ELECTRICITY
AUTHORITY
TE MANA HIKO

9 November 2021

Nicky McLean
By email: fyi-request-17178-9d73427e@requests.fyi.org.nz

Dear Nicky

Thank you for your request, received on 13 October 2021 under the Official Information Act 1982 (the Act): Please see **“Appendix 1: OIA Request Nicky McLean 13/10”** on the following pages.

Thank you for bringing to our attention the issues with the process of publishing wholesale market price data to EMI datasets.

The underlying data held by Authority is, to the best of our knowledge, accurate. We use this data to produce the files which we publish on EMI datasets. Following the migration of our many data collections to a cloud-based data lake, we have been working through the re-coding of the various data publishing tasks. We recently created two new price datasets, which can be found at:

- https://www.emi.ea.govt.nz/Wholesale/Datasets/FinalPricing/EnergyPrices, and
- https://www.emi.ea.govt.nz/Wholesale/Datasets/FinalPricing/ReservePrices respectively.

The legacy collection of prices found at https://www.emi.ea.govt.nz/Wholesale/Datasets/Final_pricing, i.e., the price files that you commented on, will soon be decommissioned. Please make use of the newly published price files.

The Authority considers these new price data sets reliably reflect the accurate underlying data. However, we would be happy to hear from you at info@ea.govt.nz if you continue to notice any issues with the prices we publish.

You have the right to seek an investigation and review by the Ombudsman of this decision. Information about how to make a complaint is available at www.ombudsman.parliament.nz or freephone 0800 802 602.

If you wish to discuss this decision with us, please feel free to contact us by emailing oia@ea.govt.nz.

Yours sincerely

Sarah Gillies
GM Legal Monitoring and Compliance

Appendix 1: OIA Request Nicky McLean 13/10

"Dear Electricity Authority,

Via its website https://www.emi.ea.govt.nz/Wholesale/Datasets, each month the Electricity authority publishes collections of data files containing half-hourly data on various aspects of the workings of the electricity system, and in particular, data on "Final prices" and "Final Reserve prices" – though with variations on the file names from time to time. Those data files (and others) had been published via a DVD issued twice a year, with also an associated website offering monthly updates. The last files provided in this way were called fp2013m09.csv and rp2013m09.csv.
The new arrangement involved ftp://ftp.emi.ea.govt.nz (which has since been abandoned) and initially, data for October 2013 and onwards were provided, though under different folder names and with varying file names that might be replaced later, such as file 201404_Final_pricing.csv vanishing in favour of 201404_Final_prices.csv as just one example. Automated systems neither create nor follow such flounderings, but no matter.
The key development was the later re-supply of all the earlier data files in this new scheme, all the way back to October 1996 for the final prices (199610_Final_prices.csv), yet oddly, only to November 1996 for the reserve prices (199611_Reserve_prices.csv) – this is shown via your website https://www.emi.ea.govt.nz/Wholesale/Datasets/Final_pricing with its two sub-directories Final_Prices and Reserve_prices, and is not to be confused with your website https://www.emi.ea.govt.nz/Wholesale/Datasets/FinalPricing that contains other matter of no interest here.
Thus arises the first request: for the revised data file for the reserve prices of October 1996. This file is omitted from your offering, when previously it had been available as file rp1996m10.csv. Why has file 199610_Reserve_prices.csv and the data it contains been suppressed?
And, just out of curiosity, why are there no reserve price data files for April 1997 to March 2004? As with Pokemon characters, "Gotta catch them all!"

There was a mass resupply of these data files in October 2018 that repaired a number of basic errors, such as truncated files - e.g. file 199912_Final_prices.csv was of length 2,216 characters (not 10,292,760), being truncated in line 80; file 200312_Final_prices.csv was of length 9,928,277 (not 10,133,442) as it omitted many time slots, e.g. half-hour 1 on 1/12/2003 for all names; and others, but further details of now-replaced files would be supererogatory.
When a collection of data files is replaced by a new set of data files, whether with different file names or not, an immediate question arises: are there changes to the data offered by these revised files? Indeed there are.
First, the positive. Amongst the millions of matching values, two new ones appear:
24/3/2004, hh 7: 7.03 BDE0111
25/3/2004, hh15:53.14 BDE0111
In the original data files, there was no mention of a value for that name at those times. Demonstrating lacunae lead to problems in epistemology because any given collection of items also omits an arbitrary number of other items (for instance, a schedule of winning lotto numbers identified by draw for the year to come), thus a putative list of omitted items would be infinitely large, further, any such list must be incomplete via Georg Cantor's "diagonal argument" (1891) that shows how to construct a new item that is not in the list. Should that be added to the list, then another new item can be generated in the same way, and, this process can proceed indefinitely.
Happily, in this context we need not consider transfinite collections and can deal with rather less elevated concepts. Going by the date, the expectation would be that somewhere in the original file fp2004m03.csv would appear data for name BDE0111 corresponding to half-hour 7 on 24/3/2004 and half-hour 15 on

