BAB 2

KAJIAN PUSTAKA

2.1 Otomatisasi DCC di BSN

Otomatisasi dokumen adalah proses penggunaan perangkat lunak untuk mengekstrak data dari satu sumber dan secara sistematis menyusunnya ke dalam dokumen terformat, dengan tujuan utama untuk mengurangi intervensi manual dan meningkatkan efisiensi [3]. Dalam konteks Badan Standardisasi Nasional (BSN), otomasi menjadi krusial untuk menjawab tantangan strategis dalam proses pembuatan Sertifikat Kalibrasi Digital (DCC). Sesuai tujuan BSN untuk mengadopsi standar internasional [1], langkah fundamental pertama adalah mengotomatisasi pembuatan DCC Tingkat 1 (*human-readable*). Tantangan utama dalam tahap ini adalah tingginya variasi tata letak pada lebih dari 40 templat dokumen yang berbeda antar laboratorium [4].

2.2 Document Layout Analysis (DLA) sebagai Bidang Masalah

Tantangan dalam memahami dokumen secara otomatis berada dalam bidang masalah Document Layout Analysis (DLA). Secara definitif, DLA adalah proses untuk mengidentifikasi dan mengsegmentasi halaman dokumen menjadi wilayah atau blok logis, seperti paragraf, tabel, gambar, dan judul [5]. Tujuan DLA adalah untuk memahami "peta" atau struktur sebuah dokumen sebelum isinya dapat diekstraksi secara akurat.

Untuk menyelesaikan masalah DLA, terdapat berbagai pendekatan yang dapat digunakan. Pemilihan pendekatan sangat bergantung pada karakteristik masalah. Mengingat dokumen BSN memiliki keteraturan pola internal yang dapat diprediksi (misalnya, penggunaan konsisten simbol ':' pada form), penelitian ini mengadopsi pendekatan berbasis aturan sebagai solusi yang paling efisien.

2.3 Pendekatan Berbasis Aturan

Pendekatan yang diimplementasikan dalam penelitian ini adalah sistem ekstraksi informasi berbasis aturan. Dalam pendekatan ini, aturan-aturan dirancang secara manual oleh seorang ahli berdasarkan analisis mendalam terhadap struktur dokumen target. Hal ini sejalan dengan kerangka kerja yang diusulkan oleh Chen,

An, dan Zeng [2], di mana mereka juga menggunakan aturan buatan manusia (*hand-made rules*) untuk mengekstrak data dari dokumen ilmiah semi-terstruktur.

Prinsip dasar dari pendekatan ini adalah pemanfaatan "petunjuk visual" (*visual cues*) yang ada di dalam dokumen untuk memahami strukturnya. *Paper* oleh Chen *et al.* [2] memperkenalkan konsep analisis dua tingkat:

- Identifikasi *Regions-of-Interest* (ROI) Merupakan proses segmentasi dokumen menjadi "wilayah" atau blok-blok informasi yang lebih besar dan bermakna secara logis [2].
- Identifikasi *Entities-of-Interest* (EOI) yang Merupakan proses ekstraksi dan populasi data spesifik atau "entitas" dari dalam ROI yang telah diidentifikasi sebelumnya [2].

Secara konseptual, ini identik dengan algoritma yang dikembangkan dalam penelitian ini, yang pertama-tama melakukan segmentasi data menjadi "blok" logis, kemudian mengklasifikasikan dan mengekstrak data dari dalam blok tersebut.

2.3.1 Implementasi Algoritma DLA Berbasis Aturan

Algoritma yang dikembangkan dalam penelitian ini bekerja dalam dua fase utama yaitu fase pertama menangani ekstraksi data mentah dan fase kedua berfungsi untuk mempopulasi data ke dalam dokumen *template*.

A Ekstraksi dan Klasifikasi Blok

Fase ini bertugas menganalisis tata letak file Excel dan mengekstrak informasinya. Alur kerjanya adalah sebagai berikut:

1. Menerima Input Data Algoritma menerima input berupa file Excel, sebuah praktik yang umum di validasi dalam penelitian sejenis [6, 7].
2. Segmentasi Blok (Identifikasi ROI) Algoritma membaca lembar kerja dan menggunakan setiap baris kosong sebagai penanda heuristik untuk melakukan segmentasi, yaitu memisahkan data menjadi "blok-blok" informasi. Langkah ini mengacu pada proses identifikasi *Regions-of-Interest* (ROI) yang dilakukan pada penelitian berjudul "A Rule-based Information Extraction System for Human-readable Semi-structured Scientific Documents" [2].

3. Klasifikasi Blok (Pengenalan Pola) Untuk setiap blok, algoritma menerapkan aturan heuristik untuk mengenali polanya dan mengklasifikasikannya sebagai Form, Tabel, atau Paragraf. Ini merupakan implementasi dari teknik *Pattern Recognition* [7].
4. Ekstraksi Data (Identifikasi EOI) Setelah klasifikasi, data spesifik di dalam setiap blok diekstrak (misalnya, pasangan kunci-nilai dari form). Langkah ini setara dengan proses identifikasi *Entities-of-Interest* (EOI) [2] dan disimpan dalam sebuah struktur data *dictionary* [3].

B Fase 2: Populasi Dokumen Word

Fase ini mengambil data terstruktur dan memasukkannya ke dalam dokumen akhir.

