MIND&MATTER

UK GAS DATA INFRASTRUCTURE: GEMINI, GAS DATA PORTAL & DSI

AGREED UPON DELIVERABLES

WP1 DELIVERABLES - PROGRESS

- Stakeholder Identification & Engagement
    - *Achieved – Stakeholders identified; Knowledge elicitation process is ready for distribution post-feasibility study phase*
- Data Inventory & Mapping
    - *Achieved - Clear data inventory produced for both portals*
- Gap Analysis
    - *Achieved – Thorough analysis of current state of Gemini and Gas data portal systems; clearly laid out key features and strengths of both systems; identified gaps and opportunities; Need for Data sharing infrastructure (combining strengths and opportunities of both systems)*

PROJECT VISION: UNIFIED GAS DATA ACCESS

The Challenge

National Gas currently operates two separate data systems Gemini (secure, closed commercial platform) and the Gas Data Portal (open, public-facing). These systems serve different user groups but create duplicated data flows, fragmented ownership, and inconsistent experiences. Additionally, Gemini data is integrated into internal platforms such as D&I and GCP, further increasing duplication risks.

The Goal

This project aims to develop a unified data access framework the Gas Transmission Data Sharing Infrastructure (GTDSI) , that bridges both external systems (Gemini and the Gas Data Portal) and internal data platforms under a single, governed access layer. This unified approach will eliminate data duplication, ensure consistency, and enable controlled access for all stakeholder types.

Connection to NESO DSI

The GTDSI is not a standalone or separate infrastructure. It is National Gas's Data Preparation Node (DPN) , the building block that enables NGT to connect to the NESO-led Data Sharing Infrastructure. Through this connection, gas sector data can be shared with the wider energy ecosystem, including electricity, enabling cross-sector optimisation, coordinated balancing, and whole-system planning for Net Zero

STAKEHOLDER CAPTURE

INTERNAL STAKEHOLDERS

- **IT Architecture Team**
  - Ensures DSI aligns with enterprise architecture and technical standards
- **Data & AI Team**
  - Provides analytics and AI expertise, especially for data insights and access modelling
- **Data Governance Team**
  - Oversees data quality, standards, and governance policies
- **Steering Committee / Project Sponsors**
  - Provides oversight, decision-making, and sign-off on deliverables

EXTERNAL STAKEHOLDERS

- **Existing Industry Users**
    - Gas shippers, distribution network operators, energy traders, commercial customers (users of Gemini and Gas Data Portal)
- **Regulators & Policy Bodies**
    - Ofgem (regulatory alignment), Department for Energy Security & Net Zero
- **Innovators & Third Parties**
    - Tech vendors, data analytics firms, academic researchers, app developers
- **NESO & Other Network Operators**
    - National Energy System Operator (DSI coordination), electricity and gas network operators (standards alignment)
- **Wider Stakeholder Community**
    - New market entrants, researchers, and the public (future users of DSI)

GTDSI – MIDDLE GROUND APPROACH: MAPPING USER REQUIREMENTS

- **Objective:** Develop National Gas's Data Preparation Node (GTDSI) as the foundation for connecting to the NESO-led Data Sharing Infrastructure, balancing robust security with open, self-service data access for a wide range of users.
- **User Requirements Mapping:**
  - Due to the large and diverse external user base, a **questionnaire method** is recommended during the development stage.
  - This approach enables systematic capture of stakeholder needs, preferences, and access requirements.
- **Feasibility Study Outcome:**
  - Specific questions have been identified to guide requirement gathering in the development phase.
  - These questions will inform the design of access controls, data classification, and user experience, ensuring the GTDSI meets both regulatory and stakeholder expectations.
  - Questionnaire ready for distribution post-feasibility study

GTDSI – MIDDLE GROUND APPROACH: MAPPING USER REQUIREMENTS

Section 2: Data Needs and Use Cases

National Gas is currently evaluating the development of a new Gas Transmission Data Sharing Infrastructure (GTDSI). This platform would unify access to gas system data currently available through the Gemini system and the Gas Data Portal, while also enabling broader, more flexible access for a wider range of stakeholders.
The GTDSI aims to:

- Improve discoverability and usability of gas data
- Support real-time or near real-time access
- Enable integration with external systems via APIs
- Balance openness with appropriate access controls
- Align with regulatory frameworks such as Ofgem's Data Best Practice and NESO's federated GTDSI model

The following questions seek to understand your organisation's current experience and future expectations.

6. What is your primary use of gas system data?
   (Select all that apply)
   - o Operational decision-making
   - o Market analysis or trading
   - o Regulatory or compliance reporting
   - o Research or forecasting
   - o Application development or innovation
   - o Other: __________
7. Which enhancements to the existing Gemini system and/or Gas Data Portal would most improve your ability to access and use National Gas data? *(Select as many as applicable)*
   - o More real-time or near real-time data

Example questions for internal stakeholders

Section 0: Respondent Information

1. Please indicate your role in relation to the Gemini system and/or the Gas Data Portal. (Select all that apply)
   - o Data Architecture
   - o Cloud Infrastructure / DevOps
   - o API Development / Software Engineering
   - o Data Governance / Compliance
   - o Cybersecurity / IT Risk
   - o Other: ______________________________

Section 1: Current Systems & Operational Challenges

2. National Gas is assessing the potential requirement for a GTDSI, that could consolidate access to operational and commercial data currently available through Gemini and the Gas Data Portal. The aim would be to reduce duplication, improve consistency, and streamline workflows across teams.
   In your view, which operational processes or tasks would benefit most from such a platform, if it were to be implemented?
   (Free text response)

Section 2: Platform Architecture & Integration

3. Which (if any) existing systems (e.g. Gemini, Gas data portal) present technical limitations for integration, scalability, or security, and should be considered for modernisation or decommissioning? *(Free text response)*
4. What architectural model should the GTDSI adopt (e.g. centralised, federated, hybrid), and

Example questions for external stakeholders

GEMINI SYSTEM

GEMINI SYSTEM – PURPOSE AND STRENGTHS

Gemini is the core commercial system for Great Britain's gas market, managing capacity, nominations, trading, balancing, and invoicing (National Gas Transmission, 2025). Operated by Xoserve under the Data Services Contract (DSC) (Xoserve, 2024), it provides secure, compliant market access. The Gemini Sustain Plus upgrade modernised the platform into a cloud-native Azure microservices system with ~99.9 % uptime and enhanced scalability and resilience

Category	Description
Core Functionality	Central platform supporting all commercial processes including capacity management, trading, balancing, and invoicing (National Gas Transmission, 2025).
Governance	Joint structure with National Gas Transmission as owner and Xoserve as operator under the DSC, ensuring accountability and transparent operations (Xoserve, 2024).
Technology Stack	Modernised to Gemini Sustain Plus using Java microservices, Azure APIM, and JWT authentication for secure API-based integration (Xoserve, 2024).
Performance & Resilience	Delivers ~99.9 % availability with tested failover, incident response, and disaster-recovery measures (Correla, 2023).
User Base	Access for licensed shippers, storage operators, and transmission users through secure web and API channels (National Gas Transmission, 2025)
Compliance & Security	Aligns with UNC requirements and ISO 27001 principles, safeguarding data integrity and regulated market operations (Ofgem, 2024).

GEMINI – CURRENT LIMITATIONS AND INTEGRATION CHALLENGES

Legacy Constraints: Pre-upgrade platform used obsolete Oracle modules nearing end-of-support (*National Gas Transmission, 2025*).

Complex User Experience: Prior interface required specialist training; limited browser/device support.

Integration Barrier: Closed environment; non-participants could not directly access operational data for analysis.

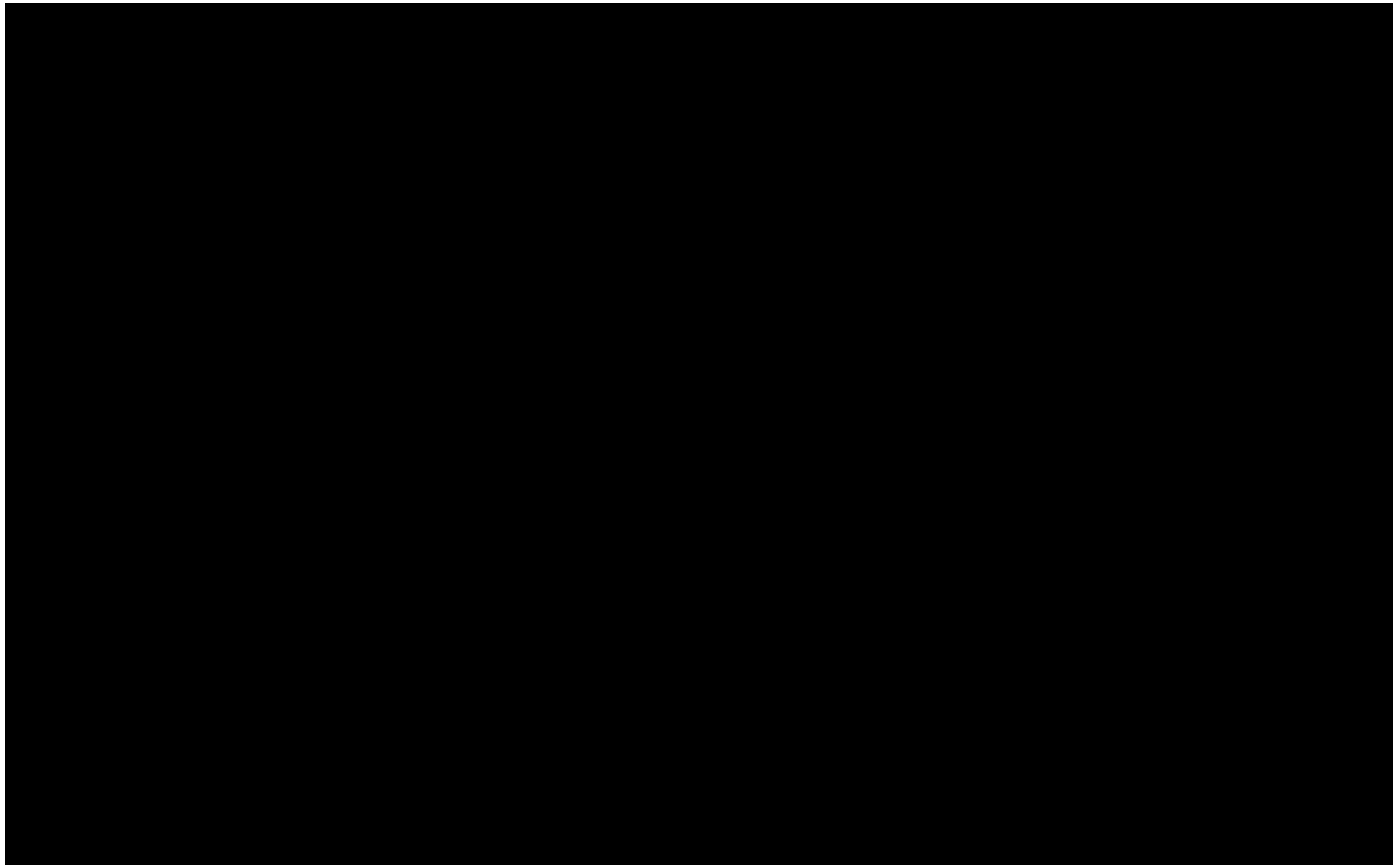
Operational Risk: Outage or maintenance affects national balancing and capacity trading despite back-up procedures (*Xoserve, 2024*).

API Transition Gaps: Early APIs focused on user processes, not on machine-to-machine data publishing for wider ecosystem integration.

GEMINI – EXTERNAL INTEGRATIONS

Illustrates the interaction between the Gemini system and its surrounding services, including API Management, authentication layers, and internal National Gas Transmission systems. The diagram shows how Gemini data products are securely exposed via the Azure API Management layer for integration with other market systems and participants (Xoserve, 2024; National Gas Transmission, 2025).

Gemini API Data Flow



GEMINI – DATA INVENTORY

Data Attribute	Data Type	Source System
[redacted]		

GEMINI – DATA INVENTORY (cont)

Data Attribute	Data Type	Source System
[redacted]		

GAS DATA PORTAL

MIND&MATTER

GAS DATA PORTAL – KEY FEATURES AND STRENGTHS

The Gas Data Portal provides open, transparent access to real-time and historical gas system information, including pressures, flows, capacity, linepack, and market notices (National Gas Transmission, 2024).It serves as the primary public-facing data service for stakeholders across the gas sector, delivering near-real-time operational data through a free online portal, APIs, and a mobile app (Xoserve, 2024). Handling over 70 million annual data requests, it supports market transparency and regulatory compliance under Ofgem's open-data framework (Correla, 2023; Ofgem, 2024).

Category	Description
Accessibility	Free and open to all users, with web portal, REST API downloads (launched November 2024), and mobile app access for live system information
Transparency	Promotes equal access to operational data, supporting Ofgem's transparency and competition objectives
Data Scope	Provides live and historical datasets on capacity, flows, balancing, and storage, supporting research and industry analysis
Performance	Manages more than 70 million annual data requests with responsive delivery and scalable infrastructure
Usability	Offers search tools, CSV downloads, and interactive dashboards that enhance user experience and data discoverability
Regulatory Alignment	Fully compliant with Ofgem's open-data policy, ensuring fair market visibility and consistent publication standards

GAS DATA PORTAL – CURRENT GAPS AND OPPORTUNITIES

Focus: On Transmission-level data

Data Granularity: Aggregated data protects commercial confidentiality but restricts advanced analytics.

User Differentiation: Lacks tiered access (for regulated vs public data).

Metadata Weakness: Inconsistent naming and units reduce discoverability.

Enhancement Potential: Unified search engine, personalised dashboards, and data standardisation (DESNZ, 2025).

INTEGRATION AND OVERLAP – CURRENT STATE AND RISKS

Gemini and the Gas Data Portal currently operate as separate but complementary platforms, with Gemini managing the core commercial processes; capacity, nominations, balancing, and settlement, and the Portal publishing open, near real-time datasets drawn partly from Gemini (National Gas Transmission, 2025). While this separation supports compliance and security, it creates duplication, fragmented ownership, and data-lineage gaps as outputs are re-published through the Portal. A move toward a unified, API-driven data exchange with clear publication points would reduce overlap and prepare both systems for integration into the Data Sharing Infrastructure (DSI) model

Area	Description
System Separation	Gemini handles transactional, role-restricted data; the Gas Data Portal focuses on open publication, creating siloed workflows (National Gas Transmission, 2025).
Duplication Risks	Some datasets are exported from Gemini and re-ingested into the Portal, introducing timing and version discrepancies (Correla, 2023).
Data Lineage	Lack of a single metadata catalogue makes it difficult to trace the origin, timestamp, and transformation of data between systems (DESNZ, 2025).
Integration Readiness	Both systems now offer REST APIs; introducing secure inter-platform messaging or event triggers would automate synchronisation (Xoserve, 2024).
DSI Alignment	Implementing a Gas Data Preparation Node would connect both systems seamlessly to the future federated data exchange model (Energy Systems Catapult, 2024).

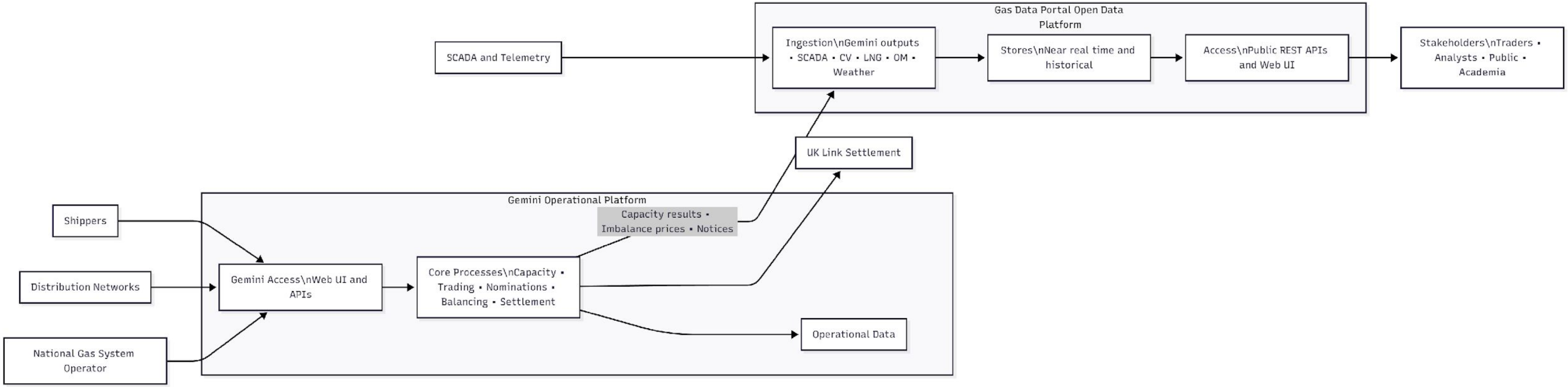
MIND&MATTER

GAS DATA PORTAL – DATA INVENTORY

Data Category / Item	Description	Source System	Update Frequency	Data Format(s)	Access Method	Historic Range	Key User Groups
Supply – Forecast Flow	Aggregate forecast of gas entering the NTS for the current day	Forecasting Systems (DFNs, SFNs)	Hourly	CSV, XML, JSON	Web Portal, SOAP API, REST API	5 years	Shippers, Operators, Regulators
Supply – Physical Flow	Actual aggregated gas flow into the NTS in real time	SCADA Telemetry	Real-time (~12-min)	CSV, XML, JSON	Web Portal, SOAP API, REST API	5 years	Shippers, Analysts, Regulators, Public
Demand – Forecast	Projected total gas demand on the NTS	Forecasting Models, DNO inputs	Hourly & Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Operators, Traders, Regulators, Analysts
Demand – Actual	Actual gas demand outturn on the NTS	SCADA & Allocation Systems	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	All users
Linepack	Total volume of gas in the transmission system	SCADA & NG Calculations	Hourly	CSV, XML, JSON	Web Portal, APIs	5 years	Operators, Traders, Regulators, Analysts
Capacity – Entry/Exit	Booked and available capacity on the NTS	Gemini Capacity Management	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Shippers, Regulators, Analysts
Interruption Events	Records of interruptible capacity curtailment	Gemini / Control Centre	Event-driven	CSV, XML, JSON	Web Portal, APIs	5 years	Regulators, Affected Shippers, Analysts
Storage	Gas storage operations data	Storage Operators	Daily & Within-day	CSV, XML, JSON	Web Portal, APIs	5 years	Traders, Regulators, Analysts, Public
LNG	LNG import terminal data	LNG Terminal Operators	Real-time & Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Traders, Regulators, Analysts, Public
Balancing	System balancing and market operation data	Gemini & Market Systems	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Traders, Regulators, Analysts
Prices (Balancing Market)	Balancing market price data	Gemini/Market Systems	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Traders, Regulators, Analysts, Public
Calorific Value (CV)	Energy content per unit volume	Gas Chromatographs & DNOs	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Distribution Companies, Regulators, Analysts
Shrinkage & Ops Usage	Gas used or lost by the system	NG Internal Systems	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Regulators, NG Ops, Analysts
Operating Margins & Notices	Contingency gas and Margin Notices	Control Centre & Planning	Event-driven	Web, XML, JSON	Web Portal, APIs	Several years	Market, Regulators, Public
Weather Data	Weather metrics relevant to gas demand	Met Office / Forecast Systems	Daily	CSV, XML, JSON	Web Portal, APIs	5 years	Analysts, Regulators, Public

CURRENT DATA FLOW BETWEEN GEMINI AND THE GAS DATA PORTAL

This diagram shows how data currently moves between the Gemini Operational Platform and the Gas Data Portal. Gemini handles the market-critical processes, capacity, trading, nominations, balancing, and settlement, fed by inputs from shippers, distribution networks, and the National Gas System Operator. Operational and SCADA telemetry data flow from Gemini and other internal systems into the Gas Data Portal, where they are aggregated, stored as near real-time and historical datasets, and made publicly accessible via web and REST APIs. This setup supports transparency but highlights the current duplication of publication pathways and the need for a unified API-driven model under the future Data Sharing Infrastructure (DSI) framework.



