Ministry of Technology,
Communication and Innovation

Building Datasets

Preparing Data for Publication – The Data Wrangling Process

Adapted partly from: New York State Open Data Dataset Submission Guide v3.0 Open NY

Ver. 1.0, April 2018

Table of Contents

1. Glossary

Serial	Term	Description/Explanation
1	**API**	Application Programming Interface: Software that allows machine to machine communication over the internet. For data, APIs allow apps to read just the data they need directly, without downloading an entire dataset, saving bandwidth and ensuring that the data used is the most up-to-date available.
2	**CSV**	Comma Separated Values: A standard format for spreadsheets where data is stored in a plain text file, with each data row on a new line and commas separating the values on each row. As a simple open format it is easily read by computers and is widely used for publishing open data.
3	**Dataset**	A dataset is any organized collection of data. The most basic dataset is composed of data elements in a table. Each column represents a particular variable. Each row corresponds to a given value of that column's variable. A dataset may also present information in a variety of non-tabular formats, such as an extended mark-up language (XML) file, a geospatial data file, or an image file. Dataset is a flexible term and may refer to an entire database, a spreadsheet or other data file, or a related collection of data resources.
4	**Datastore**	DKAN Datastore bundles a number of modules and configuration to allow users to upload CSV files, parse them and save them into the native database as flat tables, allowing users to query them through a public API. To get the fullest functionality possible out of your datasets, you should add your CSV resources to the datastore.
5	**Database**	Can be a software system for processing and managing data, including features to update, transform and query the data. Examples are PostgreSQL (open source) and Microsoft Access (proprietary). A database can also refer to a set of data.
6	**Data Resources**	The individual tables, data dictionaries, and visualizations that comprise a dataset, along with the associated metadata to make them findable and usable.
7	**Data File**	Tabular. Data must be machine-readable, and formatted according to uniform technical standards for import to od.govmu.org
8	**Data File Format**	Comma-Separated Value (CSV) or tab-separated value (TSV); UTF-8 character encoding is required; Data file must be saved in CSV
9	**DKAN**	DKAN is the Drupal-based open source data platform (Drupal Knowledge Archive Natwork) that the portal uses for its open data efforts. DKAN allows governments to publish data to the public, provide visualizations and data stories and create internal analytics dashboards.
10	**Historical Data and**	For example, if a particular dataset's methodology changed in 2002, the historical data for the years 1970-2001 can be a standalone dataset, capped at the year of the old methodology (with explanatory documentation). A new

	Changes in Methodology	dataset would display the data beginning with the effective date for the new methodology (e.g., beginning in 2002).
11	**JSON**	JavaScript Object Notation A simple format for data that can describe complex data structures, is both machine-readable and somewhat human-readable, is independent of platform and programming language, and has become a format for data exchange between apps, programs and computer systems.
12	**Machine Readable**	Machine readable is information formatted in a standard computer language that can be read automatically by a web application or computer system such as spreadsheets with header columns that can be exported as comma separated values (CSV).
13	**Metadata**	Metadata provides important structural and contextual information about the data; it describes characteristics and attributes of the data (e.g., who, what, where, why, how) . Metadata makes finding content and data faster and easier. Metadata facilitates data discovery and linkage across relevant and different data sources.
14	**Open Data**	Data is open if it can be freely accessed, used, modified and shared by anyone for any purpose. Open Data is data put in machine readable form that can be used, reused and redistributed freely.
15	**Visualizations**	A visual representation of data, such as a chart, graph or dashboard, is often the easiest way of communicating with data, bringing out its key features. Many visualization tools exist such as Google Charts, Excel, ArcGIS, Tableau, and PowerBI. Creating a dataset's visualisation requires careful attention to the meaning of the variables, the relations between them and the stories inherent in the data, to design a visual representation that lets the message of the data shine through.
16	**XML**	Extensible Markup Language, is a flexible file format designed to store, transport and share data over the Internet. XML is both human- and machine-readable.

2. Preparing Data for Publication in the Datastore

You must upload your dataset to the Portal’s datastore in order to publish it. This will enable the public to preview the dataset as well as access it for use in an external web application. The datastore bundles a number of modules and configurations with your CSV files, parses them and saves them into the native database as flat tables. Once properly uploaded to the datastore, your dataset will be available to access via a public API.

