

# Data curation: The case of Call Reports

Ian D. Gow

29 January 2026

I recently (Gow 2026) proposed an extended version of model of the data science “whole game” of in [R for Data Science](#) (Wickham, Çetinkaya-Rundel, and Grolemund 2023). In Gow (2026), I used Australian stock price data to illustrate the data curation process and in this note, I use “call report” data for United States banks as an additional illustration of my proposed extension.

My extension of the data science “whole game”—depicted in Figure 1 below—adds a *persist* step to the original version, groups it with *import* and *tidy* into a single process, which I call *Curate*. As a complement to the new *persist* step, I also add a *load* step to the *Understand* process.<sup>1</sup>

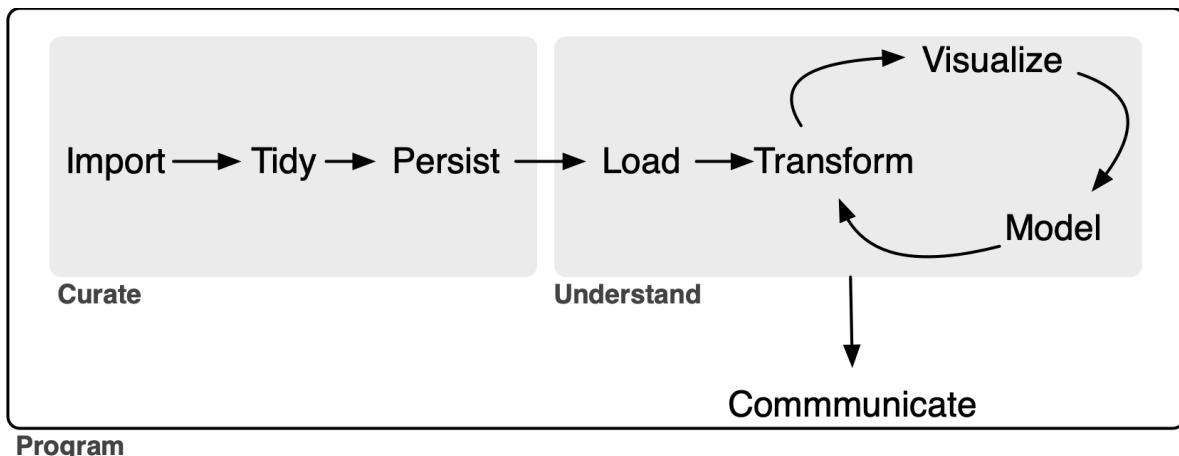


Figure 1: A representation of the data science workflow

In this note, as in Gow (2026), I focus on the data curation (*Curate*) process. My rationale for separating *Curate* from *Understand* is that I believe it clarifies certain best practices in the curation of data. the notion of a service-level agreement to delineating roles and responsibilities in the preparation and analysis of data. As will be seen, my conception of *Curate* will encompass some tasks that are included in the *transform* step (part of the *Understand* process) in [R for](#)

<sup>1</sup>As will be seen this, *load* step will not generally be an elaborate one. The inclusion of a separate *load* step serves more to better delineate the distinction between the *Curate* process and the *Understand* process.

[Data Science](#). The current version of this note uses daily data on Australian stock prices as an application.

While I will argue that even the sole analyst who will perform all three processes can benefit from thinking about *Curate* separate from *Understand*, it is perhaps easiest to conceive of the *Curate* and *Understand* processes as involving different individuals or organizational units of the “whole game” of a data analysis workflow. In Gow (2026), I used the idea of a service-level agreement to delineate the division of responsibilities between the *Curate* and *Understand* teams. In effect, I will act as a self-appointed, single-person, unpaid *Curate* team and I imagine potential users of call report data as my *Understand* clients.

This note was written using [Quarto](#) and compiled with [RStudio](#), an integrated development environment (IDE) for working with R. The source code for this note is available [here](#) and the latest PDF version is available [here](#).