TOWARD A UNIFIED ARCHITECTURE (DSI MODEL)

1. Unified & Secure Access

- Single platform for all stakeholders
- Combines openness (Portal) with security (Gemini)
- Identity & access management (login/API key)

2. Intelligent Data Management

- AI-driven data classification by sensitivity
- Dynamic, risk-based access controls

3. Interoperability & API-First Design

- Standardised APIs for all data
- Aligns with NESO's Digital Spine & Open Energy standards
- Seamless integration across the energy sector

4. User Experience & Scalability

- One-stop shop for gas data
- Cloud-based, modular, and futureproof (ready for hydrogen, new vectors)
- Reduces manual requests, improves efficiency

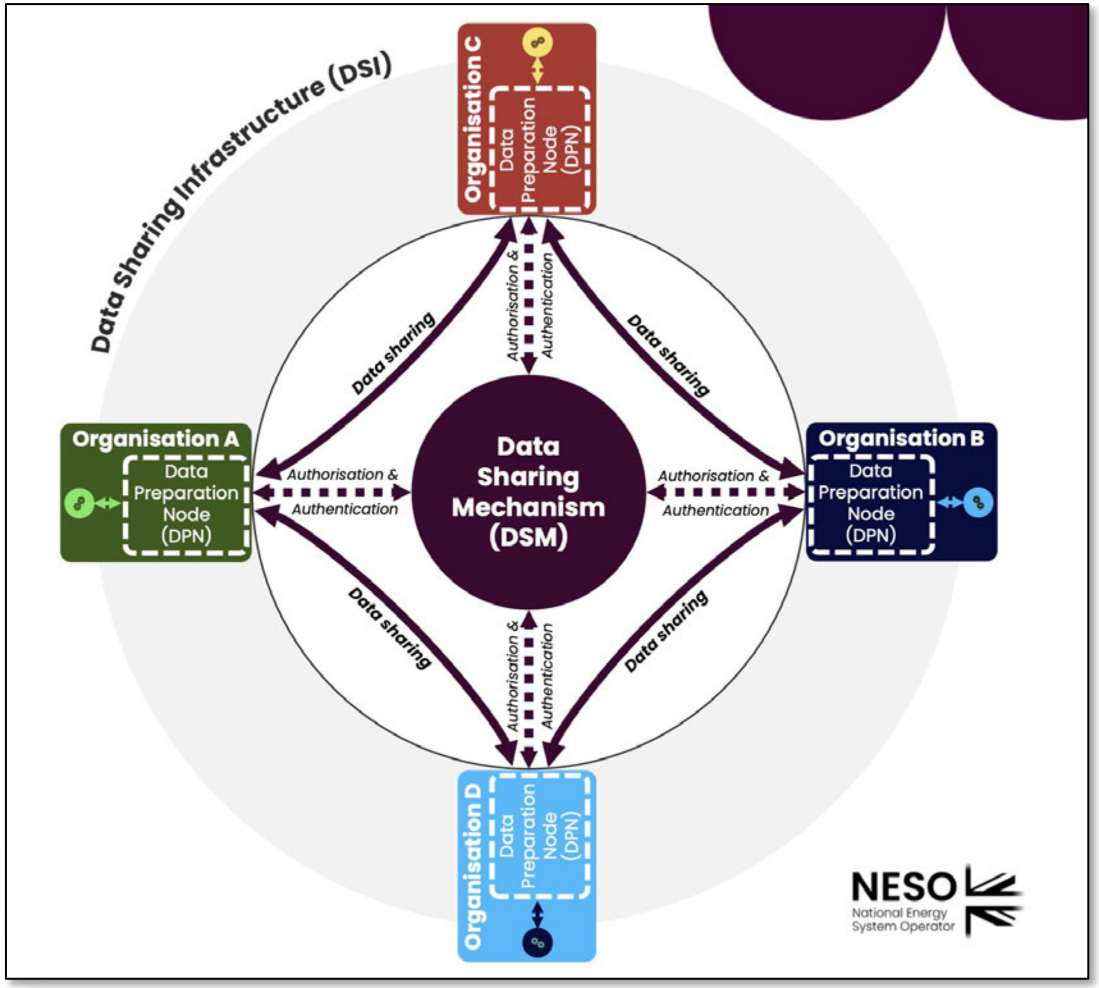
5. Governance, Compliance & Net Zero

- Robust governance and continuous monitoring
- Privacy/security by design
- Supports Ofgem's Data Best Practice and net zero goals

SCALABILITY – USAGE AND FUTURE GROWTH

- Current Load: The Gas Data Portal now handles over 70 million annual requests, providing near-real-time and historical datasets to shippers, analysts and the public, while Gemini runs 24 hours a day to support core market transactions such as nominations, capacity and balancing (Correla 2023; National Gas Transmission 2025).
- Growth Drivers: Forthcoming demands include hydrogen blending and transport, tighter gas-electric data coupling, and adoption of digital-twin simulations for forecasting and operational optimisation
- Infrastructure Readiness: The Gemini Sustain Plus platform uses Azure Kubernetes Service (AKS) microservices with serverless APIs managed through Azure API Management, allowing dynamic scaling during market peaks
- Performance & DSI Synergy: A distributed architecture aligned with the planned Data Sharing Infrastructure (DSI) eliminates single bottlenecks and improves resilience, throughput and recovery
- Operational Considerations: Continuous tracking of API latency, bandwidth for high-frequency telemetry, and data-retention costs is needed to sustain efficiency as dataset volumes expand (DESNZ 2025).

NESO DSI – PILOT STAGE



- Federated access via Trust Framework
- RESTful APIs, JSON, OAuth2, OpenID Connect
- Accredited participants via Energy Sector Trust Framework
- Centralised catalogue with metadata tagging
- Risk-based classification, role-based access
- Cloud-native, modular microservices

https://www.neso.energy/document/361651/download

KEY LESSONS FROM NESO'S DSI TRIAL

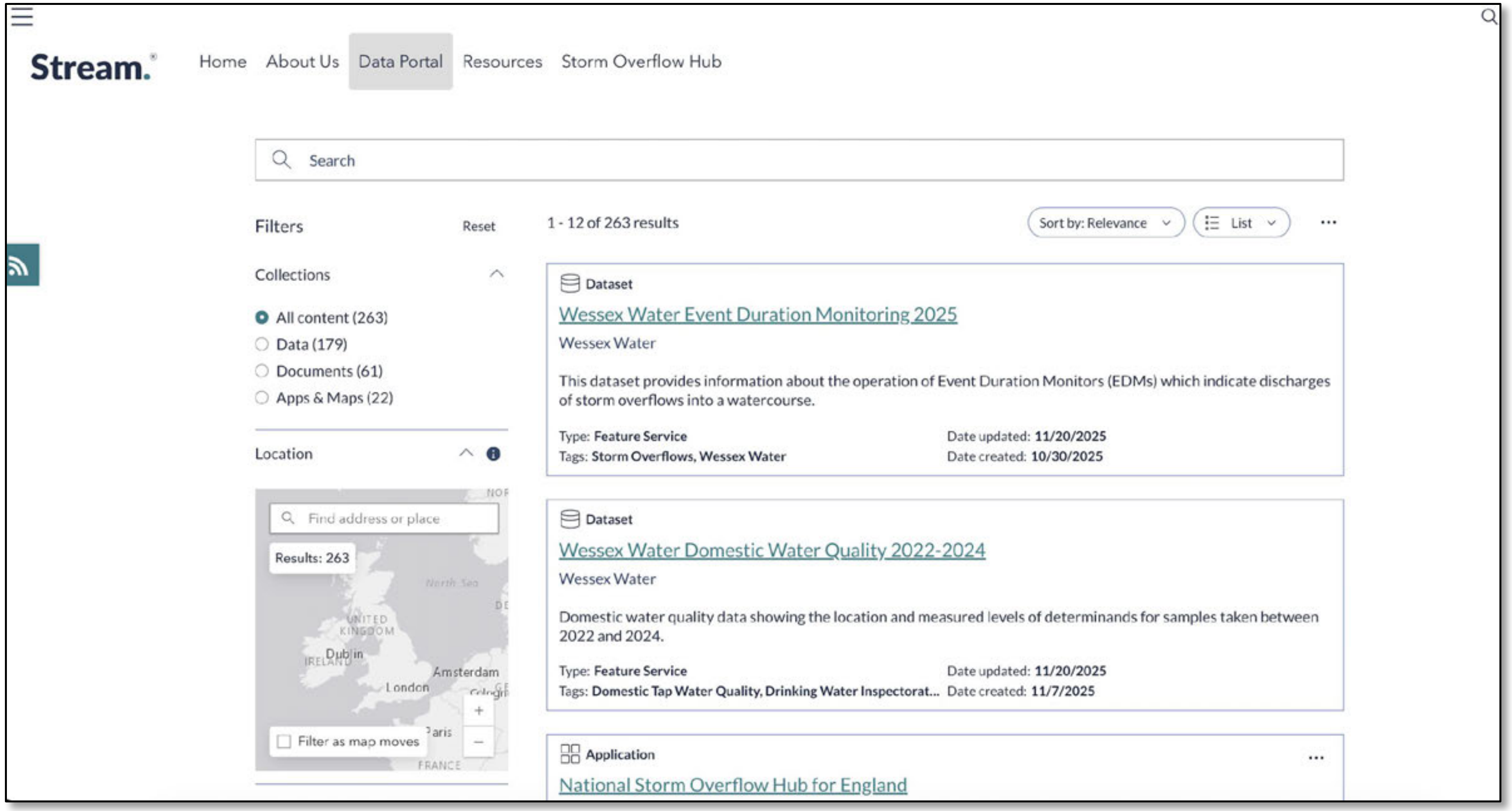
1. **Simplify Deployment & Configuration**
   - Use modular scripts, containerisation (Docker), and orchestration (Kubernetes).
2. **Build Reliability into Core Components**
   - Add retry logic, health checks, fallback mechanisms; monitor staged rollouts.
3. **Plan for File Size & Data Volume**
   - Support large datasets via chunking, streaming APIs, or cloud-native storage.
4. **Secure Technical Resources Early**
   - Form cross-functional teams (data engineers, architects, security experts).
5. **Invest in UI, Automation & Monitoring**
   - Provide user-friendly UI, automate workflows, and implement health dashboards.
6. **Design for Realistic Test Environments**
   - Mirror production; use synthetic data and sandboxed APIs.

NEED FOR GAS TRANSMISSION DATA SHARING INFRASTRUCTURE

- **Bridges the Gap:**
    - Gemini is secure but closed and siloed; Gas Data Portal is open but limited in scope and control.
    - Current dual approach leads to duplicated data flows, inefficiencies, and inconsistent user experiences.
- **Unlocks Data Value:**
    - Enables broader, controlled access to high-quality gas data for a wider range of stakeholders (not just market participants).
    - Supports innovation, research, and new market entrants by making more data available in a secure, standardised way.
- **Meets Regulatory & Future Demands:**
    - Aligns with Ofgem's Data Best Practice and "Presumed Open" policy.
    - Prepares for future requirements (Digital Spine, net zero, cross-sector integration).
- **Improves Efficiency & Governance:**
    - Reduces manual data requests and duplicated publication pathways.
    - Enhances traceability, security, and data quality through unified governance and AI-driven access controls.
- **Supports Net Zero & Whole-System Optimisation:**
    - Facilitates integration with electricity and other energy vectors.
    - Provides the foundation for digitalisation, transparency, and decarbonisation in the UK energy system.

SIMILAR DATA SHARING INFRASTRUCTURE PORTALS

STREAM WATER DATA

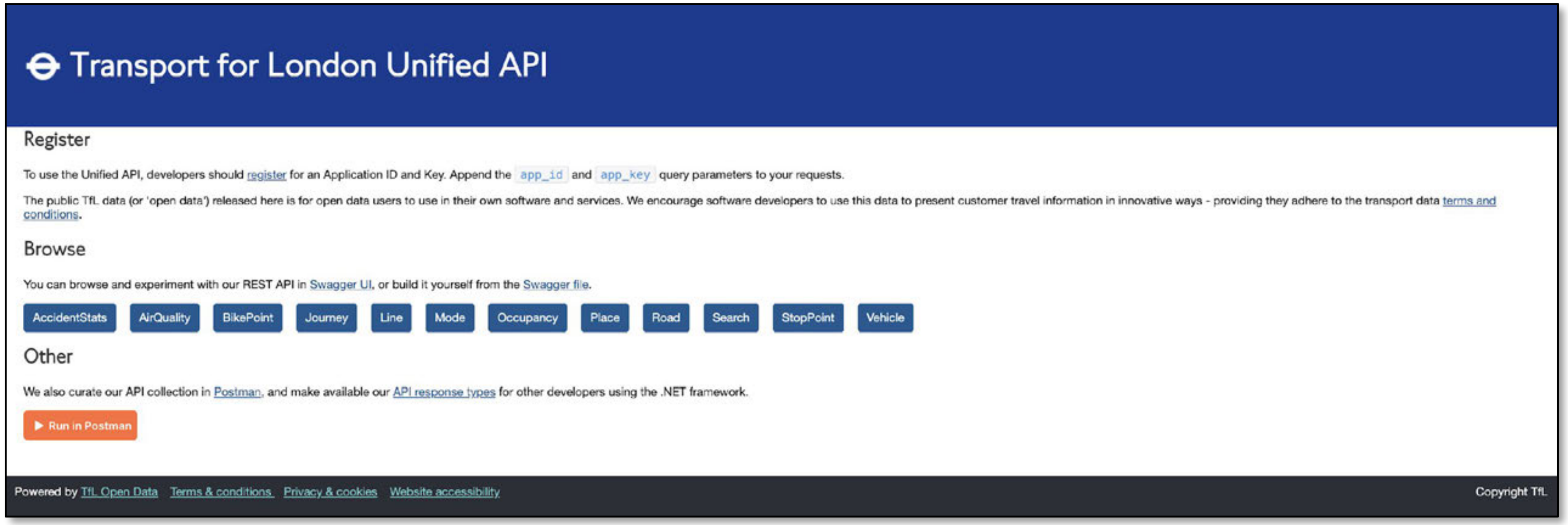


https://www.streamwaterdata.co.uk/api/search/definition/

- Open and Shared tiers via sector roadmap
- REST APIs, Open
- API spec (planned)
- Sector-wide coordination, IAM in development
- Metadata standards under Open Data Framework
- Data triage process, manual classification
- Centralised platform with decentralised data nodes

MIND&MATTER

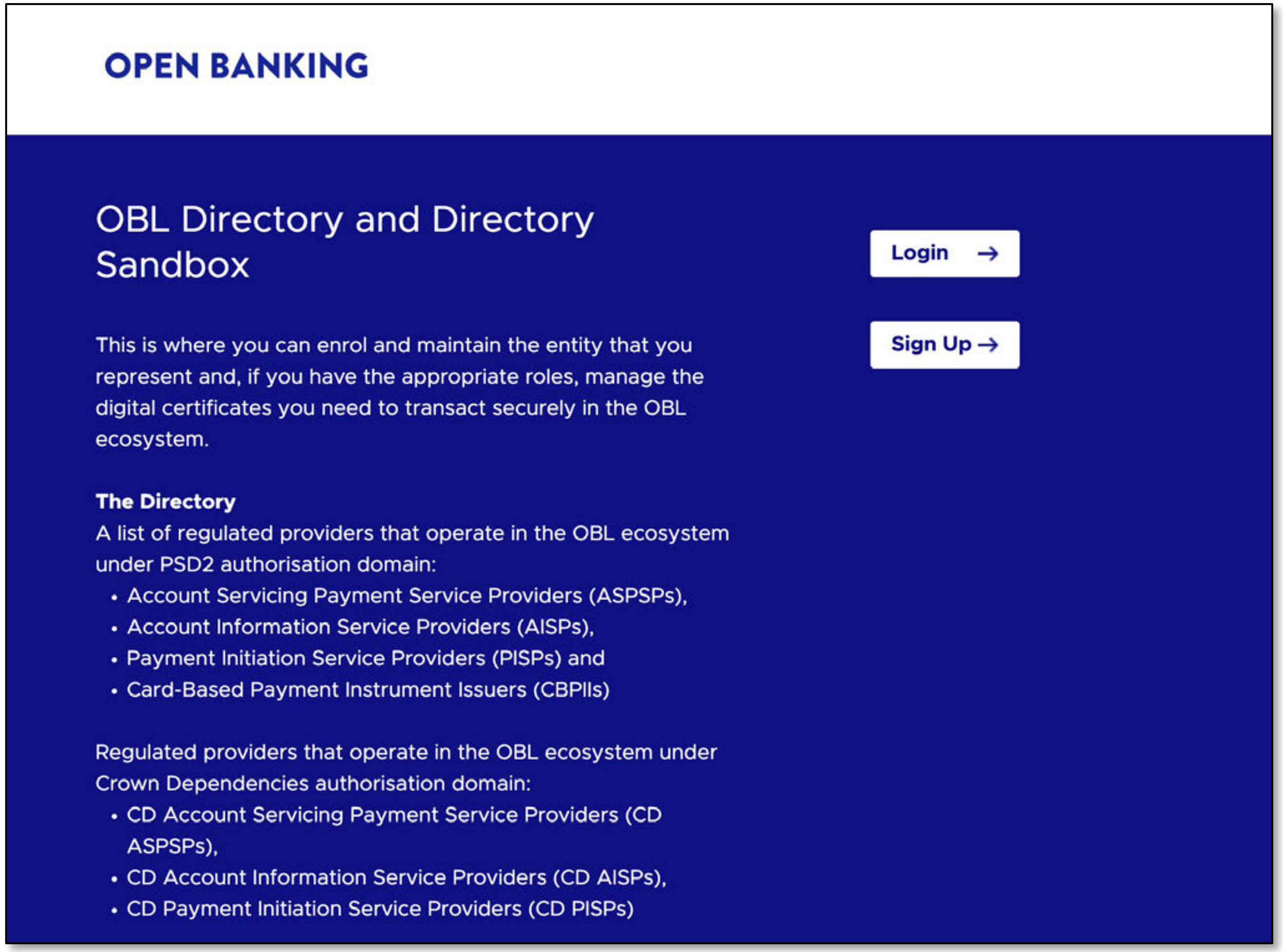
TFL UNIFIED API



https://api.tfl.gov.uk/

- Open access with API key registration
- RESTful APIs, XML, JSON
- API key-based access, no federated IAM
- Static and dynamic metadata per feed
- Only non-sensitive data published
- Centralised API gateway, cloud-hosted

OPEN BANKING INITIATIVE


OPEN BANKING

OBL Directory and Directory Sandbox

Login →

Sign Up →

This is where you can enrol and maintain the entity that you represent and, if you have the appropriate roles, manage the digital certificates you need to transact securely in the OBL ecosystem.

The Directory

A list of regulated providers that operate in the OBL ecosystem under PSD2 authorisation domain:

- Account Servicing Payment Service Providers (ASPSPs),
- Account Information Service Providers (AISPs),
- Payment Initiation Service Providers (PISPs) and
- Card-Based Payment Instrument Issuers (CBPIIs)

Regulated providers that operate in the OBL ecosystem under Crown Dependencies authorisation domain:

- CD Account Servicing Payment Service Providers (CD ASPSPs),
- CD Account Information Service Providers (CD AISPs),
- CD Payment Initiation Service Providers (CD PISPs)

https://www.openbanking.org.uk/directory/

- Consent-based sharing via accredited apps
- RESTful APIs, Open Banking API specOAuth2, OpenID Connect, central directory of providers
- Rich metadata, versioned schemas
- Customer consent, strict sensitivity rules
- Federated architecture, secure cloud APIs

THE COMMON EUROPEAN ENERGY DATA SPACE FRAMEWORK - Ongoing

- CEEDS is a European Union initiative to create a unified, secure, and interoperable data space for energy data across Europe.
- It aims to enable seamless sharing of energy-related data (e.g., flows, tariffs, operational metrics) between countries, companies, and sectors.
- Projects like INT:net and ENERSHARE are developing the technical and governance blueprints for CEEDS, focusing on standards, trust frameworks, and federated architectures.

