Motivation: Make Tidy Datasets Easier to Release Exchange and Reuse

Version 0.1.6 https::/doi.org/10.5281/zenodo.6969653

Daniel Antal

The aim of the dataset package is to make tidy datasets easier to release, exchange and reuse. It organizes and formats data frame R objects into well-referenced, well-described, interoperable datasets into release and reuse ready form. We apply a subjective, R language compatible interpretation of the W3C Data Cube Vocabulary based on the statistical SDMX data cube model[1]. We enrich these statistical data exchange definition with similar definitions from information science and the exchange of scientific resources on open science repositories by applying the Dublin Core and DataCite metadata specifications for R attributes to improve the findability, accessibility, interoperability and reusability of the datasets.

We organize and format standard tabular data R objects (data.fame, data.table, tibble, or even well-structured lists like json and json-ld) become highly interoperable and can be placed into relational databases, semantic web applications, archives, repositories. This puts the FAIR principles into the practice of an R environment workflow: they make these tabular data objects more findable, more accessible, interoperable and reusable.

While the tidy data standards make reuse more efficient by eliminating unnecessary data processing steps before analysis or placement in a relational database, the application of DataSet definition and the datacube model with the information science metadata standards make reuse more efficient with exchanging and combining the data with other data in different datasets.

In the first step, we organize tidy datasets according to the data cube model originally developed by the SDMX for exchanging statistical data, by clearly stating which columns are to be used for grouping, filtering and for calculations. In this step, we clearly state which tidy columns are to be used as "measurements", "dimensions" and "attributes" of the dataset. This step will greatly simplify how we obtain tidy subsets from a dataset, or how we reduce the dimensions of a dataset to a triad or a quad to be placed on a knowledge graph.

In the second step, we add metadata is attributes to the R object (regardless if it is a data.frame or an inherited tibble/data.table object) which make the finding, understanding and the exchange of data (including refreshing data from source), its placement in a database or on a knowledge graph easier. This step will make the dataset easier to understand both for human analysis and for computers, and greatly increase their reuse potential, or the ability to review and replicate them in reproducible research.

A mapping of R objects into these models has numerous advantages: it makes data importing easier and less error-prone; it leaves plenty of room for documentation automation; and it makes the publication of results from R following the FAIR principles much easier.

Data semantics

The tidy data principles are seemingly very simple: each variable is a column, each observation is a row, and place all new units of measurement into a new table. However, the steps of tidying up a messy dataset

[1]RDF Data Cube Vocabulary, W3C Recommendation 16 January 2014 https://www.w3.org/TR/vocab-data-cube/, Introduction to SDMX data modeling https://www.unescap.org/sites/default/files/Session_4_SDMX_Data_Modeling_%20Intro_UNSD_WS_National_SDG_10-13Sep2019.pdf

require practice, like solving a Rubik cube—-take a long time to do quicky.

The tidy data principles frame Codd's relational algebra, particularly the third normal form (3NF) into statistical language, and make the storage and reuse of data in relational databases optional. It suggests a format for tabular data and data semantics that makes a very important, and usually the most time (resource) consuming part of data analysis far more efficient, because the analyst does not have to reinvent the wheel all the time.The data cube model can be seen as an extension of the tidy model that makes the storage, reuse, and exchange of data among different database, or their placement on a knowledge graph optimal.

The data cube model

The tidy data is designed to be easily handled in relational databases and it is an ideal format for analysis in a closed system – for example, when all your data for the analysis is stored in a well-designed relational database. This semantics is however insufficient for exchanging data with several data sources, a problem that has been attracting solutions for a long time by statistical agencies and designers of hyperlinked, web-based products.

The dataset package is inspired by the way statisticans have been for a long-time organizing data for publications: the data cube model. The data cube model has additional, non-normative prescriptions for the structure of the (tidy) dataset: columns must be grouped into measurements, dimensions, and attributes. The dimensions are non-measured variables, such as reference geographical areas, reference time periods that are usually used to filter out subsets (slices) from a datacube. Attributes provide information about the measurements, such as unit of information, estimation method, and usually can be seen as metadata for modelling purposes: they contain observations (measurements) level metadata that is also often used for filtering (selecting actual, imputed or missing cases, for example.) Measurements are in fact the observations that we want to analyze.