## 1 Call Reports

According to [its website](#), the Federal Financial Institutions Examination Council (FFIEC) “is an interagency body ... focused on promoting consistency in examination activities [related to United States financial institutions]. The FFIEC”does not regulate financial institutions [and its] jurisdiction is limited to prescribing uniform principles, standards, and report forms for the federal examination of financial institutions and making recommendations to promote uniformity in the supervision of financial institutions.”

One of the services provide by the FFEIC is its “Central Data Repository’s Public Data Distribution web site. Through [this site](#) you can obtain Reports of Condition and Income (Call Reports) for most FDIC-insured institutions. The information available on this site is updated to reflect the most recent data for both prior and current periods. The earliest data provided is from March 31, 2001.”

Many academic studies use Call Report data. For example, Kashyap, Rajan, and Stein (2002, 52) states “our data come from the ‘Call Reports,’ the regulatory filings that all commercial banks having insured deposits submit each quarter. The Call Reports include detailed information on the composition of bank balance sheets and some additional data on off-balance-sheet items. These data are reported at the level of the individual bank.”

While raw data are offered by the FFIEC, some work is required to arrange the data into a form amenable to further analysis. In other words, some curation of the data is required.

Several commercial data providers of the FFIEC data in a curated form, but accessing those requires an institutional or individual subscription. It appears that the Federal Reserve Bank of Chicago provided a curated data set on its [website](#), but no longer does so.

Other services provide “API wrappers” that allow users to access the data via the FFIEC API, but this approach is best suited to real-time feeds of small amounts of data. Collation of the data into a single repository using this approach seems unfeasible.

Regardless of the availability of paid curation services, in this note I will imagine that such sources either do not exist or are unavailable to my hypothetical *Understand* clients and illustrate how one can curate Call Reports data from the FFIEC source.

## 1.1 Getting the raw data

The FFIEC [Bulk Data Download site] provides the Call Report data in two forms. The first is as zipped tab-delimited data files, one for each quarter. The second is as zipped XBRL data files, one for each quarter. At the time of writing, getting the complete data archive amounts to point-and-clicking to download each of the 99 files for each format.

The FFIEC data sets are not as amenable to automated downloading as those offered by other government agencies such as the SEC (see my earlier [note on XBRL data](#)), the PCAOB (see my [note on Form AP data](#)), or even the Federal Reserve itself (I used data from the [MDRM site](#) in preparing this note). However, a group of individuals has contributed the Python package `ffiec_data_collector` that we can use to collect the data.<sup>2</sup>

To install this Python package, you first need to [install Python](#) and then install the `ffiec_data_collector` using a command like `pip install ffec_data_collector`.

As discussed in [Appendix E](#) of Gow and Ding (2024), I organize my raw and processed data in a repository comprising a single parent directory and several sub-directories corresponding to various data sources and projects. For some data sets, this approach to organization facilitates switching code from using (say) data sources provided by Wharton Research Data Services (WRDS) to using local data in my data repository. I will adopt that approach for the purposes of this note.

As the location of my “raw data” repository is found in the environment variable `RAW_DATA_DIR`, I can identify that location in Python easily. The following code specifies the download directory as the directory `ffiec` within my raw data repository.<sup>3</sup>

```
import os
from pathlib import Path
print(os.environ['RAW_DATA_DIR'])
```

/Users/igow/Dropbox/raw\_data

---

<sup>2</sup>I discovered this package after writing most of this note. In my case, I pointed-and-clicked to get many of the files and [Martien Lubberink](#) of Victoria University of Wellington kindly provided the rest.

<sup>3</sup>I recommend that readers follow a similar approach if following along with note, as it makes subsequent steps easier to implement. A reader can simply specify `os.environ['RAW_DATA_DIR'] = "/Users/igow/Dropbox/raw_data"`, substituting a location where the data should go on his or her computer.

```
download_dir = Path(os.environ['RAW_DATA_DIR']) / "ffiec"
```