Additional notes – Common European Dat Spaces is also being rolled out for 14 sector/domains – Agriculture, heritage, Health, Media, Mobility etc.

https://digital-strategy.ec.europa.eu/en/policies/data-spaces

MIND&MATTER

GTDSI WP2

AI-Driven Data Classification & Access Modelling

Contents

WP1 Wrap-Up: Stakeholder Personas & Requirements

Mapping current and prospective data users

Industry Users

Gas shippers, DNOs, traders seeking real-time data and easier API access.

Regulators

Ofgem expecting alignment with data openness policies.

Innovators

Tech firms and researchers as new users needing greater data availability.

Current Data Models: Open vs. Closed Trade-offs

Strengths and limitations of current approaches

Gas Data Portal (Open) 🌐

Public web portal, no login required. Approximately 12,000 data series accessible, enabling high diverse usage that fosters broad innovation and reuse. Lacks user-level control or tracking.

Gemini (Closed) 🔒

Secure system, login required for access. Hosts critical operational data for authorized market participants only, with intensive 24/7 usage and fine-grained access control that can stifle broader innovation.

Detailed Model Comparison: Openness vs. Control

Analysis of how current dual approach impacts data access, security, and innovation potential across energy sector stakeholders.

Gas Data Portal

- **Public web access — ~70M requests/year, open to broader ecosystem.**
- **12,000+ open series available for public use and analysis.**
- **High innovation potential driven by public ecosystem and third-party developers.**
- **Limitations: No authentication — cannot authenticate users, limited user control.**

Gemini System

- **Secure login required — ~2M calls/day from operators, access controlled.**
- **Operational data is restricted and intended for authorized users only.**
- **Fine-grained access control enabling role-based and data-level permissions.**
- **Limitations: Rigid access model — excludes new users and slows onboarding.**

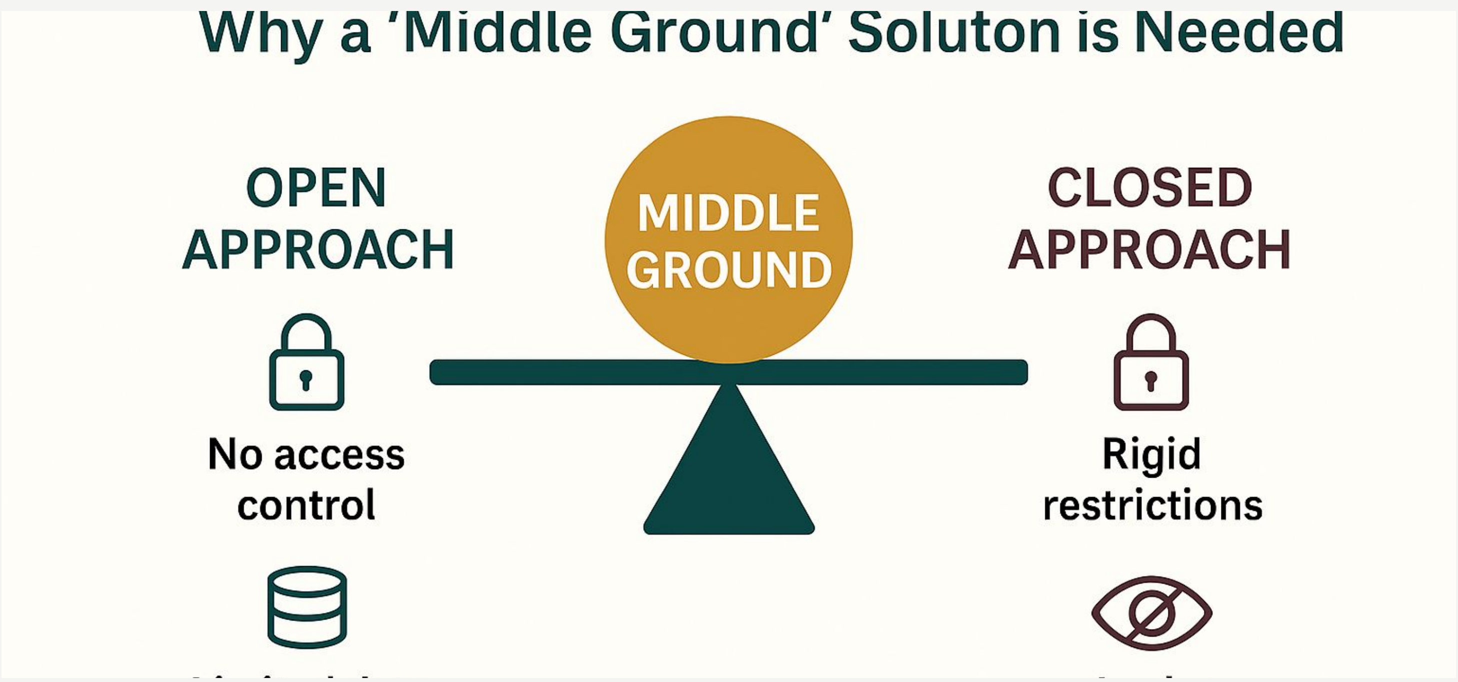
Why a Middle Ground Solution is Required

Limitations of Exclusively Open Approach

Cannot serve sensitive or granular data due to lack of authentication; no ability to restrict or track usage; excludes data stakeholders who need access control; innovation limited by data availability constraints.

Limitations of Exclusively Closed Approach

Rigid silos prevent data sharing with legitimate new users; excludes researchers, innovators, and new market entrants; stifles broader innovation and ecosystem growth; difficult to onboard new stakeholders; limits transparency and data openness objectives.



Identified Data Gaps & Critical User Needs

Information unavailable under current models

Granular Real-Time Data

Industry users (shippers, DNOs, traders) request more real-time, detailed operational data than currently available via open portal.

User-Specific Data Views

Need for customized data subsets based on user role (regulator vs. operator vs. researcher) not supported by one-size-fits-all portal.

Sensitive Commercial Data

Certain commercial and operational sensitive data locked in Gemini, inaccessible to legitimate broader stakeholders and researchers.

API Access Ease

Request for easier, more intuitive API access with better documentation and authentication flows.

Actionable Next Steps: WP2 Preparation Framework

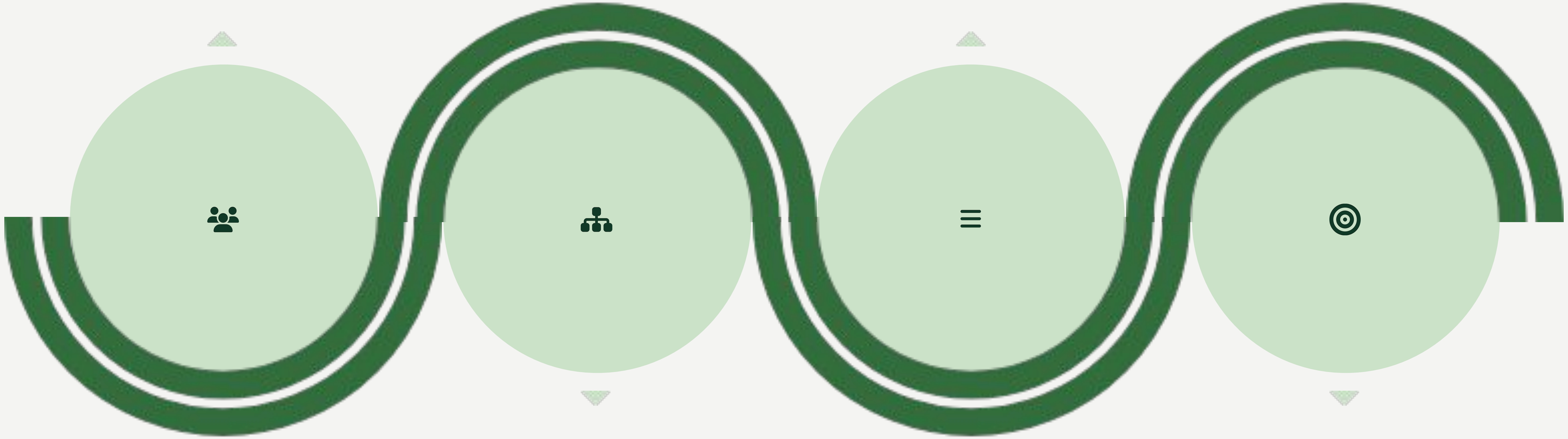
Key deliverables for bridging the gap

Milestone 1 - Document

Stakeholder Insights

Milestone 3 - Extract

WP1 Recommendations



Milestone 2 - Map

Data Categorization

Milestone 4 - Define

Access Model

How to Navigate This Technical Deck

For All Audiences

The summary tables (slides 10-13) provide high-level overview of all methods, approaches, and frameworks evaluated in this feasibility study. These tables are designed to be self-contained for readers who need to understand the 'what' and 'why' without deep technical detail.

For Technical Readers

Parts A through F contain detailed feasibility analysis, including worked examples with Gemini and Gas Data Portal datasets, ML model architectures, feature engineering approaches, and implementation considerations. This content can also serve as the basis for a standalone technical paper.

Structure

- Part A (AI Classification Workflow)
- Part B.1 (Supervised ML)
- Part B.2 (Unsupervised)
- Part C (Dynamic Access Control)
- Part D (Data Sensitivity Levels)
- Part E (Risk-Based Access Matrix)
- Part F (Governance & Compliance)

MIND&MATTER

Summary: AI Classification Approaches

This table summarises the four classification strategies evaluated in this study. The hybrid approach is recommended for NGT. Detailed analysis and worked examples follow in Parts A and B.

Approach	Best For	Accuracy	Pros	Cons
Rule-Based (Pattern Matching)	Structured data with clear sensitivity markers	High for known patterns	Transparent, auditable, fast to deploy, no training data needed	Cannot handle novel patterns, brittle to schema changes
Supervised ML (Labelled Data)	Datasets with existing classification history	High (with quality labels)	Learns complex patterns, adapts to new data types, scalable	Requires labelled training data, ongoing retraining, less interpretable
Unsupervised (Clustering/NLP)	Unstructured documents without labels	Moderate	No labels needed, discovers hidden patterns, scales to large volumes	Requires human validation, cluster interpretation can be subjective
Hybrid (Recommended)	NGT's mixed data landscape	Highest overall	Combines rule precision with ML adaptability, human-in-the-loop for risk mitigation	More complex to implement and maintain, requires governance framework

Summary: Access Control Methods

Comparison of access control mechanisms evaluated for the Gas DSI. The risk-based model is recommended to balance security with data openness. Details in Parts C–E.

Method	Mechanism	Use Case	Pros	Cons
Role-Based Access Control (RBAC)	Permissions assigned by user role	Gemini market participants	Simple, well-understood, easy to audit	Coarse-grained, cannot adapt to context
Risk-Based Access Model	Role + sensitivity + context attributes	Recommended for Gas DSI	Fine-grained, context-aware, supports regulatory compliance	More complex to implement, needs classification engine
Dynamic Access Control (Non-NER)	Schema & metadata-driven rules	Structured datasets without PII detection	No NLP dependency, works with existing metadata	Cannot detect embedded sensitive content
Federated Identity (OIDC/OAuth 2.0)	Cross-organisational SSO tokens	DSI cross-org access	Standards-based, scalable, supports MFA	Requires identity provider infrastructure setup

MIND&MATTER

Summary: Regulatory & Compliance Framework

Overview of regulatory obligations relevant to the Gas DSI. Full compliance analysis including incident response and audit trail requirements is provided in Part F.

Regulation	Scope	Key Requirement for DSI	Status
GDPR	Personal data protection	Breach notification (72hr), access logs, data minimisation	Partially compliant
NIS2 Directive	Critical infrastructure cybersecurity	Cryptography, access controls, incident detection & reporting	Gap identified
Ofgem Data Best Practice	Energy sector data governance	Data visibility, presumed open principle, controlled access rationale	Aligned in principle
REMIT	Energy market integrity	Market data transparency, insider trading prevention	Compliant via Gemini
EU AI Act	AI system regulation	Transparency, human oversight, risk classification of AI models	Future requirement
Cyber Resilience Act	Digital product security	Security-by-design, vulnerability management	Future requirement

Summary: Technical Methods Evaluated

All ML/AI techniques assessed during this feasibility study with maturity level and deployment recommendation. Detailed walkthroughs are provided in the corresponding Parts.

Technique	Category	Application	Maturity	Recommended
Keyword/Regex Detection	Rule-Based	Detecting known sensitive fields	High	Yes - Phase 1
Metadata Inference	Rule-Based	Source/schema-based classification	High	Yes - Phase 1
Random Forest / Gradient Boosting	Supervised ML	Multi-class sensitivity classification	High	Yes - Phase 2
BERT / Transformers	Deep Learning	Semantic sensitivity analysis	Medium	Evaluate
K-Means / DBSCAN	Unsupervised	Document grouping by sensitivity	Medium	Yes - Phase 2
Topic Modelling (LDA)	Unsupervised	Theme extraction from documents	Medium	Evaluate
NLP Entity Detection	NLP	PII / sensitive entity flagging	High	Yes - Phase 1
Graph Neural Networks	Advanced ML	Relational data classification	Low	Future
Weak Supervision (Snorkel)	Semi-supervised	Label generation from heuristics	Medium	Evaluate

Incident Response & Regulatory Compliance Framework

Compliance Alignment

Shield — NIS2 Directive

Cryptography compliance, precise access controls, incident detection and reporting within 72h, regular security testing.

File-contract — GDPR

Personal data protection, audit trails for GDPR breach notification, access logs retention, and support for data subject rights.

Check-circle — Ofgem Data Best Practice

Data visibility and controlled access principles, clear rationale for classifications, auditability via logs, and federation support.

Incident Response Workflow

1. Detection

Dynamic monitoring flags suspicious activity.

2. Alerting

CSIRT receives immediate alert.

3. Response

Suspend accounts and access via IAM.

4. Recovery & Reporting

Contain incident and notify stakeholders within regulatory timelines.

> ***Compliance is embedded in architecture, not bolted-on.***

Part A - AI-Driven Classification

A Feasibility Study

Transitioning to AI-Driven Classification

The path forward for data governance

Feasibility

Using AI to classify data from Gemini and Gas Data Portal into risk-based sensitivity levels (Public, Internal, Confidential, Restricted) is feasible and beneficial.

Data Characteristics

Both systems have structured data with useful metadata; Gemini is transactional and secure, while Gas Data Portal is aggregated and public.

Regulatory Framework

AI must operate within GDPR requirements and Ofgem's Presumed Open policy for energy network data.

Human Oversight Essential

AI acts as an assistant; highly sensitive categories always require human review for accuracy and compliance.

Strategic Impact: NGT's Industry Leadership Position

NGT's position as innovator

Industry Innovation

NGT's adoption of AI for data classification sets a precedent for other network operators. Positions NGT as a thought leader in energy data governance.

Competitive Advantage

Automated classification enables faster, more accurate data sensitivity triage. Reduces manual overhead for data governance teams and scales as data volumes grow.

Alignment with Trends

Reflects the broader industry movement toward automation and AI-driven governance. Supports Ofgem's data transparency and openness objectives and enables rapid response to evolving regulatory requirements.

Security & Governance Considerations

Protecting sensitive data throughout the process

AI Process Security

Ensure sensitive data is not inadvertently leaked to third parties or through logs. If using cloud services (Azure, AWS), evaluate data governance carefully. Prefer in-cloud analysis to keep data within a secure tenant.

Regulatory Alignment

GDPR compliance requires AI to reliably detect PII and label it as highly sensitive. Ofgem Presumed Open principle demands AI consistently triage data into Open/Shared/Closed categories. Integration with the overall data governance framework is critical for success.

Implementation Framework: AI Classification Workflow

Integration with NGT governance

Step 1: Data Preparation

Create training dataset, tag metadata

Step 2: Rule Definition

Encode PII rules, pattern matching

Step 3: ML Training

Train model, validate accuracy

Step 4: Hybrid Workflow

Route cases: rules → ML → human

Step 5: Governance

Feed to catalogs, continuous learning

Mitigating Misclassification Risk

Ensuring accuracy and regulatory compliance

Err on Side of Caution

False negatives (missing sensitivity) are more damaging than false positives; when uncertain, classify as more sensitive to reduce harm and regulatory exposure.

Continuous Learning Loop

Human feedback is continuously fed back into ML retraining so the system improves over time as more examples are reviewed and labeled accurately.

Edge Case Documentation

Context-dependent sensitivity and dataset combinations are carefully documented; edge cases may require manual intervention or advanced rule handling.

Explainability & Audit Trail

AI must output understandable reasons for classification to enable auditing, refinement, and to build regulatory trust through transparent decision records.

MIND&MATTER

AI Classification Approaches: Feasibility & Trade-offs

Evaluating options for implementation

Rule-Based

High Precision ↔ Inflexible

- ✓ Fast
- ✓ Transparent
- ✗ Novel patterns

Supervised ML

Interconnected nodes → learning

- ✓ Adapts
- ✓ Complex patterns
- ⚠️ Needs labels

Cloud AI

Cloud with lightning — powerful, managed

- ✓ Quick deploy
- ✓ Scalable
- **$ Usage costs**

Data Source Characteristics for AI Classification

Gemini System

Millions API calls/day

Commercially Sensitive

Rich Metadata

- Structured operational and market data, including capacity bookings, nominations, and flows, suitable for tabular ingestion and relational joins.
- Semi-structured outputs such as logs and reports that are mostly numeric or categorical and require light parsing or normalization.
- High volume: millions of API calls per day, continuous streaming and frequent incremental updates.
- SENSITIVITY: Commercially sensitive (individual shipper positions and contracts) and operationally sensitive (real-time network status); no direct personal data present.
- METADATA: Rich metadata including dataset name, source module, update frequency, unit measures, and data owner tags to support provenance and governance.

Gas Data Portal

12,000+ Series

Public by Design

Aggregated Data

- Structured, aggregated datasets designed for system-wide analysis, typically anonymized and aggregated at cohort or region level.
- Consistent formats such as time-series, tabular exports, and APIs/CSV endpoints enabling straightforward automated ingestion.
- Examples include daily demand series, aggregate flows, and day-ahead forecasts provided as regular feeds.
- SENSITIVITY: Public by design with little or no confidential information, though risk exists of re-classification if granularity increases or datasets are combined to reveal sensitive details.
- METADATA: Descriptive metadata including title, field definitions, update frequency, and historically known public status to aid discoverability.

Part B.1 - Supervised ML Classification

Classification with Labelled Data

Supervised Machine Learning Classification Strategy

MIND&MATTER

100+

Manually Labeled Datasets

Rule-Based System

(high precision)

Labeled Examples

(training data)

ML Model

(ambiguous cases)

Prepare

Data mapping & labeling

Train

Model training

Validate

Held-out testing

Deploy

Proof-of-concept

Refine

Iterate with new labels

MIND&MATTER

ML Feature Engineering: What the Model Learns

Metadata Features

Source system (Gemini vs Portal), update frequency (real-time vs monthly), number of fields, key words in field names.

Content Features

Numeric ranges (aggregated vs individual data), entry distribution patterns, value magnitude analysis, granularity indicators.

NLP Features

Text vectorization from descriptions, sample content analysis, domain-specific terminology, sensitivity keyword detection.

Example: Model learns "source=Portal" nearly always indicates Public; contains field "Postcode" likely indicates Restricted (confidential). Shows how model picks up combinations of factors that rule-based systems miss.

ML Model Validation & Interpretability

Model Performance & Validation

- Hold out known test cases for accuracy measurement.
- Expect high accuracy for clearly Public vs clearly Restricted cases.
- Grey areas require human review.
- Continuous refinement as new data types appear.

Interpretability & Explainability

- Use interpretable algorithms (decision trees, explainable boosters) for transparency.
- Apply SHAP values or LIME to explain complex model predictions.
- Ensure Data Governance Board can understand classification reasoning.
- Document model decisions for audit trails.

Feasibility is good - computational tools standard (Python scikit-learn), problem is classic multi-class classification, main effort is preparing quality training data.

