
Energy sector data: Suggestions for improving data quality and usability

BROOKINGS INDIA

QUALITY. INDEPENDENCE. IMPACT.

Prepared By:
Sahil Ali (sahil@brookingsindia.org)
Rahul Tongia (rtongia@brookingsindia.org)

DISCUSSION NOTE

BROOKINGS INDIA

SEPTEMBER 2018 | 092018

This document is the outcome of collective brainstorming sessions around energy sector data and how to improve its reliability and usability that were held by the Energy Network of Scholars, an informal group of scholars, practitioners and experts on issues of energy and sustainability in India. The network serves as a platform for engagement, enabling better research, sharing of ideas and lessons. The network also provides a space for coordination between the research community and policy stakeholders to improve effectiveness of research and policy-making. Brookings India, New Delhi serves as the secretariat for the network.

Introduction

More energy and power sector data is available in India than ever before, especially through a combination of websites and portals. We propose that issues in locating, procuring and acquiring data be ironed out for researchers and practitioners to conduct more evidence-based policy research and contribute to the national discourse. Researchers use energy and related data from a variety of government departments and ministries, but its synchronisation and accuracy vary across the respective departments, making the prescriptions subjective to the source of information used. Moreover, in the absence of critical data, researchers end up relying on rules-of-thumb passed down from year to year, with no basis for re-evaluating the underlying assumptions. This may lead to an unwarranted convergence in the outcomes of analyses, in that they may reinforce a bias against changing dynamics on the ground as they are unable to fully capture them.

At the same time, there is a spurt of public data in the form of models, portals and other initiatives by ministries, departments and enterprises in the public sector. While the necessity of quality control in the form of a centralised information management system in the energy sector has never been stronger than before, we realise it is a massive institutional undertaking that requires a considerable amount of time to plan and implement. Therefore, we propose that as a first step, we have higher standards and consistency in the availability and reporting of public data on energy.

Background and motivation

This note offers suggestions for owners of energy sector public data to update their methods and reporting so that the data they disseminate can be more meaningful to researchers and other stakeholders for evidence-based analysis.

A number of new initiatives have been taken in the form of online government portals, including some that give real-time data on demand-supply, access, distribution and infrastructure, such as the NITI Aayog's India Energy Dashboards, GARV Dashboard, MERIT, URJA, UJALA, GARV, GTG India, etc. Each of these focusses on specific aspects: from a generalised view on energy demand and supply and its constituents, to performance of DISCOMS, to status on electrification and household access, to disbursement of energy efficient lighting, to more technical aspects of financial and operating performance of power plants and renewable integration.

The spurt in the reporting of such data to supplement traditional reports (often annual) from the Central Electricity Authority (CEA), Ministry of Coal (MoC), Coal Controller's Office (CCO), Ministry of Petroleum and Natural Gas (MoPNG), Ministry of New and Renewable Energy (MNRE) etc. is necessary, useful and welcome in analysing past and current performance and coming up with evidence-based policy research. This can be further improved with some standardisation and offering choices to the user to customise the data, unpack the assumptions and study trends.

We have attempted to highlight some suggestions for improving data dissemination, and assess how some of these can be incorporated in the ongoing efforts. We also give examples of websites/data compilations already following the particular suggestion/s.

Suggestions for improving data dissemination

1. Make data available for downloading

Several dashboards have great data but are geared towards online visualisations. Downloading allows researchers and other stakeholders to undertake their own efforts in analysis and visualisation, reducing the effort on data owners to try and “show it all”.

Examples: PPAC’s Oil and Gas Marketing Data¹, NITI Aayog’s India Energy Portal²

2. Provide historical data (as best available)

Several dashboards give great data but are limited to a few days or months. It would help to have archives available online. As of now, some scholars and institutions run scraping tools to get data regularly. This is a step that can lead to errors and is avoidable by providing archived data. The use of archives also allows historical data to reflect any corrections/updates, which instantaneous data may not capture properly.

Examples: MoSPI’s Energy Statistics³, CCO’s Coal Statistics⁴, MoPNG’s Petroleum and Natural Gas Statistics⁵

3. Enable long-term archiving with appropriate nomenclature

Online repositories should standardise nomenclature of data sets, files, etc. to enable easy identification and differentiation. Instead of calling a file “ouput.xls” or “Annual_Report.pdf” it would help to have these segmented into folders and have names such as “load_ouput_20.09.17” or “Annual_Report_2016”. This is particularly important because when a researcher lists an Annual Report as a reference in a study, the web address may lead to a different version of the Annual Report.

Examples: CEA’s Load Generation and Balancing Reports⁶, MoSPI’s annual publications⁷

4. Make data available in the right format(s)

i) There are two parts to this. In some cases, there is no download available, and so just having HTML data is difficult to work with. There can be good visualisations online, but downloadable data is amenable for analysis. Second, where data are available for download, they should be more than PDFs. In fact, some PDFs are images, despite the underlying data being generated in numerical/database/excel format. Excel formats are a neutral mode, since databases can be proprietary.

ii) A good starting point, especially in case of online portals, would be to archive data in the periodicity chosen into downloadable excel sheets. This will supplement (i) and enable more granular analysis.

iii) Where hard copies are available, have soft copies available as well, which avoids printing and mailing costs. Data providers can continue to charge for such data as required, but in the long run, all public data should be available online and free.

