# Access to official statistics from R: an overview

Olav ten Bosch, Statistics Netherlands, o.tenbosch@cbs.nl
Edwin de Jonge, Statistics Netherlands, e.dejonge@cbs.nl

Providing access to output data is an essential task for Statistical Institutes. This is reflected in the awesome list of official statistics software [1] and the CRAN Task View: Official Statistics & Survey Statistics [2]. The awesome list was created in 2017 and grew over time [3] with many contributions from the official statistical community, partly from uRos workshops. Currently the vast majority of packages on the list is R software. The list is organised according to the Generic Statistical Business Process Model (GSBPM) [4], see Figure 1. The largest category on the list is "Access to official statistics" which contains over 30 software packages that help users access official statistics data or metadata from International or National organisations. In this presentation we take a closer look at the R-packages on this list to describe the current state of access to official statistics from R and we suggest potential improvements to this software landscape.

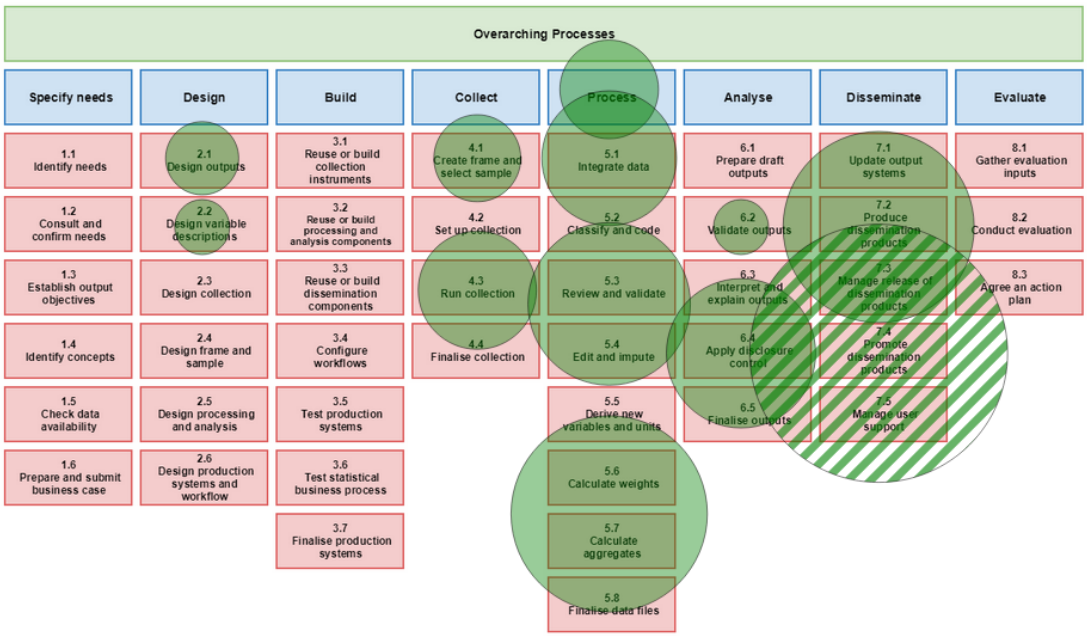


Figure 1: Awesome packages organised by GSBPM with largest category 7.4 "access to official statistics".

There are 28 R-packages in this category. Some are more generic and targeted at standardised access to multiple data providers, others contain detailed and dedicated functionality to access just one national or international organisation. Figure 2 provides an overview. It shows the dependencies between the packages on the list, the statistical data providers and the standards being used. These relationships were derived from the R-packages documentation, the pages they link to and running some of the packages, such as *rdsmx* which offers a list of pre-configured data-providers. Only main data providers on (inter)national level were considered. We note that this is an abstraction of reality as for example different versions of standards and endpoints are not taken into consideration, which would complicate the figure considerably.

From this network we can identify standards-based packages such as *rsdmx*, *readsdmx* using the SDMX standard [5], *rjstat* offering access to the JSON-STAT data format [6] and various *px** packages targeted at the PC-Axis format [7]. The ODATA [8] standard is used by one organisation: statistics Netherlands. Others packages such as *inegiR*, *readabs* and *statcanR* provide dedicated access to just one specific data provider. Note that it could be that a package does use a standard internally but this is not mentioned in the documentation as the package writer tries to hide these details from the user. In that case the use of standards is underestimated here. This work in progress can be followed here: [https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r](https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r)


Figure 2: R-packages for data access, statistical data providers and standards.

