D2.5.1.2

MAPPING ENERGY DATA GAPS AND REUSE BARRIERS: EVIDENCE AND IMPLICATIONS FOR NFDI4ENERGY

13.01.2026

MAPPING ENERGY DATA GAPS AND REUSE BARRIERS: EVIDENCE AND IMPLICATIONS FOR NFDI4ENERGY

Christina Speck (https://orcid.org/0009-0004-7386-8334)[1], Thimo Schulz (https://orcid.org/0009-0009-5548-9898)[1], Christof Weinhardt (https://orcid.org/0000-0002-7945-4077)[1]

[1] Karlsruhe Institute of Technology, Institute for Information Systems, Kaiserstraße 93, 76133 Karlsruhe,

Published by

Karlsruhe Institute of Technology, Institute for Information Systems, Kaiserstraße 93, 76133 Karlsruhe

Acknowledgements

The authors of this article have used various preparatory works from the NFDI4Energy to create this portrait, and references have been made where possible. Thanks to all those who are not named.

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. The work was funded by the German Research Foundation (DFG) – 501865131 within the German National Research Data Infrastructure (NFDI, www.nfdi.de).

License

Table of Contents

General Information 3

Summary 3

Deliverable within NFDI4Energy 3

Deliverable 4

Introduction 4

Research Background 5

Research Design 9

Results 10

Implications for NFDI4Energy 15

References 16

General Information

Summary

The energy transition is increasingly a data challenge: models, monitoring, and public narratives all depend on datasets that are not only available, but usable – granular, timely, well-documented, and interoperable. Yet even in the comparatively data-rich European and German landscape of official statistics, regulated transparency platforms, registries, and research repositories, stakeholders repeatedly run into blind spots and frictions. The practical problem is not just “non-FAIR data,” but also data that are missing in different ways: not collected at the needed resolution, inaccessible due to confidentiality or technical barriers, or hard to trust and reuse when figures diverge across portals and processing steps are not traceable.

This deliverable offers a structured, evidence-based “gap map” that helps practitioners and researchers see where the bottlenecks sit – and what kinds of remedies they imply. It synthesises three inputs: researcher surveys on missing reference datasets, and on missing elements and missing values in industrial datasets, and contextual evidence from both the Destatis colloquium in 2024 and the SciCAR conference in 2025. Using a pragmatic domain structure and a typology of gap mechanisms, the report consolidates heterogeneous observations into a coherent picture of where the energy data ecosystem works well, where it breaks down, and which infrastructure levers can accelerate progress.

Deliverable within NFDI4Energy

This deliverable advances NFDI4Energy by consolidating consortium evidence – including surveys and internal inputs as well as TA2-related stakeholder contexts such as the Destatis colloquium and the SciCAR conference – into a structured map of energy-research data gaps across domains and missingness mechanisms. By translating dispersed stakeholder “pain points” into infrastructure-relevant problem classes, it provides an auditable basis for prioritising NFDI4Energy actions (e.g., reference dataset curation, harmonised metadata and provenance standards, and tooling that lowers practical reuse barriers) while clarifying where progress depends on coordination with official statistics, regulators, and platform operators.

Deliverable

Introduction

Energy system modelling through computer-based numerical methods is very common in energy research (Hirth, 2020). Energy policy is heavily relying on model assessments for policy advice (Hirth, 2020; Pfenninger et al., 2014; Süsser et al., 2021), and the private sector uses comparable analyses for operational tasks such as energy trading and the dispatch of power stations (Hirth, 2020). At the core of these applications is the data going into models and the data produced as outputs for interpretation (Pfenninger et al., 2018). Accordingly, energy system analysis is increasingly data-intensive, relying on heterogeneous datasets spanning physical infrastructure (generation, grids, storage), operational time series (dispatch, flows, outages), markets (prices, bids, transactions), and socio-economic or behavioural variables that drive demand and technology adoption. At the same time, the research community faces persistent challenges of limited transparency and reproducibility in modelling and data workflows (Wiese et al., 2019), which can hinder scientific scrutiny and policy credibility. These challenges have been widely documented in the energy modelling and open science literature, which emphasises that verifiable results require traceable datasets, explicit assumptions, and reusable computational workflows (Morrison, 2018; Pfenninger et al., 2018).

Against this background, open modelling has gained momentum, combining open-source models with openly available data and traceable research outputs (Hirth, 2020). From a research data management perspective, the FAIR principles provide a widely accepted normative benchmark: data and associated artefacts should be findable, accessible, interoperable, and reusable by both humans and machines (Wilkinson et al., 2016). In energy research, however, FAIR-aligned reuse is often difficult because confidentiality constraints, fragmented responsibilities, and heterogeneous publication practices complicate discovery, access, and harmonisation even where data exist (Wiese et al., 2019).

Importantly, the practical problem encountered by stakeholders is not limited to “non-FAIR data,” but also includes data that are missing in different ways. A FAIR diagnosis presupposes that data exist and can be stewarded into reusable form. In contrast, stakeholders frequently report (i) data that are not collected or not released in usable granularity, (ii) data that exist but are not comparable because definitions and classifications differ across sources, (iii) data that conflict without resolvable provenance, and (iv) data that cannot be interpreted or trusted because assumptions, processing steps, or responsibilities remain opaque. Making these mechanisms explicit is essential because they imply different remedies: some require upstream action (measurement, reporting, publication practices), while others can be addressed through stewardship, standardisation, metadata, and interoperable access pathways.

