

Instructions to update COVID-19 data on statistics.gov.scot and GitHub

Author: Victoria Avila (victoria.avila@gov.scot)

Last update: 30/11/2020

Contents

PRODUCING NEW FILES	1
UPLOADING NEW FILE TO statistics.gov.scot.....	4
QUALITY ASSURANCE FOR statistics.gov.scot	6
PROJECT SETUP	10
UPLOADING FILES ON TO GitHub	11
REPORTING ISSUES	12

PRODUCING NEW FILES

1. [Step 1 is only necessary if you want to use Git. You can skip it if you are going to upload files on to GitHub manually]

If you aren't working in a project cloned from GitHub, please follow the instructions in the [PROJECT SETUP](#) section before proceeding.

Open the project in RStudio, and get the latest version of the code. In the RStudio Terminal tab, run **git pull origin master** (this will update all files in the project to the latest version uploaded to GitHub)

```
dsap01@OS06 ~/Documents/COVID-19-Management-Information (master)
$ git pull origin master
From https://github.com/DataScienceScotland/COVID-19-Management-Information
* branch      master      -> FETCH_HEAD
Already up to date.
```

2. Go to the SG webpage
<https://www.gov.scot/publications/trends-in-number-of-people-in-hospital-with-confirmed-or-suspected-covid-19/>
3. Copy URLs for Scotland and Health Boards data files.
4. Compare them to yesterday's URLs in the [R code](#). Some days they change it, some days they don't. If they change, update lines 23 and 24 with the new URLs and line 7 with the date for the last update.

It has been agreed with the web team that the URLs will be as follow modifying the bits in black accordingly to the date:

Scotland:

<https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2020/04/trends-in-number-of-people-in-hospital-with-confirmed-or-suspected-covid-19/documents/trends-in-number-of-people-in-hospital-with-confirmed-or-suspected-covid-19/trends-in-number-of-people-in-hospital-with-confirmed-or-suspected-covid-19/govscot%3Adocument/Trends%2Bin%2Bdaily%2BCOVID-19%2Bdata%2B9%2BApril%2B2020.xlsx>

Health Boards:

<https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2020/04/trends-in-number-of-people-in-hospital-with-confirmed-or-suspected-covid-19/documents/covid-19-data-by-nhs-board/covid-19-data-by-nhs-board/govscot%3Adocument/COVID-19%2Bdata%2Bby%2BNHS%2BBoard%2B9%2BApril%2B2020.xlsx>

5. Run the R code with the updated links.

[If you double click on the code file, RStudio will open with the folder the code file is as the working directory]

6. All the output files should have been created in the folder where the code file is, run git status in the Terminal to check:

```
@sap01@DS06 ~/Documents/COVID-19-Management-Information (master)
$ git status
On branch master
Your branch is up to date with 'origin/master'.

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working directory)

    modified:   COVID19 - Daily Management Information - Data reformattting.r
    modified:   COVID19 - Daily Management Information - Scotland - Ambulance.csv
    modified:   COVID19 - Daily Management Information - Scotland - Calls.csv
    modified:   COVID19 - Daily Management Information - Scotland - Care home workforce.csv
    modified:   COVID19 - Daily Management Information - Scotland - Care homes.csv
    modified:   COVID19 - Daily Management Information - Scotland - Deathly.csv
    modified:   COVID19 - Daily Management Information - Scotland - Delayed discharges.csv
    modified:   COVID19 - Daily Management Information - Scotland - Hospital care.csv
    modified:   COVID19 - Daily Management Information - Scotland - Testing.csv
    modified:   COVID19 - Daily Management Information - Scotland - Workforce.csv
    modified:   COVID19 - Daily Management Information - Scottish Health Boards - Cumulative cases.csv
    modified:   COVID19 - Daily Management Information - Scottish Health Boards - Hospital patients - Confirmed.csv
    modified:   COVID19 - Daily Management Information - Scottish Health Boards - Hospital patients - Suspected.csv
    modified:   COVID19 - Daily Management Information - Scottish Health Boards - ICD patients.csv
    modified:   COVID19 - Daily Management Information - Tidy dataset to upload to statistics.gov.scot.csv
    modified:   COVID19 - Daily Management Information - Tidy dataset to upload to statistics.gov.scot_9999999.csv

Untracked files:
  (use "git add <file>..." to include in what will be committed)

    COVID-19-Management-Information.Rproj

no changes added to commit (use "git add" and/or "git commit -a")
```

7. Run **git add .** to stage the changes:

```
@sap01@DS06 ~/Documents/COVID-19-Management-Information (master)
$ git add .
```

8. Run **git commit -m 'type your own commit message here'** to commit the changes:

```
dsap01@DS06 ~/Documents/COVID-19-Management-Information (master)
$ git commit -m 'add 04/06/2020 files via upload'
[master af520f2] add 04/06/2020 files via upload
17 files changed, 226 insertions(+), 12 deletions(-)
create mode 100644 COVID-19-Management-Information.Rproj
```

The commit message will appear in the git history, and will show next to the files when viewing on GitHub.