The data cube model has been long the preferred format to organize data for analytical use by SDMX. Statistical agencies exchange with each other and publish data in datasets that are conforming the data cube model. By organizing your R tabular data like statisticians, you make your work easier to synchronize with statistical data sources: you can download or exchange date in this format. Following the data cube model helps you to understand more easily the dataset and bring your data to be analyzed to a tidy form in one or two steps — usually using attributes for subsetting, dimensions for grouping, and measurements to carry out arithmetic calculations like computing the group average values.

Interoperability

While tidying aims and preventing reinventing the wheel and many times repeating the work of bringing the data from a data storage, such as a database, into a form that an analyst can work with, reproducible research has further aims that add to the quality and inefficiency of the analysis: it aims to make the tidy datasets and the research output easier to reuse, and to review and replicate for quality control and peer review. Data scientists know that however simple the tidy data rules are, it takes a long time to get a feeling to easily put any messy, stored data into a tidy format. To put the data into a format that is easy to review, and which offers an easy replication is even harder. As the famous article by Monya Baker, 1,500 scientists lift the lid on reproducibility shows, the vast majority of scientists have experienced the problem that seemingly replicable, published scientific data cannot be replicated, and more worryingly, the majority of them expressed doubt about their ability to replicate their own findings. Datasets that are not easily replicated are also not easily used as building blocks for further joined, more composite data.

The lack of interoperability is often connected to the barrier of understanding a dataset. Having a simple term in the column heading, and knowing that observations are in rows, variables are in columns is not always a sufficient help to understand the meaning of the dataset. The data cube model, which explicitly states dimensions, attributes for the measurements greatly increase the understandability.

> "Data is only potential information, raw and unprocessed, prior to anyone actually being informed by it." —Jeffrey Pomerantz, Metadata

Understanding data (stored in a tabular or other format) requires metadata, information by which the data can be represented in a simpler form. The simplest and most useful metadata for this purpose is a title or a label in the RDF model, a single, human readable sentence that describes the data.

The data cube model describes attributes, which can be seen as metadata about observations, to be placed into the tabular format, and dataset level metadata, such as the title. What is missing from a tidy dataset in a data.frame, tibble or data.table format is the metadata.

Attributes

In R, attributes of an object are designed to record metadata about the object – if it is a tabular data, then about the dataset itself. For a tidy dataset, at least the names of the columns, the names of the rows, and the name of the object in memory are recorded. R places very few restrictions on how to add metadata in the attribute fields of an object, or place new fields – you can add far more metadata than the size of the actual data frame if you like.

The dataset package defines a pseudo-class that enriches data frames with standardized metadata used in the RDF and SDMX datacube model, and the Dublin Core and DataCite metadata standards of libraries that make datasets easier to find, attribute, and more interoperable and reuseable by providing more information about their contents.

The dataset class (see: `dataset()`)is a pseudo-class. Although it is defined as a class inherited from the base R data.frame S3 class, in fact, it does not alter the way you interact with the data, and for analytical purposes the object remains a data.frame. It is also a pseudo-class because the functions built in the dataset constructor work exactly as well on tibbles or data.table objects, because they only touch metadata attributes of the R object.

The `dublincore()` function and its helper functions adds metadata as attributes to any R object The Dublin Core Elements are originally library standards that facilitate finding resources (files) in libraries. The RDF datacube definition, which aims to enable machines to connect datasets automatically, refers to a part of the Dublin Core definition —however, it is a good practice to add all Dublin Core elements to a dataset.

The `datacite()` functions and its helper functions adds mandatory and optional metadata from the DataCite standard, which is often preferred by open science data repositories aimed to a public storage of datasets. The DataCite standard is newer than Dublin Core, and was specifically designed to facilitate finding, accessing, and reusing data. It greatly overlaps with Dublin Core but contains further information that is particularly useful for datasets (for example, estimated dataset object size in memory or data storage.)