This report addresses these issues by identifying data needs based on common data gap categories in energy research, with emphasis on primary data – data released in raw or minimally processed form to enable independent validation, reuse, and methodological transparency. Building on NFDI4Energy surveys and external stakeholder evidence from conferences involving diverse data users and producers, we develop a structured inventory of gap types and map them across core energy-data domains. This mapping provides two immediate benefits. First, it translates heterogeneous answers to “what data is missing?” into a consistent picture of why data are missing (or effectively missing) across domains, thereby enabling synthesis across researchers, public-sector stakeholders, and intermediaries such as data journalists. Second, it supports prioritisation by linking domains to gap mechanisms: it clarifies which needs can be mitigated through platform and stewardship services and which require coordination with upstream data holders. The purpose of this report is to provide a concise, structured overview of reported energy data gaps and reuse barriers and, on this basis, a first evidence-informed orientation for NFDI4Energy's data priorities. This mapping serves as a starting point for subsequent expert validation of gap severity and of which gaps are addressable through research data infrastructure versus requiring coordination with primary data holders.

Research Background

The energy research data ecosystem as a fragmented, intermediary-rich landscape

Energy research draws on datasets that differ in provenance, measurement intent, temporal resolution, spatial granularity, and governance. In practice, this heterogeneity often shifts effort from reuse toward generation or reprocessing: a survey among energy researchers reports that many primarily use generated data rather than reusing existing datasets (Arndt et al., 2022). For a report focused on identifying and classifying data gaps, the key implication is that stakeholders' statements of "missing data" frequently refer to different underlying mechanisms – ranging from genuinely uncollected information to data that exist but are difficult to locate, compare, or interpret reliably (Wiese et al., 2019).

A compact way to understand why these mechanisms recur is to distinguish three common production and dissemination contexts. First, official statistics and public administration publish energy-relevant indicators under legal mandates and disclosure constraints; this often supports comparability at aggregate level, while granular access may be restricted (European Parliament and Council, 2022; European Union, 2019). Second, regulated operational and market transparency provides time series and registers needed for system operation and market integrity; these sources can be highly relevant for analysis but may require specialised handling and can exhibit usability or documentation limitations for research purposes (European Parliament and Council, 2011; Hirth, 2020). Third, research and non-profit intermediaries curate and repackage data from these sources to make them easier to use; Open Power System Data is a prominent example of adding value through checking, processing, documenting, and publishing commonly needed electricity datasets (Wiese et al., 2019).

From a user perspective, these contexts are accessed through a practical "portal layer" rather than through producers directly. Official statistical portals (e.g., GENESIS-Online, Eurostat) disseminate standard aggregates; sector and registry portals (e.g., MaStR, ENTSO-E/-G, ACER/REMIT-related portals) provide operational or compliance-relevant data; open government portals (e.g., GovData.de, data.europa.eu) aggregate public-sector datasets; and research repositories (e.g., institutional repositories, Zenodo) archive project outputs. This portfolio of portals increases nominal availability but also creates dispersion: users may face multiple versions, heterogeneous formats, and uneven documentation. While intermediary curation can mitigate these issues and is closely aligned with FAIR-oriented expectations around findability and reusability (Wiese et al., 2019; Wilkinson et al., 2016), the distributed landscape makes it difficult to assess "what is missing" without a structured approach that distinguishes different gap mechanisms across domains.

Typical energy data domains as modelling input classes

Given the distributed portal landscape described above, stakeholders' "missing data" statements point to many different objects – time series, registers, statistics, and curated research datasets. To synthesise these statements across sources and stakeholder groups, therefore, a simple organising layer is needed that indicates which part of the energy system a reported need refers to. We use a pragmatic set of empirical data domains that reflects the recurring input areas of energy-system modelling and empirical evaluation. The domains below are non-exhaustive and serve as a structuring device for locating gaps.

1. **Electricity generation & grid.** Generation capacities and time series, system operation indicators, and grid infrastructure (assets, topology/parameters) are core empirical inputs for dispatch, adequacy, and expansion studies. Modeller-oriented curation targets these inputs explicitly (Wiese et al., 2019), and open European model datasets operationalise them through network and system representations (Hörsch et al., 2018). This domain is also embedded in the European transparency environment for operational electricity data (European Parliament and Council, 2013).

2. **Household energy consumption.** Residential demand by carrier and related indicators are standard empirical inputs for monitoring and model calibration (Swan & Ugursal, 2009), typically anchored in official sectoral energy statistics (European Union, 2023, 2025a).

3. **Commercial and service sector energy consumption.** Services are treated as a distinct sector in final energy consumption reporting and energy balances, providing a standard empirical basis for sectoral monitoring and analysis (European Union, 2023, 2025b).

