# General feedback

I'm honestly over the moon that Ofgem and the wider UK energy industry are putting so much work into opening up energy data! This is fantastic! Thank you for all the work you are doing in this space.

One small, general bit of feedback: For me, the main motivation for opening up energy system data is to reduce carbon emissions. In my humble opinion, we simply will not get to a net-zero energy system without a step-change improvement in data sharing and interoperability. I was surprised that the Ofgem documents do not shout more loudly about the link between decarbonisation and digitalisation :)

# Data Best Practice Guidance

## Question 1: Recommended improvements to the Principles, Explanations, Techniques or Examples

### Principles 2, 3, 5 & 8. Use common terms & data standards...

I applaud the focus on industry standards for data and metadata!

BUT! In many cases, these standards either don't exist yet, or it isn't yet clear which standard should be used.

In my humble opinion, the ideal situation would be that industry converges on a single data standard for each type of data, and then the Ofgem guidance can point to those specific recommendations (perhaps by way of the "living" Supporting Information document referenced in paragraph 1.20 of the 'condoc'). But I do think *someone* needs to make specific recommendations of specific data standards! (Although this is already discussed in the supplementary information. But are we sure the Digital Framework Task Group will move fast enough?).

As the document stands at the moment, I fear energy companies will do the bare minimum, and choose whichever "data standard" minimises the work they have to do. Which may result in data *not* complying with common standards (because each org will cherry-pick different "standards").

Furthermore, we need to converge on a **common naming convention** for assets. In my humble opinion, using the same names for the same physical assets is far more important than choosing a common data format. I understand that BEIS are working on an asset register, which is great news. But maybe the Ofgem guidance could add a paragraph about the importance of using common *names* for things (although, today, we don't have those common names!). In my view, we need to urgently converge on common names for physical

assets. (For example, even within National Grid ESO, different systems use different names to identify grid supply points (GSPs), and there isn't a clean one-to-one mapping between these different GSP naming systems).

On the topic of standards for metadata describing energy datasets: Perhaps the guidance could mention that the EDVP Alpha work (being done by Icebreaker One, Hippo Digital, and Arup) is also looking at defining a metadata schema for describing energy dataset. To be blunt, I think Dublin Core is so generic as to be almost useless! When searching for a dataset, I want a machine-readable field stating which energy technology the dataset is about, and the geographical domain of the data, etc. This isn't covered in Dublin Core in a machine-readable way. I'd propose that, as a minimum, we should use DCAT or Schema.org:Dataset as the schema for metadata describing energy datasets.

Please require people to use the ISO 8601 date format (YYYY-MM-DD) in both metadata *and* data. (I think para 3.28 of the supporting information talks just about using ISO 8601 for the data, not the metadata).

In para 3.27 of the supporting info, please add the bullet point "Timezone for datetimes (e.g. UK localtime; or UTC)".

Regarding the example JSON on page 24 of the supporting information: Please use code highlighting like this. And, sorry to be a pedant, but each line needs to end in a comma for it to be valid JSON :) :


```
{
    "title": "Solar PV cost data",
    "creator": "Department for Business, Energy and Industrial Strategy",
    "subject": "solar power station [http://www.electropedia.org/], energy cost [http://www.
electropedia.org/]",
    "description": "Experimental statistics. Dataset contains information on the cost per kW
 of solar PV installed by month by financial year. Data is extracted from the Microgeneratio
n Certification Scheme - MCS Installation Database.",
    "publisher": "Department for Business, Energy and Industrial Strategy",
    "contributor": "Microgeneration Certification Scheme",
    "date": "2019-05-30",
    "type": "Dataset",
    "format": "XLSX",
    "identifier": "19c3b55d-32dc-4bb3-8141-32bb2175affc",
    "source": "https://certificate.microgenerationcertification.org/",
    "language": "English",
    "relation": "",
    "coverage": "Great Britain (GB)",
    "rights": "Open Government Licence v3.0"
}
```


## Principle 4

"The Licensee must ensure a point of contact is provided..." This point of contact should be a generic email address for assistance with data (such as data.support@company.co.uk) not an individual's email address (such as joe.blogs@company.co.uk). In practice, the generic email address may be set up to forward to a specific individual. But this forwarding can be changed easily when that person leaves the company or goes on holiday.

Para 3.45 of supporting info: It might be worth adding a sentence to say that data must be described using DCAT or Schema.org:Dataset in order for Google Dataset Search to find it.

### Principle 7: Ensure data quality maintenance and improvement is prioritised by Data User needs

3.11 says "...the Licensee must ensure that these [data] issues are rectified as soon as practicable". But, in some instances, it will be impossible to fix data issues (e.g. if the issue is a gap in a timeseries then the missing data may be lost forever). The guidance should acknowledge that some data issues will be impossible to resolve; but that the Licensee must use their "best effort" to resolve issues.

