

OPENGRID PLATFORM COMPONENTS – DATA HUB

One open platform, powerful interoperable components

OpenGrid's foundation is robust, trusted, open data, with task-ready applications layered on top. Each component was shaped by conversations with the grid planning community and user research on modeling pain points, and each deep-dive sets out what we heard, the outcome we're working toward, and a draft product concept for the community to react to.

Each deep-dive draws on more than 65 conversations with over 55 participants globally across the grid planning ecosystem, conducted in early 2026. We group participants into six personas along two lines: whether they produce analyses or use them, and whether they sit inside or outside the organization whose decisions are at stake. Each group brings different needs, and together they shape how every OpenGrid component is designed. To protect confidentiality, interviewees are identified only by organization type. The draft product concepts are starting points for OpenGrid's Technical Committee and working groups to refine, and we expect scope, priorities, and naming to change as that work happens.

This summary describes the current vision for each component of the OpenGrid Platform. They are not definitive specifications and are still subject to ongoing refinement. Each component will be shaped by the Technical Committee and a working group that is open to anyone who wants to take part. We expect scope, priorities and naming to change as that work happens. Comments, corrections and alternative proposals are welcome. We are publishing at this stage deliberately, so that direction can be shaped in the open rather than presented once it is already fixed.

Data Hub

Pain points, from user research

Modelers spend more time finding, reconciling, and maintaining data than analyzing it, and there is no shared place to check where a number came from or how good it is. Every group we interviewed raised it, and it drew more explicit support than any other component.

What we heard

- A US renewables developer spends millions a year on multiple data sources and still has to build its own datasets. Even a dedicated team of 7 to 8 struggles to keep data current, with multiple sources reporting different numbers. About 50 percent of the developer's modeling time goes to reconstructing how other stakeholders are likely to model the same system. If financiers trusted standardized open data, developers could stop convincing 300+ lenders individually.
- A US IOU said maintaining and maturing data takes more time than running analyses. RTOs publish models on different timelines and formats, with the same substation spelled differently across datasets. Most network data is CELL, and data sharing is still sometimes faxed.
- A large European DSO has about 20 people on data quality across millions of customer connections and called it a "fight against windmills." The root cause is 20+ years of records entered by staff who didn't know the data would later feed grid models. The worst errors are in customer contract data, such as connected renewables, larger customers, and heat pumps.
- A US RTO reported data that is 99.9 percent accurate, but the remaining 0.1 percent creates significant problems (identifying errors is a "needles in haystack" problem). Economic data has no compliance standards and no audits, and smaller members fill gaps with purchased datasets or similar-facility assumptions.
- A commercial software provider named entity resolution across datasets as the highest-value data problem. Fuzzy matching is constant, some portion is always manual, and one state model failed QA due to database mismatches.

- A US state PUC saw network upgrade costs for a major transmission proposal grow far beyond initial estimates once detailed analysis was done. The commission had no independent reference data on transmission costs to catch it earlier.
- An open resource planning initiative noted that storage ELCC studies yield 3 percent to 90 percent depending on assumptions, which are often opaque, making data transparency one of the lowest-hanging fruits.
- Where operational data can't be released, an academic research team built a synthetic test system of its country's grid from public data that matched historical curtailment patterns, but because it is synthetic, it doesn't yet carry sufficient weight in formal regulatory proceedings. The team sees international backing and coordinated validation for nationally developed and open-source network datasets as a way to build that standing.
- A Southeast Asian think tank said data held by the state utility is available only as PDF.

What users asked for

- A US RTO asked for quality grading and a predictable refresh cadence in addition to coverage: "How do you know what's quality?"
- A US IOU said consistent common base models would be transformational, even if organizations build their own projections. It favored full-resolution estimated networks that users can fix over time over zonal approaches that are "too black boxy."
- A US industry association asked for commonly referenced open datasets as reference points, especially 8760 load shapes, cost, performance, and avoided costs, and a US state PUC wanted independent reference cost data.
- A US university asked for a centralized pointer to data resources, with API access for large weather datasets, and a US IOU recommended starting with generation data, since it is least likely to be CELL.

Vision

Finding, trusting, and preparing data for a grid model is fast and repeatable, and the data behind any major public study can be traced and checked.

A modeler anywhere can find the best available data for their system, see where it came from and how good it is, and get it into model-ready form quickly and the same way every time. Data preparation shrinks from the largest share of a study's effort to a small one, and that time goes back into exploring more scenarios, understanding the results, and engaging with key stakeholders. Every dataset shows its sources, how it was validated, how current it is, and how it relates to other data describing the same assets.