25/3/2004 – as well as data for other times and names. It turns out that data are scattered around in irregular clumps, but in records 270,820-270,821 there appear
BDE0111,24/03/2004,8,F,2.58,25/03/2004 10:56:52
BDE0111,24/03/2004,9,F,2.58,25/03/2004 10:56:52
and in records 270,934-270,939
BDE0111,24/03/2004,1,F,13.03,25/03/2004 10:56:52
BDE0111,24/03/2004,2,F,15.06,25/03/2004 10:56:52
BDE0111,24/03/2004,3,F,10.76,25/03/2004 10:56:52
BDE0111,24/03/2004,4,F,10.38,25/03/2004 10:56:52
BDE0111,24/03/2004,5,F,7.25,25/03/2004 10:56:52
BDE0111,24/03/2004,6,F,6.88,25/03/2004 10:56:52
so there is no entry for half-hour 7, while record 282,405 starts a sequence with
BDE0111,25/03/2004,16,F,87.26,26/03/2004 16:30:53
and record 282,515 ends another with
BDE0111,25/03/2004,14,F,29.58,26/03/2004 16:30:53
thus omitting any value for half-hour 15. Presumably your organisation has not discarded the data files of the DVD collections, so you can verify this situation yourselves, should you wish.
Conversely, the revised file 200403_Final_prices.csv offers record 265,017: 2004-03-24,7,BDE0111,7.03
record 278,401: 2004-03-25,15,BDE0111,53.14

Evidently, those two values had been omitted from the original file, yet they appear in the (re-)revised file.This is a puzzle, as one imagines that some data storage system has been commanded something like "dump all final price data for March 2004" rather than "dump all final price data for March 2004 except for BDE0111 on hh7 24/3/2004 and hh15 on 25/3/2004". How can those additional values appear, a decade after their time? And why did they not the first time?
Also troubling, and a much larger problem, is the change in data format evident between the content of the two data files. The original layout of those data is in the form of
Name,Date,Half-hour number, F-code, Datum, Timestamp.
This last interpretation was eventually confirmed by the appearance of data files in 2004 onwards with a heading line that stated "PRICE_RUN_TIME" (and in later files, perhaps "Price_Run_Time", and in the Reserve Price data files "Run_Time" for a similar appendage), so taking these data at face value, price values are calculated well after the time to which they apply.
Yet your revised data files omit all mention of such timestamps, thus the second request: supply these Final and Reserve price data with the timestamps included, as before, not suppressed.
A question also arises as to the provenance of the data in the resupplied files, as far example what "price run time" ought to be assigned to the two newly-appearing values mentioned above. The original data came from the NZ Stock Exchange and presumably derived directly from the values that were calculated at the time and were used to drive the money exchanges between the various businesses involved. A contemporaneous record, then. Where have the data in the revised files come from? The same original records of the past transactions? If so, how does new data appear?
A similar issue is raised by the Reserve Prices data files. Instead of presenting data for hundreds of different names, just two names are offered. Unfortunately, they were BEN2201 and HAY2201, these being exactly the names also appearing in the Final Prices data files for different data so one must keep track of the source file name to keep them distinct, and variations in those file names were not helpful. Anyway, each name had two values supplied for each time slot plus a timestamp appendage (usually) that vanished after September 2013 when a new format was introduced, using the names SI and NI, until the data for October 2015 when an additional two data fields appeared so that there were four for each slot, not two. Then, all the files were replaced (bearing a date of October 2018), all the way back to that for November 1996 (still not October 1996 as with the Final Prices data files) and the additional two data fields now were found back to part-way through the 21'st of July 2009.

Naturally, additional data series are additional grist for the mill, but if it is worth the effort to resupply these files with the additional data back from October 2015 to July 2009, why not do so all the way back to the start, since the earlier files are being resupplied. Alas again, with the timestamps that were originally supplied suppressed.