Having specified a location to put the downloaded files, it is a simple matter to adapt a script provided on the [package website](#) to download the raw data files.

```
import ffiec_data_collector as fdc
import time

downloader = fdc.FFIECDownloader(download_dir=download_dir)

periods = downloader.select_product(fdc.Product.CALL_SINGLE)

results = []
for period in periods[:4]:

    print(f"Downloading {period.yyyymmdd}...", end=" ")
    result = downloader.download(
        product=fdc.Product.CALL_SINGLE,
        period=period.yyyymmdd,
        format=fdc.FileFormat.TSV
    )
    results.append(result)

    if result.success:
        print(f" ({result.size_bytes:,} bytes)")
    else:
        print(f" Failed: {result.error_message}")

# IMPORTANT: Be respectful to government servers
# Add delay between requests to avoid overloading the server
time.sleep(1) # 1 second delay - adjust as needed
```

```
Downloading 20250930... (5,686,532 bytes)
Downloading 20250630... (6,231,006 bytes)
Downloading 20250331... (5,704,421 bytes)
Downloading 20241231... (6,605,092 bytes)
```

```
# Summary
successful = sum(1 for r in results if r.success)
print(f"\nCompleted: {successful}/{len(results)} downloads")
```

Completed: 4/4 downloads

Note that the code above downloads just the most recent four files available on the site. Remove `[:4]` from the line `for period in periods[:4]:` to download *all* files. Note that the package website recommends using `time.sleep(5)` in place of `time.sleep(1)` to create a five-second delay and this may be a more appropriate choice if you are downloading all 99 files using this code. Note that the downloaded files occupy about 800 MB of disk space, so make sure you have that available if running this code.

## 1.2 Processing the data

With the raw data files on hand, the next task is to process these into files useful for analysis. For reasons I will discuss below, I will process the data into Parquet files. The Parquet format is described in *R for Data Science* (Wickham, Çetinkaya-Rundel, and Grolemund 2023, 393) as “an open standards-based format widely used by big data systems.” Parquet files provide a format optimized for data analysis, with a rich type system. More details on the parquet format can be found in Chapter 22 of Wickham, Çetinkaya-Rundel, and Grolemund (2023) and every code example in Gow and Ding (2024) can be executed against Parquet data files created using my `db2pq` Python library as described in Appendix E of that book.

The easiest way to run the code I used to process the data is to install the `ffiec.pq` R package I have made available on GitHub. First, in R install the `pak` package using `install.packages("pak")`. Then, use `pak` to install my package using `pak::pak("iangow/ffiec.pq")`.

The easiest way to run the package is to set the locations for the downloaded raw data files from above and for the processed data using the environment variables `RAW_DATA_DIR` and `DATA_DIR`, respectively.

I already have these set:

```
Sys.getenv("RAW_DATA_DIR")
```

```
[1] "/Users/igow/Dropbox/raw_data"
```

```
Sys.getenv("DATA_DIR")
```

```
[1] "/Users/igow/Dropbox/pq_data"
```

If you don’t have these set, you can set them within R using commands like the following. You should set `RAW_DATA_DIR` to match what you used above in Python and you should set `DATA_DIR` to be where you want to put the processed files. The processed files will occupy about 2 GB of disk space, so make sure you have room for these there.

```
Sys.setenv(RAW_DATA_DIR="/Users/igow/Dropbox/raw_data")
Sys.setenv(DATA_DIR="/Users/igow/Dropbox/pq_data")
```

Having set these environment variables, we can load my package and run a single command `ffiec_process()` without any arguments to process *all* the raw data files. This takes a bit over three minutes for me:

```
library(ffiec.pq)
system.time(results <- ffec_process())
```