9. Run **git push origin master** to push the files from your local repository to GitHub:

```
dsap01@DS06 ~/Documents/COVID-19-Management-Information (master)
$ git push origin master
Counting objects: 19, done.
Delta compression using up to 4 threads.
Compressing objects: 100% (19/19), done.
Writing objects: 100% (19/19), 7.84 KiB | 573.00 KiB/s, done.
Total 19 (delta 17), reused 0 (delta 0)
remote: Resolving deltas: 100% (17/17), completed with 17 local objects.
To https://github.com/DataScienceScotland/COVID-19-Management-Information
2c88ea0..af520f2 master -> master
```

The changes should then be visible on the remote repository:

<https://github.com/DataScienceScotland/COVID-19-Management-Information>

UPLOADING NEW FILE TO statistics.gov.scot

10. Open statistics.gov.scot admin site and log in:

<https://pmd3-production-admin-sg.publishmydata.com/admin>

11. Go to the Covid-19 – Management Information dataset:

<https://pmd3-production-admin-sg.publishmydata.com/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fcoronavirus-covid-19-management-information>

12. Click on 'EDIT' tab and then on 'CLEAR DATASET CONTENTS'

Coronavirus - COVID-19 - Management Information
The data provide past data around COVID-19 for the daily updates provided by the Scottish Government.

DATA ABOUT API **EDIT** PIPELINES

EDIT DATASET METADATA

EDIT

EDIT DATASET RDF CONTENTS

These actions allow you to edit the RDF contents of the dataset

ADD RDF DATA TO DATASET

CLEAR DATASET CONTENTS

Once you make any changes to the dataset, the system will automatically create a new draft called 'Untitled'. You can rename it if you like, but since this is the only one you will have in your accounts and it's going to be used for 30 minutes, you can leave it unchanged.

13. Click on the 'PIPELINES' tab and then on the first type of pipeline – Spreadsheet to datacube

Coronavirus - COVID-19 - Management Information
The data provide past data around COVID-19 for the daily updates provided by the Scottish Government.

DATA ABOUT API EDIT **PIPELINES**

LOAD DATA INTO THIS DATASET VIA A PIPELINE

The following pipelines target exactly one dataset. Other pipelines may be available for this site.

Note that not all pipelines shown may produce data available for this dataset.

Spreadsheet to datacube
Given a spreadsheet in normalised format, produce RDF datacube observations

RUN PIPELINE

14. Select the 'Coronavirus – COVID-19 – Management Information' dataset as the target dataset and the 'COVID19 - Daily Management Information - Tidy dataset to upload to statistics.gov.scot.csv' as the input data. Run the pipeline.

Pipeline: Spreadsheet to datacube

Given a spreadsheet in normalised format, produce RDF datacube observations

TARGET DATASET

Target dataset for the pipeline

Coronavirus - COVID-19 - Management Information

Or create a new dataset.

SPREADSHEET

Spreadsheet to use as input data

Browse...

RUN

RUN PIPELINE

15. For daily updates you should only see a list with two changes:
- Reference time intervals – as you have added a new date
 - the dataset itself

Pipeline succeeded

SUMMARY OF CURRENT CHANGES IN UNTITLED

After running this pipeline, this draft set has the following changes:

- Updated Reference time intervals ontology
- Updated Coronavirus - COVID-19 - Management Information dataset

QUALITY ASSURANCE FOR statistics.gov.scot

16. Click on the dataset and go to the 'API' tab. Check the number of observations under the 'DATA LINKED RESOURCES' section.
17. Download the whole dataset as 'CSV' and compare the number of observations. Numbers should match.
18. Select a slice of the dataset and check it downloads fine.
19. Go to 'TOOLS/SPARQL Query' and run the following SPARQL queries. Make sure you check 'Validates URIs'.

Results format

☒ Validate URIs

All the queries should give no results and all the URIs should come up in green.

QUERY RESULTS

Your query ran successfully but returned no results.