Looking from the data provider side we can see that certain data providers can be accessed via multiple R packages. Eurostat (ESTAT) and supporting SDMX as well as JSON-STAT and the existence of two dedicated packages *restatapi* and *eurostat* is a clear example. For the rest, the set of providers offering JSON-STAT data is mostly disjoint from the set of SDMX providers. All PX providers provide JSON-STAT as well. Some organisations, such as Eurostat and the World Bank, provide multiple endpoints for specific domains. Some endpoints provide harmonised data on one specific domain via an dedicated R-package. Examples are *rdbnomics* offering access to economic data from many institutions and *ipumsr* providing access to census and survey data integrated across time and space. Although a special category it is useful to note the existence such *official statistics aggregator sites* and their dedicated R-packages in the official statistics open data landscape.

The list R-packages on the awesome list also provide us some insight into the functionality that is usually offered. We can see certain features reoccurring, such as:

- *endpoint hiding*: wrapping the preconfigured endpoint(s) in a R function within the package
- *catalogue retrieval*: the ability to list the availability datasets on the endpoint(s)
- *search*: the ability to search for datasets or within datasets on the endpoint(s)
- *endpoint queries*: the ability to query for subsets on the endpoint(s) side
- *local queries*: the ability to easily slice or filter the retrieved data on the client
- *caching*: preventing unnecessary roundtrips to the endpoint(s) by caching results
- *cartographic queries*: retrieve a (cartographic) map to be used with the data

A category, not covered so far, is access to *statistical metadata*. Many organisations, mostly international, offer access to definitions, classifications and code lists in metadata registries. These predominantly use SDMX. Access from R to SDMX metadata has proven to be useful for statistical operations, such as checking data against internationally harmonised code list in the validation process [9].

Some organisations offer metadata in the form of *linked data*. Examples are Eurostat and Statistics Netherlands [10]. Linked data has the promise to make it easier to link and re-use statistical content with other open data sources, aligning to the FAIR principles [11]. Linked data can be accessed from R via generic software, such as *rdflib*, *jsonld,* or more experimental packages such as *glitter* if the endpoint provides for a queryable linked data interface in SPARQL. All in all, we expect that the growing use of linked (meta)data in official statistics will positively influence the official statistics open data landscape in the near future.

From the above, we see that the official statistics open data landscape grows towards standardisation, but also that in many cases there is a need to develop dedicated software targeted at specific functionality or specific data providers. At this point the R user can choose from at least 28 packages to access official statistics, each offering different functionality. There is no ‘one-for-all’ R-package that provides access to all official statistics data providers. This is understandable from an organisational viewpoint, but from an end-users viewpoint a generic package would be convenient. The situation will improve with ongoing standardisation on the data providers side, however as an uRos community it might be useful to also work towards creating a generic interface to all official statistics data providers, notwithstanding the value of the dedicated packages. The analysis presented in this paper could serve as a start.

**References**

[1] Awesome list of official statistics software, https://github.com/SNStatComp/awesome-official-statistics-software
[2] CRAN Task View: Official Statistics & Survey Statistics, https://cran.r-project.org/web/views/OfficialStatistics.html
[3] Olav ten Bosch, Mark van der Loo, Alexander Kowarik, The awesome list of official statistical software: 100 ... and counting, The Use of R in Official Statistics - uRos202 (virtual)
[4] Generic Statistical Process Business Model GSBPM, version 5.1, UNECE, https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1
[5] Statistical Data and Metadata eXchange (SDMX), https://sdmx.org
[6] JSON-stat: A simple lightweight standard for data dissemination, https://json-stat.org/
[7] PC-Axis software family, https://www.scb.se/en/services/statistical-programs-for-px-files/
[8] OData (Open Data Protocol), https://www.odata.org/
[9] Olav ten Bosch, Mark van der Loo, (2021), Validation in R Using Metadata from SDMX Registries, 8th SDMX Global Conference, INEGI, Mexico (virtual)
[10] ten Bosch O, de Jonge E, Laloli H, Laaboudi-Spoiden C (2022) FAIR Digital Objects in Official Statistics. Research Ideas and Outcomes 8: e94485. https://doi.org/10.3897/rio.8.e94485
[11] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18