Very Important: To upload data to the Open Data Mauritius Portal all datasets **must be in a flat CSV file format.**

2.1 Dataset Basics – Create a Machine Readable Dataset

Machine Readable Information or data is in a format that can be easily processed by a computer - without human intervention. To be machine readable, data must be structured in an organized fashion. CSV, JSON and XML are formats that contain structured data that a computer can automatically read and process. Other materials such as photos and handwritten documents are not machine readable even when scanned. For example, a pdf document containing tables of data is digital but is not machine-readable because the tables are still simply images.

2.2 Data Wrangling Process

Data wrangling is the process of cleaning and unifying messy and complex **data** sets for easy access and analysis.

With the amount of data and data sources rapidly growing and expanding, it is getting more and more essential for the large amounts of available data to be organized for analysis.

This process typically includes manually converting/mapping **data** from one raw form into another to allow for more convenient and consumption of the data organization**.**

3. Converting a non-CSV file to CSV:

Checklists of Dataset Submission Guideline		
S/N	**Component**	**Description/Explanation**
1	Data File Format	Comma-Separated Value (CSV) or tab-separated value (TSV); UTF-8 character encoding is required; Data file must be saved in CSV
2	Record termination characters in CSV	Carriage return/line feed characters must exist one and only once at the end of each data record
3	Carriage return/line feed characters embedded in source fields	Must be removed or converted to some other separating character such as a space, comma, semi-colon, etc.

	such as multiline addresses or comments	
4	Use Vertical Rather than Horizontal Orientation	Horizontal data orientation should be restructured to vertical whenever feasible Vertical data orientation to be used for Datasets containing data by year, especially numerous years
5	Years	Years should have their own rows rather than columns in the data.
6	Header Row	Data should contain one and only one header row. Multi-row headers are not acceptable
7	Column Names	• Column names must be clear and in plain English, instead of the source system database naming conventions. • Do not use underscores in column names. • Avoid use of abbreviations, use title case for field names • Codes should not be used. However, if any codes absolutely must be used, they must be fully explained in the Metadata • Column names should be kept to less than 50 characters in length whenever practicable where shortening will not result in misinterpretation.
8	Empty Cells in a Group of Rows	A group of rows related to one entity should repeat the entity for all rows in the group.
9	Blank, “N/A” or unknown cells	• If the blank field represents zero, then the field should be zero. • N/A should be removed and kept as blank • If the blank field represents “not collected” or “unknown”, then this should be explained in the metadata or data dictionary. • Zero and N/A are not same
10	Texts - N/A or unknown	To be excluded in numeric field.

11	Subtotal or Total Rows, or Other Grouped Data	Avoid including subtotal and total rows unless absolutely necessary
12	Coded Fields	To be explained in the data dictionary document
13	Text Fields	Must be trimmed of leading or trailing whitespace
14	Numeric Fields	Do not mix text in a field that is intended to contain numeric data
15	Money	• Numeric data that represents money should be provided with either no decimal places or two decimal places; • Do not vary the number of decimal places used to format the values throughout the data – consistency is key. • Do not include currency symbols, or commas for place-separators. • Negative values should be preceded with a minus-sign, not placed within parentheses
16	Measures (Ratios, Quantities, Percentages)	Varying decimal places are acceptable. Do not include commas for place-separators. Negative values should be preceded with a minus-sign, not placed within parentheses.
17	Date Fields	Full dates must be provided in MM/DD/YYYY format Example: 09/02/2013 The importance of standardizing this format is that this is the only way to display trends over time. It is critical for conducting analyses, time series, and inform decision-making.

