolaan Lowongan Kerja di Kabupaten Balangan Development of the Lamar Bagawi Application

p-ISSN: 2775-4227
e-ISSN: 2775-4219

with the SDLC Waterfall Method for Job Vacancy Management in Balangan Regency," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 2, pp. 321–329, 2025, doi: https://doi.org/10.52436/1.jpti.658.

[12] N. N. Umami and A. Yudhistira, "Pengembangan Sistem Pelaporan Keuangan Berbasis Web Menggunakan Metode Waterfall Untuk Meningkatkan Transparansi Pengelolaan Dana di MTS MA Margodadi Fakultas Teknik dan Ilmu Komputer , Universitas Teknokrat Indonesia , Indonesia Development of a Web-Base," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 4, pp. 909–918, 2025, doi: https://doi.org/10.52436/1.jpti.725.

[13] S. S. Verdananti, K. R. Ummah, and U. P. Boyolali, "Rancang Bangun Sistem Informasi Berbasis Web untuk Pendataan Hasil Susu Sapi di Usaha Dagang Pramono Boyolali Teknik Informatika , Universitas Muhammadiyah Surakarta , Indonesia Design af A Web-Based Information System for Data Collection af Cow ' s Milk," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 7, pp. 1891–1903, 2025, doi: https://doi.org/10.52436/1.jpti.713.

[14] A. Ri and W. Widayat, "Rancang Bangun Sistem Informasi Rekam Medis Pasien Bersalin Berbasis Website Design and Development of a Web-Based Medical Record Information System for Maternity Patients," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 6, pp. 1595–1608, 2025, doi: https://doi.org/10.52436/1.jpti.839.

[15] M. Kurniasih, W. Widayat, T. Informatika, F. Komunikasi, and U. M. Surakarta, "Sistem Informasi Manajemen Stok Berbasis Web Menggunakan Framework Laravel A WEB-BASED STOCK MANAGEMENT INFORMATION SYSTEM USING THE LARAVEL FRAMEWORK," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 5, pp. 1457–1469, 2025, doi: https://doi.org/10.52436/1.jpti.816.

[16] I. Sommerville, *Software Engineering*, 10th ed. Boston: Pearson Education, 2016.