Retrieving multi-sheet XLS/XLSX resources

Excel files (XLS and XLSX) are a common form of data on the City of Toronto Open Data Portal.

In cases where the file only contains one sheet, the resource is returned as a tibble. For example, this data set on TTC Ridership Analysis from 1985 to 2018:

```
library(opendatatoronto)

list_package_resources("https://open.toronto.ca/dataset/ttc-ridership-analysis/") %>%
  get_resource()
#> # A tibble: 71 x 37
#>    TORONTO TRANSIT COMMISS~1 ...2  ...3  ...4  ...5  ...6  ...7  ...8
#>    <chr>                     <chr> <chr> <chr> <chr> <chr> <chr> <chr>
#>  1 ANALYSIS OF RIDERSHIP     <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
#>  2 1985 TO 2019 ACTUALS (00~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
#>  3 <NA>                      <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
#>  4 <NA>                      FARE~ 2019  2018  2017  2016  2015~ 2014
#>  5 WHO                       ADULT <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
#>  6 <NA>                      PRES~ 1906  N/A   N/A   N/A   N/A   N/A
#>  7 <NA>                      PRES~ 829   N/A   N/A   N/A   N/A   N/A
#>  8 <NA>                      PRES~ 1668~ 1109~ 67829 27397 13323 9862
#>  9 <NA>                      PRES~ 4340  1496  N/A   N/A   N/A   N/A
#> 10 <NA>                      PRES~ 9937  4442  N/A   N/A   N/A   N/A
#> # i 61 more rows
#> # i abbreviated name: 1: `TORONTO TRANSIT COMMISSION`
#> # i 29 more variables: ...9 <chr>, ...10 <chr>, ...11 <chr>,
#> #   ...12 <chr>, ...13 <chr>, ...14 <chr>, ...15 <chr>, ...16 <chr>,
#> #   ...17 <chr>, ...18 <chr>, ...19 <chr>, ...20 <chr>, ...21 <chr>,
#> #   ...22 <chr>, ...23 <chr>, ...24 <chr>, ...25 <chr>, ...26 <chr>,
#> #   ...27 <chr>, ...28 <chr>, ...29 <chr>, ...30 <chr>, ...
```

When the file contains multiple sheets, the resource is returned as a named list, where the names are the names of the sheets, as in the dataset on Wellbeing Toronto Demographics:

```
library(dplyr)

wellbeing_toronto_demographics <- list_package_resources("https://open.toronto.ca/dataset/wellbeing-toro
  filter(name == "wellbeing-toronto-demographics") %>%
  get_resource()

str(wellbeing_toronto_demographics, max.level = 1)
#> List of 3
#>  $ IndicatorMetaData      : tibble [87 x 8] (S3: tbl_df/tbl/data.frame)
#>  $ RawData-Ref Period 2008: tibble [141 x 85] (S3: tbl_df/tbl/data.frame)
#>  $ RawData-Ref Period 2011: tibble [141 x 39] (S3: tbl_df/tbl/data.frame)
```

To access the relevant sheet, pull out the list element:

```
wellbeing_toronto_demographics[["IndicatorMetaData"]]
#> # A tibble: 87 x 8
#>    PROVENANCE               SHORT_NAME     LONG_NAME DESCRIPTION URL
#>    <chr>                    <chr>          <chr>     <chr>       <chr>
#>  1 Statistics Canada Census Total Populat~ Total Po~ For Refere~ http~
#>  2 Statistics Canada Census Pop - Males    Total Po~ For Refere~ http~
#>  3 Statistics Canada Census Pop - Females  Total Po~ For Refere~ http~
#>  4 Statistics Canada Census Pop 0 - 4 yea~ Total Po~ For Refere~ http~
#>  5 Statistics Canada Census Pop 5 - 9 yea~ Total Po~ For Refere~ http~
#>  6 Statistics Canada Census Pop 6-12 years Total Po~ For Refere~ http~
#>  7 Statistics Canada Census Pop 10 - 14 y~ Total Po~ For Refere~ http~
#>  8 Statistics Canada Census Pop 15 -19 ye~ Total Po~ For Refere~ http~
#>  9 Statistics Canada Census Pop 20 - 24 y~ Total Po~ For Refere~ http~
#> 10 Statistics Canada Census Pop  25 - 29 ~ Total Po~ For Refere~ http~
#> # i 77 more rows
#> # i 3 more variables: CURRENCY <dttm>, DATE_UPDATED <dttm>,
#> #   DOMAIN <chr>
```

There are also cases where the file contains multiple sheets and it would be helpful to have them all together as a single data set. For example, the 2019 TTC Bus Delay Data:

```
ttc_bus_delays_2019 <- search_packages("TTC Bus Delay Data") %>%
  list_package_resources() %>%
  filter(name == "ttc-bus-delay-data-2019") %>%
  get_resource()
```