Oh, and data for the 31'st October 2010 are omitted from the Reserve Price data, yet are present for the Final Price data. Has someone forgotten "Thirty days hath September..." ?

But there are more serious problems with the revised Reserve Price data files. Many are corrupt. Thanks to the daylight savings changeover days, there are two days a year that do not have twenty-four hours: one has twenty-three and the other has twenty-five. Correspondingly then, there are two days that do not have forty-eight half-hourly values: one has forty-six values and the other has fifty values. Any plan based around a constant twenty-four hours in a day will be disrupted, and dealing with this constitutes a problem whose difficulty frequently surpasses the competence of data suppliers. So it is here.

To take an example, consider the year 2010. The daylight saving rule for that year places the stretch day on Sunday the 4'th of April and the shrink day on Sunday the 26'th of September. Now consider file 201304_Reserve_prices.csv – it is easy enough to see that the half-hour numbers run from 1 to 48 only, not to 50, while in the original version of that file (i.e. prior to it being re-supplied with different names and omitting the timestamps) the half-hours run from 1 to 50 as is proper.

Similarly, consider file 201009_Reserve_prices.csv for the shrink day. The half-hour numbers run from 1 to 46, as is proper, but as well, half-hours 5 and 6 are omitted, which is not proper. Thus, that day offers data for twenty-two hours – it has been shrunk twice. And again, the originally-supplied version (that contains timestamps) also has the half-hour numbers running from 1 to 46, but with no omissions.

This miss-numbering means that data are misaligned. For example, some of the original data: (Alas, this webpage interface does not allow the specification of a fixed-spacing fount such as Courier, and multiple spaces are converted to single spaces. Thus, underline characters have been inserted to reduce the damage to the layout)

```
26/ 9/2010  0·01  0·01  0·01  0·01   A VOID!   0·01  0·00  0·01  0·01  0·00  0·00
Sunday__    0·00  0·01  0·01  0·01  0·01  0·09  0·40  0·47  0·49  0·45  0·40  0·02
-2 h.hrs!!_  0·02  0·02  0·02  0·02  0·01  0·01  0·01  0·02  0·01  0·01  0·01  0·01
________    0·02  0·01  0·29  0·33  0·50  0·91  0·41  0·42  0·01  0·01  0·01  0·01
```

compared to the revised data:

```
26/ 9/2010   0·01  0·01  0·01  0·01   A VOID!    __?_  __?_   0·01  0·00  0·01  0·01
Sunday__     0·00  0·00  0·00  0·01  0·01  0·01  0·01  0·09  0·40  0·47  0·49  0·45
-2 h.hrs!!_   0·40  0·02  0·02  0·02  0·02  0·02  0·01  0·01  0·01  0·02  0·01  0·01
________     0·01  0·01  0·02  0·01  0·29  0·33  0·50  0·91  0·41  0·42  0·01  0·01
```

Here, twelve values are shown to a line, so the standard forty-eight values require four lines. For the shrink day, to maintain alignment the two slots for the non-existent times are filled with "A VOID" (quite so) and here the ? where a number should appear signifies that no datum has been supplied so there is no number to show so one isn't. As distinct from showing a zero.

For the stretch day there are fifty values to be shown, and alignment is maintained by a fifth line that is suitably offset to show the overlap:

```
 4/ 4/2010  7·00  5·00  4·00  4·00  4·00  30·00 A JOLT!
Back 2!__  __  __ __ __ __ __ 16·19  7·00  4·00  5·00  5·00  5·00  4·00  3·00
Sunday__   3·00  3·50  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02
+2 h.hrs!!_  0·02  0·02  0·02  0·03  0·03  0·03  3·00  0·03  0·02  0·02  0·01  0·01
________   0·01  0·01  0·01  0·01  0·01  0·01  0·01  0·02  0·02  0·02  3·50  4·00
```

Whereas the revised data file offers

```
 4/ 4/2010  7·00  5·00  4·00  4·00  30·00  7·00 A JOLT!
Back 2!__ __ __ __ __ __ __  4·00  5·00  5·00  5·00  4·00  3·00  3·00  3·50
Sunday__    0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02  0·02
+2 h.hrs!!_   0·02  0·03  0·03  0·03  3·00  0·03  0·02  0·02  0·01  0·01  0·01  0·01
```

________ 0·01 0·01 0·01 0·01 0·01 0·02 0·02 0·02 3·50 4·00 __?_ __?

So again, values have been incorrectly placed and two omitted.

