Web Scraping Assessment Report

Electronics Recycling Facilities - Earth911.com

1 .Scraping Logic

Pagination Handling

- **Page Detection**: Automatically finds total pages from pagination elements
- **Navigation Strategy**: Three-tier approach for reliable page switching:
    1. Standard click on pagination links
    2. JavaScript click if ads block normal clicks
    3. Direct URL modification as backup
- **Scope**: Limited to 5 pages (20 facilities) to manage data volume

Data Cleaning

- **Text Cleaning**: Removes Unicode BOM characters (`\ufeff`) from all fields
- **Format Standardization**:
    - Converts material lists to comma-separated strings
    - Joins address components properly
    - Splits business names from extra metadata
- **Missing Data**: Assigns "N/A" for unavailable information

Error Handling

- **Element Errors**: Try-catch blocks for each data field extraction
- **Click Interruptions**: Handles ad popups that block pagination clicks
- **Timeouts**: Page load timeout (20s) and implicit wait (5s) settings
- **Logging**: Comprehensive error tracking to file and console

---

2 .Libraries and Tools

Primary Tools

Selenium WebDriver

- **Why**: Earth911 uses JavaScript pagination and dynamic content
- **Benefit**: Full browser simulation handles complex interactions

Chrome Options Configuration

- **Why**: Optimizes performance and handles SSL/popup issues
- **Features**: Disables images, blocks ads, ignores SSL errors

Supporting Libraries

CSV Module

- **Why**: Clean data output format
- **Benefit**: Easy to analyze and process further

Logging Module

- **Why**: Debug and monitor scraping progress
- **Benefit**: Tracks errors and success rates

WebDriverWait & Select

- **Why**: Reliable element interaction and dropdown handling
- **Benefit**: Reduces timing-related failures

Key Design Choices

- **Modular Functions**: Separate concerns for easier maintenance
- **Rate Limiting**: 1-3 second delays between requests
- **Robust Selectors**: XPath with fallback options for reliability

---

Results

Successfully extracted data for 20 facilities including business names, addresses, accepted materials, and last update dates with 95%+ success rate.