4. **Industrial energy consumption.** Industry is a standard domain in sectoral consumption reporting and is frequently needed for intensity analysis and transition monitoring (Fujimori & Matsuoka, 2011). Official sectoring provides the baseline empirical structure, often supplemented by subsector detail in applied work (European Union, 2023, 2025c).

5. **Energy storage systems.** Storage capacities and asset attributes form an empirically trackable infrastructure class that is increasingly relevant for flexibility analysis (Schill & Zerrahn, 2018). Storage is explicitly covered in administrative registration practices, indicating the availability (in principle) of registry-based empirical inputs (BNetzA, n.d.).

6. **Energy market data (prices and transactions).** Wholesale price time series (day-ahead, intraday, balancing) and traded volumes are standard empirical inputs for market analysis and for validating modelling assumptions (Ventosa et al., 2005). In Europe, key system- and market-relevant time series are disseminated under the transparency regime (European Parliament and Council, 2013). Additional transparency signals relevant to integrity and transactions are shaped by REMIT-related disclosure practices (European Parliament and Council, 2011).

7. **Spatially resolved energy data.** Many network and regional analyses require geolocated assets and spatial representations of infrastructure and demand. Benchmark datasets for grid analysis illustrate this need for reusable empirical network data (Meinecke et al., 2020), and spatially resolved open model datasets show how such inputs are operationalised in modelling workflows (Hörsch et al., 2018).

8. **Climate and energy policy databases.** Policy evaluation and scenario work often require empirical representations of enacted policies and legal commitments (Nachtigall et al., 2024). Structured policy corpora and databases provide machine-readable access to policy texts and attributes used in comparative analysis (e.g., IAE, Climate Policy Radar).

9. **Societal attitudes towards energy and climate.** Survey-based measures of perceptions, preferences, and acceptance provide empirical context for transition feasibility and policy analysis (Verschoor et al., 2020). Large survey infrastructures supply standardised indicators reused in energy and climate attitude research (*European Social Survey*, 2017).

This domain structure serves as the organising device for reporting and comparing stakeholder needs in the Results section. It is deliberately non-exhaustive but designed to be stable enough to locate where missingness concentrates and to enable consistent cross-domain interpretation.

Gap types as a coding frame for "what is missing" across domains

To consolidate heterogeneous stakeholder statements in an analytically transparent way, the report distinguishes not only which datasets are reported as missing, but also how they are missing. For this purpose, we adopt a six-part gap typology that differentiates between (i) primary information not collected, (ii) limited comparability, (iii) inconsistencies across providers, (iv) accessibility barriers, (v) non-partisan concerns, and (vi) missing transparency of assumptions and methods (McWilliams, Ben et al., 2026). The typology functions as a coding frame to organise evidence across domains and stakeholder groups and to structure the Results section.

The typology is particularly useful because it captures distinct mechanisms that are often conflated in responses to "what data is missing?" For example, "missing" may refer to genuinely uncollected information, but it may also refer to data that exist yet cannot be compared across jurisdictions due to incompatible formats and definitions, data that conflict across providers because of differing calculation conventions, or data that are formally published but remain difficult to process without specialised tooling. In addition, some reported gaps concern the institutional context of evidence production – such as whether scenarios or infrastructure roadmaps are produced by actors with potential conflicts of interest – or the transparency of model inputs and assumptions required for verification (McWilliams, Ben et al., 2026). Because several of these mechanisms also affect whether datasets can be discovered, accessed, integrated, and reused, the gap typology provides a practical bridge to FAIR-oriented infrastructure discussions (Wilkinson et al., 2016). The illustrative examples used for each gap type in Table 1 are intended to clarify meanings and coding decisions; they should not be interpreted as an exhaustive empirical accounting of each gap class.

Gap type	The problem	Illustrative example	Additional References
Primary information	Certain information is not collected at all.	The EU targets manufacturing 40% of clean-tech demand by 2030, but there is no comprehensive data on current manufacturing capacities or demand.	Industrial decarbonisation modelling often requires site-level industrial energy and emissions data that are "rarely available" from open sources, motivating reconstruction/estimation approaches (Zazzera et al., 2025). Method papers for industrial/energy tools explicitly note that limitations of underlying data sources create incomplete coverage, i.e., "data gaps" in sectoral supply chains (Chen et al., 2020). For power-system infrastructure, researchers have demonstrated that key grid/system datasets are sufficiently incomplete/unavailable that predictive mapping using open data is needed as a workaround (Arderne et al., 2020). Hirth et al. are also noting the incompleteness of load data on the

