Department for Transport

PUBLIC
A SOLITA COMPANY

API Development at the DfT - User Research

Final Report

May 2026

Department for Transport
Great Minster House
33 Horseferry Road
London
SW1P 4DR

This publication is also available on our website at
www.gov.uk/government/organisations/department-for-transport

Any enquiries regarding this publication should be sent to us at
www.gov.uk/government/organisations/department-for-transport

Contents

Foreword

Executive summary

Purpose and scope

Under the Data Action Plan and Transport Data Strategy, the Department for Transport (DfT) aims to improve the discoverability, accessibility, interoperability and quality of transport data. This will support innovation, enable effective decision making, and allow more information to be available to transport users. The Data Action Plan commits to promoting “suitable APIs as the expected method for sharing transport data where appropriate". Whilst existing initiatives, such as the Rail Data Marketplace, and the NaPTAN and NPTG APIs, provide a strong foundation for using data to improve the transport system, API readiness remains uneven across the sector and many organisations still rely on manual data-sharing.

In November 2025, DfT commissioned PUBLIC to deliver the “DfT API Development - User Research” project, to understand the value of releasing more DfT data via APIs and the implications for API delivery and sustainability.

This Executive Summary summarises findings and recommendations from our Final Report, drawing on research between November 2025 and February 2026. It begins with the headline implications for DfT, followed by key findings and the recommended approach.

What this means for DfT

Collectively, the evidence points to four implications for DfT’s near-term approach:

1. **Prioritisation should focus on where APIs materially reduce user cost and friction.** This is most visible for high-demand, time-sensitive or high-frequency datasets; large statistics assets where queryability reduces manual wrangling; and foundational reference datasets that underpin multiple downstream uses.

2. **Adoption will depend as much on data quality as on new endpoints.** Users repeatedly described discovery, documentation, metadata clarity, and data quality as binding constraints, meaning that, for some datasets, improving usability and consistency may deliver disproportionate value even before (or alongside) API conversion.

3. **DfT should treat charging as a targeted sustainability lever, not a default gate.** Interviewees indicated that pricing/licensing design strongly affects uptake, particularly for smaller organisations, and preferred charging models that scale with use/value where charging is required. This reinforces the case

for protecting baseline public value while identifying objective triggers for cost recovery (e.g., very high-volume use, premium service requirements).

4. **DfT should build the evidence base needed to manage sustainability decisions.** Defensible tiering or marginal cost recovery depends on consistent evidence on usage intensity, concentration, and marginal cost-to-serve, which is not yet systematically available.

Summary of findings

DfT's open datasets vary in accessibility, availability and quality

We reviewed 85 publicly available DfT dataset entries to assess data coverage, access routes, update patterns, and metadata usability. Some entries are derivative releases of a main dataset. Our findings include:

- **Data concentration:** DfT publishes a wide range of open data, heavily concentrated on roads and operational activities. There is also substantial coverage of travel behaviours and safety, public transport (including bus/NaPTAN), and freight, logistics and licensing. Coverage of non-land transport and environmental data is limited.

- **Data access and licensing:** Most datasets are released as downloadable files, usually spreadsheets or ZIPs. APIs and real-time feeds are exceptions but deliver high value where available (e.g., Street Manager and the Bus Open Data Service (BODS)), enabling automated and time-critical operational use cases. Most datasets are licensed under the Open Government Licence (OGL), though licensing information is inconsistently present.

- **Update frequency:** Update cycles range from annual statistical releases to daily or real-time feeds, and are often inconsistently described, limiting automation and comparability.

- **Structure and metadata:** Datasets are generally well structured, with clear geographic and temporal fields. However, uneven and incomplete metadata often requires significant cleaning and validation before reuse.

Benchmarks suggested that APIs should be use-case led

We benchmarked APIs from the Department for Education, Transport for London, and the Driver and Vehicle Licensing Agency. The findings show that APIs should be treated as a portfolio of service types, rather than a single "publish data" decision. This approach should support prioritisation based on where APIs reduce user burden or enable time-critical use, supported by predictable change practices and proportionate access controls:

- **Across these API benchmarks, standardisation emerges as the primary adoption driver.** Predictable responses, consistent naming and error handling,

and practical documentation and guidance often reduce engineering effort and support demand.

- **Openness is typically combined with lightweight governance rather than strict gating.** Access is generally open, with registration or API keys used primarily for operational monitoring and service protection.