Example 1: Classifying Gemini Daily Nominations

Dataset Details

- Dataset name: Daily Nominations by Shipper
- Content: Lists each shipper's nominated gas volume per entry point per day, recorded hourly where applicable
- Columns: Shipper ID, Entry Point, Hourly Volume, Date, Timestamp, Unit of Measure, Remarks
- Records: Approximately 50 records (10 shippers × 5 entry points), with multiple hourly entries per day
- Update frequency: Daily or real-time updates depending on source feed

Expected Sensitivity & Training Label

- Expected sensitivity: Confidential — contains company proprietary gas flow information that could affect market fairness if shared openly
- Access: Restricted to the relevant shipper and authorized regulators only; not for broad distribution
- Training label: Confidential — "Contains company-specific operational data"

MIND&MATTER

Example 1: ML Feature Analysis & Model Output

Feature Values for Gemini Dataset

- Source system = Gemini
- Update frequency = Daily (real-time)
- Number of fields = 5
- Field keywords = {Shipper, Volume}
- Number of records ≈ 50
- Distribution: Relatively granular (multiple entries per shipper)
- Description: "each shipper's nomination"

Model Recognition Pattern

- Decision tree first node: "if field name contains 'Shipper' → predict not Public"
- Deeper node: "if source=Gemini AND records > 1 → predict Confidential"
- Logistic regression: Positive weights for 'Gemini' and 'Shipper field', negative weights for 'Portal' and 'Aggregate'

SHAP Explanation

- Top factors: "Contains 'ShipperID' (+0.8 towards Confidential), Source=Gemini (+0.5), Description contains 'each shipper' (+0.3)"
- Note: "Aligns with expert expectations, provides clear justification."

Example 1: Production Outcome & Benefits

[redacted]

1. ML Classification Output

"Confidential" (high confidence). Not published openly on portal, restricted to owners and regulators.

[redacted]

2. Production Access Control

Model tags similar nominations datasets automatically every day without expert manual intervention, saving significant effort while maintaining correct protection level.

[redacted]

3. Operational Benefit

Summary: outcomes feed into automated monitoring and processing pipeline; key benefits shown in the metrics below.

80%

Reduces manual review time per dataset

100%

Maintains security compliance

24/7

Enables daily automated processing

Example 2: Classifying Gas Data Portal Dataset

Dataset Details

- Dataset name: Daily Linepack Summary — a daily aggregated record of system linepack values for each day of the year.
- Source: Gas Data Portal — published by the system operator for public transparency and reference.
- Content: Shows total linepack (gas volume in system) recorded at the end of each day as a single aggregated value for that day.
- Columns: Date and Linepack (in GWh), representing the day and the corresponding aggregated system volume measurement.
- Records: Approximately 365 entries, typically one record per day covering a full year of observations.
- Update frequency: Daily updates, with a new aggregated linepack value published for each day.
- Data type: Single aggregated figure per day, not containing per-customer or granular confidential details.

Expected Sensitivity & Training Label

- Expected sensitivity: Public — this dataset contains aggregate system-level data with no confidential or identifiable individual-level details, suitable for public release.
- Training label: Public — reasoning: "Aggregate system data, safe for open publication" because values are system totals without customer-level granularity.
- Pre-vetted status: Already pre-vetted as open and published by NGT for transparency; serves as an example of openly shareable dataset content.

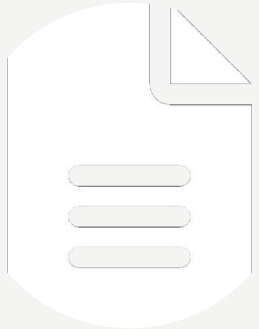
Example 2: ML Feature Analysis & Model Output

PREDICTION: PUBLIC

Confidence

95%

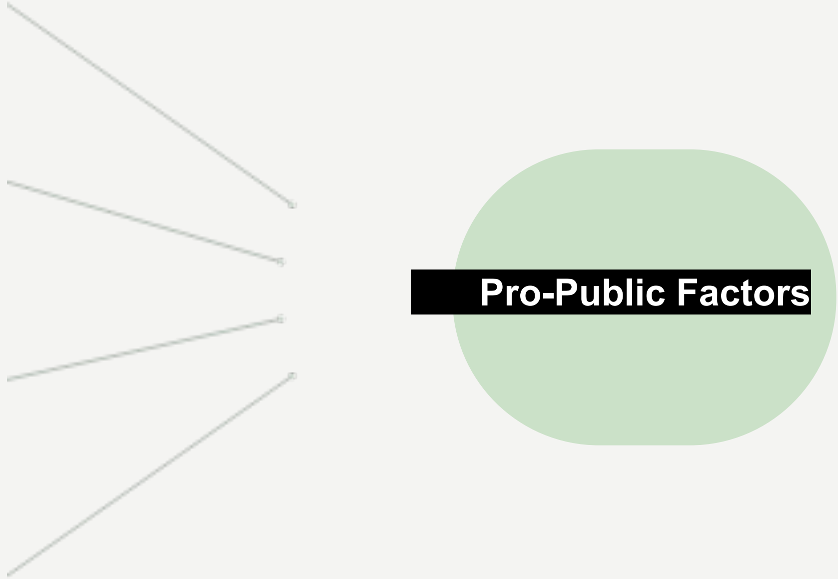
Portal
Source



365
Records

2
Fields

[redacted]	**Source: Portal** ✔ Strong pro-Public signal (green)
[redacted]	**Date field only** ✔ Non-identifying, supports Public
[redacted]	**Summary title** ✔ Aggregated/summary semantics (pro-Public)
[redacted]	**No identifiers** ✔ No personal/company identifiers detected



Example 2: Production Outcome & Validation

[redacted]

ML Classification Output

"Public" (95% confidence). Dataset confirmed safe for open portal publication (already published). ML "agrees" with current state, reinforcing correct classification.

[redacted]

Production Confirmation

Secondary insight: If Portal dataset currently withheld, ML might flag Public if no sensitive content → triggers review opportunity to identify over-protected data.

[redacted]

Validation Insight

Ensures consistency: anything intended as public is recognized by model, preventing mistaken lockdown of open data.

Feasibility: ML Model Integration Approaches

Hybrid Approach (Recommended)

- Rule-based system classifies with high confidence.
- ML model handles remainder and ambiguous cases.
- Maintains high precision as rules catch straightforward labels while boosting recall because ML catches subtle patterns.
- Example: Obvious cases like "filename contains 'Confidential'" are caught by rules, ML focuses on nuanced patterns.
- Training: Rule-agreed cases serve as initial labeled data for the ML model.

Double-Check Approach

- ML model runs on all datasets as a verification layer.
- If ML and rules agree: proceed.
- If disagreement: flag for human review.
- Benefit: Over time, as the model improves, disagreements drop.
- Eventually lean more on ML with rules maintained as a safety net.

Computational feasibility: Very good - modest load on standard hardware (scikit-learn on laptop trains in seconds even with 1000 examples).

MIND&MATTER

Feasibility: Model Maintenance & Explainability

🔁 Model Maintenance

Quarterly

- Central cadence for retraining and review

1

🔎 New Data Types

Add representative examples (e.g., hydrogen blend) for coverage

2

🧑‍⚖️ Expert Labels

Domain experts label a few representative examples for accuracy

3

🔄 Model Update

Incremental updates incorporate novel patterns into the model

4

📈 Monitor Output

Production outputs are reviewed; anomalies are fed back to training

5

"Model rationale mirrors expert reasoning, executed faster and consistently."

💡 Explainability Flow

1

🤖 AI Decision

Model produces a classification or decision in production

2

🔬 SHAP Analysis

Generate per-instance explanations showing feature contributions

3

🛡️ Governance Review

Present explanations to governance and auditors to build trust

Part B.2 - Unsupervised Classification

Classification Without Labels

Classification Without Labels: Unstructured Data Strategy

Determining sensitivity when labeled examples are unavailable

1 Rule-Based Heuristics

Create explicit rules based on known sensitivity patterns and indicators.

2 Leverage Metadata

Use source system, file type, update frequency, and author information as features.

3 Apply NLP Entity Detection

Scan text for sensitive entities (PERSON, ORG, financial data, custom identifiers).

4 Clustering & Topic Modelling

Use unsupervised learning to group documents and identify sensitivity traits.

5 Human-in-the-Loop

Route uncertain classifications to human reviewers based on confidence thresholds.

Key Success Factors for Label-Free Classification

MIND&MATTER

Ensuring accuracy and compliance without pre-labeled data

Expert Rule Definition

Domain expertise is crucial for creating effective rules. Energy sector knowledge required. Regular review and updates to the rule set ensure relevance and accuracy.

Quality Metadata Usage

Invest in rich metadata collection and tagging. Source, owner, frequency, and sensitivity hints are critical. High-quality metadata drives inference accuracy.

Confidence Thresholds

Set realistic thresholds balancing automation versus manual review. Err on the side of caution with higher sensitivity classification and monitor false positive and negative rates.

Human Expertise

Domain experts are essential for edge cases. Build consensus on sensitivity criteria and provide ongoing training and calibration for reviewers.

Continuous Feedback

Use human classifications to improve rules and plan transition to supervised ML. Document learnings and recurring patterns to refine the system.

MIND&MATTER

Integrating No-Labels Approach with Overall Strategy

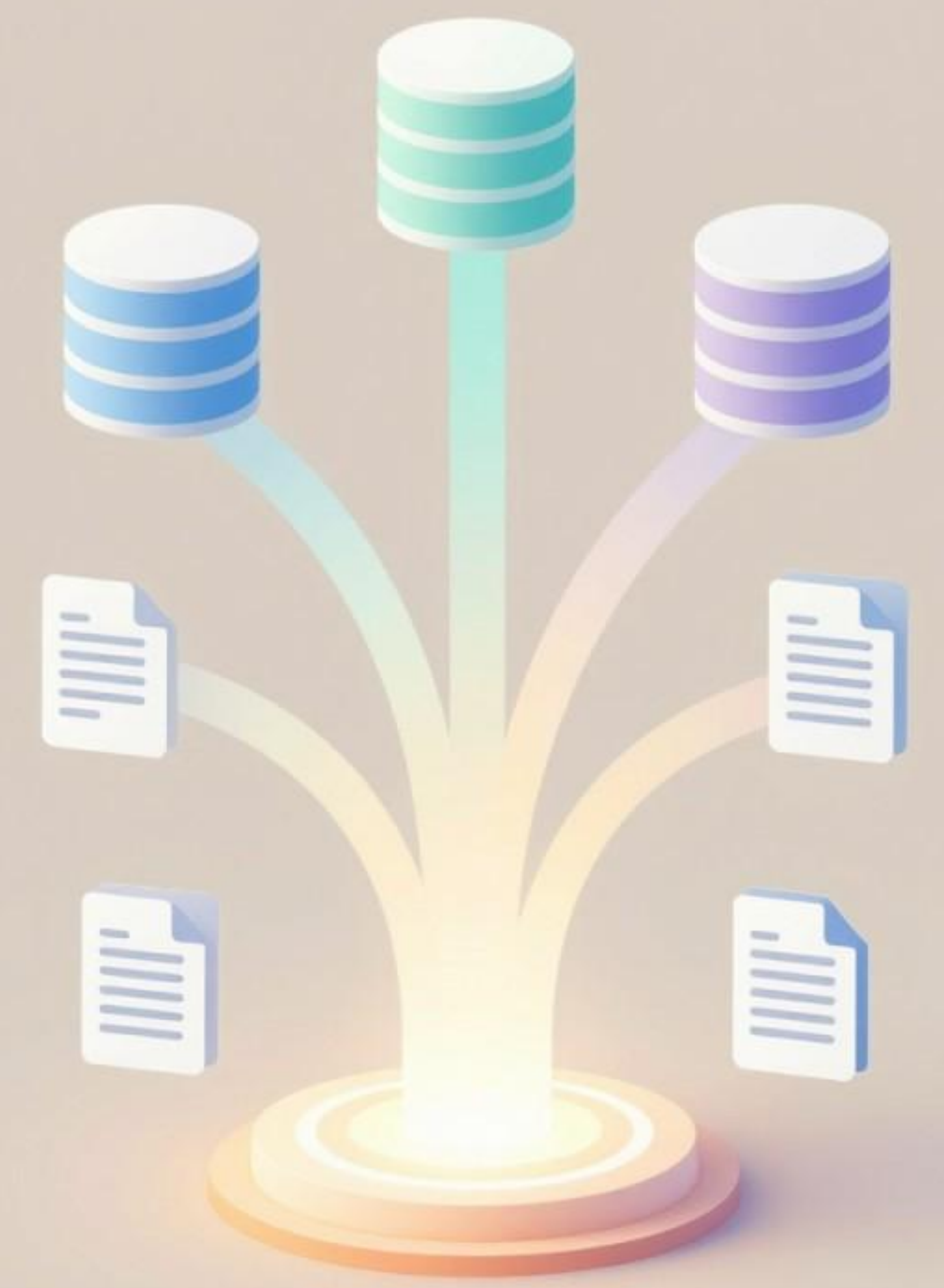
Combining structured and unstructured data classification

Structured Data Path

Sources: Gemini, Gas Data Portal. Leverage metadata-rich features and pre-existing system tags and documentation. Apply rule-based classification for deterministic mapping. Low need for human review due to clear structure and explicit metadata.

Unstructured Data Path

Sources: reports, logs, documents. Use rule-based heuristics combined with NLP entity detection. Apply clustering and topic modelling for discovery. Higher need for human review and validation; create labeled datasets for future ML training and progressively transition to supervised models as labels accumulate.



Implementation Workflow: No-Labels Classification

End-to-end process for unstructured data

Phase 1: Rule Definition & NLP Setup

Define sensitivity indicators and construct the initial rule set. Set up NLP entity recognition using tools such as spaCy and AWS Comprehend. Extract required metadata fields and prepare unstructured data sources for processing. Establish baseline rules and thresholds to guide automated detection.

Phase 2: Initial Automated Classification

Run rule-based detection across the document corpus and execute NLP entity scanning. Assign confidence scores to each classification result. Separate high-confidence items for automatic classification and isolate uncertain or low-confidence cases for further review.

Phase 3: Human Review & Labeling

Subject matter experts review uncertain and low-confidence documents and assign correct sensitivity labels. Document the reasoning for edge cases and ambiguous items. Create a labeled dataset from reviewed documents to support future supervised training.

Phase 4: Refinement & Scaling

Update rules and thresholds based on human feedback and labeled examples. Consider training supervised ML models on the labeled subset to expand automated coverage. Automate an increasing portion of classifications over time while continuously monitoring accuracy and planning iterative improvement cycles.

Human-in-the-Loop Review Process

Balancing automation with expert oversight

1

Step 1: Set Confidence Thresholds

Define high, medium, and low confidence levels using a combination of rule-based detection signals and ML/NLP scoring metrics. High confidence above 85% will be auto-classified without human intervention, medium confidence between 50% and 85% will be flagged for human review, and low confidence below 50% will be routed directly to manual classification workflows.

2

Step 2: Rule-Based Classification

Apply deterministic rule-based filters to capture obvious cases and strong pattern matches that can be reliably categorized. When these rules produce high confidence outcomes they move directly to final classification, reducing the volume of documents that require further ML scoring or human attention.

3

Step 3: ML/NLP Scoring

Run NLP pipelines for entity detection, text classification, and other model-based analyses on documents not resolved by rules. Aggregate model outputs into confidence scores for each candidate classification; lower confidence results trigger flags for human review while higher scores proceed toward automated decisions.

4

Step 4: Human Review & Validation

Domain experts review the flagged or uncertain cases, provide final classification decisions, and validate model outputs. Approved classifications are added back into the training dataset, and reviewers supply feedback to iteratively refine rule sets and retrain models to improve future automation accuracy.

Clustering & Topic Modelling for Large Volumes

Unsupervised learning approaches

Topic Modelling (LDA)

Group documents by themes and topics. Identify themes that consistently appear with sensitive data. Manually label representative clusters. Infer sensitivity of new documents based on cluster assignment.

Document Clustering

Find similar documents using similarity metrics. Group documents that share sensitivity traits. Manually label a few representative clusters. Use cluster properties to classify new documents. Effective for discovering hidden groupings without labels.

NLP for Sensitive Entity Detection

Automated scanning for content-based sensitivity indicators

Named Entity Recognition

PERSON, ORG, LOC entities may indicate personal or commercial sensitivity. Extract and flag named entities in unstructured text to detect potential privacy or commercial exposure risks.

Financial & Operational Entities

DATE, MONEY, QUANTITY entities may indicate financial or operational data. Flag datasets containing financial terms, contract values, or transactional amounts for further review and protection.

Custom Domain Entities

Add custom entities specific to the energy sector such as Shipper ID, Contract Number, Nomination Data, and Capacity Booking. Use domain-specific pattern matching to improve recall.

Text Classification

Use NLP models to classify documents by sensitivity level. Identify paragraphs with restricted information and flag mixed-sensitivity documents for appropriate handling.

Leverage Metadata for Inference

Available Metadata Cues

- **Source system — distinguish between Gemini and Gas Data Portal as different provenance contexts that affect expected sensitivity and access patterns.**
- **File type — identify whether files are PDF reports or CSV exports, since formatted reports often contain narrative detail while CSVs contain raw numeric records.**
- **Update frequency — note if data is real-time versus monthly, with real-time feeds more likely to include current, sensitive operational information.**
- **Author or owner — consider whether the author is an internal team or a public-facing group, as internal authors often signal restricted or operational content.**
- **Dataset size and scope — assess the breadth and granularity of the dataset, since larger, more granular datasets can increase re-identification risk and sensitivity.**

Inference Examples

- **Source = Gemini + real-time + numeric data → likely Confidential, because real-time numeric streams from Gemini often contain operationally sensitive measurements.**
- **Source = Gas Data Portal + monthly + aggregated → likely Public, since monthly aggregated exports are designed for broad distribution and reduced sensitivity.**
- **Author = Internal ops team + financial terms present → likely Internal, given the combination of internal authorship and finance-related content.**
- **Author = Public communications → likely Public, when the authoring source is intended for external audiences and lacks restricted details.**

Using available metadata to infer sensitivity levels

Human Oversight & Continuous Learning in Practice

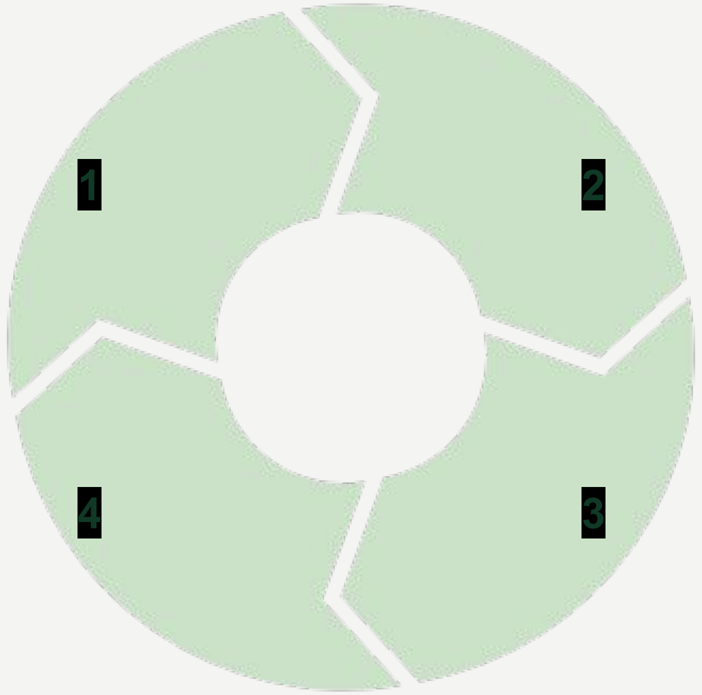
How feedback loops improve AI classification over time

Stage 1: Initial Classification (AI-Driven)

AI processes new Gemini reports using rules, metadata, and NLP. It outputs a classification with a confidence score and flags high-stakes cases (anything Confidential or Internal) for review. Audit logs of all decisions are maintained for traceability.

Stage 2: Human Verification & Feedback

A data steward reviews flagged decisions and verifies whether Gemini auction data is truly confidential (YES in real scenarios). If the AI misclassified an item, the human correction is recorded. Feedback is logged into the system for learning.



Stage 4: Precision Improvement Loop

Over time there are fewer false positives and negatives, and humans handle only genuinely uncertain cases. Logs are kept for audit and regulatory compliance. The system becomes smarter while remaining under expert supervision, and routine cases are eventually fully automated.