Examples: PPAC’s Petroleum and Natural Gas Price Statistics⁸, NITI’s India Energy Portal

5. Give multiple options and end-user-controlled user interfaces for visualisations

To complement real-time visualisations, which likely will be data-owner driven, having a set of choices for data visualisation will help users. Examples range from the ability to choose time periods, locations, granularity, show comparisons, etc.

Example: NREL and MoP’s Greening the Grid⁹

6. Clarify/Standardise/Improve the headings and meta-data

i) This starts with having units properly listed, but extends to making formulae transparent when numbers are shown. CEA does this in places, where they show, for example, column $P = [F - G]/B$. In the long run, we suggest standardisation of headings, even to the extent of .xml schemas. We (as scholars/researchers) can help the government with such an exercise.

¹ Full link: http://ppac.org.in/content/5_1_ReportStudies.aspx

² Full link: <http://www.indiaenergy.gov.in/edm/#electricity>

³ Full link: http://www.mospi.nic.in/sites/default/files/publication_reports/Energy_Statistics_2017r.pdf

⁴ Full link: https://www.coal.nic.in/sites/upload_files/coal/files/coalupload/Provisional1213_0.pdf

⁵ Full link: http://petroleum.nic.in/sites/default/files/pngstat_0.pdf

⁶ Full link: <http://www.cea.nic.in/annualreports.html>

⁷ Full link: http://mospi.nic.in/download-reports?main_cat=NzI2&cat=All&sub_category=All

⁸ Full link: http://ppac.org.in/content/149_1_PricesPetroleum.aspx

⁹ Full link: <https://maps.nrel.gov/IndiaGTG/>

- ii) Several times researchers have to compare between different sources to obtain a sense of the information that is being conveyed. In case of MoSPI and CEA data on electricity generation, the scope of the exercise (whether captive generation is included in the numbers or auxiliary consumption is netted out) is unclear to compare and it can at best be inferred that the higher numbers in one case include captive generation.

Even so, it is highly likely that it does not get fully accounted. With the spurt of new data in the public domain, we should be careful to ensure there is no difference between what we intend to convey and what we end up conveying. Such standardisation must be an integral part of the data maintenance protocol on new data sources such as tools and dashboards.

Example: CEA's General Reviews¹⁰

7. Be transparent in data sources and references

- i) This not only helps in tracking down primary data and sources, but even to link to underlying methodologies. A weblink is an ideal reference, except in the case where a link may not be available. In both cases, detailed information is helpful. For example, instead of just saying "Source: CEA" it would be helpful to explicitly state which CEA report or document was the source.
- ii) References are especially important for webpages, online reports, etc. This is sometimes done, but often the presentations by staff/officers found online lack this information.

Example: NITI's India Energy Portal

8. Share methodologies and assumptions

- i) To the extent possible, through a note or additional explanation, clarify relevant assumptions and methodologies. For example:
 - (1) Is the number for gross or net?
 - (2) Where is this number calculated (for what location or node)? At the unit, plant, bus-bar, boundary meter, etc.?
 - (3) Is a particular number instrumented, reported, or calculated via some methodology?
 - (4) As an example, for stating power shortfall, it should be specified if it is instrumented or calculated.
- ii) This is particularly important for sources that intend to convey a systems perspective (especially models and tools) or draw from multiple assumptions (empirical, thumb-rules, etc.). This should be clarified and stated upfront, along with the units of analyses, any rounding off, etc. For example, it would help if NITI's Energy Dashboard also provided a basis for arriving at consumer category-wise connected load.

In relation to (4) above, the methodology and nomenclature should also be standardised across the board. So the Aggregate Technical & Commercial (AT&C) and Transmission and Distribution (T&D) losses reported in MoSPI and CEA should conform to their strict definitions to enable comparisons.

Example: NITI's India Energy Security Scenarios¹¹

9. Give a date for all materials {both dissemination date and relevant data period date(s)}

- i) For reports and data sets, the date of release and the date of underlying data are both helpful. For presentations and reports, date is important as the material presented otherwise can lack context. This is particularly useful for archived data.
- ii) Sometimes, presentations made by government officials at various forums are made available online. These often contain previously unpublished data but nevertheless can be traced to a credible source. These sources should be accompanied with dates so that it becomes easier to use them for analysis.

Example: Indiatat¹²

10. List a point of contact for data queries

There should be a means to contact someone (who may or may not be the data "owner") for clarifications and queries.

Examples: NITI's India Energy Security Scenarios and India Energy Portal

¹⁰ Full link: <http://www.cea.nic.in/reports/others/planning/pdm/contents.pdf>

¹¹ Full link: <http://indiaenergy.gov.in/>

¹² Full link: www.indiastat.com