- **Charging is not the default: where it exists, it is rules-based and narrow.** Baseline access is usually free. Any fees are typically linked to defined transaction type or exceptional usage pattern. This means that cost recovery is most defensible when it is tightly scoped and transparently explained.

- **The definition of "good" depends on the data type.** Operational APIs prioritise reliability and stability; statistical APIs prioritise queryability, metadata and reproducibility; and reference APIs prioritise being authoritative, stable, and easy to integrate.

User demand for API access is high but depends on data quality

We conducted 55 user interviews to understand where API-enabled access adds the most value and what conditions are required for success. The research developed **five user personas**, who use DfT data in different ways and have distinct needs: Local Transport Authority Analyst, Transport Researcher, Operational Controller, Data Engineer & Integrator, and Developer & Innovator. Across personas, seven consistent themes emerged:

- **Multimodal interoperability needs a clear mission and should be considered from the start.** DfT has an opportunity to act as a system integrator and standard-setter across its Arm's-Length Bodies (ALBs) and transport modes, particularly to address gaps in freight, maritime and aviation.

- **Data quality is a prerequisite.** Users consistently emphasised that quality comes first. Gaps in accuracy, completeness, provenance and documentation undermine trust regardless of whether access if file-based or via API.

- **Discoverability remains a major barrier.** Many users are unaware of what DfT data exists and find it difficult to locate relevant datasets. This highlights the need for a single, authoritative catalogue setting out ownership, coverage, quality notes, access routes and update schedules.

- **APIs should be use-case led, not universal:** Users see greatest value in targeted queries (such as specific fields, subsets or time ranges) and high-frequency or near-real-time access. Users do not support default API provision for all datasets.

- **Pricing and licensing affect uptake:** Smaller organisations are sensitive to high upfront or blanket fees and prefer usage-based models. Local authorities also highlighted that potential value of coordinated procurement of high-value commercial datasets (e.g., mobile network data) to reduce duplicated purchases and improve value for money.

- **Users want outcome-based prioritisation.** They want clarity on DfT's intended outcomes from sharing data, and clear accountability and quality standards for maintaining and improving each dataset.

- **Incremental delivery is preferred:** Users favour iterative implementation over a waterfall delivery approach, starting with the most widely used datasets (e.g., BODS, Street Manager, traffic data, NTS, STATS19) and expanding over time.

A five-factor framework helps evidence-based dataset prioritisation

Drawing on the findings above, we prioritised DfT datasets for API development using a structured assessment that combines user demand, the need for timely updates, GDS alignment, dataset size, and data quality. Full details of the prioritisation framework and scoring are provided in Appendix 4.

This assessment groups datasets into three broad priority bands:

- **High-priority datasets** (Bands 1-2): The top-scoring datasets are the Roadworks Service API (Street Manager) and BODS, both already available via API and demonstrating clear value. Other high-priority datasets cover daily domestic transport use by mode, LGV vocational driving tests, BODS bus locations, NaPTAN, and the Transport Statistics API, and road traffic data (largely available via API), as well as vehicle licensing/EV charging, travel behaviour, and road safety series.

- **Medium-priority datasets** (Band 3): This group is dominated by analytical and monitoring datasets, notably safety and transport statistics, and some travel time measures.

- **Lower-priority datasets** (Bands 4-6): These datasets are mainly periodic publications or datasets least prioritised by users. They are more suitable for inclusion in longer-term roadmap items or targeted discovery in the future.

This prioritisation also aligns closely with user feedback. Interviews with 55 users identified clear demand for API-enabled access, particularly for:

- Buses and public transport operations

- Road events/restrictions/disruption

- Road traffic demand

- Travel behaviour statistics

- Safety and compliance

- Vehicle licensing and EV charging; active travel; fares, ticketing and payments; advanced products (e.g. digital twins)

This framework provides a practical, evidence-based approach for API development.

API access creates value by reducing friction and enabling decisions through five primary benefit mechanisms

Based on user interviews, we identified five primary mechanisms through which improved access to DfT data, including via APIs, creates value. Collectively, these support the Government's growth agenda by increasing productivity, both by enabling innovation and new value-add services to be developed, and by improving the efficiency and reliability of transport decision-making:

- **Faster time-to-insight:** reducing manual downloading, cleaning, and reconciliation effort, enabling analysts to spend more time on interpretation and decision support.
- **Automation and system integration:** enabling stable machine-to-machine workflows (pipelines, dashboards, system integrations) that scale with frequency of use rather than analyst capacity.
- **Improved timeliness and responsiveness:** reducing access friction and refresh delays, supporting time-sensitive monitoring and operational decisions where timeliness is binding.
- **Cost savings and reduced dependency:** reducing duplicated local processing and reliance on bespoke solutions or alternative paid data sources where DfT data can meet the need.
- **Better consumer-facing services and user experience:** improving consistency and reliability of passenger-facing information, reducing end-user uncertainty and the need to cross-check multiple sources.