Stage 3: Refinement & Rule Updates

Well-understood cases, such as Gemini auction data, become routine as rules are refined. AI precision improves through accumulated feedback. New or unusual data types are routed for human decision, and edge cases are documented and potentially added as new rules.

Advanced ML & Deep Learning for Data Classification

Leveraging modern AI techniques for intelligent sensitivity detection

Transformer Models (BERT)

Effective for text and content classification. Excel at understanding context and capturing relationships between words. Interpret ambiguous domain terms and handle limited data through transfer learning.

Large Language Models (LLMs)

Fine-tune on labeled examples to inject domain knowledge. Use few-shot prompting for scarce-label tasks and ingest metadata and schema as additional text. Capture language nuances such as "for internal use only" and apply domain-adaptive pre-training on energy-sector corpora.

Graph Neural Networks (GNNs)

Incorporate structure and relationships in data to model connections within and between datasets. Handle ambiguous variable names via linkages and integrate metadata and relational information for schema-aware learning capabilities.

Hybrid Approaches

Combine Transformer and GNN strengths: Transformers handle textual nuance while GNNs manage metadata, relationships, and source systems. Leverage complementary capabilities to achieve enhanced context-aware classification.

Ensemble Systems

Combine multiple independent models to reduce single points of failure. Provide multiple checks and consensus, enable compliance through hard-coded constraints, and improve overall robustness and accuracy.

Training Paradigms & Learning Strategies

Selecting the right approach based on data and constraints

Supervised Learning

Ideal approach with complete labeled examples. Transformer models often achieve the highest accuracy. Requires substantial labeled training data, which can be costly to obtain from experts. Best when the budget allows for comprehensive labeling effort.

Semi-Supervised Learning

Leverages both labeled and unlabeled data. Self-training lets the model predict and learn from unlabeled examples; label propagation spreads labels through similar instances; weak supervision uses multiple noisy rules to label data. Significantly reduces the labeling burden.

Unsupervised Learning

Used for feature extraction and data understanding. Techniques like clustering discover similar data groups; topic modeling identifies thematic patterns. No labeled examples needed, making it suitable for exploratory phases and pattern discovery.

Active Learning

The model identifies the most informative examples for expert review, focusing human effort on impactful cases. Strategic selection of data to label maximizes learning under a limited labeling budget and provides an efficient path to high-accuracy models.

Continuous Learning

Incrementally improves as new labeled data arrives through feedback loops from human reviews. Adapts to an evolving data landscape with regular retraining cycles so the system becomes smarter over time.

Weak Supervision & Limited-Label Scenarios

MIND&MATTER

Leveraging noisy rules to label large datasets automatically

Weak Supervision Concept

Use multiple noisy rules as teachers rather than perfect labels. Rules don't need to be 100% accurate. Different rules can be imperfect and conflicting. Aggregate votes to create probabilistic labels.

Snorkel Framework

Modern weak supervision platform for limited-label scenarios. Define programmatic labeling rules. Automatically label large unlabeled datasets. Train powerful models on weakly-labeled data. Iteratively improve rules.

Energy Sector Application

Domain experts write sensitivity detection rules (e.g. 'if Gemini + market terms = likely Confidential'). Rules capture domain heuristics (partial patterns, probabilistic signals). Auto-label thousands of datasets using rule votes. Train transformer model on weakly-labeled examples. Reduce need for manual expert labeling.

Benefits for NGT

Reduces expensive manual labeling requirements. Works with incomplete or imperfect rules. Scales to large data volumes. Iteratively refine rules as system learns. Bridges gap between rule-based and ML approaches.

Hybrid ML Approaches: Rules Plus Models

MIND&MATTER

Combining precision of rules with contextual judgment of ML

Rule-Based Layer (Hard Overrides)

Must-confidential rules for absolute matches. Known patterns that always indicate sensitivity. Policy-mandated classifications. Blocking layer to prevent false negatives. Transparent, auditable decision points.

ML Model Layer (Contextual Judgment)

Transformer and GNN models for nuanced classification. Handles ambiguous cases and edge scenarios. Contextual sensitivity determination. Learns patterns beyond explicit rules and improves with feedback and examples.

Ensemble Integration

Rule layer acts as first filter with high precision, while the ML layer provides broader coverage. Mechanisms to review ML predictions lacking identifiable tokens. Reconciliation processes when rule and ML disagree. Multiple independent checks increase reliability.

Adaptive Confidence

Rules yield high-confidence direct classifications. ML predictions provide confidence scores for review and flag low-confidence outputs for human validation. Escalation path for uncertain cases and threshold-based auto-classification versus manual review.

Graph Neural Networks for Relational Data

Understanding connections and structure in energy data systems

Structure Modeling

Model data as graphs with nodes representing datasets and fields and edges representing relationships. Capture API endpoint connections and represent field lineage and data flows. Incorporate schema hierarchies and dependencies to reflect structural context.

Connection Intelligence

Resolve ambiguous field names using connected context across the graph. Understand how fields relate to system components and infer semantics, e.g., a field linked to Shipper systems is likely commercial. Propagate sensitivity signals through connected nodes.

Metadata Integration

Naturally incorporate source system information and add access control metadata to graph nodes. Include field ownership and custodian details and leverage knowledge graph aspects to map the broader data landscape.

Enhanced Classification

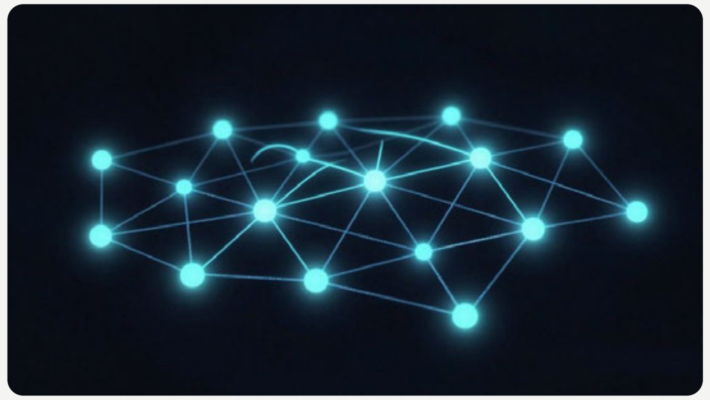
Combine textual signals with structural signals by using both content and relationships for classification decisions. Detect implicit sensitivity through network analysis and improve accuracy on ambiguous edge cases.

Why Graph Neural Networks for the Gas DSI?

Gas transmission networks are inherently graph-structured - GNNs can exploit this topology for classification, anomaly detection, and access control.

Relational Learning

Infer from connections



Network Graph

Central visualization of datasets and flows

Semi-Supervised

Label only 5-10%

Propagation

Auto-flag downstream

GNN Use Cases for Gas DSI: Real-World Examples

MIND&MATTER

Recommendation: Prioritise Use Cases 1 and 2 for the MVP phase (data sensitivity classification and access anomaly detection). Use Cases 3-5 are candidates for Phase 2 and beyond.

Use Case	How GNNs Apply	Real-World Evidence	Maturity
Data Sensitivity Classification	Model Gemini/Portal datasets as graph nodes connected by data lineage edges. GNN classifies sensitivity by learning from dataset features AND lineage context.	GNNs achieve state-of-the-art node classification in semi-supervised settings, requiring only 5-10% labelled nodes.	Medium
Access Anomaly Detection	Build bipartite graph of users → datasets. GNN learns normal access patterns and flags anomalies (e.g., shipper accessing internal balancing data).	GNNs used for intrusion detection in IoT/network graphs. ENX energy platform uses graph analytics for risk assessment.	Medium-High
Cross-System Data Lineage	Represent Gemini → D&I → GCP → Portal data flows as directed graph. GNN tracks sensitivity propagation through transformations.	Knowledge graph embeddings used for data provenance tracking. Graph-based lineage tracking adopted by major cloud platforms.	Medium
Power Flow & Grid Forecasting	Model gas transmission network topology as graph. GNN predicts flow patterns, pressure anomalies, and demand.	Multi-Task GNN achieves superior power flow prediction at German transmission grid transformers vs standard NNs.	High
Federated DSI Link Prediction	As organisations join NESO DSI, GNNs predict which datasets from different DPNs should be linked or cross-referenced.	Link prediction is core GNN capability. Open Banking's federated directory uses graph-based matching.	Low-Medium

Transformer Architecture for Sensitivity Classification

Why BERT and variants excel for energy data governance

Context Understanding

Transformer attention mechanisms capture word relationships. Understand multi-word phrases and domain terms. Interpret context-dependent sensitivity (e.g. 'confidential' vs generic use). Handle nuanced language variations.

Metadata Integration

Ingest API documentation as additional text input. Process field names, descriptions, lineage metadata. Incorporate source system information. Include access control hints and data ownership tags.

Transfer Learning Power

Pre-trained models leverage broad language understanding. Fine-tune on smaller domain-specific labeled sets. Reduced data requirements compared to training from scratch. Faster adaptation to energy sector vocabulary.

Domain Adaptation

Domain-adaptive pre-training on energy industry corpora. Learn sector-specific terminology and patterns. Understand Gemini and Gas Data Portal conventions. Recognize energy-sector sensitivity markers.

Rule-Based Heuristics for Sensitivity Detection

Pattern-based classification indicators

Rule Examples	Classification	Example
Contains personal names, emails, phone numbers	Confidential	John Smith, john.smith@company.com
Mentions internal systems, passwords, IP addresses	Internal	Access via VPN at 10.0.0.1
Aggregated statistics, public reports, summaries	Public	Total demand for Q1: 12,000 GWh
Company identifiers (Shipper ID, Contract ID)	Confidential	Shipper ID: 8723, Volume: 5000
Phrases like 'confidential', 'not for public release'	Confidential	This report is confidential

AI Classification of Unstructured Gas Data: Worked Example

Real-world scenario demonstrating classification without prior labels

The Scenario

We have unstructured data outputs from Gemini system (closed, market operations) and Gas Data Portal (open public platform). Need AI to automatically classify as Public, Confidential, or Internal WITHOUT any pre-labeled examples. Must rely on heuristics, metadata, and NLP.

Classification Approach

- **AI processes sample datasets from each source.**
- **Detects features and patterns.**
- **Applies rule-based logic.**
- **Highlights where human review helps for edge cases.**

Expected Outcomes

- **Demonstrate practical AI decision-making.**
- **Show how metadata and content cues drive classification.**
- **Illustrate when human intervention needed.**
- **Validate feasibility of label-free approach.**

Key Takeaways: Unstructured Data Classification Without Labels

MIND&MATTER

Inherent Signals Work

Gemini and Gas Data Portal have built-in sensitivity signals. Metadata (source, access level), content (keywords, identifiers), and structure all matter. AI does not require pre-labeled examples; patterns are evident. Both systems are designed with data governance in mind.

Rules Are Domain-Driven

Rules are based on policy (GDPR, Ofgem) and domain knowledge. Energy sector context (Shipper IDs, auctions, linepack) is captured. Rules are transparent, auditable, and explainable. Easier to maintain than black-box ML in early stages.

Human Loop Mitigates Risk

High-stakes classifications (Confidential/Internal) are flagged for review. Humans provide ground truth for feedback loops. False negatives (missing sensitivity) are caught by expert review. The system gets smarter incrementally.

Regulatory Compliance Built-In

GDPR: PII detection is automatic and verifiable. Ofgem: Presumed Open principle is consistently applied. An audit trail shows reasoning for every classification. A governance framework is integrated from the start.

Scalable & Improvable

Label-free approach works now without delay. As examples accumulate, the system can transition to supervised ML. Feedback loops continuously refine rules and confidence thresholds. Performance improves over time with real-world use.

Why this approach is robust and ready for implementation

Sample Classification Features & Decision Matrix

How different characteristics drive classification outcomes

Sample Data & Source	Detected Features (AI Reasoning)	AI Classification	Access Rules
Gemini: Daily Auction Bids Summary (JSON)	Source = Gemini (closed system), Market-sensitive metrics, Implied company-specific info, No public markings	CONFIDENTIAL	Share only with authorized shippers & regulators, not public
Gemini: System Audit Log (Text)	Content shows internal IT details, Marked 'Confidential' in header, No external use; security-related	INTERNAL	Strictly internal NGT use only - do not share externally
Gas Data Portal: NTS Demand Report (CSV)	Source = Gas Data Portal (open data), Aggregated system-level values, Known licensing: must be open, No sensitive keywords	PUBLIC	Openly publishable, share freely with public
Gas Data Portal: Maintenance Schedule (PDF)	Source = Portal (implies public), Mostly non-sensitive, Phone number flagged by NLP (checked by human)	PUBLIC	General info, personal data verified as public helpdesk, no restrictions

MIND&MATTER

Comparative Analysis: Gemini vs Gas Data Portal

How identical AI processes yield different classifications

🔒 CONFIDENTIAL

Source: Gemini (closed)

- Market-sensitive auction data
- Authenticated access only
- Commercial terms detected
- Flagged for human review

Result: Restricted to authorized users

🔓 PUBLIC

Source: Portal (open)

- Aggregate system data
- Open license, no login
- No sensitive identifiers
- Auto-classified

Result: Freely publishable

Key Insight

Same AI engine → Different outcomes. Source metadata drives classification bias (Gemini→confidential, Portal→public). Context-aware rules adapt without prior labels.

How AI spots sensitivity signals in Gemini auction data

Example A Deep Dive: NLP & Keyword Detection

Domain-Specific Terms Detected

- "Shipper" = signals a company entity operating in the gas market context and implies commercial counterparties
- "Auction" = term often treated as confidential until final results are published, indicating sensitive process data
- "Bids" = reveals competitive actions and pricing intent that could impact market behavior if leaked
- No public indicators like "Summary for public" or "Press release" present in the document
- Absence of public markings itself is a signal that the content may be internal or restricted

Explicit Watermarks & Headers

- If data includes a header such as "Confidential" (common in Gemini specs), this is an explicit tag the AI detects
- Such explicit markings cause the system to automatically mark the document as Internal/Confidential
- Document-level metadata (e.g., classification fields) will override other inferred signals when present

Edge Case: Internal vs Confidential

- If the data is a Gemini system audit log containing user IDs, IP addresses, and error codes, it is classified as INTERNAL rather than merely Confidential
- Presence of IP address formats and dense technical jargon signals "Internal – NGT eyes only" handling
- Security-related data, such as credentials or access logs, receives the highest restriction level and escalated protection

Example B Deep Dive: Generic vs Sensitive Content Signals

How AI distinguishes public from sensitive in portal data

Generic Terms & Units

'Forecast' and 'Actual' are typically generic, non-sensitive context indicators used in reporting. Units such as GWh or mcm, when presented in aggregate or summary form in public reports, are considered non-sensitive. By contrast, explicit trading metrics like price expressed in £/therm more directly signal commercial trading data and should be treated as potentially sensitive. Content cues like these guide automated classification systems toward appropriate handling decisions.

Edge Case: Maintenance Data

The Gas Portal publishes maintenance schedules that are generally non-sensitive, including site names, outage dates, and helpdesk contact information. However, AI systems may flag phone numbers as potential PII via NLP heuristics. A human reviewer must verify whether the contact is a general helpdesk entry intended for public use or a personal contact that would be sensitive. The portal team enforces policies to avoid personal data, with AI serving as a backstop to catch anomalies.

Confidence & Auto-Classification

Open portal data such as aggregated demand and supply metrics typically flow through automatically as Public when they match known dataset patterns. AI models are well calibrated for established portal datasets, while newly onboarded dataset types receive occasional human spot checks. Clearly innocuous content that matches existing templates passes without manual review, reducing operational burden while maintaining safeguards.

MIND&MATTER

AI Classification System Benefits & Outcomes

Automated Efficiency

AI-driven classification scales to thousands of datasets without manual intervention. Continuous monitoring and dynamic updates ensure classifications stay current and relevant.

Transparency & Trust

Explainable decisions, audit trails, and governance overrides build user confidence. Stakeholders understand why data is classified and can trust the system's reasoning.

Risk Mitigation

Conservative ensemble logic and hard guardrails prevent data leaks. Behavioral controls catch misuse patterns early. System remains compliant with GDPR and UK standards.

Innovation Enablement

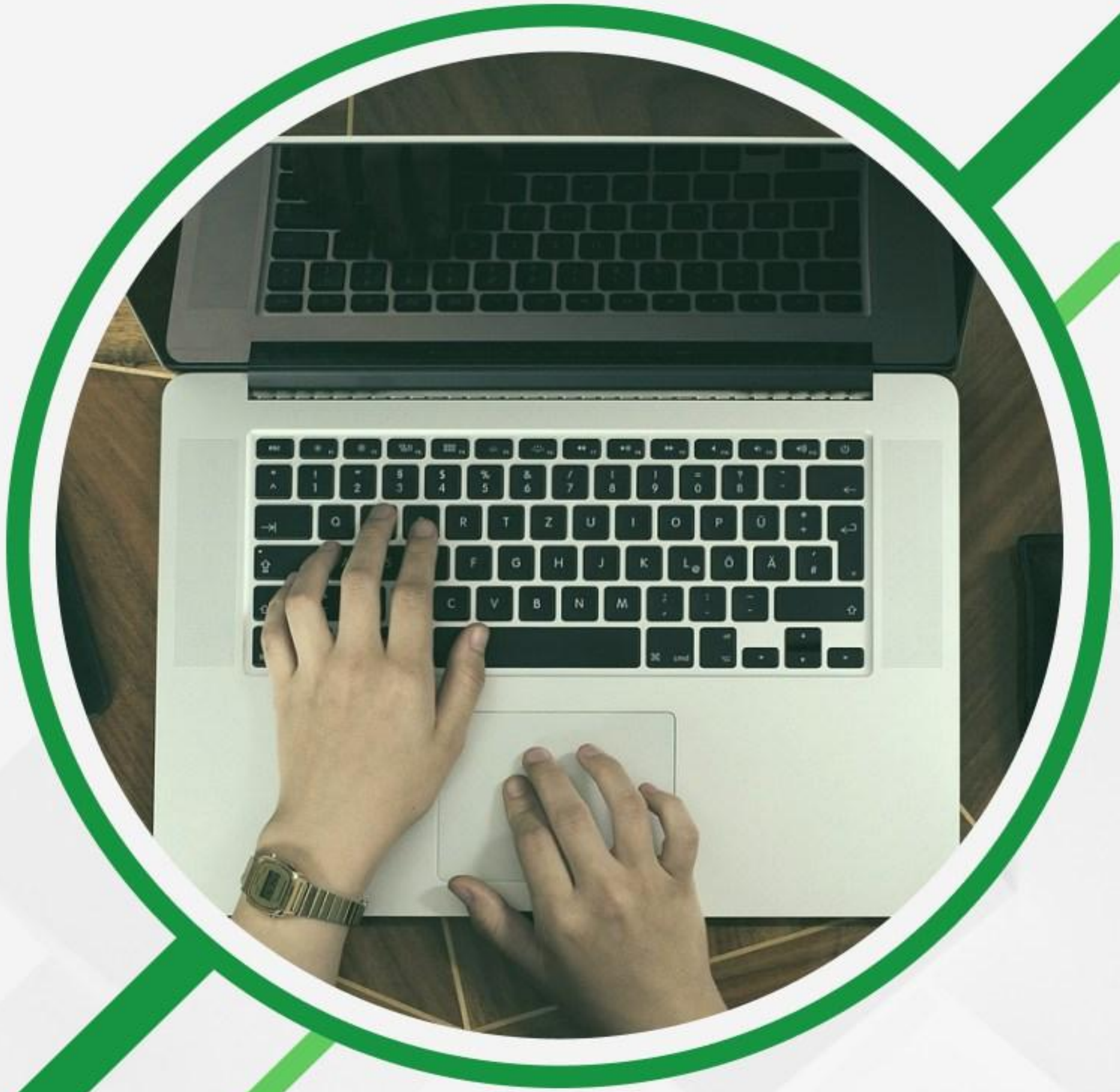
Open access to safe data fosters innovation and research. Researchers and external stakeholders can leverage public datasets while sensitive data remains protected for intended users.

GTDSI WP3:

Interoperability & Standards Alignment

Mind & Matter Tech

UK



Part A: NGT Current Systems Assessment

Evaluating Gemini and Gas Data Portal architecture, capabilities, and operational characteristics.