A list of the corrupt files would further expand this communication. Checking the data for the daylight savings changeover days can be done easily enough via a command something like

for # in ~Reserve~ do list # when HoursInDay(day) ¬= 24 The specific errors in the Reserve Price data can also be found, with something like

for # in ~.Reserve~ do dump # when IsBad(#) provided you have access to a suitable system. I emphasise that the omitted data are omitted from the data files that you have re-supplied, and are not necessarily absent from whatever data storage system had been accessed to produce those files: searching that will not find omissions from them. Indeed, because the originally-supplied data files do contain the values that have been omitted from the re-supplied files, the data storage system presumably still contains them and so could supply them afresh given a suitable set of commands to do so. Considering that your organisation has emitted regulations requiring that data suppliers are to supply data that are correctly arranged, it should be so commanded.

As for the Final Prices data files (which are much largeer than the Reserve Prices files), they also are corrupt in parts, even if good in other parts, although omitting the timestamps.

Since these files contain historical data, how can it be that the re-supplied files contain data series not previously supplied? A specific example is the series named HWA1001 which offers data for 21/7/2009 to 19/8/2009, which data are equal to those supplied for both HWA1101 and HWA1102 over that date span. At a guess, HWA1001 is a mistype for HWA1101, but, how can it have acquired data? If data were missing for HWA1101 for that date span, a mistype could be supposed in the incoming data, but it would be better if it wasn't.

Similarly, the files mention some twenty-nine names with two letter name codes, such as AN2201 and TR0331. It turns out that somehow, the first letter, a M, had been omitted: those names should be MAN2201 and MTR0331 respectively as all names have a three-letter code at their start. They have data for some five months, and the data for the corresponding three-letter names have a hole five months wide. Although the system I use has long had provision for collating deviant names into one so that for this case it slots those twenty-nine five-month pieces into their corresponding normal series, it would have been nice if the revised files had corrected the problem. Persons working only from the supplied data files will be stuck.

In general, there are three possibilities. First, the deviant name is associated with data that slots neatly into a hole and clearly belongs with its neighbouring data. Secondly, the deviant name's data does not fill a hole but instead exactly matches the other data, or, the differences seem small enough to sweep under the rug of "rounding error" or the like. But thirdly, they are different, even if similar.

On the other hand, the revised files do suppress an annoyance in the originally-supplied files. From March to December 2003 there appeared some 944,484 records with a code "V" each of which was followed by another record (for the same name, date, and half-hour) with a code "F" but a different value. For instance,

record 11187: ABY0111,2/3/2003,1,V,57.64,3/3/2003 7:46:09 AM

record 22371: ABY0111,2/3/2003,1,F,57.93,4/3/2003 9:45:18 AM Happily, all such revisions bear a later timestamp: the revisions in the data file are chronologically ordered by the "price run time", and this applied even when an additional F-coded record appeared, though this is rare: nine occasions have been noted. The V and F data turn out to be similar, with daily correlation coefficients varying between 0·56 and 0·99976 or so.

In the revised data file, record 11186 has 2003-03-02,1,ABY0111,57.93 which is at least the same as the F value. These revised files do not supply a timestamp, nor a code letter (either F or V) and the stuttering does not occur. But even aside from the absence of the mysterious V values, this is not all to the good.

When confronted by a large collection of numbers one could just gloat over the existence of data files and move on to other matters, or, make some attempt to characterise the data by collecting various statistics such as minimum, median, average, maximum, standard deviation, skewness, kurtosis, etc. as an initial analysis. Although an "average price" is rather dubious, an average is still part of a description of a distribution. But for these data, any descriptive parameters are deformed by the frequent presence of bizarre values such as 100000, -100000, and -9999 when the more ordinary value is around 60, as in the above examples. Alas, because these prices are determined well after the event (as shown by the Price_Run_Time data) the obvious ploy of switching on load when the price is negative will be precluded, nor will generators be discouraged thereby.

Looking at those usages, it appears that they indicate situations where a “final price” could not be concocted by the process that produces them, for instance before a location is in use, or after it ceases being used, and so they stand for “no value here” instead of being absent. Similarly, further inspection shows that the Final Price values range not just over reasonable positive values, but also into similarly-sized negative values, and so it is possible that zero is a proper Final Price value: I would like to buy gold on the spot market at that price... Yet there can also be long sequences of zero values at the start and end of a series, also presumably indicating “no value here” instead of being absent from their data file. This is particularly vexing as there is no discernible difference in value between a proper zero value and a “no value” zero value. Attempts at assessing the distribution of values are wrecked. As a further example, the data series for Te Kaha has many zero values in the interior of its date span: which are zero and which are a different type of zero? Analysis might show that they tend to occur around the time of high tide, say: if so, plans could be laid...