18	Address Data	Clean address data is very valuable as it can add another dimension to your data; addresses must be broken into four columns: street address, city, state and postal code
19	Converting Excel (a non-CSV file) Data to CSV	**Mandatory:** * Create headers for each column (first row/column names) *(see Example 1 and 3 below)* * Header values *(see Example 1 below)* * Values for fields *(see Example 1 below)* Excel files (xls, xlsx) will not be accepted since they can contain features that cause the import to fail such as merged cells, macros, data spanning tabs, and formulas.
20	Importing into Excel	Format the cells of the blank workbook in the 'Number' tab of the 'Format Cells' menu from the default value 'General' to 'Text' format
21	Blank Rows and Columns	It is critical to check for and remove any inadvertently created blank rows or columns. Care must be taken to ensure that any blank rows and columns have been removed prior to creating your CSV An easy way to determine whether such blank rows or columns are present in Excel is to press [Ctrl]+[End] inside the spreadsheet and see if this takes you beyond your data in the spreadsheet
22	Merged Cells	**Not Allowed:** Merged cells are not acceptable, and cannot be reproduced in a CSV

23	Empty Rows and Columns	**Not Allowed:** Empty rows and empty columns among the data is not acceptable; If blank rows are present the data should be cleansed to ensure that any blank rows and columns have been removed prior to creating your CSV
24	Calculated Fields	These data fields should be expanded to include each data component especially when the creation of visualizations will rely upon this data
25	Multiple Data Items in a cell	A cell may contain only one item of information; multiple lines within a cell will cause the import process to fail; Alternative: Data may be presented as being highly vertical
26	Commas, Backslashes and Quotation Marks	Commas indicate the separation between field values and quotation marks indicate where text values begin and end; To signal that a quotation mark is a part of the text value and not an indicator of the beginning or end of a text value, you must immediately precede the quotation mark with a quotation mark, and surround the text value with quote marks; The backslash is an escape character, which indicates that the next character has some special meaning (e.g. “\n” is not the letter “n”, but is the newline character).
27	Blank/Null Values	Every column must be accounted for in a CSV or TSV, regardless if the source value is blank or null for a particular row. That is, if a dataset consists of ten columns, every row in the dataset must contain ten columns. This is accomplished in a CSV or TSV file by including the separating commas or tab characters with nothing in between
28	Superscript	Eg. 1946^1 1 to be removed; explain in Metadata
29	Header/Footer	To be included in the Metadata

30	Commas in figures	To select in Excel and view as thousand separator
31	Total/Subtotal	To be removed
32	Headers	Multiple headers to be reviewed
33	Naming Convention	The name of Data files prefixed with DATA; Metadata files to be prefixed with METADATA
34	Merged columns	**Not Allowed:** To unmerge and rename
35	Thousands, millions values	Eg Value (millions) - to be kept unchanged
37	Thousand separator	"," to be removed
38	Decimal places	To be consistent
39	Decimal zero lost	Select, click on decrease decimal place and then increase by 1
40	Excel Files	Unfreeze Panes

3.1 Mandatory

1. Create headers for each column (first row/column names) *(see Example 1 and 3 below)*
2. Header values *(see Example 1 below)*
3. Values for fields *(see Example 1 below)*

3.2 Not allowed

1. Merged cells *(see Example 3 below)*
2. Multiple tables *(see Example 3 below)*
3. Notes/descriptions/footnotes (add this information to your metadata) *(see Example 3 below)*
4. Non-data elements *(see Example 3 below)*
5. Blank row or columns within the data *(see Example 1 and 3 below)*
6. Aggregate (sum of values) rows

3.3 Formatting best practices

1. Dates (i.e. years, or actual dates) should be stored as rows *(see Example 2 below)*

2. Each type of numeric field (i.e. percentages or totals) value should be its own row in a single column in numerical format *(see Example 1 and 3 below)*

3. Table should be tall and narrow vs short and wide, 30 columns max is suggested *(see Example 2 below)*

4. IDs and abbreviations should be changed to full names or values

5.Columns should be data fields, Rows should be where the values for individual entities are stored *(see Example 2 below)*

3.4 Examples

Example 1 demonstrates cleaning up a dataset from improper date format to machine-readable date format by first removing blank cells in rows and columns, then creating a date column, then replacing the old format with the new, single column with a header of "date" at the top of the column.