			ENTSO-E Transparency platform (Hirth et al., 2018).
Comparable information	Information is not comparable across sources.	A few countries provide daily natural gas demand data but publish this on different websites in different formats.	Comparative monitoring of the EU gas system relies on multiple transparency platforms and associated tooling, underscoring how heterogeneous platform conventions/structures shape cross-country comparability (Jung et al., 2024). Energy accounting work highlights that different conventions/definitions can make energy statistics difficult to compare across contexts, even when “the same” indicator label is used (Giampietro & Sorman, 2012).
Consistent information	Information is not consistent across providers.	Eurostat reported 406 bcm EU gas imports in 2022; DG ENER (2023) reported 334 bcm—different methods lead to conflicting totals.	A dedicated review of energy and CO_2 datasets documents that different organizational practices and methodological choices lead to discrepancies across published statistics (Macknick, 2011). Large electricity data platforms are reviewed with emphasis on data-quality and definitional issues, including inconsistencies that require careful handling for research use (Hirth et al., 2018).
Accessible information	High technical barriers to access/processing.	The ENTSO-G transparency platform provides gas flow data, but extracting aggregate volumes by country is difficult.	Even when data are nominally public, energy-modelling inputs are frequently dispersed and tedious to process, motivating “frictionless data” packaging and standardised processing pipelines (Wiese et al., 2019). A review of Europe’s electricity transparency platform highlights practical reuse barriers (documentation/structure/quality constraints) that make access and processing non-trivial for researchers (Hirth et al., 2018). EU gas transparency platforms are valuable archives, but their effective use in analysis depends on specialised tooling/workflows, illustrating the “technical barrier” dimension of accessibility (Jung et al., 2024).
Non-partisan information	Data or scenarios are provided by actors with	ENTSO-E/ENTSO-G publish 10-year grid investment roadmaps; as TSO	Empirical transitions research shows grid operators are embedded incumbents who may simultaneously enable change while also engaging in

	potential conflicts of interest.	associations, their analyses are not institutionally independent.	activities that maintain regime structures, motivating scrutiny of actor-produced evidence (Galeano Galvan et al., 2020). Transition/political-economy literature documents regime resistance by incumbent actors in low-carbon transitions, supporting the general concern that actor interests can shape narratives and evidence bases (Geels, Frank W., 2014). Work on transmission-grid planning demonstrates the importance of transparent stakeholder involvement and governance principles, reflecting that planning processes are contested and can be perceived as interest-laden if not robustly designed (Komendantova et al., 2018).
Transparent information	Missing open reference models or assumptions prevent verification.	EU impact assessments (e.g., a 90% emissions-reduction target by 2040) rely on PRIMES; model inputs/parameters are not public, limiting scrutiny.	A major modelling review identifies "uncertainty vs. transparency" as a core challenge for energy-system analysis, explicitly linking transparency to credibility and interpretability (Pfenninger et al., 2014). Model-capability assessments highlight that lack of transparency and standardisation makes it difficult to evaluate model suitability for policy questions, i.e., it constrains verification and scrutiny (Savvidis et al., 2019). In an EU policy context, PRIMES is described as a private model with limited transparency, directly supporting the claim that non-open assumptions/parameters can limit scrutiny of impact assessments (Szabó, 2023).

Table 1 - Data Gap Types according to McWilliams, Tagliapetra, Zachmann (2025), enriched with additional references from literature.

Research Design

Building on Section 2, which introduces (i) a pragmatic set of empirical energy-data domains and (ii) six recurring gap types, we apply these categories as an analytic coding frame to consolidate heterogeneous evidence on "missing data" into a comparable structure. Because data gaps are multi-causal (legal, technical, organisational) and are expressed across diverse artefacts (survey responses, conference notes, presentations), we use an exploratory qualitative synthesis based on directed qualitative content analysis with an a priori category system (Mayring & Fenzl, 2019). The resulting domain-by-gap-type matrix is a synthesis artefact to structure reported needs and not intended to quantify gap prevalence.

Data sources. Evidence was compiled from three inputs: (1) notes and selected presentation material from the 33rd Wissenschaftliches Kolloquium “Energiewende und Energiepreiskrise – zur Rolle der Daten” (Destatis/Deutsche Statistische Gesellschaft; Wiesbaden, 28–29 Nov 2024); (2) documented contributions from the SciCAR conference in 2025 within a dedicated energy data session, focusing on statements by data journalists and researchers on access and transparency barriers; and (3) two NFDI4Energy internal surveys capturing missing reference datasets for modelling and validation (n=7) and missing elements/missing values in industrial datasets used in research (n=37).

Extraction and coding. The unit of analysis is a “gap statement”: a text segment describing missing or insufficient data for a given purpose. All identifiable gap statements were extracted, consolidated to reduce redundancy, and coded to (i) one or more of the six gap types and (ii) one or more domains defined in Section 2. Statements were additionally marked by the actor group(s) most plausibly able to address the gap (public administration/official statistics; industry/regulated operators; research/community). Because our data sources differ in format and intent – two instruments directly eliciting “what is missing” and two settings providing contextual elaboration – we do not interpret the evidence as a quantitative measurement of gap prevalence. Instead, we apply the domain structure and gap-type typology introduced in Section 2 as a diagnostic coding frame to consolidate heterogeneous statements into a common representation. The resulting domain-by-gap-type matrix serves two purposes: it locates reported needs in empirical domains and it distinguishes mechanisms of missingness (e.g., not collected, not accessible, inconsistent, not transparent). This distinction is analytically important because the same “data are missing” statement can imply different underlying constraints and, consequently, different classes of follow-up measures. Blank cells indicate that the current evidence base did not surface corresponding statements; they should not be interpreted as evidence that no gaps exist.