Both the Dublin Core Metadata Elements definition 1.1 and DataCite Metadata Scheme 4.4 refer to standard information elements, properties, for example, the ISO standard of describing the size of an object in computers, or the ISO standard of describing reference time or reference geographical areas for the data collection. Some of the helper functions of `dublincore_add()` and `datacite_add()` already validate the correct formatting of this information, some will be added in later versions.

```
library(dataset)
iris_datacite  <- datacite_add(
  x = iris,
  Title = "Iris Dataset",
  Creator = person("Anderson", "Edgar", role = "aut"),
  Publisher = "American Iris Society",
  Identifier = "https://doi.org/10.1111/j.1469-1809.1936.tb02137.x",
  PublicationYear = 1935,
```

```
  Description = "This famous (Fisher's or Anderson's) iris data set gives the measurements in centimete
  Language = "en")
```

In R, objects can have arbitrary attributes. For example, a data.frame has a class attribute that tells functions to treat the object as a data.frame.Under the hood, you still keep your data frame object of choice, the good old base r ?data.frame, or the more modern data.table or tibble.

We add descriptive metadata conforming the Dublin Core and DataCite standard as data frame attributes, because they must clearly describe the dataset. The dataset-level attributes do not interfere with the tidy data concept, because the tidy data concept relates to the contents of the data frame.

```
datacite(iris_datacite)
#> $names
#> [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
#>
#> $Title
#>          Title titleType
#> 1 Iris Dataset     Title
#>
#> $Creator
#> [1] "Anderson Edgar [aut]"
#>
#> $Identifier
#> [1] "https://doi.org/10.1111/j.1469-1809.1936.tb02137.x"
#>
#> $Publisher
#> [1] "American Iris Society"
#>
#> $Issued
#> [1] 1935
#>
#> $publication_year
#> [1] 1935
#>
#> $Type
#>   resourceType resourceTypeGeneral
#> 1      Dataset             Dataset
#>
#> $Description
#> [1] "This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the
#>
#> $Geolocation
#> [1] NA
#>
#> $Language
#> [1] "eng"
#>
#> $Rights
#> [1] NA
#>
#> $Size
#> [1] "12.72 kB [12.42 KiB]"
```

Planned functionality

The dataset package is aiming to make tidy datasets easier to release, exchange and reuse. The motivation of the ecosystem of the dataset, statcodelists, and the dataobservatory packages is to provide a comprehensive toolkit that helps many reproducible research tasks.

Our real goal is to facilitate efficient reproducible research workflows that perform flawlessly tasks that most R users, even scientific researchers, are unfamiliar with—or if they are familiar with them, they find it boring, because they will usually not get credited for it as analyst or researchers

Our datasets:

- ☒ contain Dublin Core or DataCite (or both) metadata that makes the findable and easier accessible via online libraries
- ☒ are easily reduced to linked open data formats to be serialized to, or synchronized with semantic web applications
- ☒ follow the datacube model of the Statistical Data and Metadata eXchange, therefore allowing easy refreshing with new data from the source of the analytical work
- ☒ contain processing metadata that greatly enhance the reproducibility of the results, and the reviewability of the contents of the dataset, including metadata defined by the [DDI]
- ☒ relatively lightweight in dependencies and easily works with data.frame, tibble or data.table R objects.

The datacube model is used by the W3C consortium for describing datasets, tabular data resources that can be easily connected via the internet infrastructure. RDF , Because the SDMX standards pre-date the definition of RDF, it took a long time sufficiently harmonize RDF and SDMX to become a standard for statistical data that can be easily exchanged by humans and computers alike, therefore making it ideal for reproducible research.

We also aim to replace the `survey` in the retroharmonize survey harmonization package to an inherited dataset that is optimalised to contain social sciences survey data.

The connecting statcodelists packages facilitates further reproducability with standardized, natural language independent codelists for categorical variables. Such categorical variables are correctly interpreted in a wide array of statistical applications and can be easily joined with data from many countries irrespective of the primary sources' language.

See: Linked SDMX Data

References

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604): 452–54. https://doi.org/10.1038/533452a.

Capadisli, Sarven, Sören Auer, and Axel-Cyrille Ngonga Ngomo. 2015. "Linked SDMX Data: Path to High Fidelity Statistical Linked Data." *Semantic Web* 6 (2): 105–12. https://doi.org/10.3233/SW-130123.

Core, Dublin. 2020. "DCMI Metadata Terms." http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/.

Group, DataCite Metadata Working. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4." DataCite e.V. https://doi.org/10.14454/3w3z-sa82.

Pomerantz, Jeffrey. 2015. *Metadata*. The MIT Press Essential Knowledge Series. Cambridge, MA, USA: MIT Press.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.