Not Acceptable: Empty Rows and Empty Columns

Company			Work Related Injury
ABC	2017 January		Skin Disorders
ABC	February		Respiratory Conditions
ABC	2018 January		Poisoning
XYZ	2017 January		Skin Disorders
Stat Job	February		Respiratory Conditions
XYZ	2018 January		Poisoning

Acceptable:

Company	Date	Work Related Injury
ABC	1/1/2017	Skin Disorders
ABC	1/2/2017	Respiratory Conditions
ABC	1/1/2018	Poisoning
XYZ	1/1/2017	Skin Disorders
XYZ	1/2/2017	Respiratory Conditions
XYZ	1/1/2018	Poisoning

- Remove Blank Row
- Remove Blank Column
- Add missing Title as ‘Date’
- Change Date format

Example 2 is comparing a long formatted dataset – Vertical data orientation to a wide formatted dataset – Horizontal data orientation. A long formatted dataset is machine-readable, a wide-formatted dataset is not.

Not Acceptable – horizontal data orientation

Company	Year	Skin Disorders	Respiratory Conditions	Poisoning	Other Illnesses
ABC	2009	0	1	0	2

Acceptable – vertical data orientation

Company	Year	Work Related Injury	Number of Cases
ABC	2009	Skin Disorders	0
ABC	2009	Respiratory Conditions	1
ABC	2009	Poisoning	0
ABC	2009	Other Illnesses	2

Example 3 shows a multi-table dataset with metadata being converted to several machine-readable tables. The data has also been converted from wide to long formatting, a date column is created for each table and headers are added to the top of the table. The description at the footnote of the table will be moved to a metadata document.