The result of is a list with an element for every month of data, each of which is a tibble:

```
str(ttc_bus_delays_2019, max.level = 1)
#> List of 12
#>  $ Jan 2019  : tibble [6,743 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Feb 2019  : tibble [6,958 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Mar 2019  : tibble [5,712 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Apr 2019  : tibble [5,144 x 11] (S3: tbl_df/tbl/data.frame)
#>  $ May 2019  : tibble [5,023 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ June 2019 : tibble [5,232 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ July 2019 : tibble [5,113 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Aug, 2019 : tibble [4,354 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Sept 2019 : tibble [3,894 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Oct 2019  : tibble [4,283 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Nov 2019  : tibble [5,436 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Dec 2019  : tibble [4,484 x 10] (S3: tbl_df/tbl/data.frame)
```

Note that the data for for the element `Apr 2019` has one more variable than the rest (11 versus 10):

```
sapply(ttc_bus_delays_2019, colnames)
#> $`Jan 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#>
#> $`Feb 2019 `
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Mar 2019 `
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Apr 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident ID" "Incident"    "Delay"
#>  [9] "Gap"         "Direction"   "Vehicle"
#> 
#> $`May 2019 `
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`June 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Delay"       "Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`July 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Aug, 2019 `
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Sept 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Oct 2019 `
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Min Delay"   "Min Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Nov 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Delay"       "Gap"
#>  [9] "Direction"   "Vehicle"
#> 
#> $`Dec 2019`
#>  [1] "Report Date" "Route"       "Time"        "Day"
#>  [5] "Location"    "Incident"    "Delay"       "Gap"
#>  [9] "Direction"   "Vehicle"
```

It seems that the `Apr 2019` data has gained a variable `Incident ID`, and that the variables `Min Gap` and `Min Delay`, present in all the other months, have been renamed to `Gap` and `Delay`, respectively.

We can rename these two variables:

```r
ttc_bus_delays_2019[["Apr 2019"]] <- ttc_bus_delays_2019[["Apr 2019"]] %>%
  rename(`Min Gap` = Gap, `Min Delay` = Delay)
```

and combine all of the elements into a single tibble using `dplyr::bind_rows()`:

```r
ttc_bus_delays_2019_combined <- bind_rows(ttc_bus_delays_2019)

ttc_bus_delays_2019_combined
#> # A tibble: 62,376 x 13
#>    `Report Date`       Route Time                Day     Location
#>    <dttm>              <dbl> <dttm>              <chr>   <chr>
#>  1 2019-01-01 00:00:00    39 1899-12-31 00:13:00 Tuesday NECR
#>  2 2019-01-01 00:00:00   111 1899-12-31 00:15:00 Tuesday Eglington
#>  3 2019-01-01 00:00:00    35 1899-12-31 00:18:00 Tuesday Finch
#>  4 2019-01-01 00:00:00    25 1899-12-31 00:30:00 Tuesday Don Mills Rd~
#>  5 2019-01-01 00:00:00    36 1899-12-31 00:40:00 Tuesday Humberwood
#>  6 2019-01-01 00:00:00    45 1899-12-31 00:51:00 Tuesday Kipling stn
#>  7 2019-01-01 00:00:00    32 1899-12-31 01:55:00 Tuesday Royal York a~
#>  8 2019-01-01 00:00:00    53 1899-12-31 02:19:00 Tuesday FSTN
#>  9 2019-01-01 00:00:00   112 1899-12-31 02:33:00 Tuesday Kipling Stat~
#> 10 2019-01-01 00:00:00    85 1899-12-31 02:57:00 Tuesday DONS
#> # i 62,366 more rows
#> # i 8 more variables: Incident <chr>, `Min Delay` <dbl>,
#> #   `Min Gap` <dbl>, Direction <chr>, Vehicle <dbl>,
#> #   `Incident ID` <dbl>, Delay <dbl>, Gap <dbl>
```

Unfortunately, it looks like the `Time` variable got Excel™ed, and will need some data cleaning.

For interests sake, it appears that `Incident ID` is a lookup ID for the type of incident – only present in the `Apr 2019` data, but interesting nonetheless!

```r
ttc_bus_delays_2019_combined %>%
  filter(!is.na(`Incident ID`)) %>%
  distinct(`Incident ID`, Incident)
#> # A tibble: 11 x 2
#>    `Incident ID` Incident
#>            <dbl> <chr>
#>  1             5 Investigation
#>  2             1 Mechanical
#>  3             4 Utilized Off Route
#>  4             3 Diversion
#>  5             9 <NA>
#>  6             8 General Delay
#>  7             6 Emergency Services
#>  8            10 Late Leaving Garage - Operator
#>  9            11 Late Leaving Garage - Mechanical
#> 10             7 Vision
#> 11            12 Late Leaving Garage - Management
```