NGT's Dual System Architecture

Integrated Energy Infrastructure

National Gas Transmission operates two distinct but complementary systems serving different market participants and use cases, from secure transactional operations to open public data access.

Gemini: Secure Transactional

Proprietary system for gas market operations with rigorous access controls, role-based authorization, and approximately two million API calls daily.

Gas Data Portal: Open Public Access

Public-facing platform publishing aggregated gas system data with open REST/HTTP APIs, no authentication requirements, and approximately 500,000 daily API calls.

Fundamental Divide

Gemini represents closed, proprietary infrastructure, while the Gas Data Portal represents fully open, public data sharing with no access restrictions.

NGT System Performance and Scale

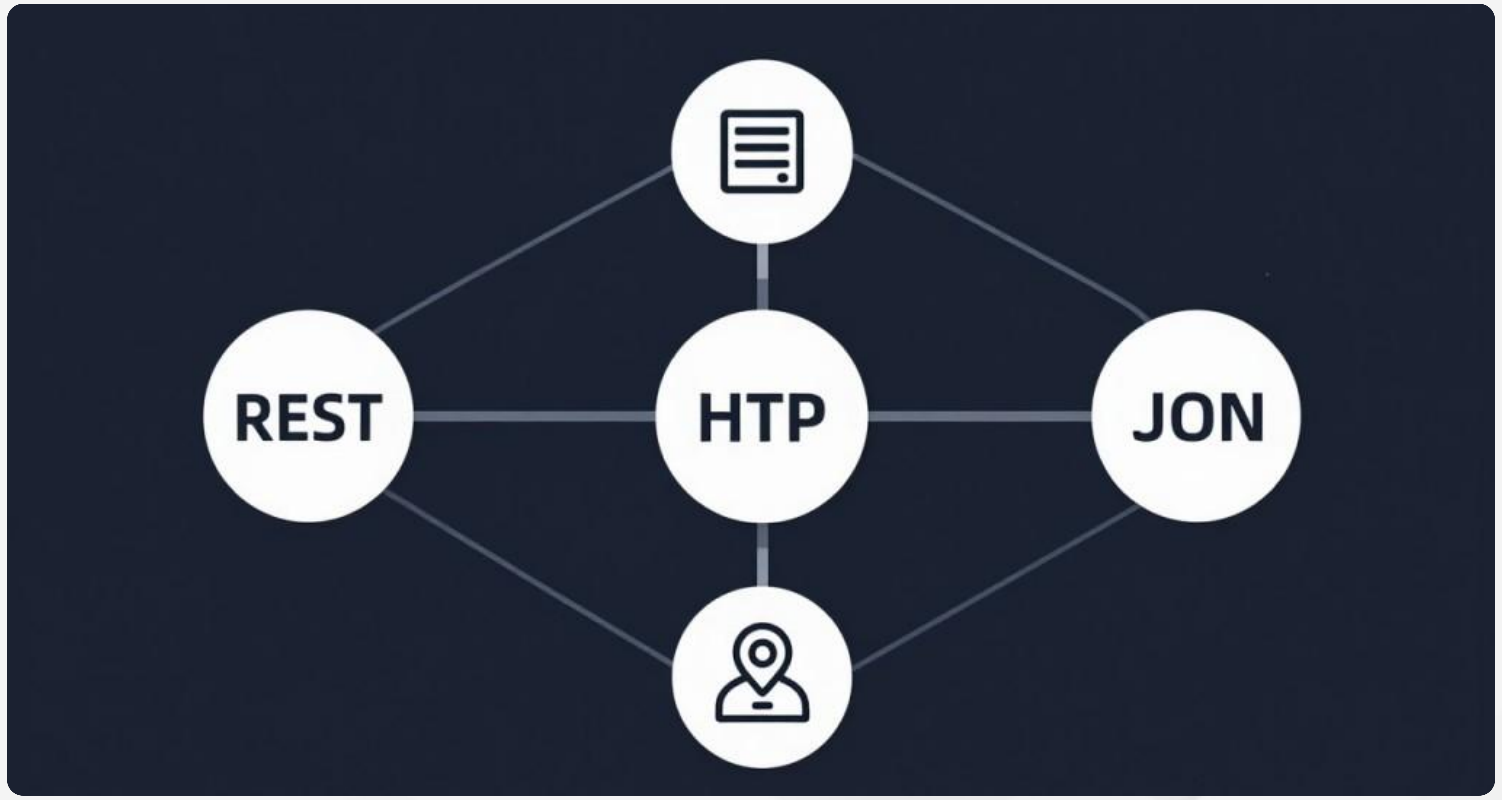
Gemini Infrastructure

Supporting high-throughput transaction processing with horizontal scaling within a private, secure environment.

Gas data Portal Infrastructure

Cloud-based platform serving approximately 70 million requests annually with elastic scaling for public consumption patterns and diverse client types.

API Architecture and Data Formats



Gemini: Proprietary Interfaces

Historically used proprietary SOAP/XML protocols with complex relational schemas

Now modernized through Azure API Management with RESTful JSON, but access remains tightly controlled and limited to registered market participants only.

Gas Data Portal: Open Standards

Built on modern web standards using REST/HTTP APIs delivering JSON and CSV responses with human-readable field names.

Designed for open consumption by any user without authentication, promoting transparency but lacking differentiation or granular access restrictions.

Authentication and Authorization Capabilities

NGT systems employ disparate access control mechanisms reflecting their fundamentally different design principles and operational requirements for market participants and the public.

Security Feature	Gemini	Gas Data Portal	Access Model
Authentication	Registered market participants	Anonymous, no authentication	Divergent
Authorization	Strict Role-Based Access Control	Read-only, no user roles	Divergent
Identity	SSO (Xoserve), proprietary tokens	No identity tracking	Divergent
Multi-Factor	MFA enabled	Not applicable	Divergent
Federated IAM	None - proprietary	None - open access	Gap

Data Integration Challenges and Current Gaps

Current State

- NGT systems operate with manual integration between Gemini and Gas Data Portal
- Data flows through periodic batch processes (typically daily) rather than real-time synchronization
- Changes to Gemini data structures require manual transformation logic updates in Gas Data Portal pipelines
- Different data models, naming conventions, and aggregation approaches create semantic confusion

Impact & Consequences

- External consumers struggle to correlate data due to inconsistencies and naming differences
- No data lineage documentation explaining calculated metrics or transformations
- Metadata requires domain expertise and is buried in internal documentation
- Data quality issues difficult to identify due to lack of monitoring
- Integration bottlenecks limit real-time insights into gas market conditions

Operational Maturity and Reliability Assessment

MIND&MATTER

99.9%

System Availability

Gemini demonstrates high operational maturity meeting critical infrastructure requirements

100%

Proven Track Record

Consistent performance managing complex gas market operations

High

Architecture

Microservices provides scalability and resilience

Strengths

- Regular maintenance windows scheduled proactively to minimize disruption
- Comprehensive monitoring and alerting infrastructure detecting issues
- Gas Data Portal with elastic cloud infrastructure supporting variable demand
- Established operational procedures and incident response protocols

Challenges

- Aging data models and proprietary interfaces limit extensibility
- Tight coupling within Gemini complicates enhancements and modernization
- Manual integration processes introduce human error risk and operational overhead
- Lack of automated data quality validation requires manual monitoring
- Staff expertise concentrated in proprietary systems requires capability building

Overall assessment: Operationally mature but architecturally constrained, requiring modernization for future demands.

DSI Enabled Use Cases and Innovation Opportunities

Real-time cross-sector optimization

Standardized APIs enable coordinated operations between gas and electricity systems, supporting Net Zero transition through integrated balancing and demand response.

Third-party analytics and applications

Developers building innovation on standardized data access (visualization tools, forecasting models, market analysis platforms).

Supply chain transparency

Standardized data enables tracking of gas flows, renewable gas integration, and supply security monitoring.

Research and analysis

Academic institutions and consultants access comprehensive energy data for system optimization research and analysis.

Market efficiency

Transparent, standardized data reduces information asymmetry enabling competitive market operations and fair pricing.

Flexibility services marketplace

Standardized data enables cross-sector flexibility service offerings coordinating gas and electricity resources.

Emergency response

Rapid data access during system emergencies enables coordinated response across organizations.

Regulatory reporting

Standardized data simplifies compliance reporting and enables automated validation of regulatory requirements.

Global Data Sharing Initiatives and Benchmarking

European Context

- ENTSO-E (electricity) and ENTSOG (gas) developing similar federated data initiatives
- EU regulations increasingly mandate open data access and standardization
- Common Information Model (CIM) widely adopted in European electricity systems

UK Alignment

- UK NESO DSI aligns with European standards enabling cross-border data sharing
- NGT early adoption positions UK gas sector as innovation leader

S W
T O

Future Outlook

- CEEDS project advancing harmonized data standards
- Interoperability enables innovation platforms leveraging cross-jurisdictional data
- Standards alignment ensures future-proofing against global requirements

Global Adoption

- International Data Charter principles increasingly adopted globally
- Energy sector moving toward standardized data practices
- Utilities in US, Australia, and Canada implementing similar federation models

MIND&MATTER

1. **Governance Framework**
   NESO provides governance framework for DSI evolution ensuring standards remain current as technology and requirements change

2. **Participant Organizations**
   Network operators, regulators, market participants participate in steering committees and technical working groups guiding DSI direction

3. **Standards Development**
   Collaboration with IEC, CENELEC, and international bodies

4. **Data Governance**
   Establish common rules for sharing while preserving organizational sovereignty

5. **Dispute Resolution**
   Mechanisms address conflicts between organizations

6. **Review Cycles**
   Regular assessments (annually or as needed) to evaluate standard effectiveness and propose evolution

7. **Transparency**
   Openness in standards development with public consultation periods

8. **Community Support**
   Accelerated contributions through open source implementations, tools, libraries and support programs

9. **Training & Certification**
   Programs supporting implementation and interoperability

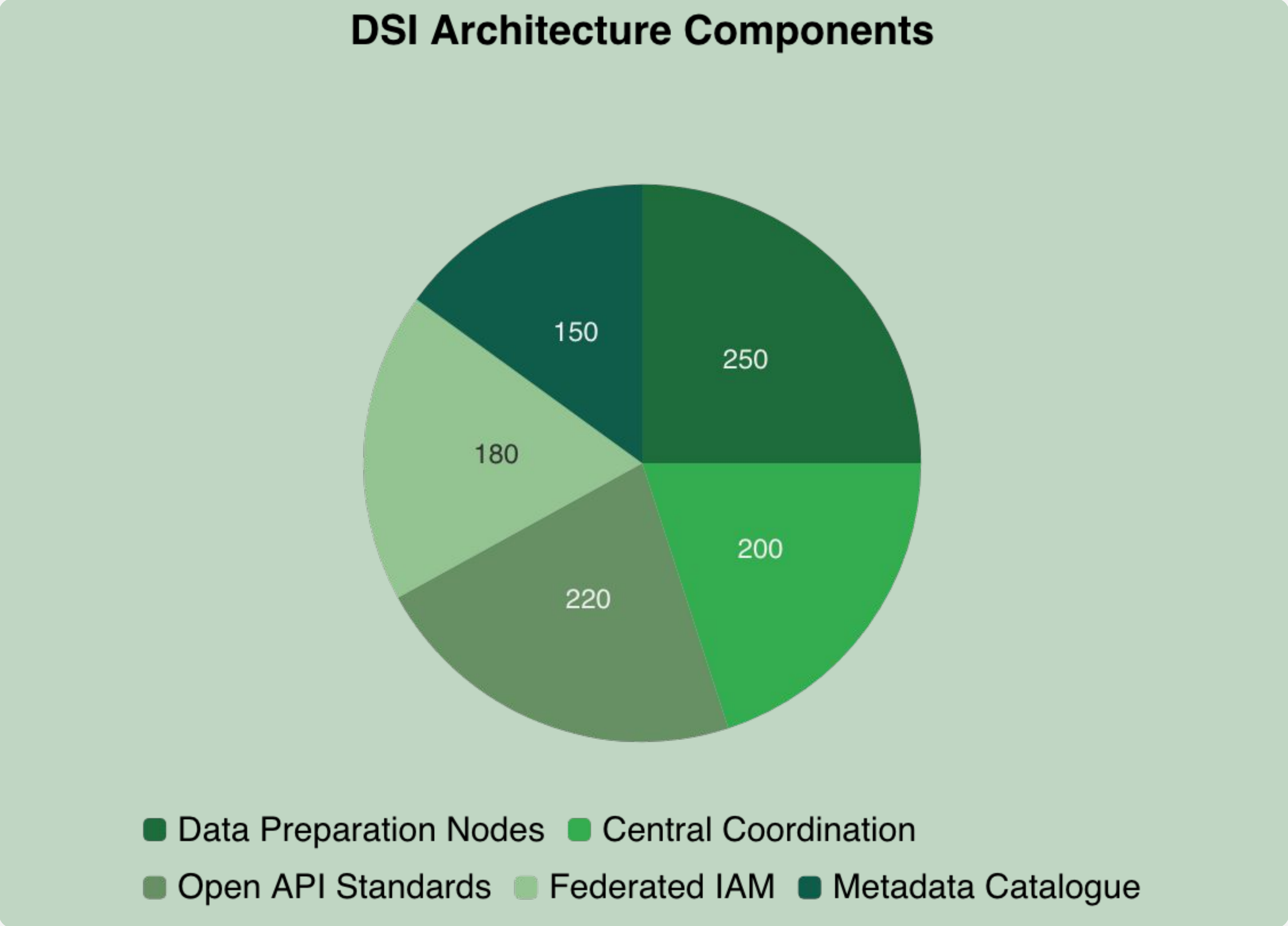
10. **NGT Participation**
    Enables influence on energy sector digital infrastructure direction and establishes credibility through early leadership positioning

Federated Architecture: Core Principles

Distributed Data Ownership

NESO DSI implements a federated model where each organization operates its own Data Preparation Node (DPN) to publish data. Each organization maintains data at its source while enabling secure, standardized access through common interfaces.

> **Data remains with its source, accessed through unified mechanisms, eliminating data duplication while preserving organizational control and data governance.**



MIND&MATTER

Risk Assessment and Mitigation Strategies

Operational Risks

System disruption during modernization mitigated through parallel operation, extensive testing, rollback procedures, conservative sequencing starting with non-critical systems

S

W

Technical Risks

Complexity and obstacles mitigated through prototyping, external expertise, phased rollout, architectural reviews

Schedule Risks

Timeline delays mitigated through aggressive project management, resource allocation, dependency management

O

Organizational Risks

Staff resistance, key person dependencies mitigated through change management, early involvement, retention strategies, knowledge capture

Financial Risks

Cost overruns mitigated through detailed budgeting, contingency reserves (10-20%), periodic reviews and adjustments

Market Risks

Customer adoption lag mitigated through engagement, pilots, graduated rollout enabling feedback incorporation

T

Regulatory Risks

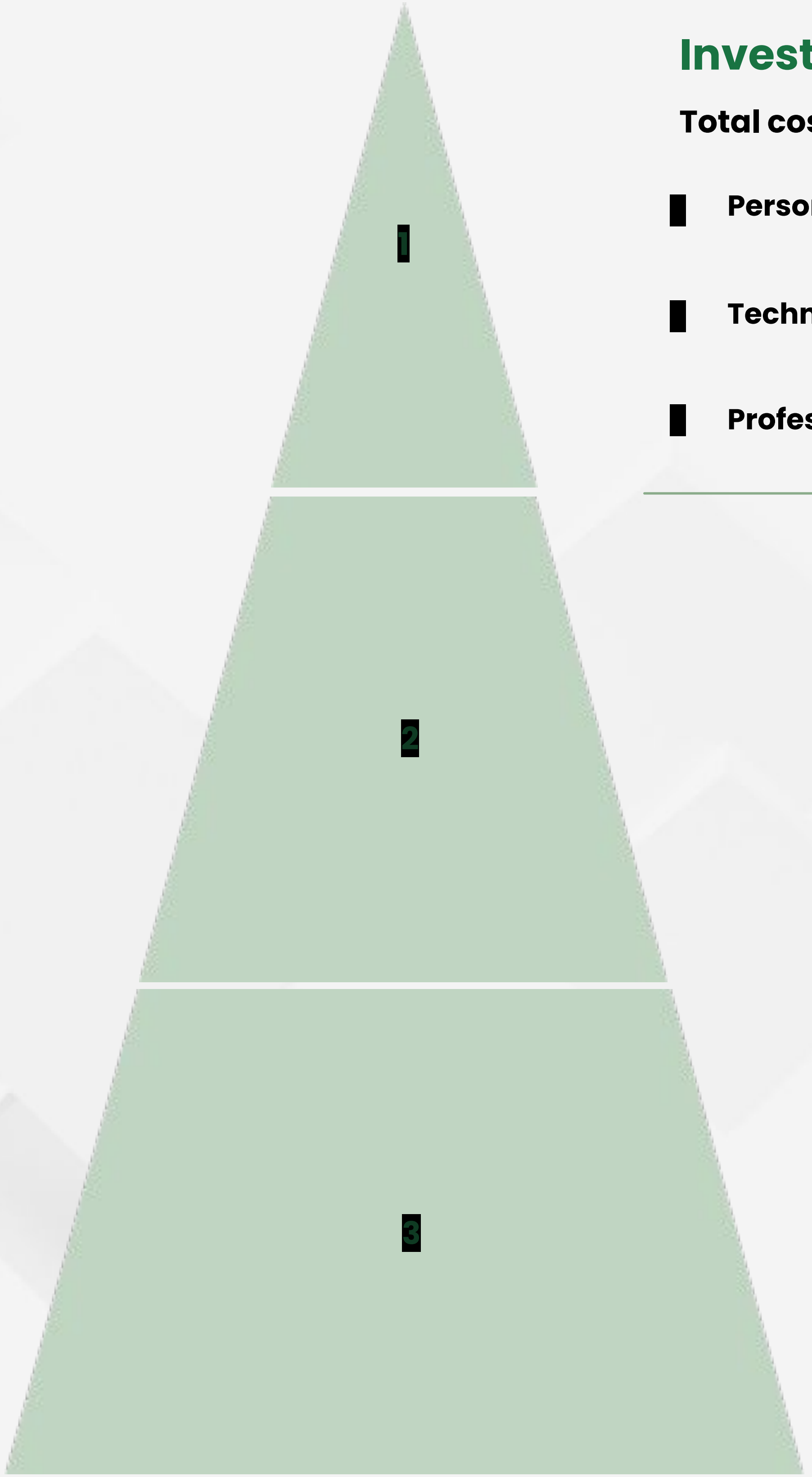
Compliance gaps mitigated through early engagement with Ofgem/NESO, regulatory impact assessment, control framework design

Vendor Risks

Platform dependencies mitigated through multi-vendor strategies, contractual protections, vendor monitoring

Top 5-10 risks prioritized for detailed mitigation planning with ownership assignments and monitoring mechanisms

Business Case and Cost-Benefit Analysis



Investment Requirements

Total cost of ownership over 3 years estimated at £2-5 million covering:

- **Personnel (40-50%): executive sponsor, program manager, architects, developers, QA, change management**
- **Technology (30-40%): identity platforms, API management, ETL, metadata catalogs, cloud expansion**
- **Professional Services (20-30%): systems integration, data architecture, IAM, change management**

Benefits Realization

- **Operational Efficiency: £200K-400K annually from reduced data duplication and manual integration**
- **Innovation Enablement: £500K-1M+ in revenue opportunities through new capabilities**
- **Regulatory Compliance: Reduced Ofgem compliance risk**
- **Market Competitiveness: Data leadership positioning**

ROI Analysis

- **Benefits exceed costs by year 3-4 with sustained operational benefits**
- **ROI positive by year 4 with 200%+ cumulative ROI by year 5**
- **Break-even analysis shows payback within 3 years including transformation costs**

Phased Implementation Roadmap and Timeline

1

Phase 1 (Months 1-6)

Foundation building with executive sponsorship, business case analysis, stakeholder engagement, maturity assessment, program planning. Quick wins include Gas Data Portal metadata publication using DCAT schema and OpenID Connect pilot implementation.