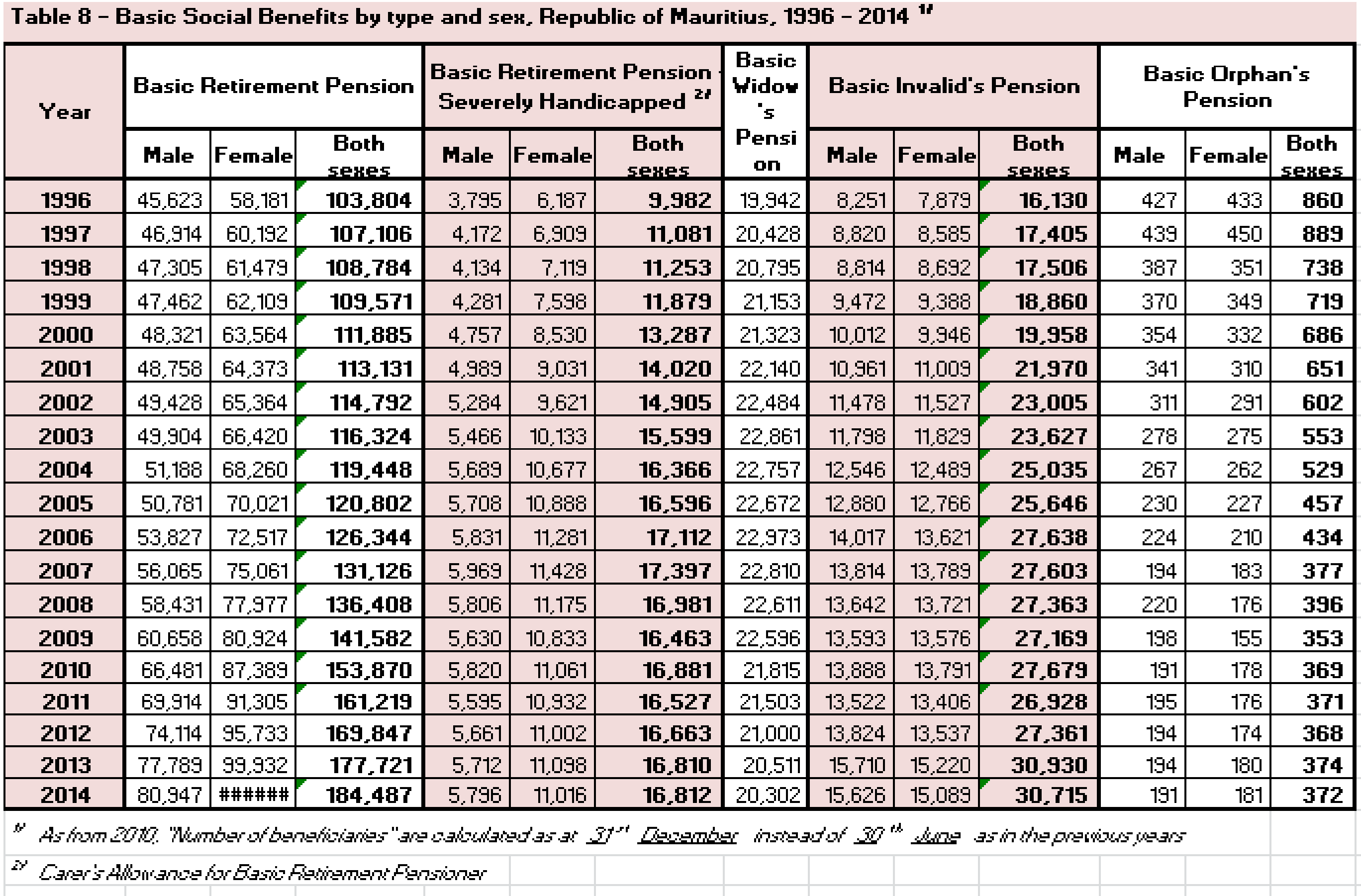


Table 8 – Basic Social Benefits by type and sex, Republic of Mauritius, 1996 – 2014 [1/]

Year	Basic Retirement Pension			Basic Retirement Pension - Severely Handicapped [2/]			Basic Widow's Pension	Basic Invalid's Pension			Basic Orphan's Pension		
	Male	Female	Both sexes	Male	Female	Both sexes		Male	Female	Both sexes	Male	Female	Both sexes
1996	45,623	58,181	**103,804**	3,795	6,187	**9,982**	19,942	8,251	7,879	**16,130**	427	433	**860**
1997	46,914	60,192	**107,106**	4,172	6,909	**11,081**	20,428	8,820	8,585	**17,405**	439	450	**889**
1998	47,305	61,479	**108,784**	4,134	7,119	**11,253**	20,795	8,814	8,692	**17,506**	387	351	**738**
1999	47,462	62,109	**109,571**	4,281	7,598	**11,879**	21,153	9,472	9,388	**18,860**	370	349	**719**
2000	48,321	63,564	**111,885**	4,757	8,530	**13,287**	21,323	10,012	9,946	**19,958**	354	332	**686**
2001	48,758	64,373	**113,131**	4,989	9,031	**14,020**	22,140	10,961	11,009	**21,970**	341	310	**651**
2002	49,428	65,364	**114,792**	5,284	9,621	**14,905**	22,484	11,478	11,527	**23,005**	311	291	**602**
2003	49,904	66,420	**116,324**	5,466	10,133	**15,599**	22,861	11,798	11,829	**23,627**	278	275	**553**
2004	51,188	68,260	**119,448**	5,689	10,677	**16,366**	22,757	12,546	12,489	**25,035**	267	262	**529**
2005	50,781	70,021	**120,802**	5,708	10,888	**16,596**	22,672	12,880	12,766	**25,646**	230	227	**457**
2006	53,827	72,517	**126,344**	5,831	11,281	**17,112**	22,973	14,017	13,621	**27,638**	224	210	**434**
2007	56,065	75,061	**131,126**	5,969	11,428	**17,397**	22,810	13,814	13,789	**27,603**	194	183	**377**
2008	58,431	77,977	**136,408**	5,806	11,175	**16,981**	22,611	13,642	13,721	**27,363**	220	176	**396**
2009	60,658	80,924	**141,582**	5,630	10,833	**16,463**	22,596	13,593	13,576	**27,169**	198	155	**353**
2010	66,481	87,389	**153,870**	5,820	11,061	**16,881**	21,815	13,888	13,791	**27,679**	191	178	**369**
2011	69,914	91,305	**161,219**	5,595	10,932	**16,527**	21,503	13,522	13,406	**26,928**	195	176	**371**
2012	74,114	95,733	**169,847**	5,661	11,002	**16,663**	21,000	13,824	13,537	**27,361**	194	174	**368**
2013	77,789	99,932	**177,721**	5,712	11,098	**16,810**	20,511	15,710	15,220	**30,930**	194	180	**374**
2014	80,947	######	**184,487**	5,796	11,016	**16,812**	20,302	15,626	15,089	**30,715**	191	181	**372**

[1/] *As from 2010, "Number of beneficiaries" are calculated as at 31st December instead of 30th June as in the previous years*