So what have we just done? Some answers come from inspection of the data frame `results` returned by the `ffiec_process()`.

```
library(ffiec.pq)
library(tidyverse)
library(farr)
library(dbplyr)
library(DBI)
```

```
results
```

```
# A tibble: 3,674 x 8
  type     kind   date      parquet      zipfile n_parts repairs inner_files
  <chr>    <chr> <date>    <chr>        <chr>      <dbl> <list>   <list>
1 por       por   2001-03-31 por_20010331.p~ FFIEC ~      1 <chr>   <chr [1]>
2 schedule ci   2001-03-31 ci_20010331.pa~ FFIEC ~      1 <chr>   <chr [1]>
3 schedule ent  2001-03-31 ent_20010331.p~ FFIEC ~      1 <chr>   <chr [1]>
4 schedule gi  2001-03-31 gi_20010331.pa~ FFIEC ~      1 <chr>   <chr [1]>
5 schedule gl  2001-03-31 gl_20010331.pa~ FFIEC ~      1 <chr>   <chr [1]>
6 schedule leo 2001-03-31 leo_20010331.p~ FFIEC ~      1 <chr>   <chr [1]>
7 schedule narr 2001-03-31 narr_20010331.~ FFIEC ~      1 <chr>   <chr [1]>
8 schedule rc   2001-03-31 rc_20010331.pa~ FFIEC ~      1 <chr>   <chr [1]>
9 schedule rca  2001-03-31 rca_20010331.p~ FFIEC ~      1 <chr>   <chr [1]>
10 schedule rcb 2001-03-31 rcb_20010331.p~ FFIEC ~      1 <chr>   <chr [1]>
# i 3,664 more rows
```

Each row in `results` represents a Parquet file created by `ffiec_process()`.

```
results |>
  count(type)
```

```
# A tibble: 2 x 2
  type      n
  <chr>    <int>
1 por        99
2 schedule  3575
```

Let's use DuckDB.

The `results` data frame is also stored as `ffiec_process_data` in the `ffiec` directory.

No doubt you are thinking “OMG! There are 3674 files? How can I use these?” It turns out that this structure is quite easy to work with given the right tools.

### 1.3 Using the data with R

Most of the files are listed as having type equal to "schedule" and the rest have type equal to "por". The latter files represent the [“Panel of Reporters” \(POR\) data](#) and they provides details on the financial institutions filing in the respective quarter. **TODO:** Need to mention that the data are quarterly.

```
db <- dbConnect(duckdb::duckdb())

results <- load_parquet(db, "ffiec_process_data", "ffiec")
```

To start with let's focus on the last available date, which is 2025-09-30 at the time of writing.

We can access the latest POR file using the `load_parquet()` function from my `farr` package.

```
por <- load_parquet(db, "por_20250930", "ffiec")
por
```

```
# A query: ?? x 13
# Database: DuckDB 1.4.4-dev6 [root@Darwin 25.3.0:R 4.5.2/:memory:]
  IDRSSD fdic_certificate_number occ_charter_number ots_docket_number
  <int> <chr>           <chr>           <chr>
1     37 10057           <NA>           16553
2     242 3850            <NA>           <NA>
3     279 28868           <NA>           2523
4     354 14083           <NA>           <NA>
5     457 10202           <NA>           <NA>
6     505 6959             1253           <NA>
7     1155 17639           <NA>           <NA>
```

```

8    1351 9392          <NA>          <NA>
9    1454 19184          15359          <NA>
10   1669 3384          13541          <NA>
# i more rows
# i 9 more variables: primary_aba_routing_number <chr>,
#   financial_institution_name <chr>, financial_institution_address <chr>,
#   financial_institution_city <chr>, financial_institution_state <chr>,
#   financial_institution_zip_code <chr>,
#   financial_institution_filing_type <chr>,
#   last_date_time_submission_updated_on <dttm>, date <date>

```

The remaining files with type equal to "schedule" represent files on various schedules and the applicable schedule is indicated by the column kind.

If we were experts in Call Report data, we might know that domestic total assets is reported as item RCFD2170 on Schedule RC for banks reporting on a consolidated basis and as item RCON2170 for banks reporting on an unconsolidated basis.<sup>4</sup> We can load the latest Schedule RC data by combining the schedule as lower case (rc) with an underscore followed by the date in yyyyymmdd form. **TODO: Guide readers to this point.**

We can make a balance sheet table using the following code.

```

rc_20250930 <- load_parquet(db, "rc_20250930", "ffiec")

bs_data <-
  rc_20250930 |>
  mutate(total_assets = coalesce(RCFD2170, RCON2170)) |>
  select(IDRSSD, date, total_assets)

```

Let's say we're interested in the top 10 banks by total assets on this date. This is easy enough to produce.

```

top_5 <-
  bs_data |>
  window_order(desc(total_assets)) |>
  mutate(ta_rank = row_number()) |>
  filter(ta_rank <= 5)

top_5 |> collect()