Results

Across sources, reported needs cluster in a small set of empirically central domains (see detailed gap reporting in Table 2): electricity grids and reference networks, decentralised consumption and heat, industrial datasets used for modelling and ML, and spatially resolved consumption/impact information. The coding frame shows that missingness is most often reported as (i) missing reference-grade primary information and (ii) limited usability due to access and completeness constraints, while consistency and traceability issues arise particularly where data are repackaged or aligned across portals.

3.1 Reported missingness is frequently a lack of reusable reference infrastructures (primary information)

A first pattern that becomes explicit through the coding is that many primary information statements refer not to isolated missing measurements, but to the absence of reusable reference infrastructures needed for benchmarking and validation. In the researcher statements, the most prominent requests are for standardised reference networks and associated reference parameters: reference district heating networks and gas-hydrogen distribution networks (including reference values for insulation parameters and pipeline thicknesses), as well as standard “district” networks for evaluating low-voltage decentralised grid supply. Related requests concern “more grid data” (topology, loads, generators) and richer time series, noting that existing reference packages can be too limited for evaluation (e.g., reference time series constrained to short horizons).

The industrial-data survey points in the same direction but operationalises the missingness as absent core system descriptors in datasets researchers work with: respondents most frequently report missing grid topology (29.73%), load profiles (27.03%), generation profiles (24.32%), and line parameters (21.62%). Several free-text responses further indicate that missingness can affect entire datasets rather than individual values, especially in pilot-site contexts. When mapped to domains and gap types, these statements jointly characterise a primary gap best described as missing reference-grade system representations – networks, topology/parameter data, and accompanying time series suitable for method comparison and model validation – rather than simply "more data points." The added value of the classification is to separate the issue of missing primary information from issues of access or documentation; without this separation, the same statements could be misread as a generic request for more publication rather than a specific need for benchmark-ready reference inputs.

3.2 "Data exist" does not imply "data are usable": accessibility and completeness are dominant constraints in applied datasets

A second pattern concerns the difference between nominal availability and practical usability. In the industrial-data survey, respondents explicitly cite limited access to data (21.62%) and privacy restrictions (21.62%) as reasons why values are missing, complemented by free-text explanations such as confidentiality/criticality constraints, provider-side gaps, and operational disruptions (e.g., malfunctioning measurement devices, sensor errors, outages). In addition, several responses describe accessibility constraints through aggregation practices: data may be available only as totals over multiple measurement points or at daily resolution, which constrains analyses that require granular or disaggregated time series.

Importantly, the survey also documents that missingness is not consistently quantified or documented: respondents note that information about missing or erroneous data is rare and that awareness of missing values is often limited to longer-term disconnects; reported missingness varies substantially across datasets, with some indicating values up to 20%, others reporting lower ranges (e.g., >5%), and pilot-site contexts describing substantially higher missingness. When coded into the gap frame, these statements fall into both accessible and transparent mechanisms: missingness stems from access constraints and aggregation, but also from limited metadata on completeness and error mechanisms. The classification therefore clarifies that "unusable data" is not a single gap: some barriers concern obtaining granular data at all (accessibility), while others concern the ability to assess validity and missing-value mechanisms once data are obtained (transparency). This distinction is consequential for downstream uses noted by respondents (e.g., validation studies and supervised ML), which are sensitive both to completeness and to documented handling of missingness.

3.3 Consistency and traceability issues arise when portals repackage data and apply methodological alignment

A third pattern is that missingness is sometimes experienced as contradictory values and limited traceability rather than a lack of publication. The colloquium notes capture that widely used public-facing services and dashboards frequently rely on upstream ENTSO-E/TSO data, yet discrepancies can occur because derivative products apply methodological choices to improve alignment (e.g., for consistency with other series, definitions, or desired resolutions). The key issue raised is that these reconciliation choices are not always understandable for users. In addition, the colloquium discussion highlights the public relevance of such differences, for example when import and export narratives require careful interpretation and when proxies are used in place of flow-consistent representations.

Within the coding frame, such statements map to consistent and transparent information gaps: users encounter multiple representations of nominally the same indicator across platforms, and the processing steps that produce these differences are not consistently documented in a way that supports verification and reuse. The analytical contribution of the classification is to show that these are not primary collection gaps: the reported problem is the traceability and comparability of reconciliation steps across derivative products.

3.4 End-use and decentralised heat are repeatedly surfaced as salient blind spots in the consumption space (primary and comparability mechanisms)

Finally, across the colloquium notes, the most salient gaps are located not in electricity generation data per se, but in end-use and decentralised heat, where the information needed for planning, evaluation, and interpretation is less robust. Specifically, the notes indicate that while there is a strong data basis for energy generation and origin, it is considerably more difficult to assess what the energy is used for – especially for heat consumption. Examples include differing collection methods and temperature correction approaches, insufficient knowledge about consumption associated with heat pumps, and limited differentiation by socio-economic or building characteristics. The colloquium also identifies limited availability of timely firm-level information on how companies are affected by energy costs, and missing renovation rates.

In the coding frame, these statements map primarily to primary information gaps (data not available at the required disaggregation, or not collected systematically) and, where methodological differences are emphasised, to comparability mechanisms (inconsistent definitions and correction approaches hindering integration). By locating these statements in the domain structure, the synthesis makes explicit that the most salient "blind spots" surfaced in the colloquium are concentrated in end-use and decentralised heat rather than in generation-origin data – an important scoping signal for subsequent prioritisation and follow-up elicitation.