2

Phase 2 (Months 7-18)

Pilot implementation developing metadata catalogue, implementing OIDC/OAuth 2.0 infrastructure, piloting CIM mapping with subset of data, developing REST APIs alongside proprietary interfaces, establishing governance framework. Focus on less critical data with pilot customer feedback informing broader rollout.

3

Phase 3 (Months 19-36)

Full-scale production rollout with complete CIM migration, REST API standardization, federated identity integration for all systems, unified Gas DPN establishment, real-time data synchronization, automated data quality validation. Careful sequencing maintains operational continuity for critical systems like Gemini. Parallel operation with rollback procedures if issues emerge. Dependencies and critical path activities highlighted for program management.

NESO DSI Standards and Technical Framework

REST/GraphQL
API Standards

CIM/DCAT
Data Models

OAuth 2.0
Identity

DCAT-AP
Metadata

Open Standard APIs

Standardized REST/JSON endpoints with optional GraphQL and SPARQL support, building on Open Energy API specifications for consistent resource structures across all organizations.

- Consistent API design patterns and documentation standards.
- Support for multiple query formats and data serialization.
- Version management and backward compatibility.

Common Data Models

Adoption of sector-wide semantic standards including Common Information Model (CIM), DCAT for metadata, and schema.org Dataset definitions for comprehensive interoperability.

- Common Information Model (CIM) for energy data.
- DCAT-AP for standardized metadata representation.
- Alignment with Open Energy API specifications.

Federated Identity

Unified trust framework enabling single sign-on across organizations using OpenID Connect (OIDC) and OAuth 2.0 for secure token exchange and attribute-based access.

- OpenID Connect (OIDC) for authentication.
- OAuth 2.0 for authorization and token management.
- Attribute-based access control across domains.

Metadata Governance and Data Lifecycle Management

Unified Metadata Catalogue

- DCAT-AP standardized metadata format for all datasets.
- Centralized searchable registry enabling discoverability.
- Rich semantic information supporting intelligent queries.

Common Governance Framework

- Standard licensing terms agreed across participants.
- Quality controls and audit trails for compliance.
- Data lineage tracking and lifecycle management.

Integrated Data Ecosystem

- Minimal data duplication through unified publishing.
- Real-time or near-real-time data access patterns.
- Cross-domain data sharing with preserved ownership.

Part B: Gap Analysis and Strategic Alignment

Comprehensive evaluation of differences between NGT current systems and NESO DSI requirements, with recommendations for strategic roadmap.

MIND&MATTER

Critical Gap Areas: NGT to DSI Alignment

Data Standards Misalignment

NGT uses proprietary and non-standard schemas, while DSI will mandate industry-standard models like CIM, DCAT, and Open Energy API specifications for universal interoperability.

Access Control Fragmentation

Gemini enforces closed proprietary RBAC, while Gas Data Portal offers no differentiation. DSI will require federated IAM with tiered access (open, shared, restricted) using OIDC/OAuth 2.0.

Metadata and Discoverability

Lacks centralized metadata cataloguing; discoverability is poor. DSI will require comprehensive DCAT-AP metadata catalogue with full lineage, enabling semantic search and intelligent data discovery.

Modularity and Integration

Gemini remains monolithic despite microservices efforts with tight coupling; no unified API gateway exists. DSI will mandate modular, decoupled services with unified DPN as single integration point.

Strategic Recommendations for DSI Alignment

Unified Integration Strategy

Establish a comprehensive roadmap for NGT to align with DSI principles through phased implementation of open standards, federated identity management, and unified data preparation infrastructure.

Adopt Open Standards

Implement Common Information Model (CIM), DCAT metadata framework, and Open Energy API specifications across both Gemini and Gas Data Portal systems.

Federated Identity Layer

Deploy OpenID Connect (OIDC) and OAuth 2.0 infrastructure for unified authentication, single sign-on, and tiered attribute-based access control.

Unified Gas DPN

Establish single Data Preparation Node as integration layer between Gemini, Gas Data Portal, and DSI ecosystem, serving as unified API endpoint for external consumers.

Part C: Mapping Data Practices and Creating a Compliance Checklist

Section C.1: Map NGT's Data Practices to NESO's Standards

Assessing current data governance and alignment with emerging regulatory frameworks.

MIND&MATTER

NESO's Digital Spine Framework

Prepare

Standardized metadata and data classification across all energy systems.

- Consistent data formats and schemas.
- Quality checks and validation rules.
- Data provenance documentation.

Trust

Establish authentication and authorization frameworks for secure access.

- Role-based access control.
- Identity verification mechanisms.
- Secure authentication protocols.

Share

Ensure interoperable logging and complete traceability across systems.

- Audit trail logging.
- Data access tracking.
- Compliance reporting.

Comprehensive Event Logging and Audit Requirements

Core Logging Principles

All AI-related events and data transactions must be logged with complete metadata including timestamps, user identification, model versions, inputs/outputs, and system context for full auditability.

Event Classification

Training, deployment, inference, updates, decommissioning must be distinctly logged.

Model Traceability

Link audit entries to specific model versions for complete decision traceability.

Explainability Logging

Log interpretable explanations or proxy indicators for AI decision transparency.

Security, Storage, and Compliance Controls

Data Protection and Access

Implement least privilege principles with role-based access control and tamper-proof, immutable audit log storage using append-only formats or cryptographic hashing.

Access Control

Audit who accessed or modified AI systems with sensitive logs protected.

Immutable Storage

Use append-only formats or blockchain-style timestamping for critical logs.

Real-Time Monitoring

Integrate with SIEM tools to detect anomalies and trigger alerts.

Regulatory Framework Alignment

UK Regulatory Approach

UK regulators emphasize transparency, accountability, and self-governance with clear audit trails for regulatory reporting.

EU Regulatory Approach

EU moving towards prescriptive, formal regulations for high-risk systems with detailed record-keeping obligations.

GDPR Requirements

Audit personal data processing, log data access, and demonstrate compliance for automated decision-making.

Ofgem Standards

Data presumed open and auditable. Companies must document restrictions and maintain clear data provenance.

Part C.2 : Create a Draft Compliance Checklist

Detailed requirements for NGT and Gas Data Portal compliance with NESO and Ofgem standards.

Compliance Checklist: Core Requirements

Essential compliance items for NGT and Gas Data Portal alignment with NESO Digital Spine and Ofgem Data Best Practice standards.

Compliance Category	Requirement	NGT Status	Gas Portal Status	Priority
Event Logging	Log all AI/data events with metadata	Partial	Partial	Critical
Model Versioning	Maintain version-controlled model registry	In Progress	Not Started	Critical
Access Control	Implement role-based access control	Implemented	Partial	High
Data Provenance	Document data source and transformations	Partial	In Progress	High
Audit Trail Storage	Use immutable, append-only formats	Not Started	Not Started	Critical
Real-Time Monitoring	Integrate with SIEM tools	Not Started	Not Started	High

Extended Compliance Checklist: NGT Systems Focus

Detailed compliance requirements specific to NGT data systems and audit logging implementation.

NGT System Component	Compliance Requirement
Event Logging Infrastructure	Capture all data transactions with metadata
Model Version Registry	Maintain AI model version history
Access Control System	Role-based access for NGT systems
Data Quality Framework	Implement validation rules and checks
Audit Log Storage	Deploy immutable append-only storage
Monitoring Dashboard	Real-time system performance tracking

Extended Compliance Checklist: Gas Data Portal Focus

Detailed compliance requirements specific to Gas Data Portal systems and regulatory reporting.

Portal Component	Compliance Requirement
Data Publication Logging	Log all data publication events
Access Audit Trail	Track all portal user access events
REMIT Compliance	Ensure fair access to inside information
Data Validation Rules	Implement input validation and checks
Breach Notification System	72-hour incident reporting capability
Archival Procedures	Secure long-term log retention

MIND&MATTER

Compliance Gaps and Remediation Plan

Critical Gap Analysis

Six critical compliance gaps identified across NGT and Gas Portal require immediate remediation. Prioritized execution plan developed for sequential implementation with clear accountability.

Gap 1: Immutable Storage

Action: Deploy append-only database. Timeline: 4 weeks. Owner: Infrastructure Team.

Gap 2: Incident Response

Action: Develop 72-hour notification process. Timeline: 2 weeks. Owner: Security Team.

Gap 3: Model Registry

Action: Build version control system. Timeline: 8 weeks. Owner: Data Science Team.

Compliance Testing and Validation Framework

Comprehensive testing framework ensures all compliance requirements meet NESO standards and Ofgem expectations. Multi-layer validation approach with both automated and manual verification.

Unit Testing

Test individual components (logging, access control, archival) for correct functionality and compliance.

Integration Testing

Verify all components work together correctly and audit trails are complete across systems.

Compliance Auditing

Conduct full compliance verification against NESO, Ofgem, and GDPR requirements before production.

Compliance Risk Assessment Matrix

Risk Severity Assessment

Risk matrix identifies critical vulnerabilities requiring immediate attention. Audit Logging and Data Storage present highest risk due to current gaps and regulatory importance.

- **Critical Risk: Could result in regulatory penalties or operational failure.**
- **High Risk: Significant compliance gaps requiring prompt remediation.**
- **Medium Risk: Important but non-critical items with longer timelines.**

Compliance Audit and Sign-Off Procedures

Formal audit procedures ensure compliance is verified, documented, and signed off by appropriate stakeholders. Multi-phase validation with evidence collection for regulatory demonstration.

Stakeholder Review

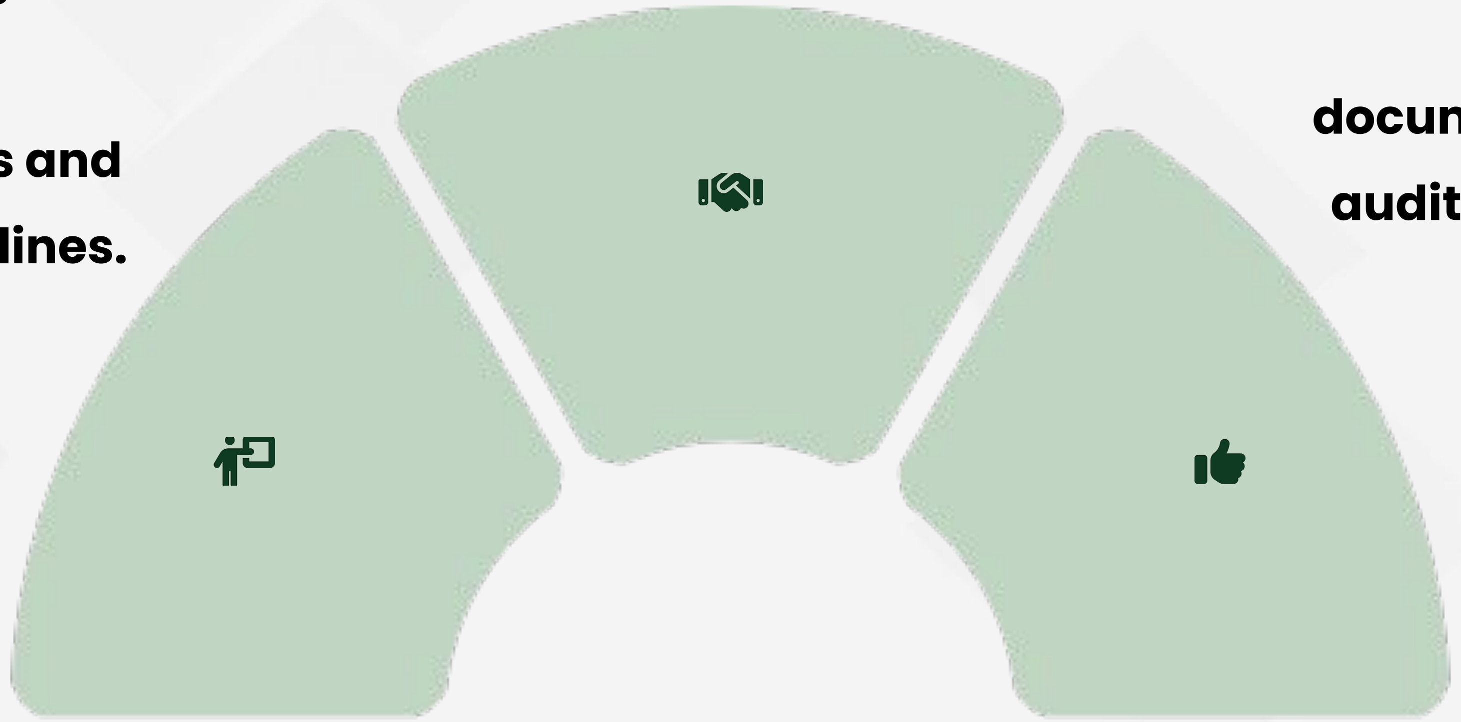
IT, Security, Data, and Business teams review and approve compliance evidence and testing results.

Internal Audit

Compliance team verifies implementation against requirements. Document gaps and remediation actions with timelines.

Regulatory Sign-Off

Executive approval and documentation prepared for external auditors and regulatory authorities.



Data Quality and Provenance Requirements

Data Source Documentation

Document all source systems and data origins with timestamps.

Transformation Logging

Log all data transformations and processing steps applied.

Quality Assurance

Record quality checks, validation results, and data acceptance.

Training Data Auditability

Log complete provenance of training datasets including source systems, transformations, quality checks, and consent/usage rights for sensitive data.

Human Oversight and Regulatory Alignment

Human Review Documentation

Document all instances where human reviewers override or validate AI decisions, including rationale and reviewer identity for accountability.

Framework Alignment

Align logs with ISO 42001, NIST AI RMF, and Ofgem Guidance.

Regulatory Reporting

Support external audits and regulatory compliance reporting.

Retention Policies

Define periods based on GDPR, REMIT, and Ofgem requirements.

Ethical AI Governance Framework

AI Governance and Transparency

Ofgem's 2025* ethical AI guidance encourages establishing comprehensive AI governance frameworks with transparent decision-making processes and maintained audit trails for all AI models.

Transparency in Decisions

Document and communicate how AI systems make decisions affecting customers and stakeholders.

Governance Framework

Establish clear policies, oversight mechanisms, and accountability structures for AI deployment.

Performance Monitoring

Track AI system performance metrics and maintain documentation of model improvements.

* - Ofgem 2025 ethical AI guidelines

Data Retention and Archiving Best Practices

Retention Policies by Regulation

Define data retention periods based on regulatory requirements from GDPR (6 years typical), REMIT (5 years), Ofgem Data Best Practice, and NIS regulations to ensure compliance and auditable records.

Regulatory Timeframes

GDPR: 6 years, REMIT: 5 years, OFGEM: per guidance, NIS: forensic period.

Secure Archiving

Store archived logs in tamper-proof systems with restricted access and encryption.

Retrieval Procedures

Maintain procedures for rapid retrieval of archived logs for audits and investigations.

Incident Response and Breach Notification Procedures

NIS2 Compliance and Incident Management

Incident Detection

Establish robust incident response procedures compliant with NIS2 Directive requirements, including mandatory 72-hour breach notification to authorities and affected parties.

Forensic Analysis

Maintain detailed logs for forensic investigation and root cause analysis of incidents.

Notification & Remediation

Document incident detection, escalation, containment, and remediation steps clearly. Report cybersecurity incidents to authorities within 72 hours of discovery.

Model Registry and Version Control Implementation

Maintain a comprehensive, version-controlled registry of all AI models with complete metadata linking each audit log entry to the specific model version that generated decisions.

Model Metadata

Track model name, version, deployment date, training date, and performance metrics.

Decision Traceability

Link each decision to model version, timestamp, and input data for complete audit trail.

Version History

Maintain complete history of model changes, updates, and retirement with documentation.

Additional Regulatory Frameworks

NGT must comply with multiple regulatory frameworks covering market integrity, cybersecurity, and AI governance standards.

REMIT (Market Fairness)

Audit logs must show data publication timing and prevent advance access to inside information.

NIS/NIS2 Regulations

Log management, incident reporting for essential services, and 72-hour breach notification.

EU AI Act (Draft)

Specific record-keeping for high-risk AI systems in critical energy infrastructure.

Key Recommendations for Success

Systematic approach to achieving full NESO and Ofgem compliance

Strategic Implementation

Achieve full NESO and Ofgem compliance through systematic, well-planned implementation of audit, logging, and governance frameworks.

Prioritize Critical Items

Focus on event logging and immutable storage as highest priority items.

Continuous Improvement

Implement regular audits and updates to maintain compliance standards.

Stakeholder Alignment

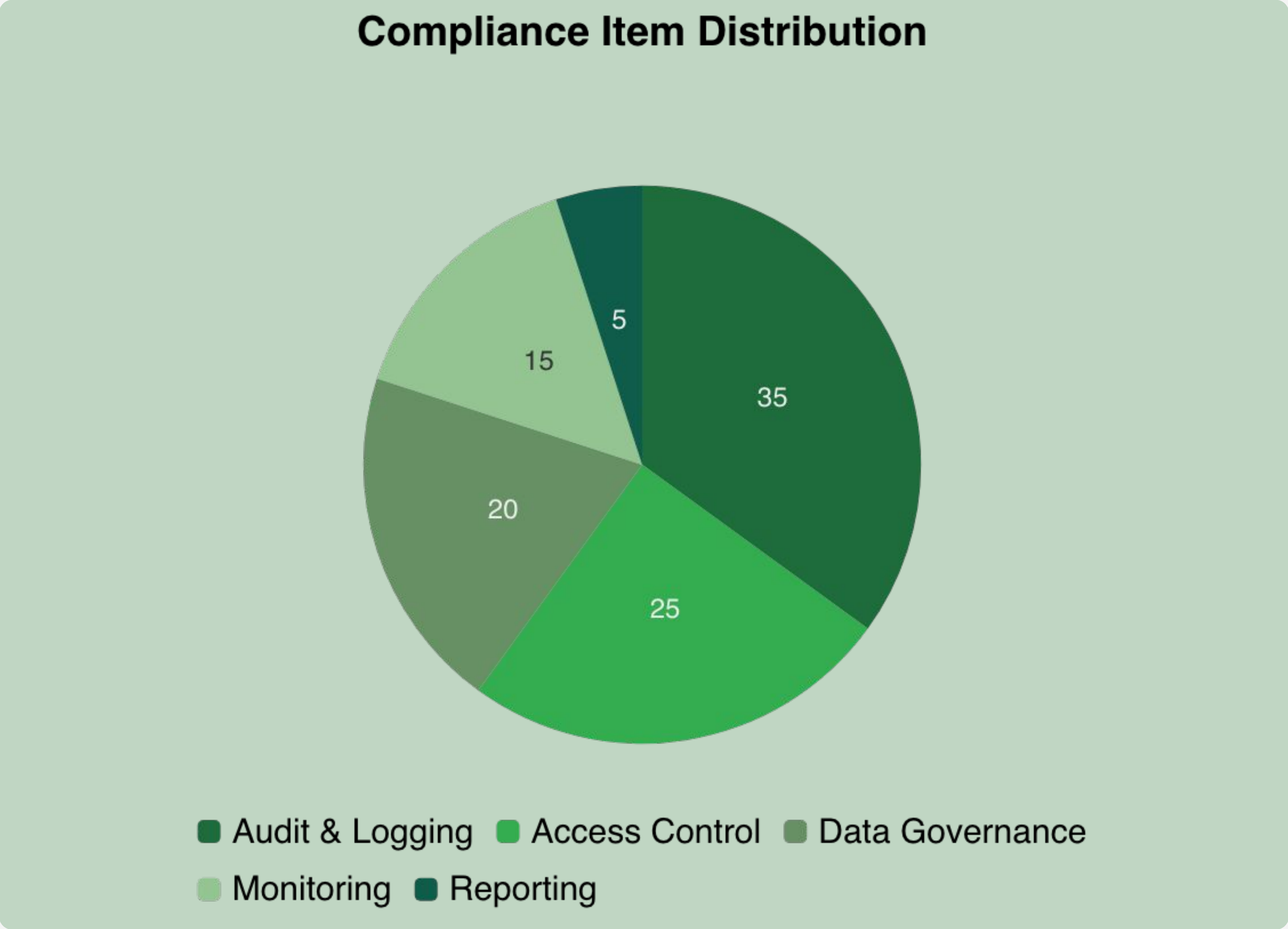
Ensure coordination across NGT, Gas Portal teams, and regulatory bodies.

