BASICS of our data requirements

WEBSITE FOR DATA

We need to scrape records from an Irish website of obituary records. Records are added daily to this site. We would like to harvest the data on a quarterly basis. From this site [www.rip.ie](http://www.rip.ie).

Typically there are about 33k records on this site per year spread over 5 target counties (Dublin, Kildare, Meath, Cork, Galway and Wicklow).  The target counties represent about 8k records in Dublin, 1k in Kildare, 1k in Meath, 4k in Cork, Galway 2k and 1k in Wicklow (so 17k out of about 35k in Ireland).

The landing page of the target website (rip.ie) has a search bar for each county and a date a range.

A screenshot of a computer

Description automatically generated

PURPOSE OF SCRAPED DATA

We are basically using this data in Excel or Power BI to slice and dice and monitor trends over time.

Previously we were receiving data in a csv format that we could import directly to our spreadsheet.  A sample set of records were typically like those in the table below, organised in a file for each county.  The data was normally reported each quarter one month in arrears.

DELIVERY FORMAT OF SCARPED DATA

I attach a sample CSV format with the column headers (Sample CSV.xlsx)

We would like the data supplied in 1 csv file per county per quarter. So we need 6 files per quarter for each Dublin, Kildare, Meath, Cork, Galway and Wicklow as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| From | To |  | Dublin | Kildare | Meath | Wicklow | Cork | Galway |  |
| 01/01/2023 | 31/03/2023 | 1 Quarter | 1 file | 1 file | 1 file | 1 file | 1 file | 1 file | Now |
| 01/04/2023 | 30/06/2023 | 1 Quarter | 1 file | 1 file | 1 file | 1 file | 1 file | 1 file | Now |
| 01/07/2023 | 30/08/2023 | 1 Quarter | 1 file | 1 file | 1 file | 1 file | 1 file | 1 file | Now |
| 01/10/2023 | 31/12/2023 | 1 Quarter | 1 file | 1 file | 1 file | 1 file | 1 file | 1 file | Future |

FIELD DATA FORMAT AND NAMES

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field Ref in Image | Field / Column Name | Current Order | Data Type | Data Description |
| Field 1 | name | 1 | Text | First Name |
| Field 2 | url | 2 | URL | Url of record |
| Field 3 | county | 3 | Text | County Name |
| Field 4 | town | 4 | Text | Town Name |
| Field 5 | funeral\_director | 5 | Text | Funeral Director Name |
| Field 6 | death\_notice | 6 | Text | Text of Death Notice |
| Field 7 | church\_parsed | 7 | Text | Church name extracted from text (If possible) |
| Field 8 | church\_from\_map | 8 | Text | Church name from "Map" |
| Field 9 | cemetery\_parsed | 9 | Text | Cemetery name extracted from text (If possible) |
| Field 10 | cemetery\_from\_map | 10 | Text | Cemetery name from "Map" |
| Field 11 | date\_of\_death | 11 | Date | Date field data |
| Field 12 | date\_published | 12 | Date | Date field data |
| Field 13 | last\_name | 13 | Text | Name 2 field data |
| Field 14 | first\_name | 14 | Text | Name 1 field data |
| Field 15 | nee | 15 | Text | Name 3 field data |
| Field 16 | noticeid | 16 | Number | Unique record ref from website |
| Field 17 | error | 17 | Text |  |
| Field 18 | is\_duplicate | 18 | True or False | True if record has duplicate noticeid in dataset False if record has unique noticeid |
| Field 19 | county\_prio | 19 | Number | County Priority |

NAMING CONVENTION OF SCARPED DATA

This is the naming convention for each file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| County Name | Begin Date | End Date | File Name | File Name Convention |
|  |  |  |  |  |
| Dublin | 01/01/2023 | 31/03/2023 | Dublin\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |
| Kildare | 01/01/2023 | 31/03/2023 | Kildare\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |
| Meath | 01/01/2023 | 31/03/2023 | Meath\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |
| Cork | 01/01/2023 | 31/03/2023 | Cork\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |
| Galway | 01/01/2023 | 31/03/2023 | Galway\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |
| Wicklow | 01/01/2023 | 31/03/2023 | Wicklow\_2023\_01\_01\_to\_2023\_03\_31 | County\_Begin\_Date\_to\_End\_Date |

The typical record page with each of the required data field items highlighted in red below.

THIS IMAGE IS A COUNY TABLE

A screenshot of a computer

Description automatically generated

THIS IS AN IMAGE OF A FULL RECORD

A screenshot of a computer

Description automatically generated

THIS IS THE TOP HALF OF A RECORD

A screenshot of a website

Description automatically generated

THIS IS THE BOTTOM HALF OF A RECORD

A screenshot of a computer

Description automatically generated