

## ABSTRACT

This project primarily exists as my personal desire to create a web crawler application for browsing the web and store the information without much interaction from human. And Java is chosen as development base because it runs very well on many platform ( including my Linux rig ) and it has a good collection of network or Internet ready API which fulfill my needs. The great threading ability of Java also becoming a major advantage for this project because web crawler ability to process multiple link simultaneously must be implemented using threads.

Web crawler means any application which capable of automatically process any hyper links found on a HTML page and process it to gather more links which will be processed again, inside the application this behavior will take place until a stop condition is met ( usually a certain link depth or number of files retrieved ). Although there are many purpose of a web crawler, the main purpose of using it is for information gathering or archiving. The contents within this paper will describe the process of developing a web crawler used for archiving files gathered from HTML paged to local storage for off line browsing.

And for the final words, this project has meets it's goal though not 100% perfect because there's still some minor glitch in few of the code. Although rare, this problem will cause inability to retrieve some type of relative links. In the end I just going to say thanks to all who has support this project, I'm looking forward to fix the glitches and make the application better.

**Keywords:** Web Crawler, Java Network Programming, Spider.

## DAFTAR ISI

|  |    |
|--|----|
| BAB 1: Pendahuluan.....                | 12 |
| 1.1. Latar Belakang.....               | 12 |
| 1.2. Tujuan.....                       | 12 |
| 1.3. Gambaran Umum.....                | 13 |
| 1.4. Pembatasan Masalah.....           | 13 |
| 1.5. Detail Teknis.....                | 14 |
| 1.6. Definisi dan Singkatan.....       | 15 |
| 1.7. Struktur Laporan.....             | 17 |
| BAB 2: Spesifikasi Produk.....         | 19 |
| 2.1. Perspektif Produk.....            | 19 |
| 2.1.1. Antarmuka Sistem.....           | 19 |
| 2.1.2. Antarmuka Pengguna.....         | 19 |
| 2.1.3. Antarmuka Perangkat Keras.....  | 20 |
| 2.1.4. Antarmuka Perangkat Lunak.....  | 20 |
| 2.1.5. Antarmuka Jaringan.....         | 21 |
| 2.1.6. Batasan Memori.....             | 21 |
| 2.1.7. Persyaratan Adaptasi.....       | 21 |
| 2.2. Fungsi Produk.....                | 21 |
| 2.3. Karakteristik Pengguna.....       | 22 |
| 2.4. Asumsi dan Ketergantungan.....    | 22 |
| 2.5. Penundaan Persyaratan.....        | 22 |
| 2.6. Batasan Performa.....             | 22 |
| 2.7. Fitur Produk Perangkat Lunak..... | 22 |
| 2.7.1. Download File.....              | 23 |
| 2.7.2. Parsing Link.....               | 24 |
| 2.7.3. Membuat Site Map.....           | 25 |
| 2.7.4. Pengecekan Broken Link.....     | 25 |

|   |    |
|---|----|
| 2.8. Atribut - Atribut Perangkat Lunak..... | 26 |
| 2.8.1. Keandalan.....                       | 26 |
| 2.8.2. Ketersediaan.....                    | 27 |
| 2.8.3. Keamanan.....                        | 27 |
| 2.8.4. Pemeliharaan.....                    | 27 |
| 2.8.5. Perpindahan.....                     | 27 |
| BAB 3: Desain perangkat lunak.....          | 28 |
| 3.1. Identifikasi.....                      | 28 |
| 3.2. Konsep Eksekusi.....                   | 29 |
| 3.2.1. Use Case.....                        | 29 |
| 3.2.2. Keterangan Use Case.....             | 30 |
| 3.2.3. Sequence Diagram.....                | 34 |
| BAB 4: Implementasi desain.....             | 40 |
| 4.1. Perencanaan Implementasi.....          | 40 |
| 4.1.1. Implementasi Komponen.....           | 40 |
| 4.1.2. Gambaran Implementasi Komponen.....  | 42 |
| 4.2. Detail Implementasi Komponen.....      | 44 |
| 4.2.1. Package crawler.....                 | 44 |
| 4.2.2. Package crawler.conf.....            | 46 |
| 4.2.3. Package crawler.core.....            | 52 |
| 4.2.4. Package crawler.core.job.....        | 54 |
| 4.2.5. Package crawler.core.link.....       | 58 |
| 4.3. Sequence Diagram Proses.....           | 60 |
| 4.3.1. Ekstrak Link.....                    | 60 |
| 4.3.2. Download File.....                   | 61 |
| 4.3.3. Web Crawling.....                    | 62 |
| 4.4. Keterangan Implementasi Komponen.....  | 64 |
| 4.4.1. Package crawler.....                 | 64 |
| 4.4.2. Package crawler.conf.....            | 66 |
| 4.4.3. Package crawler.core.....            | 69 |
| 4.4.4. Package crawler.core.job.....        | 71 |
| 4.4.5. Package crawler.core.link.....       | 74 |