5. Menerima Input Templat Algoritma menerima input kedua berupa file templat Word sebagai target populasi data [8, 6].
6. Pencocokan dan Populasi Data Algoritma mengiterasi elemen di dalam templat Word dan memetakannya dengan data yang telah diekstrak berdasarkan pencocokan judul atau label, lalu mengisi data tersebut.
7. Menyimpan Output Dokumen Word yang telah terisi disimpan sebagai hasil akhir dari proses [7].

2.4 Pengujian Sistem

Pengujian sistem dilakukan untuk memverifikasi fungsionalitas dan memastikan sistem otomatisasi dapat menghasilkan dokumen secara akurat sesuai dengan data sumber. Proses pengujian ini dilakukan dengan menggunakan 10 dokumen kalibrasi yang berasal dari 3 jenis laboratorium yang berbeda, yaitu laboratorium Kelistrikan, Pengukuran Panjang, dan Kalibrasi Suhu.

Setiap dokumen yang dihasilkan oleh sistem akan dibandingkan dengan data pada file Excel untuk memeriksa kebenaran dan akurasi pemetaan elemen-elemen kunci seperti *form* dan tabel. Akurasi pemetaan data diukur secara kuantitatif menggunakan metrik *precision* dan *recall* Hasil dari pengujian ini menjadi tolok ukur untuk menentukan tingkat akurasi dan keandalan fungsional sistem dalam memproses dokumen [6].

Untuk mengukur efektivitas algoritma yang diusulkan, dirancang sebuah skema evaluasi kuantitatif. Evaluasi ini bertujuan untuk mengukur performa pada dua tingkat granularitas yang berbeda, selaras dengan pendekatan yang diulas oleh Chen *et al.* [2], yaitu pada tingkat wilayah (ROI) dan tingkat entitas (EOI). Untuk mengukur performa sistem ekstraksi informasi dan klasifikasi dokumen, digunakan metrik evaluasi kuantitatif yang standar. Dua metrik yang paling umum digunakan adalah presisi (*precision*) dan daya panggil (*recall*) [9].

2.4.1 Presisi (Precision)

Presisi didefinisikan sebagai persentase dokumen relevan yang berhasil diambil dengan benar oleh sistem, relatif terhadap jumlah total dokumen yang diambil oleh sistem tersebut. Metrik ini mengukur tingkat keandalan atau kualitas dari hasil yang diklaim berhasil oleh sistem [10]. Rumus untuk menghitung presisi terdapat pada Rumus 2.1:

$$Precision = \frac{TP}{TP + FP} \quad (2.1)$$

Rumus 2.2: Rumus Presisi

Di mana TP (True Positive) adalah jumlah dokumen yang diklasifikasikan dengan benar, dan FP (False Positive) adalah jumlah dokumen yang salah diklasifikasikan.

2.4.2 Daya Panggil (Recall)

Daya panggil didefinisikan sebagai persentase dokumen relevan yang berhasil diambil dengan benar oleh sistem, relatif terhadap jumlah total dokumen relevan yang seharusnya ada. Metrik ini mengukur tingkat kelengkapan atau kemampuan sistem untuk menemukan semua target yang relevan [10]. Rumus untuk menghitung *recall* terdapat pada Rumus 2.3:

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

Rumus 2.4: Rumus Recall

Di mana FN (False Negative) adalah jumlah dokumen yang seharusnya diklasifikasikan ke dalam kategori tertentu tetapi terlewatkan oleh sistem [10].

Kedua metrik ini digunakan secara bersamaan untuk memberikan gambaran performa yang seimbang, karena presisi yang tinggi dengan recall yang rendah (atau

sebaliknya) tidak dapat menunjukkan sistem yang baik secara keseluruhan.

A Pengukuran Kinerja ROI

Pada tingkat ini, evaluasi berfokus pada kemampuan algoritma untuk melakukan segmentasi dan klasifikasi blok data dengan benar. Dalam konteks penelitian ini, ROI adalah sebuah blok logis ('Form', 'Tabel', atau 'Paragraf'). Performa diukur dengan metrik Presisi dan Recall, yang didefinisikan sebagai berikut:

- True Positive (TP) yang merupakan sebuah blok data diidentifikasi dengan batasan (baris awal dan akhir) yang benar DAN diklasifikasikan ke dalam kategori yang tepat [2].
- False Positive (FP) terjadi jika sistem mengidentifikasi sebuah blok dengan batasan yang salah (misalnya, memecah satu tabel menjadi dua) atau dengan klasifikasi yang salah (misalnya, mengklasifikasikan tabel sebagai form) [2].
- False Negative (FN) Terjadi jika sistem gagal mendeteksi sebuah blok data yang seharusnya ada [2].

B Pengukuran Kinerja EOI

Setelah sebuah ROI diidentifikasi dengan benar, evaluasi dilanjutkan pada tingkat EOI untuk mengukur keakuratan ekstraksi data di dalamnya. Dalam konteks ini, EOI adalah unit data individual (misalnya, satu pasangan 'kunci:nilai' atau satu baris data tabel). Metrik utama yang digunakan adalah Presisi dan Recall, yang didefinisikan sebagai:

- True Positive (TP) Sebuah entitas data spesifik berhasil diekstrak dengan nilai yang benar dari ROI dan dipetakan dengan benar [2].
- False Positive (FP) Sebuah entitas data diekstrak dan dipetakan, namun nilainya salah atau tidak lengkap. Kasus kegagalan pada form multi-baris adalah contoh dari FP pada tingkat EOI [2].
- False Negative (FN) Sebuah entitas data yang seharusnya ada di dalam sebuah ROI yang benar terlewatkan dan gagal diekstrak oleh sistem [2].