```

```
# A tibble: 5 x 4
```

---

<sup>4</sup>From inspection of the data, we can confirm that banks report one or the other.

```
IDRSSD date      total_assets ta_rank
<int> <date>      <dbl>    <dbl>
1 852218 2025-09-30 3813431000     1
2 480228 2025-09-30 2651090000     2
3 476810 2025-09-30 1844189000     3
4 451965 2025-09-30 1767105000     4
5 504713 2025-09-30 679293260      5
```

However, it's hard for us to know who these banks if we haven't committed the table of IDRSSD values to memory. Fortunately, we can get the information we need from por.

```
top_5_names <-
  top_5 |>
  inner_join(por |>
    select(IDRSSD, date, financial_institution_name),
    join_by(IDRSSD, date)) |>
  arrange(ta_rank) |>
  select(IDRSSD, financial_institution_name, ta_rank) |>
  rename(bank = financial_institution_name) |>
  compute()

top_5_names |> collect()
```

```
# A tibble: 5 x 3
IDRSSD bank                      ta_rank
<int> <chr>                     <dbl>
1 852218 JPMORGAN CHASE BANK, NATIONAL ASSOCIATION 1
2 480228 BANK OF AMERICA, NATIONAL ASSOCIATION      2
3 476810 CITIBANK, N.A.                  3
4 451965 WELLS FARGO BANK, NATIONAL ASSOCIATION      4
5 504713 U.S. BANK NATIONAL ASSOCIATION            5
```

Already you might be feeling frustrated. "I don't want to access data for just one quarter. Do I have to make tables for every quarter like this?"

The answer is "no".

In principle the `load_parquet()` function can take **glob** entries, such as this:

```
rc <- load_parquet(db, "rc_*", "ffiec")
```

But, while that appears to work, we will quickly run into trouble because the items that appear on any given schedule can vary over times. Fortunately, we can use the function `ffiec_scan_pqs()` to combine all schedules into a single table that we can use easily.

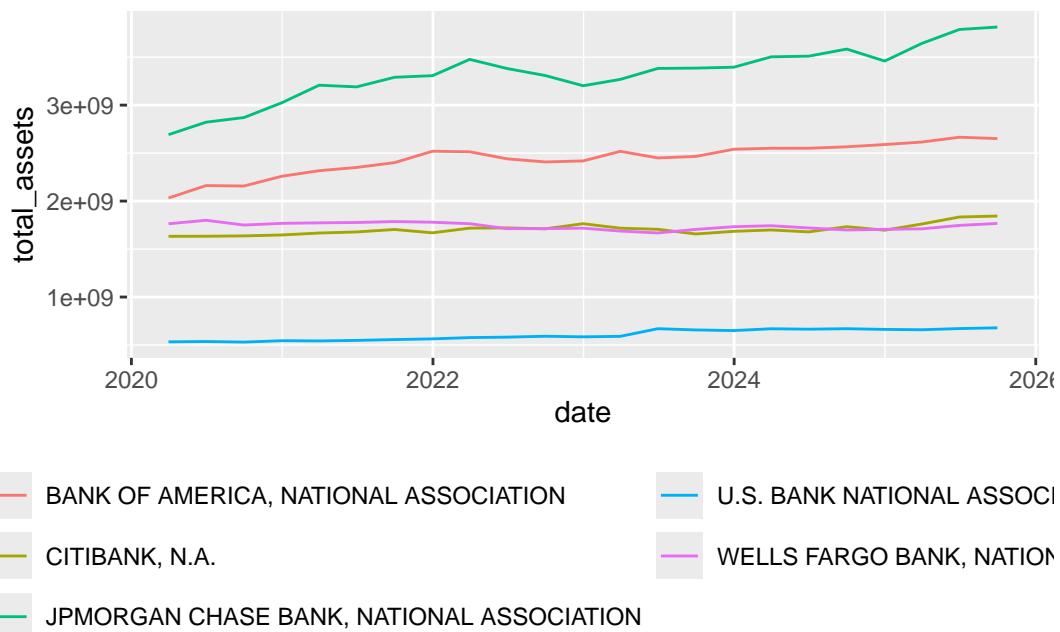
```

rc <- ffiec_scan_pqs(db, schedule = "rc")

bs_panel_data <-
  rc |>
  mutate(total_assets = coalesce(RCFD2170, RCON2170)) |>
  select(IDRSSD, date, total_assets) |>
  inner_join(top_5_names, join_by(IDRSSD)) |>
  collect()

bs_panel_data |>
  filter(date >= "2020-01-01") |>
  ggplot(aes(x = date, y = total_assets, colour = bank)) +
  geom_line() +
  theme(legend.position = "bottom") +
  guides(colour = guide_legend(nrow = 3))