|  |    |
|--|----|
| 4.5. Keterkaitan Antar Komponen.....                   | 75 |
| 4.5.1. Core Components.....                            | 75 |
| 4.5.2. Configuration Components.....                   | 77 |
| 4.5.3. User Interface Components.....                  | 78 |
| 4.6. Perjalanan Tahap Implementasi.....                | 79 |
| 4.6.1. Model Implementasi.....                         | 80 |
| 4.6.2. Realisasi Fungsionalitas.....                   | 80 |
| 4.6.3. Realisasi Antar Muka Pengguna.....              | 81 |
| BAB 5: Testing dan Evaluasi.....                       | 84 |
| 5.1. Test Case.....                                    | 84 |
| 5.2. Metodologi Pengujian.....                         | 85 |
| 5.2.1. Test case link extractor.....                   | 85 |
| 5.2.2. Test case konfigurasi.....                      | 88 |
| 5.2.3. Test case downloader.....                       | 90 |
| 5.2.4. Test case crawling.....                         | 91 |
| 5.2.5. Ulasan Hasil Evaluasi.....                      | 92 |
| BAB 6: Kesimpulan dan saran.....                       | 93 |
| 6.1. Keterkaitan Kesimpulan Dengan Hasil Evaluasi..... | 93 |
| 6.2. Keterkaitan Saran Dengan Hasil Evaluasi.....      | 93 |
| 6.3. Saran Pengembangan.....                           | 93 |

## DAFTAR GAMBAR

|   |    |
|---|----|
| Gambar 3.1: Use Case Web Crawler.....             | 30 |
| Gambar 3.2: Sequence Diagram Start Crawling.....  | 34 |
| Gambar 3.3: Sequence Diagram Stop Crawling.....   | 34 |
| Gambar 3.4: Sequence Diagram Pause Crawling.....  | 35 |
| Gambar 3.5: Sequence Diagram Create Project.....  | 35 |
| Gambar 3.6: Sequence Diagram Load Project.....    | 36 |
| Gambar 3.7: Sequence Diagram Save Project.....    | 37 |
| Gambar 3.8: Sequence Diagram Create Site Map..... | 37 |
| Gambar 3.9: Sequence Diagram Config Crawler.....  | 38 |
| Gambar 4.1: Component Diagram Web Crawler.....    | 42 |
| Gambar 4.2: Interface Launcher.....               | 43 |
| Gambar 4.3: Kelas AppLauncher.....                | 43 |
| Gambar 4.4: Kelas MacLauncher.....                | 44 |
| Gambar 4.5: Kelas JWebCrawlerApp.....             | 44 |
| Gambar 4.6: Enum Field.....                       | 45 |
| Gambar 4.7: Enum Type.....                        | 45 |
| Gambar 4.8: Kelas ConfigParser.....               | 46 |
| Gambar 4.9: Kelas ConfigFactory.....              | 47 |
| Gambar 4.10: Kelas ConfigValidator.....           | 47 |
| Gambar 4.11: Kelas ValueException.....            | 48 |
| Gambar 4.12: Kelas Config.....                    | 49 |
| Gambar 4.13: Kelas Downloader.....                | 50 |
| Gambar 4.14: Kelas Spider.....                    | 51 |
| Gambar 4.15: Kelas RobotsProtocol.....            | 51 |
| Gambar 4.16: Enum CapacityStatus.....             | 52 |
| Gambar 4.17: Enum JobStatus.....                  | 52 |
| Gambar 4.18: Kelas Job.....                       | 53 |

|  |    |
|--|----|
| Gambar 4.19: Interface JobControl.....           | 53 |
| Gambar 4.20: Kelas JobManager.....               | 54 |
| Gambar 4.21: Kelas UrlQueue.....                 | 55 |
| Gambar 4.22: Kelas LinkExtractor.....            | 56 |
| Gambar 4.23: Kelas LinkUtil.....                 | 56 |
| Gambar 4.24: Class Diagram SiteTree.....         | 57 |
| Gambar 4.25: Sequence Diagram Extract Links..... | 58 |
| Gambar 4.26: Sequence Diagram Download.....      | 59 |
| Gambar 4.27: Sequence Diagram Crawling.....      | 60 |
| Gambar 4.28: Core Components.....                | 70 |
| Gambar 4.29: Config Components.....              | 71 |
| Gambar 4.30: GUI Components.....                 | 72 |
| Gambar 4.31: Tampilan Utama.....                 | 74 |
| Gambar 4.32: Tampilan New Project.....           | 75 |
| Gambar 4.33: Tampilan Konfigurasi.....           | 75 |
| Gambar 4.34: Command Line Interface.....         | 76 |
| Gambar 5.1: LinkExtractorDemo 1.....             | 78 |
| Gambar 5.2: LinkExtractorDemo 2.....             | 79 |
| Gambar 5.3: LinkExtractor Demo3.....             | 79 |
| Gambar 5.4: Test Case Config 1.....              | 80 |
| Gambar 5.5: Test Case Config 2.....              | 80 |
| Gambar 5.6: Test Case Config 3.....              | 81 |
| Gambar 5.7: Test Case Config 4.....              | 81 |
| Gambar 5.8: Test Case Downloader 1.....          | 82 |
| Gambar 5.9: Test Case Downloader 2.....          | 82 |