SUBAK

DATE: 20.02.23

State of Climate Data: Power Plants

Climate Data Research Initiative No. 1

Abstract:

12 data professionals working for leading global energy organisations and institutions were interviewed to support one of Subak's core objectives: to identify, aggregate and index high quality climate data. We wanted to understand their perspectives on the power plant data landscape - the who (has the data?), the what (are we missing?), and the why. Unsurprisingly, there's a lot to be improved upon. There are large data gaps in South and South-East Asia and a need for better and more persistent data assets. There is a clear need to build better relationships between data users and data producers. We highlight these findings in this report, the first of many to come, and offer some key conclusions for how we can get there.

Contents

Introduction

There is a need to identify, aggregate, and index climate data from around the world to make climate data visible, searchable, trusted, and connected, so it can be used to more rapidly ramp down consumption of fossil fuels. To support this aspect of Subak's mission, the data team has launched research initiatives into specific climate data topics that are important to get right. We seek to determine where people access data, the availability and quality of the information, what it is being used for, and how it is creating impact. Two primary research topics have been selected to start off this research initiative by the Data Cooperative members, entitled Climate Data Research Initiatives (CDRI). We plan to run five of these research topics per year, of which this report, focused on power plants, is the first.

Methodology

Energy is an extremely broad and complex research topic. Therefore, this report focuses on specific aspects, the location, operational status, installed capacity and generation output of transmission connected (the connected voltage of which varies but is typically above 33kV) fossil fuel power plants.

Desk based research was conducted using web search, industry articles, academic papers and social media to identify globally:

High profile, trusted, good quality data sources
Organisations and their active role in the collection and publication of data, their position in the topic space and their climate objectives
Significant or influential people active in the production or use of datasets to be interviewees.

We conducted 12, one-hour long, interviews between November and December 2022. Interviewees were selected through both existing relationships in the energy sector and search, as well as further recommendations that came about in interviews; a 'snowball' method. Interviews were predominantly with organisations and individuals based in the UK, US and Europe, although some organisations also had regional offices in regions outside of the Global North. This led to a global research scope, however, a broader scope of countries would have been desirable but was not possible due to unsuccessful outreach to participants in various regions. Interview questions were semi-structured and tailored based on the

type of organisation, the person being interviewed, and their respective remit both sectorally and geographically.

Interview responses and desktop research were then coded according to recurring themes (of which 15 were identified overall). This report presents the outcomes along with the associated, non-exhaustive, list of data sources and a data ecosystem map which mapped relationships between actors in the data space.

Finding 1. Give me the Data

Main Findings

- Energy producers all collect some level of generation and installed capacity data, but this is not always published.
- Interviewees have unorthodox workarounds to get data in places where traditional data sources are not available, but does not substitute for better transparency.
- Unique identifiers for power plants are a key recommendation to achieve better data standardisation.

1.1 Where the data is (and isn't)

Power plant's potential generation, known as installed capacity, and its net or gross generation over specific periods of time (e.g. hourly) were the most identified datasets used by individuals and organisations. These datasets are commonplace for every energy grid globally as grids require at least some baseline of data collection in order to be operated, meaning that in only rare cases did data not exist at all. Problems identified with these datasets and their respective in-country producers were that those who held the data a) did not wish to give it to the interviewees, b) assumed the cost of access (both financial or time-wise) was too high, or c) they did not know how to technically deliver it to the interviewees. In some cases there were a combination of the above factors. These three problems extend to data other than generation and installed capacity around power plants too, and were often cited as reasons for lack of access to data by data users.

1.2 Data workarounds

The lack of data led to occasions where interviewees had to find alternative solutions to such data gaps. While this led to higher margin of error with the estimated data, interviewees were nonetheless confident in their data and demonstrated their resilience and ingenuity in data use. The following examples give evidence to this point, but also establish key areas of data gaps.

Many interviewees highlighted significant geographic variation in power plants data, especially in South and Southeast Asia, and many have had to take innovative approaches to harvest or infer data from proxy sources. For example, energy transition think tank, Ember, worked around this by taking aggregated fossil generation data from these regions and estimating a disaggregation of fossil fuel breakdown (oil, gas, coal) to some success. In data-scarce regions, Catalyst Cooperative, a US-based data wrangler, purpose-built tools which crawl websites and find such data. This was similar to what Global Energy Monitor (GEM), also a non-profit data aggregator, did to find data specifically on Chinese power plants, sometimes drawing data on power plants from job listings on hiring websites, when other more traditional sources were not available.