```



## 1.4 Using the data with Python

## 2 The boring details

## 3 The service-level agreement

**Nothing from here up has been updated from original note**

It is perhaps helpful to think of dividing the data science workflow into processes with different teams being responsible for the different processes. From this perspective, the *Curate* team manufactures data that are delivered to the *Understand* team. While I won't discuss transfer pricing (i.e., how much the *Understand* team needs to pay the *Curate* team for the data), we might consider the analogy of a [service-level agreement](#) between the two teams.

One template for a service-level agreement would specify that data from a particular source will be delivered to the *Understand* team with the following conditions:

1. The data will be presented as a set of tables in a modern storage format.
2. The division into tables will adhere to a pragmatic version of good database principles.
3. The **primary key** of each table will be identified and validated.
4. Each variable (column) of each table will be of the correct type.
5. There will be no manual steps that cannot be reproduced.
6. A process for updating the curated data will be established.
7. The entire process will be documented in some way.
8. Some process for version control of data will be maintained.

### 3.1 Storage format

In principle, the storage format should fairly minor detail determined by the needs of the *Understand* team. For example, if the *Understand* team works in Stata or Excel, then perhaps they will want the data in some kind of Stata format or as Excel files. However, I think it can be appropriate to push back on notions that data will be delivered in form that involves downgrading the data or otherwise compromises the process in a way that may ultimately add to the cost and complexity of the task for the *Curate* team. For example, "please send the final data as an Excel file attachment as a reply email" might be a request to be resisted because the process of converting to Excel can entail the degradation of data (e.g., time stamps or encoding of text).<sup>5</sup> Instead it may be better to choose a more robust storage format and supply a script for turning that into a preferred format.

One storage format that I have used in the past would deliver data as tables in a (PostgreSQL) database. The *Understand* team could be given access data from a particular source organized

---

<sup>5</sup>I discuss some of the issues with Excel as a storage format below.

as a **schema** in a database. Accessing the data in this form is easy for any modern software package. One virtue of this approach is that the data might be curated using, say, Python even though the client will analyse it using, say, Stata.<sup>6</sup>

## 3.2 Good database principles

I included the word “pragmatic” because I think it’s not necessary in most cases to get particularly fussy about **database normalization**. That said, it’s probably bad practice to succumb to requests for One Big Table that the *Understand* team might make. It is reasonable to impose some obligation to merge tables that are naturally different tables on the client *Understand* team.

## 3.3 Primary keys

The *Curate* team should communicate the primary key of each table to the *Understand* team.<sup>7</sup> A primary key of a table will be a set of variables that can be used to uniquely identify a row in that table. In general a primary key will have no missing values. Part of data curation will be confirming that a proposed primary key is in fact a valid primary key.

## 3.4 Data types

Each variable of each table should be of the correct type. For example, dates should be of type DATE, variables that only take integer values should be of INTEGER type.<sup>8</sup> Date-times should generally be given with TIMESTAMP WITH TIME ZONE type. **Logical columns** should be supplied with type BOOLEAN.