URI VALIDATION

The following URI literals were found in your query (after prefix expansion):

- <http://statistics.gov.scot/glossary/terms-and-conditions-management-information>
- <http://owl1.org/linked-data/observedObservation>
- <http://owl1.org/linked-data/observedObservedObservation>
- <http://owl1.org/linked-data/observedObservedObservedObservation>

1. Identifies any areas not in Atlas

```
# -----
PREFIX qb: <http://purl.org/linked-data/cube#>

select distinct ?area where {graph <http://statistics.gov.scot/graph/coronavirus-covid-19-
management-information> {
?obs a qb:Observation ;
<http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area .
}
OPTIONAL {?area <http://publishmydata.com/def/ontology/foi/memberOf> ?collection .}
FILTER (!bound(?collection))
}
```

2. Identifies any archived geographies

```
# -----
PREFIX qb: <http://purl.org/linked-data/cube#>

select distinct ?area where {graph <http://statistics.gov.scot/graph/coronavirus-covid-19-
management-information> {
?obs a qb:Observation ;
<http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area .
}
?area <http://statistics.data.gov.uk/def/statistical-geography#status> "Archived"
}
```

3. Identifies any observations with multiple values - count

```
# -----
PREFIX qb: <http://purl.org/linked-data/cube#>

SELECT ?DataSet ?s (count(?s) as ?NumValues)
WHERE {
?s <http://statistics.gov.scot/def/measure-properties/count> ?o.
?s qb:dataSet <http://statistics.gov.scot/data/coronavirus-covid-19-management-
information>.
}
GROUP BY ?DataSet ?s
HAVING (?NumValues>1)
```

4. Identifies any observations with multiple values - ratio

```
# -----
PREFIX qb: <http://purl.org/linked-data/cube#>

SELECT ?DataSet ?s (count(?s) as ?NumValues)
WHERE {
?s <http://statistics.gov.scot/def/measure-properties/ratio> ?o.
?s qb:dataSet <http://statistics.gov.scot/data/coronavirus-covid-19-management-
information>.
}
GROUP BY ?DataSet ?s
HAVING (?NumValues>1)
```

5. Identifies multiple labels for units

```
# -----
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?MeasureUnits (count(?MeasureUnits) as ?NumLabels)
WHERE {
?MeasureUnits a <http://purl.org/linked-data/sdmx/2009/concept#unitMeasure>.
?MeasureUnits rdfs:label ?label .
}
GROUP BY ?MeasureUnits
HAVING (?NumLabels>1)
```

6. Identifies multiple dimension values

```
# -----  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
  
SELECT ?DimensionValue (count(?DimensionValue) as ?NumLabels)  
WHERE {  
  ?DimensionValue a <http://www.w3.org/2004/02/skos/core#Concept>.  
  ?DimensionValue rdfs:label ?label .  
}  
GROUP BY ?DimensionValue  
HAVING (?NumLabels>1)
```

7. Identifies duplicate concept schemes

```
# -----  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
  
SELECT ?ConceptScheme (count(?ConceptScheme) as ?NumLabels)  
WHERE {  
  ?ConceptScheme a <http://www.w3.org/2004/02/skos/core#ConceptScheme>.  
  ?ConceptScheme rdfs:label ?label .  
}  
GROUP BY ?ConceptScheme  
HAVING (?NumLabels>1)
```

8. Identifies duplicate values in dataset

```
# -----  
PREFIX qb: <http://purl.org/linked-data/cube#>  
  
SELECT ?DataSet ?s (count(?s) as ?NumValues)  
WHERE {  
  ?s <http://statistics.gov.scot/def/measure-properties/index> ?o.  
  ?s qb:dataSet ?DataSet.  
}  
GROUP BY ?DataSet ?s  
HAVING (?NumValues>1)  
LIMIT 100
```

9. Identifies any datasets which have dropped dimensions:

```
# -----  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
  
SELECT *  
WHERE {  
  ?s <http://purl.org/linked-data/cube#dimension> ?x.  
  FILTER( !EXISTS { ?x rdfs:label ?y. } )  
}
```

10. Checks for missing Units

```
# -----  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
  
SELECT distinct ?unit ?unit_label  
WHERE {  
  ?s <http://purl.org/linked-data/sdmx/2009/attribute#unitMeasure> ?unit .  
  OPTIONAL {?unit rdfs:label ?unit_label }  
  FILTER(!bound(?unit_label))  
}
```



20. Publish the draft by clicking on Publish at the top of the window.

Current Draft: test

Submit Draft...

Publish

victoria.avila@gov.scot

 Scottish Government
Rìghdoras na h-Alba
gov.scot

STATISTICS.GOV.SCOT

ATLAS

DATA

SEARCH

DATA CART

HELP

pmd3-production-admin-sg.publishmydata.com says

This will publish your draft to the live site. Check the name of the draft and the list of datasets and vocabularies that will be changed.

Are you sure?