Utilities, labs, and researchers can publish the data and assumptions behind a study, and eventually the model itself, in a form others can inspect, cite, and replicate. When a utility and an intervenor file competing studies, or a lender reviews a grid plan, the inputs can be compared against a common reference, so the debate can focus on the assumptions that differ. Restricted and sensitive data can contribute to trusted analysis without being exposed, and where real network data can't be shared, open reference models of real systems can offer a realistic starting point for training, testing new methods, and indicative analysis. Gaps in data availability are visible, so funders and data builders can see where new data would matter most, and the organizations that maintain open data can see how their work is used and communicate their impact.

Draft product concept (for discussion)

The Data Hub is a registry for grid model input data and planning assumptions, starting with the data behind key studies. It helps a modeler find data fit for a given study, lets anyone publish a study's underlying data and assumptions so others can inspect them and rebuild the analysis in their own tools, and could, in later phases, register the models themselves and supply reference grid models for tool development, training, and indicative analysis.

It would start by registering source data, planning assumptions, model-ready data packages, and studies, with each study tagged to the data and assumptions it used. Registering the models themselves could follow once the initial foundation works. Around these, it would support four kinds of use:

- **Finding and selecting data.** Search a data domain for usable datasets, compare a dataset's quality attributes against a published rubric, see which studies used each dataset, resolve the same component across datasets that name it differently, and pull a slice of a large dataset without downloading all of it.
- **Publishing.** A utility, lab, or researcher publishes the data and assumptions behind a study in a standard format, validated at intake and issued as an immutable, versioned, citable release. Assumptions can be public while network detail stays restricted, with the restricted layer releasable bilaterally to a named party. A data provider registers a dataset once and sees every study that references it.
- **Evaluation.** A regulator sees which assumptions differ most between two studies, and by how much. An intervenor can change one filed assumption and test whether the difference is material before asking for the commission's time.
- **Stewardship.** Funders and data builders see where data is missing, by domain and geography.

A notional sequence, with each phase gated on the one before:

1. Develop data quality rubrics for each grid modeling data domain, and assess existing datasets against them.

2. Build the registry: intake validation against the schema, versioned releases, component identity resolution (shared with Open Standards & Schema), and tooling to compare study assumptions. Use an existing archival and preservation backend where possible.
3. Explore the feasibility of producing reference datasets carrying topology, nodal load, and component identifiers, built from open data, synthetic data, or both, and released geography by geography.

The Hub would point to data that partners already host well and archive only what has no good home, such as the data and assumptions behind filed studies and their versioned releases.

Open questions include whether the Hub should let users run analyses on the data it holds, much as Hugging Face and Kaggle pair hosted datasets with compute, so that a regulator, for example, could rerun a filed study without its own infrastructure. Others are whether assumptions filed against a utility's detailed network can be applied to a public reference network when the two don't match bus for bus, and whether OpenStreetMap-derived reference grids create share-alike licensing constraints for adopters.

Boundaries. The Data Hub does not define schemas (Open Standards & Schema), transform data (Translator), or run models itself; any in-place analysis would use the Compute & Solver layer. Access to restricted data stays with its owners, and the Hub can show exactly which data a study used and whom to ask for them, so access requests are faster and better targeted. It also does not duplicate data hosting that ecosystem partners already provide.

Who We Heard From

Our platform components were shaped by more than 65 conversations with over 55 participants across the grid planning ecosystem. We group them into six personas along two lines: whether they produce analyses or use them, and whether they sit inside or outside the organization whose decisions are at stake.

Producer Side — People Who Run the Models

Transmission-System Users. Engineers and planners at ISOs, RTOs, TSOs and vertically integrated utilities who model the bulk system, often chaining three to five specialized tools through a single study.

Primary Distribution-System Users. Operators and modelers of distribution networks, from Europe's largest DSOs to small municipal utilities with one- to five-person teams.

Point-Source Modelers. Developers, hyperscaler procurement teams and community energy buyers who model a single point of grid interaction, with analyses that must be defensible enough to underwrite capital commitments.

Consumer Side — People Who Use the Results

Internal Decisionmakers. People who use model outputs to allocate their own organization's capital and must trust analysis they did not run.

External Evaluators. Regulators, intervenors, advocacy organizations, think tanks and data-infrastructure bodies who use outputs to scrutinize decisions made by better-resourced entities.

Straddling Both

External Primary Modelers. Analysts at public utility commissions, reliability councils, think tanks and capacity-building organizations who run their own models to examine decisions made by others.