GEM and other research organisations had an advantage in finding data-workarounds by employing data analysts of a wide variety of language backgrounds. Language was a key tool in acquiring hard-to-get data across many regions in the Global South and outside of the Anglosphere. This was echoed by the World Resource Institute, a global research non-profit, who in cases of data scarcity would investigate blog posts in Filipino or Malay regarding power plants in the Philippines or Malaysia respectively. Many interviewees also mentioned problems with data-scarcity in South and South-East Asia, even if they weren't able to find workarounds for this. Finally, in terms of Solar Power Plants, Ember,

which is currently collecting solar export data from China to estimate Europe's solar capacity data.

1.3 Data recommendations

Where data was available, there were still questions surrounding the dataset's irregularities, consistencies and timescales. For users, there was always potential for improvement in the data they were using, and a baseline that the data should meet. For generation data, hourly historical data should be available on a suitable time lag, or at the very least at a monthly level. Instrat, a Polish energy transition think tank, have demonstrated through their work this target should be 15 minutes in Europe and other regions where data is already published hourly. Users also struggled to find updates to capacity data more granular than an annual basis. Our interviewees expressed that it is vital that capacity data being published at the very least at a monthly level. In the case of Germany, there is 400-500 MW added monthly, meaning if data is only released annually then there's a gap of 5-6 GW. Equally, both generation and capacity data should be at a unit, or specific boiler level within a power plant, level. Many interviewees cited ENTSO-E as the best current example of a data producer like this for geographic, temporal, and fuel coverage, but this varies in what is realistic across the world. Ember expressed how two things come with increased granularity. For the energy markets themselves, demand and production can be better matched, leading to less curtailment as well as better integration of renewables into the power supply. For those concerned with energy markets, increased granularity of production and consumption can help identify trends that can be better used to devise more accurate policy solutions to decarbonising energy systems.

Finally, there is a strong demand for a baseline of data standardisation. This was both in the format in which data is published, but also the need for 'unique identifiers'. The former is more difficult to achieve, and interviewees consistently brought up the example that even if a standard format was decided on, it would be unlikely if its uptake was global as others would go on to create their own standard format. However, the latter serves as a solution to the former as a standard identifier for each power plant and subsequent boilers would be easier to standardise. This would mean that despite format, datasets could be aggregated far easier going forward, something which Climate Trace, a global coalition of data aggregators, mentioned would cut their workload significantly as all data is currently matched manually. This marks one of the most important suggestions for power plant data going forward.

Finding 2. The Need for Persistent Data Assets

Main Findings

- Market intelligence are often inaccessible due to pricing and place restrictive data licences on sharing
- Issues with funding for nonprofits as they create downstream, rather than direct, impact
- Coalitions of nonprofit data aggregators pave way for persistent data assets through a product-centric approach

There were several bodies producing data on power plants across the world. There are three main groups: government sources, not-for-profit organisations, and market intelligence companies. Government sources (often departments of energy) typically have primary data in a variety of formats and granularity. Not for profits typically aim to synthesise larger quantities of information for specific use cases, such as tracking energy generation, installed capacity, or decarbonisation policy. Market intelligence companies typically collect information on capacity and generation to inform investment decisions, hedging and market participation. Generally, market intelligence sources are private with restrictive clauses on resharing, whilst government and not-for-profit sources are public or open access, although all three models differ in terms of access. For this reason, market intelligence sources have both a high barrier to entry on price, but also (mostly) cannot be shared downstream.

Nonprofits which aggregate global government data sources were tasked with a time-intensive process for their individual researchers, activists and

policy professionals, of which comprised nearly half of our interviewees. It is also a process that is ongoing, as there is work that has to be repeated frequently without a hub resource. There was a high level of agreement among interviewees that comprehensive global data sources are necessary, but core funding for initiatives is difficult to come by as it creates downstream, rather than direct, climate impact. The WRI hosts a global power plants database, but is unfortunately no longer being funded. This is despite the fact primary users include, in one interviewee's case an adjacent government department to the one providing the power plant data. Some of the most successful data assets have been created with core funding, enabling them to be enduring or persistent. In contrast assets created using project funding start to become stale when a project closes.