Note that there is an interaction between this element of the service-level agreement and the storage format. If the data are supplied in a PostgreSQL database or as parquet files, then it is quite feasible to prescribe the data types of each variable. But if the storage format is Excel files (not recommended!) or CSV files, then it is difficult for the data curator to control how each variable is understood by the *Understand* team.<sup>9</sup>

In some cases, it may seem unduly prescriptive to specify the types in a particular way. For example, a logical variable can easily be represented as INTEGER type (0 for FALSE, 1 for TRUE). Even in such cases, I think there is merit in choosing the most logical type (no pun intended) because of the additional information it conveys about the data. For example, a logical type

---

<sup>6</sup>One project I worked on involved Python code analysing text and putting results in a PostgreSQL database and a couple of lines of code were sufficient for a co-author in a different city to load these data into Stata.

<sup>7</sup>Sometimes there will be more than one primary key for a table.

<sup>8</sup>Here I used PostgreSQL data types, but the equivalent types in other formats should be fairly clear.

<sup>9</sup>SAS and Stata are somewhat loose with their “data types”. In effect SAS has just two data types—fixed-width character and floating-point numeric—and the other types are just formatting overlays over these. These types can be easily upset depending on how the data are used.

should be checked to ensure that it only takes two values (TRUE or FALSE) plus perhaps NULL and that this checking has occurred is conveyed by the encoding of that variable as BOOLEAN.

### 3.5 No manual steps

When data vendors are providing well-curated data sets, much about the curation process will be obscure to the user. This makes some sense, as the data curation process has elements of trade secrets. But often data will be supplied by vendors in an imperfect state and significant data curation will be performed by the *Curate* team working for or within the same organization as the *Understand* team.

Focusing on the case where the data curation process transforms an existing data set—say, one purchased from an outside vendor—into a curated data set in sense used here, there are a few ground rules regarding manual steps.

First, *the original data files should not be modified in any way*. If data are supplied as CSV files, then merely opening them in Excel and saving them can mutilate the original data.<sup>10</sup> I have encountered people whose idea of data curation extended to opening the original files, saving them as Excel files, and then proceeding to manually edit those files. This approach leads to a completely unreplicable set of data files, which is problematic not only in a world in which reproducibility is starting to be expected, but also when a new version of the data will be supplied by the vendor in the future.

Second, any manual steps should be extensively documented and applied in a transparent automated fashion. For example, if a data set on financial statement items of US firms contains errors that can be corrected by reviewing original SEC filings, then any corrections should be clearly documented in separate files with links to the original filings and explanations. And the corrections should be implemented through code, not manual steps.<sup>11</sup> For example, there should be code that imports the original data and the corrections and applies the latter to the former to create the final data set.

### 3.6 Documentation

The process of curating the data should be documented sufficiently well that someone else could perform the curation steps should the need arise. Often that need will arise when the vendor provides an updated data set. Perhaps the best way to understand what I have in mind here is through a case study and I provide on in `?@sec-case`.

---

<sup>10</sup>An example where this could happen is provided in `?@tip-friends`.

<sup>11</sup>Best practices here could be the subject of a separate note, as there are a lot of subtleties and my experience is that people often have bad habits here.

### **3.7 Update process**

Part of the rationale for having a well-documented process with no manual steps is that it greatly facilitates updating the data when the data vendor or other data source provides updated data. In some cases, updating the data will entail little more than downloading the new raw data and running a pre-existing script on those data. In other cases, the data may change in significant ways, such as addition of new variables, renaming of existing ones, or reorganization of data into different tables.

As future changes may be difficult to predict, the analyst might be able to do little more than describe the anticipated update process if no major changes occur. If major changes do subsequently occur, it likely makes sense for the analyst handling the update to extensively document the changes needed to process the new data, especially if earlier versions of the data remain relevant (e.g., they have been used in published research).

### **3.8 Data version control**

Welch (2019) argues that, to ensure that results can be reproduced, “the author should keep a private copy of the full data set with which the results were obtained.” This imposes a significant cost on the *Understand* team to maintain archives of data sets that may run to several gigabytes or more and it would seem much more efficient for these obligations to reside with the parties with the relevant expertise.

Unfortunately, even when data vendors provide curated data sets, they generally provide little in the way of version control. For example, there is no evidence that Wharton Research Data Services (WRDS), perhaps the largest data vendor in academic business research, provides any version control for its datasets, even though it should have much greater expertise for doing this than the users of its services.