3.5 Scope and interpretive boundaries of the matrix

The matrix synthesises reported needs from four sources with different question formats and agendas and therefore does not provide comprehensive coverage of all energy data domains. Blank cells indicate that no corresponding statements were surfaced in these inputs; they should not be interpreted as evidence of absence. Reported gaps are shaped by participant composition and elicitation focus (e.g., reference datasets; missingness in industrial datasets; colloquium discussions of consumption/heat and public interpretation). Accordingly, the matrix is used to structure and interpret this evidence base and to guide targeted follow-up, not to claim completeness.

Table 2 - Data gaps per domain, potentially addressable by: 🟦 Public administration / official statistics (Destatis, Eurostat, regulators), 🟧 Industry / regulated cooperatives (TSOs/DSOs, exchanges, platform operators, manufacturers), 🟩 Research & academic community.

Gap type Domain	**Primary information** (not collected)	**Comparable information** (not harmonised)	**Consistent information** (conflicting sources)	**Accessible information** (tech/legal barriers)	**Non-partisan information** (potential conflicts of interest)	**Transparent information** (methods /models not open)
Electricity generation & grids	No open benchmark reference datasets for LV/DH/H_2 networks; key grid descriptors (topology, line parameters) not available as reusable reference data; benchmark time series coverage insufficient for validation (e.g., SimBench time span 1y) 🟧 / 🟩	Same concepts are published with differing definitions/structures across ENTSO-E, BNetzA, and Eurostat 🟦 / 🟧	Indicators can differ between upstream ENTSO-E releases and reprocessed public dashboards (e.g., SMARD, Energy Charts), yielding conflicting values 🟧 / 🟦	Detailed grid data (especially below transmission level) restricted due to confidentiality, security, or licensing barriers 🟧 / 🟦	ENTSO-E/-G as provider–planner associations 🟧 / 🟦	Interpolation/smoothing and correction procedures insufficiently documented in public datasets 🟩 / 🟧
Household energy cons.	Systematic data on PV self-consumption and decentralised heat technologies (e.g., heat pumps/biomass) not collected or not available at required granularity 🟦	Household categories and carrier/fuel classifications differ across surveys/statistics 🟦	Varying temperature-correction and imputation methods 🟦	Smart-meter aggregates only (house-level, ≥5 HH); strict confidentiality 🟦	—	—
Commercial & service sector energy cons.	No official statistics on detailed energy use (estimates only) 🟦	Sector delineations differ between national and EU statistics 🟦	—	Microdata access restricted; few safe-centre options 🟦	—	Estimation procedures and revisions not fully documented 🟧 / 🟩
Industrial energy cons.	Lack of sub-annual (monthly/quarterly) data 🟦	Differences in branch/sector mapping (NACE levels) 🟦	Conflicting totals across providers/series 🟦	NDAs, trade secrecy; aggregates only 🟧 / 🟦	—	Missing metadata on missing-value mechanisms and edits 🟦 / 🟩

Energy storage systems	No public datasheets on technical/operational performance 🟧	—	—	Manufacturer and operator data proprietary 🟧	—	—
Energy market data (prices & transactions)	—	Different aggregation windows and products across ENTSO-E/EEX 🟧	—	Intraday/balancing prices fragmented; limited APIs/paywalls 🟧	Market/platform operators have commercial stakes 🟧 / 🟦	Limited disclosure of cleaning rules and price construction 🟧 / 🟩
Spatially resolved energy data	Fine-grained geodata linking consumption to locations (e.g., company/household-level demand proxies) are not collected or not released openly 🟦 / 🟧	Spatial units, coordinate reference systems (CRS), and identifiers differ across portals 🟦 / 🟩	—	Municipal/regional datasets are dispersed and often not machine-readable 🟦	—	—
Climate & energy policy databases	Many policy datasets capture only presence/absence, while structured attributes (coverage, stringency, timelines, enforcement) are not available comprehensively 🟩	Divergent sector mapping/coding across datasets 🟩	—	—	Potential role conflicts where providers are tied to policy advocacy or regulated entities 🟦 / 🟧	Coding rules and modelling linkages insufficiently disclosed 🟩 / 🟦

Implications for NFDI4Energy

The domain × gap-type synthesis indicates that NFDI4Energy’s most actionable contribution is to turn heterogeneous “missing data” statements into targeted infrastructure tasks. The presented matrix builds a structured foundation for our ongoing process in quantifying data gaps across the energy domain. Based on the clustered needs, four implication areas stand out:

- **Make gaps visible and attributable.** Maintain a structured, versioned gap register that links reported gaps to domains, gap mechanisms (primary/accessibility/consistency/transparency), and the institutions best positioned to address them. This converts diffuse gap narratives into actionable problem statements.

- **Provide reference-grade data packages.** Where stakeholders repeatedly request benchmarks (e.g., reference networks and associated time series), curate and steward research-ready reference packages with clear provenance, documentation, and versioning – independent of whether all upstream raw data can be centralised.