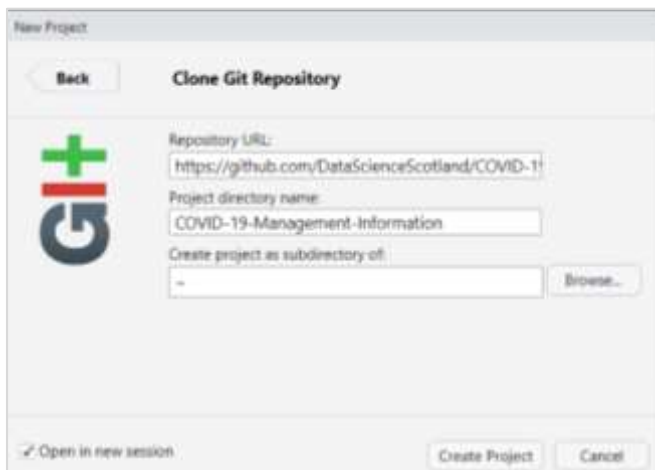
OK

Cancel

PROJECT SETUP on GitHub

If you have an existing 'COVID-19-Management-Information' project, it would probably be best to delete it before continuing, to avoid the confusion of having to projects of the same name.

1. Open RStudio.
2. From the top menu, select File > New Project > Version Control > Git
3. In the pop-up window, copy 'https://github.com/DataScienceScotland/COVID-19-Management-Information' into the Repository URL field.
4. Project directory name will auto-populate.
5. Browse to the directory you want to create the project in for 'Create project as a subdirectory of'
6. Tick open in a new session
7. Click create project



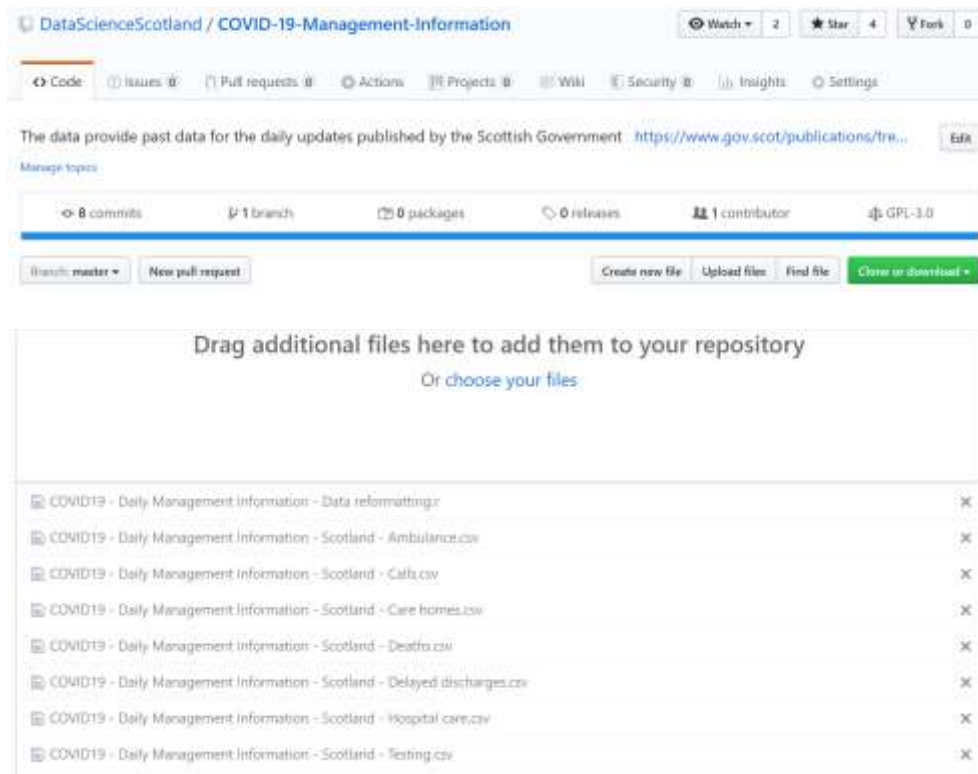
[At this point, there might be a problem with GitHub credentials. As I've already set mine up, I can't recreate that part just now, but if anyone runs into any problems with that I'm happy to help. If someone could record the steps it involves I can add them to this document too. - Tom]

UPLOADING FILES ON TO GitHub

1. Go to the GitHub folder:

<https://github.com/DataScienceScotland/COVID-19-Management-Information>

2. Click on 'Upload files' and select all the files to upload.



3. Scroll down to the bottom of the page and click on 'Commit changes'.

REPORTING ISSUES

- With the data published on **gov.scot**

Get in touch with Catriona Hayes, Neil Grant and the web team.

Catriona.Hayes@gov.scot; Neil.Grant@gov.scot; WEBSITE@gov.scot

- With uploading data on to **statistics.gov.scot**

Get in touch with Bill Roberts.

support@swirrl.com

- With **GitHub**

Get in touch with Joseph Adams.

Joseph.adams@nrscotland.gov.uk