Also, in the explanation, I'd recommend teasing out the idea that some data is almost always better than no data. Orgs shouldn't delay publishing data because they feel the data isn't of sufficient quality. Nor should they be tempted to keep some data closed because they worry that maintaining the public dataset will be too great a burden. These ideas are touched on in the existing supporting documentation, but, IMHO, should be emphasised in the explanation.

### Principle 10: Store, archive & provide access to Data Assets in ways that ensures sustained benefits.

*All* data should be archived indefinitely. Data storage is dirt-cheap. The CEDA Archive already stores 18 petabytes of data. Historical data is essential for many use-cases (such as training machine learning models; or identifying trends).

#### Open data must be accessible by machine

**Explanation**: In the near future, we want automated services to be able to consume open data without human intervention. For this to happen, open data must be published using mechanisms which allow machines to easily access the data. Specifically:

- Do not require users to sign up to access open data. Open data should be available to download directly.
- Document how the file name will change if / when the data is updated (e.g. the end of the filename will always be the date formatted as YYYY-MM-DD).

## Question 2: Additional Principles and Explanations

### Principle: "Presumed open" data must be published with an off-the-shelf open data license

[or, instead of this being a new principle, perhaps this could be an additional paragraph of principle 11]

**Recommended explanation:** If data lacks an open data license then it's not clear what users can and can't do with the data. In other words, "public data" (without a data license) is totally useless for all practical purposes. When publishing data, please provide an open data license with the data. Please use an off-the-shelf license (such as CC-BY-4.0, CC0, or OGL) unless there is a very strong reason for creating a custom license. Even the slightest modification to the license text makes it a different legal document, and so each data user will need to ask a lawyer to review the license (which is prohibitively expensive and time consuming for many small teams), and the custom license is unlikely to be machine-interpretable.

## Question 4: Data literacy, skills, and governance

In general, I agree with Ofgem's treatment of these topics. I would add a few points:

1. To be blunt: The skills gap is at the top. There are lots of young data scientists who are passionate about mitigating climate change. The bottleneck is the senior managers who haven't been outside the energy industry for decades and who simply aren't aware of how things are done now :)
2. Energy orgs mustn't rely on contractors for their data work. The data skills must be in-house.
3. Governance of *data standards* is super-important to get right. IMHO, governance of data standards should be "presumed open". Use modern approaches to engage the community (for example, many Internet standards are discussed completely openly on GitHub). I'd steer away from infrequent, lengthy consultations. Instead, please consider ways that allow the community to dip into the conversation frequently, but where each interaction is relatively quick (such as spending 15 minutes a week posting a short comment on a focused discussion thread; rather than taking up a whole day writing a 10-page response to a 100-page consultation document!)

## Question 6: Views on guidance being used as Ofgem's data & digitalisation standards.

Yes, I think the proposed content of the DBP guidance and DSAP guidance are appropriate for use as Ofgem's data and digitalisation standards.

# Question 7: "Presumed capture and published of data"

I TOTALLY agree we must do all we can to enable greater real-time visibility of the energy system at all levels.

If I may be so bold, I think we need legislation that requires all new distributed energy resources (including domestic DERs) to report their real-time power generation / demand at high temporal resolution (once per second???). This can be done using low-cost communications technologies (such as standard WiFi). It's kind of bonkers that we're expecting the ESO and DSOs to balance the grid, even though they can hardly see what the grid is doing. It's like an air traffic controller where half the aircraft are stealth (and hence invisible to radar), and half the aircraft that are visible to radar have their radios off (so they don't send or receive any additional information).

Modern DERs such as PV inverters already send real-time data back to the inverter manufacturer's website. We need this data to be shared with (at least) 2 user groups: 1) the owner of the inverter and 2) the system operator. At present, this data is kept in a silo, and is remarkably hard to get access to (even for the owner of the hardware).

Telemetry from small (< 5 MW???) DERs should be "best effort". For these small DERs, I wouldn't worry too much about defining super-strict requirements for data quality and communications up-time. In a world of millions of DERs, we need to use statistical techniques that can handle imperfect data, rather than require that each DER installs comms hardware worth tens of thousands of pounds. And it's important to note that the so-called "gold standard" comms technologies connecting transmission-connected generation to the ESO control room is laughably flakey. A £10 single-board computer with a 4G cellular modem may actually be *more* reliable, especially if it uses a sane software stack that re-tries data transmission if there's a brief network failure!

And please don't be put off by energy industry veterans whinging about having to handle a gigabyte of data per day. Modern IT infrastructure can handle *petabytes* per day.