ABSTRAK

Perpustakaan Perguruan Tinggi berfungsi sebagai pusat sumber belajar dan sumber informasi yang berkedudukan di perguruan tinggi. Selain menyediakan berbagai jenis informasi, sebuah perpustakaan juga melakukan katalogisasi, yaitu proses pengolahan data-data bibliografi yang terdapat pada bahan pustaka seperti judul, nama pengarang, nama penerbit, dan tahun terbit menjadi sebuah katalog.Pengolahan data-data tersebut jika tidak diimbangi dengan pengelolaan data yang efektif dapat menyebabkan terhambatnya kebaruan data, sehingga teknologi mulai digunakan seperti sistem informasi perpustakaan berbasis aplikasi *web* dan *mobile* yang terhubung dengan sistem *Cloud Computing*. Sistem yang telah dikembangkan tersebut masih memiliki kekurangan pada proses pemasukkan data, yaitu membutuhkan waktu yang relatif lama karena harus melakukan *input* sejumlah data ke dalam sistem dengan cara diketik secara langsung.

Salah satu solusi untuk masalah ini adalah dengan menerapkan *Optical Character Recognition* (OCR) dalam proses pemasukan data menggunakan *Tesseract* OCR. *Tesseract* OCR adalah sebuah *OCR* berbasis layanan yang dapat diakses secara gratis. Hasil akurasi *Tesseract* akan berkurang apabila terdapat objek gangguan pada citra yang akan diproses, sehingga perlu dilakukan *image preprocessing* untuk meningkatkan kualitas citra. Tahapan *image preprocessing* yang dilakukan yaitu seleksi area pendeteksian, *grayscalling*, binarisasi dan kompress ukuran citra. Data yang dihasilkan setelah proses OCR selesai dilakukan adalah sebuah string biasa atau dapat berupa *html-based* string sehingga data tersebut masih harus diproses lagi agar dapat digunakan semestinya. Salah satu pemrosesan yang dapat dilakukan adalah melakukan pengelompokkan data sesuai kebutuhan berdasarkan atribut-atribut yang dibutuhkan. Pengelompokkan data *string* dapat dilakukan dengan menggunakan algoritma *string matching* yang yaitu *Jaro-Winkler Distance*.

Berdasarkan hasil pengujian yang telah dilakukan pada 10 dokumen skripsi dan 10 dokumen buku, akurasi *Tesseract* OCR dapat ditingkatkan dengan menerapkan *image preprocessing* pada citra yang akan diproses. Nilai peningkatkan akurasi yang dihasilkan adalah 35.080% untuk pengenalan halaman pengesahan, 2,540% untuk pengenalan halaman abstrak dan 19.492% untuk pengenalan halaman penerbit. Pada pengujian klasifikasi data, Algoritma *Jaro-Winkler Distance* dapat digunakan untuk proses klasifikasi data dengan nilai rata-rata akurasi yang dihasilkan pada klasifikasi data skripsi adalah sebesar 97.871% dan klasifikasi data buku adalah sebesar 71.387%.

Kata Kunci : Perpustakaan, *Image Preprocessing*. *Optical Character Recognition, Jaro-Winkler Distance*, *Cloud Computing*.