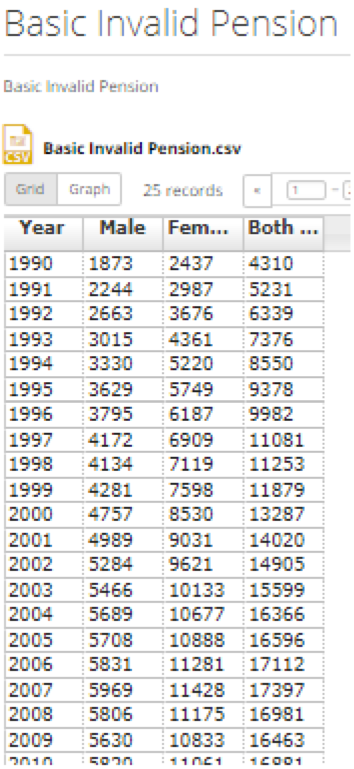
[2/] *Carer's Allowance for Basic Retirement Pensioner*

- From the table above, multiple CSV files have been created as below. Data is displayed in a vertical orientation with column 'Year' added for each CSV file.

- The footnote has been moved to another file, the Metadata file and saved as a txt file.

- The Superscript in the title has to be removed.

- Merged column (Year) has been unmerged.

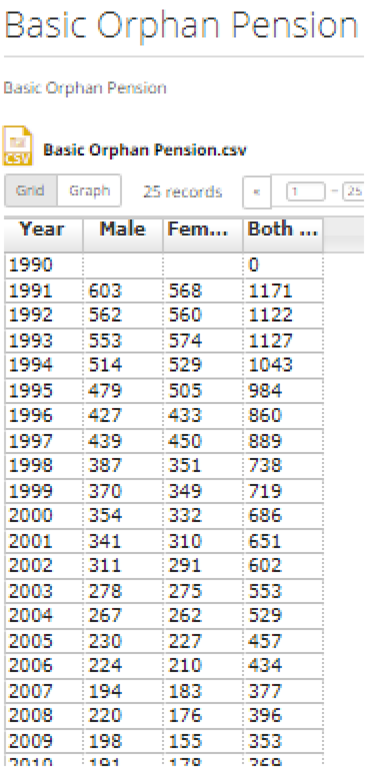


Basic Invalid Pension

Basic Invalid Pension

Basic Invalid Pension.csv

Year	Male	Fem...	Both ...
1990	1873	2437	4310
1991	2244	2987	5231
1992	2663	3676	6339
1993	3015	4361	7376
1994	3330	5220	8550
1995	3629	5749	9378
1996	3795	6187	9982
1997	4172	6909	11081
1998	4134	7119	11253
1999	4281	7598	11879
2000	4757	8530	13287
2001	4989	9031	14020
2002	5284	9621	14905
2003	5466	10133	15599
2004	5689	10677	16366
2005	5708	10888	16596
2006	5831	11281	17112
2007	5969	11428	17397
2008	5806	11175	16981
2009	5630	10833	16463
2010	5820	11061	16881

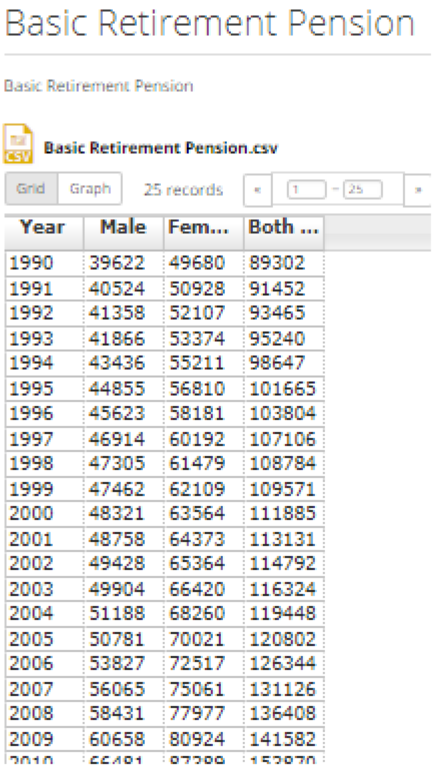


Basic Orphan Pension

Basic Orphan Pension

Basic Orphan Pension.csv

Year	Male	Fem...	Both ...
1990			0
1991	603	568	1171
1992	562	560	1122
1993	553	574	1127
1994	514	529	1043
1995	479	505	984
1996	427	433	860
1997	439	450	889
1998	387	351	738
1999	370	349	719
2000	354	332	686
2001	341	310	651
2002	311	291	602
2003	278	275	553
2004	267	262	529
2005	230	227	457
2006	224	210	434
2007	194	183	377
2008	220	176	396
2009	198	155	353
2010	191	178	369

