PMSI Scraping Estimate – Template

1. Requirements

1.1 Basic details

**Site Name**

|  |
| --- |
| Lidl UK Postcodes Promos |

**Site URL (or example if multiple URLs)**

|  |
| --- |
| <http://www.lidl.co.uk/> |

**Country / Regional Requirements**

|  |
| --- |
| Lidl UK (inc northern Ireland <http://www.lidl-ni.co.uk/> ) |

**Objective**

|  |
| --- |
| Aim is to scrape all product offer information available for each store in Lidl UK (all existing pages that we currently scape).  Offers vary per store, so we will need to loop through postcodes, select each unique store, and then scrape promo data for each unique store. |

1.2 Scrape Timings

**Frequency**

|  |
| --- |
| One off exercise |

**Timing**

|  |
| --- |
| - |

1.3 Output

|  |
| --- |
| Excel output |

1.4 Pilot (optional)

|  |
| --- |
|  |

1.5 Other

|  |
| --- |
|  |

1. Deadlines

**Estimate**

|  |
| --- |
|  |

**Pilot / Test data**

|  |
| --- |
|  |

**Production data**

|  |
| --- |
| ASAP |

1. Fields Summary

**Fields that will be scraped:**

|  |
| --- |
| -- Title  -- Price  -- BasePrice  -- Image link  -- Price Cut  -- Price without offer  -- Condition  -- Brand  -- Offer date  -- Detail Link  -- Product description |

**ID Field:**

|  |
| --- |
| -- ProductID available , need to combined storeID (address info) to distinguish same product in different store. |

**Comments:**

|  |
| --- |
| -- Site’s store locator don’t generate individual store url, for a postcode search it displays 10 nearest stores but there is no absolute url available for those links.  Example of Url available for each store:  <a href="#">Select store to see offers</a>  Once one of the link is clicked it takes to that store page, site adds stores information in cookie and each time a category is selected it sends that cookie information along with that category link. |

1. Estimates

**Development Time (Production):**

|  |  |  |  |
| --- | --- | --- | --- |
| Low: | 2 | High: | 2 |

**Development Time (Pilot):**

|  |  |  |  |
| --- | --- | --- | --- |
| Low: |  | High: |  |

**Scrape Time:**

|  |  |  |  |
| --- | --- | --- | --- |
| Low: |  | High: |  |

**Scrape Cost:**

|  |  |  |  |
| --- | --- | --- | --- |
| Low: |  | High: |  |

|  |  |  |
| --- | --- | --- |
| Site complexity: |  |  |

**Comments:**

|  |
| --- |
| Challenges:   1. Bot don’t support to update cookie for each url, for example lidl.co.uk have 612 stores that means need to generate 612 cookies and need to update cookie for each store url. 2. Need to build urls and cookies for each store 3. Need to verify that the custom urls and cookies (Urls and cookies are built from store scraped information). The problem is returned html don’t have any information that confirms our requested url returned the exact store’s html. When we use browser and select one store on browser it displays the store address (My Lidl) but this store address information doesn’t available in the source code.   Example:  Url: <http://www.lidl.co.uk/en/index.htm?ar=2&changestore=1>  Cookie: ar=ar=2;ar\_zip=London;ar\_city=NW 2 6ND;ar\_street=Edgware Road 408;nf=true  Conclusion : It is possible to scrape store based promotion sites but need to decide how to update cookies for each url (need core changes in ScrapeingHerd and WebScraper). Also need to verify the returned html to make sure bot is scraping the store it supposed to.  Proposal :  \*\*\* Need to Build cookie for each url from the store scraping data  -- We can add cookie for each url in Url name field and from that field we can update cookie for each url.  Or we can add another field in ScraperSiteUrl table like UrlSettings which will contain Json Key Value Pairs and we can convert this Json using JObject and Use it the way we want. |