An alternative model being pioneered by nonprofit data aggregator System Change Lab and Climate Trace is to develop a common baseline to build products on top of. This takes a product-centric approach but reduces overall rework across the coalition. An alternative could be an extension of open-source or community driven initiatives such as an addition to OpenStreetMap (where in a small number of countries this exists). Similar to Subak's data catalogue, this would help a broad range of stakeholders reduce the amount of duplicated work and help to align data across different applications.

Finding 3. Relationships Between Providers

Main Findings

- Transactional relationships are common between data users and primary data providers, leading to difficulty in fulfilling user needs
- User research is done, but not as much as data producers would like to
- Relationships between advocacy groups and not for profit data aggregators have been very successful through a product-centric approach to data

Interviewees told us, almost without exception, that building relationships with non-market intelligence providers of primary data sources was extremely difficult, and where money does not change hands there is often no relationship at all. These primary data providers include government sources, transparency platforms run by industry groups, network operators, and others. A common theme with these providers was that they were not taking a data product approach to fulfil specific user needs, rather they publish data for statutory or regulatory purposes. User forums did exist for certain data sources, ENTSO-E provides somewhat regular user sessions for new features to be suggested and responded to, although this is generally not for upstream data users. Even so, the ability to serve new user needs is limited by external factors such as having to aggregate data from European system operators. To build these relationships the energy data landscape would benefit from mirroring the software world, where relationships between power users and developers (through account management, forums, or even just a point of contact) are extremely important to developing new features and solving critical challenges.

Aggregators and not for profits we spoke to (Systems Change Lab, Catalyst Coop, Instrat, etc.) were actively doing much more extensive user engagement for their data than the primary data providers. Building up long term relationships in tandem with short term user research is key, many interviewees that provide aggregated data perform extensive user research to understand what use cases they were serving, although most would like to spend even more time doing so. The best relationships were in fact between advocacy groups and not-for-profits working in similar spaces, particularly where there had been funding to deliver specific joint projects, or where there is a general need for a specific piece of data. An example of this is GEM's coal plants tracker, where there is a close knit group of coal phase out advocacy groups that need access to data. This demonstrates the advantage of having relationships with user groups and understanding their needs, and the need for this to change, as the ability to serve user needs is limited by the primary data provider.

Summary of Main Findings

FInding	Call to Action and Recommendations
Energy producers all collect some level of generation and installed capacity data, but this is not always published.	Liberate data and increase transparency in data poor regions, use Subak accelerator and fellowship to support this particularly in South and South East Asia.
There are high barriers for processing data impacting the tracking of shifts in generation and capacity data from power plants.	Introduce ‘global unique identifiers’ for power plants in the form of a standard code for each power plant and subsequent boiler.
Market intelligence are often inaccessible due to pricing and place restrictive data licences on sharing.	Build relationships with market intelligence providers to increase transparency on the data held.
Open access or public datasets run by nonprofits have a tendency to become stale and cease updates due to funding for nonprofits creating downstream, rather than direct, impact.	Create persistent data assets open to all users on a community / coalition basis (similar to that of the Climate TRACE and Systems Change Lab model).
Data users lack any real relationship with primary data producers.	Increase collaboration on use cases and participating in user groups, especially using the ‘power user’ model present in the software industry. Use the Subak Data Cooperative to facilitate this.

This was a short project to explore the landscape of power plants data for the climate movement. If you would like to add to this report or fill any gaps please reach out to us at datacooperative@subak.org.

Thanks

Matt Ewen, Nicolas Fulghum, Sam Hawkins
Durand D'Souza
Wojciech Przedlacki
Jacob Bieker
Christina Gosnell, Bennett Norman
Mason Inman
Christy Lewis, Lekha Sridar
Gavin McCormick
Sultan Aliyev
Logan Byers
Tappan Parker
Abhayraj Naik, Ambar Nag
Johannes Friederich
Matteo De Felice

Get in touch

datacooperative@subak.org
subak.org
data.subak.org