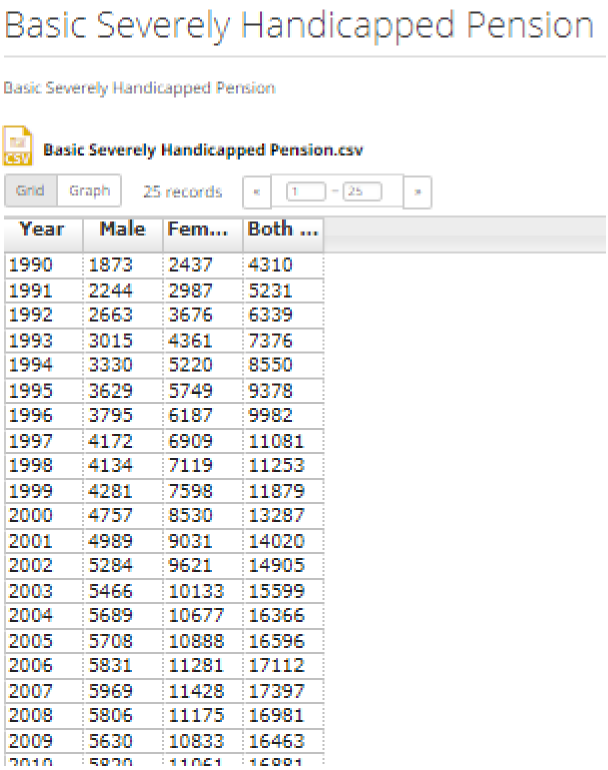


Basic Retirement Pension

Basic Retirement Pension

Basic Retirement Pension.csv

Year	Male	Fem...	Both ...
1990	39622	49680	89302
1991	40524	50928	91452
1992	41358	52107	93465
1993	41866	53374	95240
1994	43436	55211	98647
1995	44855	56810	101665
1996	45623	58181	103804
1997	46914	60192	107106
1998	47305	61479	108784
1999	47462	62109	109571
2000	48321	63564	111885
2001	48758	64373	113131
2002	49428	65364	114792
2003	49904	66420	116324
2004	51188	68260	119448
2005	50781	70021	120802
2006	53827	72517	126344
2007	56065	75061	131126
2008	58431	77977	136408
2009	60658	80924	141582
2010	66481	87389	153870

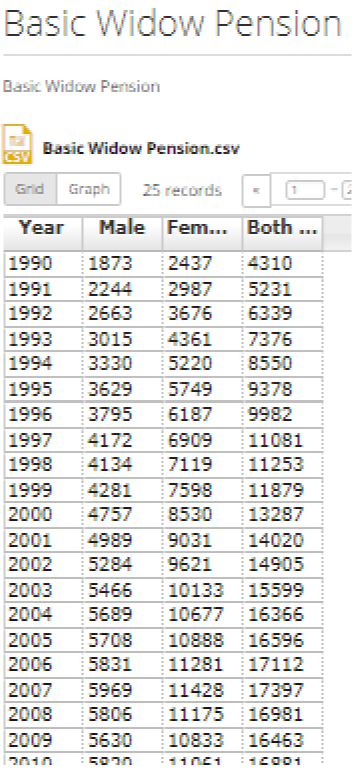


Basic Severely Handicapped Pension

Basic Severely Handicapped Pension

Basic Severely Handicapped Pension.csv

Year	Male	Fem...	Both ...
1990	1873	2437	4310
1991	2244	2987	5231
1992	2663	3676	6339
1993	3015	4361	7376
1994	3330	5220	8550
1995	3629	5749	9378
1996	3795	6187	9982
1997	4172	6909	11081
1998	4134	7119	11253
1999	4281	7598	11879
2000	4757	8530	13287
2001	4989	9031	14020
2002	5284	9621	14905
2003	5466	10133	15599
2004	5689	10677	16366
2005	5708	10888	16596
2006	5831	11281	17112
2007	5969	11428	17397
2008	5806	11175	16981
2009	5630	10833	16463
2010	5820	11061	16881