- **Standardise metadata and provenance for comparability and traceability.** Promote harmonised metadata profiles and mandatory provenance fields (definitions, resolutions, revisions, reconciliation steps) in the high-demand domains, reducing contradictions between portals and improving reuse across sources.

- **Lower practical access barriers with tooling and “research-ready” interfaces.** Support automated harvesting where permissible, standardised completeness/missingness documentation, and reusable processing pipelines that translate dispersed sources into documented, machine-actionable datasets – especially for workflows sensitive to missingness and validation (including ML).

Together, these measures strengthen interoperability between official statistics, regulated-industry data sources, and open scientific resources, while keeping the focus on the missingness mechanisms that the evidence base actually surfaced.

References

Arderne, C., Zorn, C., Nicolas, C., & Koks, E. E. (2020). Predictive mapping of the global power system using open data. *Scientific Data*, *7*(1), 19. https://doi.org/10.1038/s41597-019-0347-4

Arndt, W., Gerlich, S. C., Hofmann, V., Kubin, M., Kulla, L., Lemster, C., Mannix, O., Rink, K., Nolden, M., Schweikert, J., Shankar, S., Söding, E., Steinmeier, L., & Süß, W. (2022, December). *A survey on research data management practices among researchers in the Helmholtz Association* [Berichte]. HMC-Office, GEOMAR Helmholtz Centre for Ocean Research. https://doi.org/10.3289/HMC_publ_05

BNetzA. (n.d.). *Core energy market data register*. Retrieved 11 January 2026, from https://www.bundesnetzagentur.de/EN/Areas/Energy/CoreEnergyMarketDataRegister/start.html

Chen, X., Matthews, H., Hanes, R., & Carpenter, A. (2020). Identifying data gaps in the energy supply chains of manufacturing sectors with an input–output LCA model. *Procedia CIRP*, *90*, 494–497. https://doi.org/10.1016/j.procir.2020.02.129

European Parliament and Council. (2011, October 25). *Regulation (EU) No 1227/2011 of the European Parliament and of the Council of 25 October 2011 on wholesale energy market integrity and transparency Text with EEA relevance*. http://data.europa.eu/eli/reg/2011/1227/oj

European Parliament and Council. (2013, June 14). *Commission Regulation (EU) No 543/2013 of 14 June 2013 on submission and publication of data in electricity markets and amending Annex I to Regulation (EC) No 714/2009 of the European Parliament and of the Council Text with EEA relevance*. http://data.europa.eu/eli/reg/2013/543/oj

European Parliament and Council. (2022, May 30). *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA relevance)*. http://data.europa.eu/eli/reg/2022/868/oj

European Social Survey. (2017). Round 8 Data. https://www.europeansocialsurvey.org/news/article/round-8-data-now-available

European Union. (2019, June 20). *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast)*. http://data.europa.eu/eli/dir/2019/1024/oj

European Union. (2023). *Shedding light on energy in the EU: Share of energy products in total energy available*. Shedding Light on Energy in the EU: Share of Energy Products in Total Energy Available. https://ec.europa.eu/eurostat/cache/interactive-publications/energy/2025/01/index.html?lang=en&simple=true

European Union. (2025a). *Energy consumption in households*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households

European Union. (2025b). *Final energy consumption in services—Detailed statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Final_energy_consumption_in_services_-_detailed_statistics

European Union. (2025c). *Final energy consumption in transport—Detailed statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Final_energy_consumption_in_transport_-_detailed_statistics

Fujimori, S., & Matsuoka, Y. (2011). Development of method for estimation of world industrial energy consumption and its application. *Energy Economics*, *33*(3), 461–473. https://doi.org/10.1016/j.eneco.2011.01.010

Galeano Galvan, M., Cuppen, E., & Taanman, M. (2020). Exploring incumbents' agency: Institutional work by grid operators in decentralized energy innovations. *Environmental Innovation and Societal Transitions*, *37*, 79–92. https://doi.org/10.1016/j.eist.2020.07.008

Geels, Frank W. (2014). Regime Resistance against Low-Carbon Transitions: Introducing Politics and Power into the Multi-Level Perspective. *Theory, Culture & Society*, *31*(5), 21–40.

Giampietro, M., & Sorman, A. H. (2012). Are energy statistics useful for making energy scenarios? *Energy*, *37*(1), 5–17.

Hirth, L. (2020). Open data for electricity modeling: Legal aspects. *Energy Strategy Reviews*, *27*, 100433. https://doi.org/10.1016/j.esr.2019.100433

Hirth, L., Mühlenpfordt, J., & Bulkeley, M. (2018). The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform. *Applied Energy*, *225*, 1054–1067. https://doi.org/10.1016/j.apenergy.2018.04.048

Hörsch, J., Hofmann, F., Schlachtberger, D., & Brown, T. (2018). PyPSA-Eur: An open optimisation model of the European transmission system. *Energy Strategy Reviews*, *22*, 207–215. https://doi.org/10.1016/j.esr.2018.08.012

Jung, D., Vuillaume, J.-F., Fernández-Blanco, R., Calisto, H., Gómez, N. R., & Lavín, R. B. (2024). The European natural gas system through the lens of data platforms. *Energy Strategy Reviews*, *51*, 101297. https://doi.org/10.1016/j.esr.2024.101297

Komendantova, N., Yazdanpanah, M., & Shafiei, R. (2018). Studying young people' views on deployment of renewable energy sources in Iran through the lenses of Social Cognitive Theory. *AIMS Energy*, *6*(2), 216–228.

