Data: BMRS, EIC (station and unit sheets), DUKES

BMRS -> EIC: Registered Resource EIC code (output\*: outputDUKESToBMRS)

EIC -> DUKES: Names (strip: ‘\*’, ‘ ‘, ‘-‘, ‘'’, ‘’’, “\_”, “&”) (output**\***: outputDUKESToBMRS)

Generators -> Power Plants: STATION (output: outputBMRSUnitToStation)

**\*** These steps are combined for this output.

DUKES -> EIC:

Take closest capacity. If there’s a repeat (same EIC between the two items of equal capacity) then, while you could take the more recent time, it actually doesn’t matter, as we match with the EIC and they have the same one. What is important here, however, is that **in excel, make sure the BMRS dates are put in as ‘Short Date’s.** This is then read and outputted again internally to ensure they are all of the form ‘29/03/2005 12:00:00 AM’, so this form, or the short date format, should work. To correct, the code reads, then writes to the excel, then reads again. This was, the automatically applied change is made to standardise initial and edited excels for further processing.

**Confidence Score / Metrics (binary must be fulfilled and scores are from most to least important):**

(name selection) Binary: Name match.

(name selection) Binary: Closest capacity (if multiple name matches) / Exclusive match (closest with no competition). This feeds into the below score.

Score: % Difference of capacity (independent of competition). This is converted to an integer for the percent. 0 to 999.

Score?: Maybe incorporate a score based on the type of match. -15 to -1.

Score: Years of difference in installation year (exempt years if there’s just some generic value? eg. 1/1/2000). As the registration date might not be the installation date, and can be re-done, we’ll instead make the BMRS data = or later? Take “DINO” for an example (registered 2016, built 1984). -5 to 1.

Optional: Type (if possible, could just be ‘generation’). So give a yes (eg. Both coal), no (eg. One wind, one gas), or a neutral or NA value (eg. The BMRS is ‘generation’). Will need to map this. Maybe multiple lists of strings with ‘contain’ for both, to compare? I think using a score system may also be good, so maybe -8 if it’s a match, 0 if unsure, and 999 if they’re different. Then in the final sum this would get your score. DUKES solar is 999. -5 to 999.

*All these scores should go into the final verdict in some way. At the end give a binary for whether or not it should be included (if score < 12 it accepts):*

EG. Score = **10** -1 + 0 + 0, and if this is <10, then it is accepted. (Max accepted CapDiff% is 35.

**Future Note:**

To get a DUKES station output you must therefore combine all units into their stations (BMRS), then go from the BMRS station to associated DUKES station. So sum the BMRS units to stations, then go through the DUKES and for those with BMRS stations, take that sum.