Basic Widow Pension

Basic Widow Pension

Basic Widow Pension.csv

Year	Male	Fem...	Both ...
1990	1873	2437	4310
1991	2244	2987	5231
1992	2663	3676	6339
1993	3015	4361	7376
1994	3330	5220	8550
1995	3629	5749	9378
1996	3795	6187	9982
1997	4172	6909	11081
1998	4134	7119	11253
1999	4281	7598	11879
2000	4757	8530	13287
2001	4989	9031	14020
2002	5284	9621	14905
2003	5466	10133	15599
2004	5689	10677	16366
2005	5708	10888	16596
2006	5831	11281	17112
2007	5969	11428	17397
2008	5806	11175	16981
2009	5630	10833	16463
2010	5820	11061	16881

4. Saving Your Dataset

4.1 CSV Basics

Values in flat datasets are separated by delimiters; therefore a “csv” comma separated file is not necessarily separated with commas.

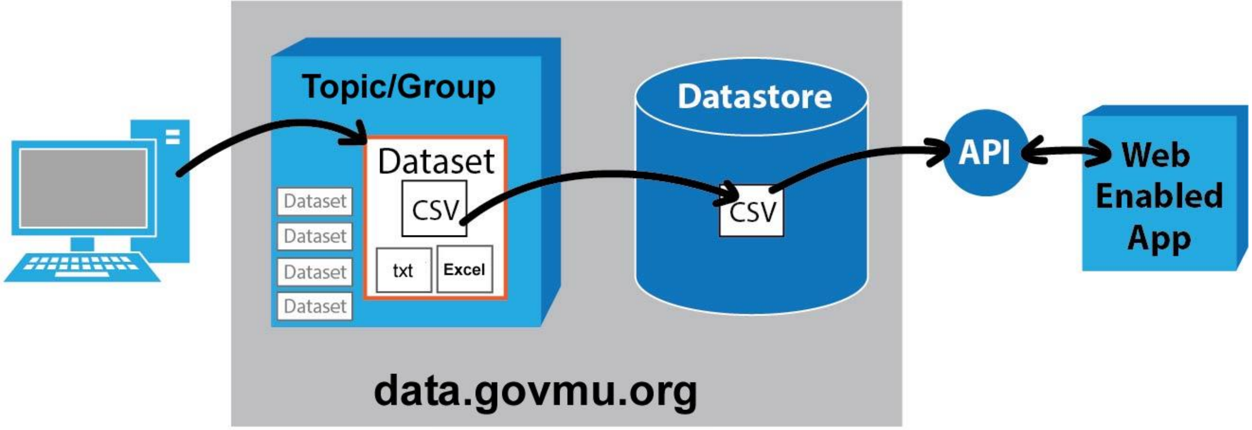
4.2 Encoding - UTF-8

Computers encode characters (i.e. “a”, “A”, “3”, “$”) in different formats. Any particular character can be encoded in many different ways, depending on which encoding is used to read or write them.

To ensure that people who download your dataset can properly understand the characters when they download it, we require that your file be encoded in UTF-8. This is the standard encoding for most systems and if you are unsure about the encoding of your file, check with the Central Open Data Team at the Ministry of Technology, Communication and Innovation.

5. What Happens with Your Machine Readable Dataset on data.govmu.org

Once your dataset is cleaned up (**the data wrangling process**), it can be uploaded to your Topic/Group on data.govmu.org. Part of your full dataset package – the resources - is the CSV that you just cleaned up, the metadata file as txt file, the source file as Excel file and could also include additional resources such as geojson files and more. The metadata is being created as txt file for the time being. (Please see the Data Publishing Process). Once these files are loaded, your CSV file is automatically pulled into the datastore, which now makes it available for web-enabled applications to access.



6. Next

The next step is creating the Metadata and data dictionary documents.

7. For more information

- DKAN Datastore documentation
- Email: opendata@govmu.org