Macknick, J. (2011). Energy and CO2 emission data uncertainties. *Carbon Management*, *2*(2), 189–205. https://doi.org/10.4155/cmt.11.10

Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds), *Handbuch Methoden der empirischen Sozialforschung* (pp. 633–648). Springer Fachmedien.

McWilliams, Ben, Tagliapetra, Simone, & Zachmann, Georg. (2026, January 9). *Europe's energy information problem*. Bruegel | The Brussels-Based Economic Think Tank. https://www.bruegel.org/policy-brief/europes-energy-information-problem

Meinecke, S., Sarajlić, D., Drauz, S. R., Klettke, A., Lauven, L.-P., Rehtanz, C., Moser, A., & Braun, M. (2020). SimBench—A Benchmark Dataset of Electric Power Systems to Compare Innovative Solutions Based on Power Flow Analysis. *Energies*, *13*(12), 3290. https://doi.org/10.3390/en13123290

Morrison, R. (2018). Energy system modeling: Public transparency, scientific reproducibility, and open development. *Energy Strategy Reviews*, *20*, 49–63. https://doi.org/10.1016/j.esr.2017.12.010

Nachtigall, D., Lutz, L., Cárdenas Rodríguez, M., D'Arcangelo, F. M., Haščič, I., Kruse, T., & Pizarro, R. (2024). The Climate Actions and Policies Measurement Framework: A Database to Monitor and Assess Countries' Mitigation Action. *Environmental and Resource Economics*, *87*(1), 191–217. https://doi.org/10.1007/s10640-023-00821-2

Pfenninger, S., Hawkes, A., & Keirstead, J. (2014). Energy systems modeling for twenty-first century energy challenges. *Renewable and Sustainable Energy Reviews*, *33*, 74–86. https://doi.org/10.1016/j.rser.2014.02.003

Pfenninger, S., Hirth, L., Schlecht, I., Schmid, E., Wiese, F., Brown, T., Davis, C., Gidden, M., Heinrichs, H., Heuberger, C., Hilpert, S., Krien, U., Matke, C., Nebel, A., Morrison, R., Müller, B., Pleßmann, G., Reeg, M., Richstein, J. C., … Wingenbach, C. (2018). Opening the black box of energy modelling: Strategies and lessons learned. *Energy Strategy Reviews*, *19*, 63–71. https://doi.org/10.1016/j.esr.2017.12.002

Savvidis, G., Siala, K., Weissbart, C., Schmidt, L., Borggrefe, F., Kumar, S., Pittel, K., Madlener, R., & Hufendiek, K. (2019). The gap between energy policy challenges and model capabilities. *Energy Policy*, *125*, 503–520. https://doi.org/10.1016/j.enpol.2018.10.033

Schill, W.-P., & Zerrahn, A. (2018). Long-run power storage requirements for high shares of renewables: Results and sensitivities. *Renewable and Sustainable Energy Reviews*, *83*, 156–171. https://doi.org/10.1016/j.rser.2017.05.205

Süsser, D., Ceglarz, A., Gaschnig, H., Stavrakas, V., Flamos, A., Giannakidis, G., & Lilliestam, J. (2021). Model-based policymaking or policy-based modelling? How energy models and energy policy interact. *Energy Research & Social Science*, *75*, 101984.

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*(8), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Szabó, Z. (2023). Biofuel Policy-Making Based on Outdated Modelling? The Cost of Road Transport Decarbonisation in EU. *Fuels*, *4*(3), 354–362. https://doi.org/10.3390/fuels4030022

Ventosa, M., Baíllo, Á., Ramos, A., & Rivier, M. (2005). Electricity market modeling trends. *Energy Policy*, *33*(7), 897–913. https://doi.org/10.1016/j.enpol.2003.10.013

Verschoor, M., Albers, C., Poortinga, W., Böhm, G., & Steg, L. (2020). Exploring relationships between climate change beliefs and energy preferences: A network analysis of the European Social Survey. *Journal of Environmental Psychology*, *70*, 101435. https://doi.org/10.1016/j.jenvp.2020.101435

Wiese, F., Schlecht, I., Bunke, W.-D., Gerbaulet, C., Hirth, L., Jahn, M., Kunz, F., Lorenz, C., Mühlenpfordt, J., Reimann, J., & Schill, W.-P. (2019). Open Power System Data – Frictionless data for electricity system modelling. *Applied Energy*, *236*, 401–409. https://doi.org/10.1016/j.apenergy.2018.11.097

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zazzera, E. B., Prina, M. G., Marchetti, R., Misconel, S., Manzolini, G., & Sparber, W. (2025). Bridging the industrial data gap: Top-down approach from national statistics to site-level energy consumption data. *Data in Brief*, *59*, 111365. https://doi.org/10.1016/j.dib.2025.111365