Compliance Priorities Overview

Priority-Based Approach

The compliance framework prioritizes critical audit and logging requirements first, followed by advanced monitoring and automation capabilities. This phased approach ensures foundational compliance before advancing to sophisticated implementations.

> **Establishing robust audit trails and immutable logging mechanisms is essential before deploying advanced SIEM integration and real-time monitoring systems.**



Part D: Digital Systems Integration (DSI): Environmental Impact Assessment

Annual Carbon Footprint Comparison

Gemini System Impact

20-30 tons CO_2e annually due to high-throughput, compute-intensive operations and legacy on-premises infrastructure.

Portal System Impact

2-5 tons CO_2e annually with lower operational intensity and potential for cloud-based optimization.

Note – numbers are estimated from research

Key Carbon Hotspots Identified

Major Emission Sources

Life Cycle Assessment reveals critical areas requiring intervention for both systems.

Always-On Servers

Gemini's continuous operation contributes significantly to baseline energy consumption year-round.

Carbon Footprint Challenge

High API Frequency

Redundant API calls across both systems increase compute demands and energy requirements.

Data Redundancy

Duplicate data storage between Gemini and Portal increases embodied energy from hardware.

Environmental Performance Benchmarks and Targets

100%
Renewable Energy

1.2
Target PUE Ratio

60%+
Utilization Rate

90%
Equipment Reuse

Energy Optimization

- Deploy DSI components in cloud regions with renewable energy sourcing.
- Target 100% renewable-powered operations.
- Leverage cloud provider efficiency (PUE ≤1.2).
- Retire legacy infrastructure as systems transition.

Data Efficiency

- Implement 'Store Once, Use Many' principle across DSI.
- Eliminate redundant data storage practices.
- Archive obsolete data in cold storage systems.
- Minimize duplicate datasets between systems.

Technical Excellence

- Deploy serverless and containerized solutions.
- Optimize API calls for lightweight operation.
- Scale compute resources on-demand only.
- Implement efficient coding practices DSI-wide.

Environmental Impact Metrics and Targets

Key performance indicators and benchmarks for DSI environmental sustainability aligned with Clean Power 2030 objectives.

Aspect	Baseline/Benchmark	DSI Goal/Recommendation
Data Center Efficiency	Enterprise DC: 1.7 PUE	Cloud Infrastructure: 1.2 PUE or better
Energy Source	UK Grid: 50% low-carbon	100% renewable-powered operations
Equipment Utilization	Baseline: 20-30% avg	Target: >60% through cloud scaling
Legacy System Retirement	Continuous power drain	Decommission as DSI transitions online
Data Storage	Duplicate datasets stored	Store once, use many principle
Equipment Lifecycle	Manufacturing carbon cost	Circular practices and recycling

Circular Data Lifecycle Practices

Single Source of Truth

Maintain one master version of each dataset accessible to all stakeholders, eliminating redundant storage and ensuring consistency across the DSI platform.

Maximize Reuse

Design datasets to be generic and reusable with rich metadata, encouraging cross-domain use and treating data as long-term strategic assets.

Lifecycle Management

Implement data retention policies with regular evaluation, archiving historical data in low-energy storage, and sunsetting obsolete datasets.

Minimize Redundancy

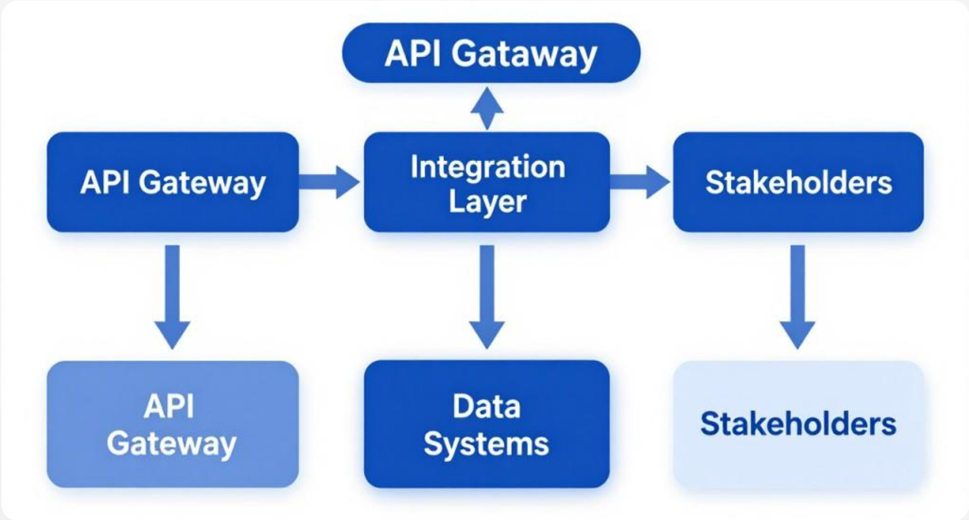
Avoid duplicate data storage in multiple formats or locations. Use references and efficient data models to prevent storage bloat.

Part D.1 : Technical Architecture and Implementation

RIIO-3 Recommendations for Sustainable Infrastructure

MIND&MATTER

DSI Technical Pillars and Infrastructure Design



Unified Cloud Platform

- Centralized cloud-based data infrastructure.
- Renewable energy-powered data centers.
- Scalable and efficient resource allocation.

API Gateway & Integration

- Robust API Gateway for efficient data access.
- Integration layer for cross-system connectivity.
- Optimized API calls with caching mechanisms.

Identity and Access Management

- Robust IAM for secure data governance.
- Role-based access controls and monitoring.
- Data transparency and audit trails.

DSI Implementation Roadmap and Timeline

Phase-Based Delivery Strategy

The DSI roadmap outlines critical milestones across three implementation phases, coordinated with National Gas Clean Power 2030 objectives.

> **2026-2027: Pilot and preparation; 2027-2028: Implementation and integration; 2030: Enable systemic decarbonization outcomes.**

Implementation Timeline: Detailed Activities and Outcomes by Phase

MIND&MATTER

2026-2027: Pilot & Preparation

- Q1-Q2: Governance Framework Setup, Technical Standards Definition, Compliance Mapping
- Q3: Pilot Deployment in 3 Units, Carbon Baseline Measurement, User Testing
- Q4: Workforce Training (500+ staff), Change Management, Initial Results Analysis
- Q1-2025: Capability Gap Analysis, Lessons Learned, Readiness Assessment

Outcomes: Compliance framework established, 15% initial efficiency gains, 500+ trained personnel

2027-2028: Implementation & Integration

- 2026: Cloud Infrastructure Deployment (AWS/Azure), Legacy Gemini Server Retirement, 40% Energy Reduction
- 2026-2027: Portal Data Feed Integration, Decommission Old Hardware, System Consolidation
- 2027-2028: Partner Onboarding (50+ organizations), Cross-Domain Data Sharing Launch, API Optimization
- 2028: Performance Tuning, Utilization Improvement (>60%), Continuous Monitoring Setup

Outcomes: 60% system utilization, 50+ integrated partners, 40% energy reduction achieved

2029-2030: Optimization & Outcomes

- 2029: Full Enterprise Adoption, Automated Scaling Implementation, All Legacy Systems Retired
- 2029-2030: Quarterly Optimization Reviews, User Feedback Cycles, New Dataset Development
- 2030: Final Impact Measurement (Target: 50-60% carbon reduction), Stakeholder Reporting, Certification & Compliance Validation, Next Generation Planning

Outcomes: 55% carbon reduction validated, 100% legacy retirement, enterprise-wide adoption completed

Part E: Sustainable Digital Roadmap for National Gas

Sustainable Digital Core Practices

Foundation of Green IT

National Gas adopts low-carbon digital practices to minimize environmental impact while maintaining critical infrastructure efficiency and data integrity.

Cloud-Native Architectures

Leveraging cloud-native designs and serverless computing for dynamic resource allocation and reduced idle capacity.

Energy-Efficient Design

Implementing API optimization, caching strategies, and data minimization to reduce computational demand and energy consumption.

Green Hosting Infrastructure

Transitioning to green hosting providers powered by renewable energy sources for sustainable long-term operations.

Short-Term Strategy (0-6 Months): Quick Wins

6
Month Duration

Immediate
Implementation

Foundation
Priority

Quick
Impact

Gemini System

Configuration adjustments and foundational infrastructure improvements for efficiency gains.

- Configuration optimizations and tuning
- Containerisation and virtualisation pilots
- API efficiency features implementation
- Caching mechanism deployment

Gas Data Portal

Rapid deployment of cloud and efficiency improvements for read-only public datasets.

- Cloud CDN migration and deployment
- REST API transition completion
- Basic rate limiting implementation
- Data pruning and archiving initiation

Common Initiatives

Shared focus on reducing carbon footprint through green infrastructure choices.

- Green hosting quick wins assessment
- Energy efficiency audits completion
- Awareness and training kickoff
- Baseline carbon metrics establishment

Mid-Term Strategy (6-18 Months): Cloud Migration & Transformation

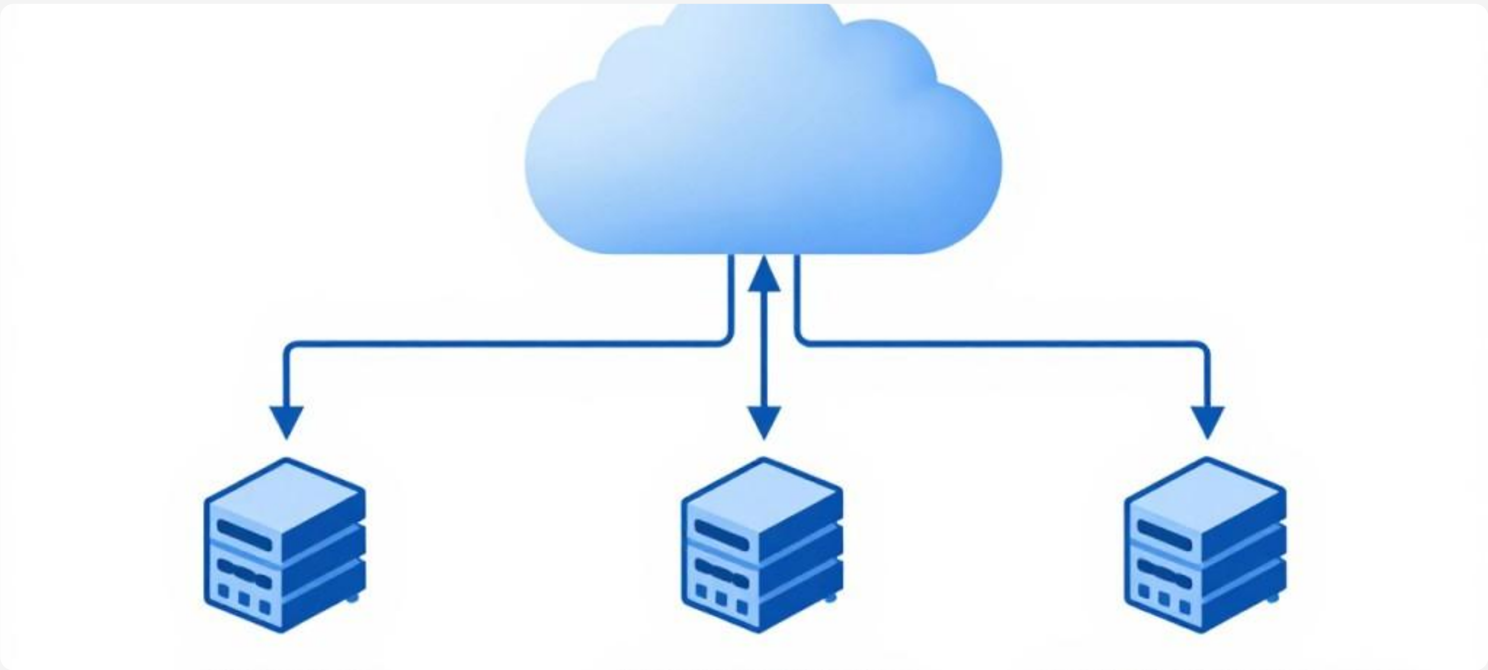
Hybrid & Cloud Deployment

Strategic cloud migration with green managed services, elastic scaling, and environmental lifecycle policies for optimized resource usage.

Architectural Refactoring

Transition to microservices and serverless architectures, especially for batch jobs, enabling dynamic resource allocation and energy efficiency.

Long-Term Vision (18+ Months): Net-Zero Digital Ecosystem



Carbon-Neutral Operations

- Full transition to carbon-neutral or renewable-powered hosting environments
- Advanced peak-shaving and auto scaling intelligence systems
- Continuous code efficiency monitoring and optimization
- Integration with sector Digital Strategy Initiative (DSI) frameworks

Fully Serverless & Automated

- Comprehensive serverless architecture with automatic scaling on demand
- Integrated carbon monitoring and real-time optimization systems
- Advanced data minimization and intelligent lifecycle management
- Community collaboration and cross-sector integration initiatives

Part E.1 : Strategic Priorities & Differentiated Approaches

System-Specific Approaches & Additional Strategic Layers

Gemini: Cautious Path

Incremental, conservative approach due to critical market functions and sensitive operational data. Focus on internal optimization, hybrid cloud moves, and risk mitigation to prevent service disruptions while achieving sustainability goals.

Gas Portal: Agile Path

Faster adoption of cutting-edge cloud and serverless solutions due to lower failure impact. Aggressive scaling optimization, advanced containerization, and rapid technology adoption for maximum efficiency gains.

Cross-Sector Integration

Inter-platform federation, shared services, experimentation with low-code/no-code frameworks, and carbon accounting integration for holistic sustainability reporting.

Edge Computing & AI

Decentralized efficiency through edge computing, AI-driven load prediction, carbon-aware scheduling, and intelligent resource optimization for reduced latency and emissions.

Stakeholder Engagement & Governance Strategy

Building Trust & Partnership

Comprehensive stakeholder engagement across internal teams (IT, Operations, Regulatory) and external partners (Ofgem, NESO, industry, research, public) aligned with technical roadmap phases.

Education & Awareness

Building understanding of sustainability goals, green IT practices, and implementation benefits across all stakeholder groups from day one.

Consultation & Co-Design

Active stakeholder participation in solution design, feedback loops, and iterative improvement cycles to ensure solutions meet diverse needs.

Governance & Adoption

Establishing governance forums, training programs, and partnership frameworks to ensure sustainable long-term adoption and continuous improvement.

Engagement Methods & Best Practice Framework

Education & Awareness

Workshops, webinars, documentation

Consultation & Co-Design

Working groups, feedback sessions, collaborative design

Feedback & Iteration

Regular surveys, suggestion channels, continuous improvement

Training & Onboarding

Technical training, adoption support, certification programs

Communications & Transparency

Monthly updates, progress reports, open communication

Partnership & Governance Forums

Steering committees, advisory boards, cross-sector collaboration

Phased Engagement Timeline: Alignment with Technical Roadmap

Short-Term (0-6 months)

Awareness & Discovery

- Building foundational understanding through executive briefings
- Conducting capability assessment workshops
- Establishing baseline knowledge via technical documentation
- Identifying pilot use cases for validation

Mid-Term (6-18 months)

Co-Design, Feedback & Onboarding

- Collaborative solution design sessions with cross-functional teams
- Structured feedback loops through prototype testing
- Comprehensive training programs for technical staff
- Gradual system integration with controlled rollouts

Long-Term (18+ months)

Adoption, Governance & Continuous Improvement

- Enterprise-wide community events showcasing success stories
- Formal governance forums for strategic alignment
- Advanced training for power users and champions
- Ongoing optimization through performance benchmarking

Key Stakeholders & Engagement Objectives

MIND&MATTER

Internal Stakeholders

IT Teams

Awareness, training, ownership

Operations Teams

Operational readiness, efficiency monitoring, cost benefits

Regulatory Affairs

Compliance alignment, sustainability reporting

External Stakeholders

Ofgem & NESO

Regulatory engagement, feedback incorporation

Industry Data Users & Research Community

Adoption, feedback, innovation

General Public

Awareness, transparency, community support

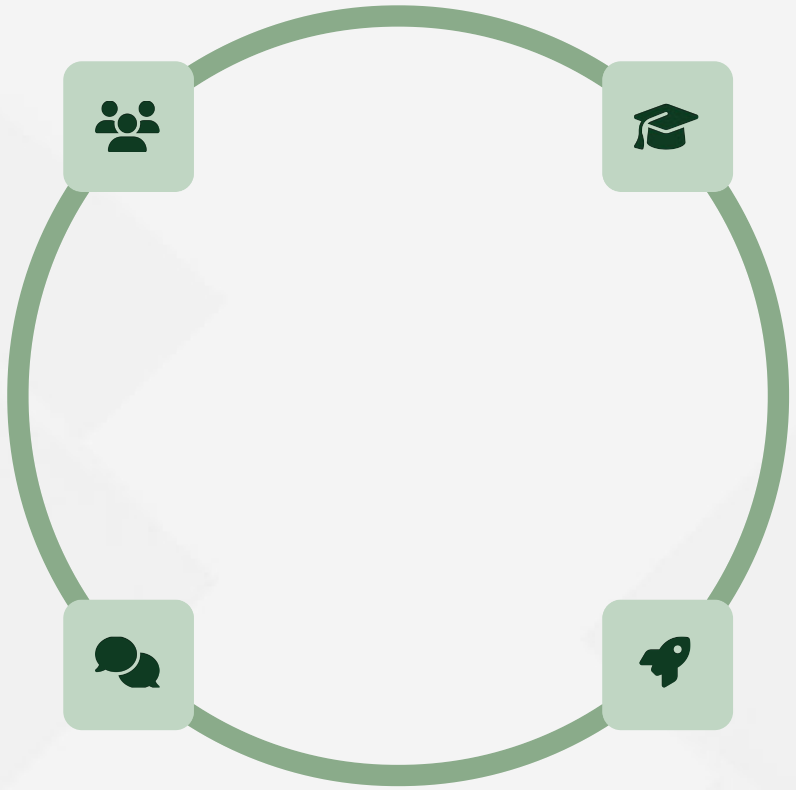
Success Metrics & Impact Measurement Framework (rough numbers)

Engagement Reach

- **Participation rates:** 85% **(target 80%+)**
- **Survey response rates:** 75% or above
- **Attendance at key events:** 92%
- **Stakeholder registration:** 95%

Knowledge Transfer

- **Assessment scores:** 88% avg.
- **Training completion:** 91%
- **Capability maturity:** Level 3.2



Feedback Integration

- **Suggestions implemented:** 142
- **Resolution time:** 2.1 days avg.
- **Stakeholder satisfaction:** 4.7/5
- **Quarterly NPS:** +48

Adoption Velocity

- **Solution adoption:** 76%
- **Feature usage:** 82%
- **Time-to-value:** 3 – 4 weeks
- **Peer benchmark:** +15%

NOTE – Numbers are an estimate only, subject to changes pending internal/external stakeholder engagement

Communication Strategy: Channels & Frequency by Stakeholder

Multi-channel approach ensures appropriate engagement cadence for each stakeholder group

Stakeholder Group	Primary Channel	Frequency	Key Messages	Responsible Team
Internal Stakeholders	All-hands meetings	monthly	Company updates, strategic direction	Executive Leadership
Regulatory Bodies	Formal reports	quarterly	Compliance updates, policy changes	Legal & Compliance
Industry Users	User forums	bi-monthly	Product updates, best practices	Product Management
Research Community	Technical documentation, annual summit	on-demand + annual	Technical specifications, research findings	R&D Team
General Public	Website updates	quarterly	Company news, public announcements	Marketing & PR

Engagement Priorities & Quick Wins: Immediate Actions

Step 1 - Establish Governance

Week 1-2

- Define steering committees
- Assign ownership
- Clarify decision rights

Step 2 - Launch Awareness Campaign

Week 2-4

- Executive briefings
- Stakeholder newsletters
- Kick-off workshops

Step 3 - Identify Champions

Week 3-6

- Recruit internal and external advocates
- Provide champion training
- Establish support network

Step 4 - Begin Co-Design

Week 5+

- Form working groups
- Launch prototype feedback
- Collect early input