Nonetheless, some notion of version control of data probably has a place in data curation, even if this is little more than archiving of various versions of data supplied to research teams.

## **4 The service-level agreement revisited**

I now return to our service-level agreement (SLA) to take stock of where we are after the above. Given that much of the focus above was on data types, I do not revisit that here and instead focus on those elements of the SLA that I did not address above.

## 4.1 Storage format

We have chosen to use parquet files for our output. `?@tbl-pq-files` provides some data on the parquet files we have produced for our hypothetical client (the *Understand* team). Assuming that the client is a group of colleagues at an institution with access to SIRCA, we (the *Curate* team) might just send a link to the Dropbox folder where we have stored the parquet files.

## 4.2 Primary keys

Table 1 provides a summary of our analysis above of primary keys.

Table 1: SIRCA ASX EOD price collection: Primary keys

Table	Primary key
si_au_ref_names	gcode, securityticker, listdate
si_au_prc_daily	gcode, date, seniorsecurity
si_au_retn_mkt	date
si_au_ref_trddays	date

## 4.3 Good database principles

In general, I think one wants to be fairly conservative in considering database principles with a data library. If the data are workable and make sense in the form they come in, then it may make most sense to keep them in that form.

The SIRCA ASX EOD data are organized into four tables with easy-to-understand primary keys and a fairly natural structure. At some level, the two primary tables are `si_au_ref_names` and `si_au_prc_daily`.<sup>12</sup> These two tables are naturally distinct, with one about companies and the other about daily security returns.

While there might be merit in splitting `si_au_prc_daily` into separate tables to reduce its size, it is actually quite manageable in its current form.

## 4.4 No manual steps

There are no manual steps in creating the parquet files except for the initial download of the CSV files from SIRCA. While some data vendors allow users to download files using scripts (e.g., the scripts I have [here](#) for WRDS), this does not appear to be an option for SIRCA. But once the data have been downloaded, the subsequent steps are automatic.

---

<sup>12</sup>It seems possible that `si_au_retn_mkt` and `si_au_ref_trddays` are generated from `si_au_prc_daily`.

While some of the checks and data-cleaning had manual elements (e.g., identifying the near-duplicate with `Gcode=="rgwb1"` in `si_au_ref_names`), the resulting code implements the fix in an automated fashion. So long as the SIRCA data remain unchanged, the fix will continue to work.

## 4.5 Documentation

A important principle here is that the code for processing the data is documentation in its own right. Beyond that the document you are reading now is a form of documentation. If the goal of this document were to provide details explaining the process used to produce the final data sets, then it might make sense to edit this document to reflect that different purpose, but in many ways I hope this document already acts as good documentation.

## 4.6 Update process

In some ways, the update process is straightforward: when new CSV files become available, download them into `RAW_DATA_DIR` and run the script. However, it would probably be necessary to retrace some of the steps above to ensure that no data issues have crept in (e.g., duplicated keys). It may make sense to document the update process as part of performing it the first time.

## 4.7 Data version control

I achieve a modest level of data version control by using Dropbox, which offers the ability to restore previous versions of data files. As discussed earlier, version control of data is a knotty problem.

## References

- Gow, Ian D. 2026. "Data Curation and the Data Science Workflow." [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=6149488](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6149488).
- Gow, Ian D., and Tongqing Ding. 2024. *Empirical Research in Accounting: Tools and Methods*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781003456230>.
- Kashyap, Anil K., Raghuram Rajan, and Jeremy C. Stein. 2002. "Banks as Liquidity Providers: An Explanation for the Co-Existence of Lending and Deposit-Taking." *Journal of Finance* 57 (3): 33–63. <https://doi.org/10.1111/1540-6261.00415>.
- Welch, Ivo. 2019. "Editorial: An Opinionated FAQ." *Critical Finance Review* 8 (1-2): 19–24. <https://doi.org/10.1561/104.00000077>.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. *R for Data Science*. Sebastopol, CA: O'Reilly Media. <https://books.google.com/books?id=TiLEEAAAQBAJ>.