Knowing what these code value signify would be helpful, further, are there any other code values lurking in the data? Perhaps with varying usage over time, as with -100000 and 0.

Using numbers as code values as well as numbers is not a good idea, especially when their types are not easily distinguished. Suppose the convention of representing Male or Female as 1 or 0 is followed, and then 0 was also used to represent any of Don’t know/Won’t say/Not recorded/De-sexed/Trans-sexed/Undecided/Hermaphrodite, etc. The resulting data would be next to useless.

A more systematic approach would be to employ an auxiliary code, as in the F and V usage above, so that the different types of value (or, non-value) could be identified, a scheme more flexible than not supplying any value at all, or supplying a “null” value or a ? or even NaN (“Not-a-Number” as has become a modern fashion) for all odd situations. This is not a new notion, and is in frequent use in the other data files your organisation publishes, for example in name sequences such as

TKH0111,HEDL,GN,TPNZ,KWh,X,I
TKH0111,HEDL,GN,TPNZ,KWh,X,F

Which both provide data for the same series. Perhaps F stands for Final, or Firm, or Fixed, while I stands for Interim, or Interpolated, or Implicit – it would be good to know. Perhaps they would stand for different data series, as with the V and F codes. Whichever, there would be clarity. And for the zero values in the Final Prices, a F code would signify a proper Final Price, while a Z code (say) would signify a place-holder signifying “no value here” and it could be treated accordingly, and statistics on the distribution of actual Final Price values would not be deformed.

Note that for the two name sequences, the F (or I) code adjoins the rest of the name sequence rather than being placed somewhere apart. Thus it would be a convenience if the Final Price data (and the Reserve Price data) were to have their code letters immediately after their name sequence as well, as in

ABY0111,V,2/3/2003,1,57.64,3/3/2003 7:46:09 AM Which is to say: Name&Code, Date, Half-Hour Number, Data, Timestamp.

The Final Price data files have a single datum, while the Reserve Price data files have two, and later, four data per line, aside from the other fields.

And it helps to have slashes in dates rather than hyphens, because hyphens are also used to signify negative numbers and so mixups are easier. Entering numerical data as rational values (such as 1/3 for one third) is not at all common, so if slashes (especially two of them) are found in an undescribed data field, then it almost certainly contains a date. Being able to recognise dates is very helpful when data files may arrive with their columns jumbled, as has happened often.

So, in short, I'm arguing for the the original data format as had been originally supplied, with dates in the dd/mm/yyyy form, timestamps on the end, and to include the F code (and other codes where appropriate) but that shifted to follow the name field as a part of a compound name.

---In Summary---
The Electricity authority operates a website https://www.emi.ea.govt.nz, that offers access to these data files, and it has the statement "A fundamental requirement of competitive and efficient electricity markets is access to reliable data and performance metrics." - in bold bluish text.
This brings to mind Simon Hoggart's Law of the Ridiculous Reverse "If the opposite of a statement is plainly absurd, it was not worth making in the first place."
The above has shown that it is a statement worth making. To attain success would be a matter of basic competence in data administration and information technology, manifested in a resupply of these data files in a form that is complete, comprehensive, coherent, and correct, thereby replacing a collection that is not. This should not be difficult nor require much effort. Your organisation has already re-supplied these data files, and done so more than once, so another re-supply is possible. This time correctly? Hopefully, not omitting October's Reserve Price file for 1996, etc. And not suppressing the timestamp information. Even presenting data correctly on the daylight savings changeover days should be easy, because you have already dealt correctly with this maddening annoyance, as in 2014 for example, though not for earlier files that were resupplied later on. Providing dates in the dd/mm/yyyy form is also possible, because file 201912_Reserve_prices.csv turned up in that style (which broke a system expecting to convert from the hyphen style), though it was later replaced by a file in the hyphen style.
Introducing a Z code (or whatever suits) along with the F code might be slightly more difficult, but it would enable the bizarre code numbers to be kept separate from the proper price numbers, and then the statistics will be good.
Humm. Is it worth saying that "Good statistics are good to have."?
There have been mass re-supplies of these data files in October 2015 and October 2018. Perhaps in October 2021?

Yours faithfully,

Nicky McLean"