APIs do not create new data: they reduce the friction of accessing and integrating existing data. As a result, APIs derive their value from the quality and decision value of the underlying data. If the data does not inform decisions, improving access will not create meaningful value. Where data materially reduces uncertainty, lowering access and integration barriers amplifies realised value by increasing usability and uptake.

Charging should be targeted, transparent and protect public value

We developed an indicative, benchmark-informed framework for marginal cost recovery associated with API access. It is not a proposal to introduce charging: rather, it sets out the circumstances in which charging could be considered proportionate, alongside the safeguards needed to protect baseline public value and maintain low-friction access.

- **Charging is not the default.** Benchmarking shows that free baseline access is standard for public-value datasets; where fees exist, they are narrow and rules-based.
- **Protect baseline public value.** Any charges should preserve low-friction access for public-interest, research, and low-volume users.

- **Be targeted and evidence-led.** Any charges should reflect observed usage intensity, concentration and clear evidence of marginal cost-to-serve.

- **Link charges to objective triggers.** Where charging is considered, it should be triggered by sustained high-intensity usage and/or premium service, rather than a general paywall.

- **Ensure transparency and predictability.** Thresholds and pathways should be clearly published and stable over time, with reinvestment in data quality, documentation, and operational reliability.

Recommended approach

Overall, the value of API access is driven by usability, not quantity. It is realised when users can find the right dataset, trust it, and use it with minimal friction. Because transport datasets vary in how they are collected, updated, and used, improving usability will require different approaches across datasets: a one-size-fits-all model is unlikely to work.

This has a practical implication: DfT cannot improve everything at once, and some changes only deliver value once the basics are in place. Prioritisation is therefore essential - deciding which datasets to tackle first and which foundations (e.g., quality, documentation, access route, enablement) to establish ensures that later investments, including APIs, are adopted and sustained.

The **recommendations** are structured to reflect that sequencing and to support decision-making using the evidence and tools developed in this study, alongside ongoing demand signals (e.g., recurring FOI/EIR requests and usage analytics) and user feedback. Our recommendation outlines what “good” looks like, while the practical next steps suggest a manageable starting point that can be expanded over time.

1. Reliability and quality assurance are prerequisites

Data reliability has two linked aspects: data quality, meaning data is accurate enough to reuse and comes with clear provenance and known limitations, and stability over time, meaning updates and changes are predictable and do not disrupt reuse.

Improvement to data discoverability, cataloguing, and APIs will only lead to adoption if users trust the underlying data and changes are predictable. For this reason, reliability is an enabling layer for all other recommendations and should be addressed through a phased approach.

Recommended approach

We recommend two complementary levers to improve data reliability:

- **Reduce upfront uncertainty:** Make it easier for users to decide whether a dataset is suitable for their use case by publishing clear provenance (where data comes from), known limitations, and basic quality indicators (e.g., completeness, timeliness).

- **Reduce disruption for frequent users:** Adopt predictable release and change management. This could include publishing refresh expectations where possible, maintaining versioned releases with a short changelog for changes to fields, codes, or methods, and retaining snapshots/versioned extracts to support longitudinal analysis and reproducibility.

Practical next steps

To keep delivery manageable and avoid excessive upfront maintenance, DfT and data owners could start by:

1. Selecting 2-3 high-use datasets where quality or change-related issues most constrain reuse, drawing on the evidence and tools from this report alongside ongoing demand signals.

2. Publishing provenance, known issues, and refresh expectations for those datasets as a minimum baseline.

3. Implementing basic versioning/changelog practice and change notices for material updates.

4. Extending to additional datasets once the approach is working in practice and data owners have capacity to keep these basics up to date.

2. Access and standardisation work best on the most reliable data

Usability improves most when users can both trust the data and find it easily. Yet users often struggle to locate the right dataset, verify it is authoritative, and understand its content without additional follow-up.

A **central catalogue will help reduce this fragmentation,** but it only works if maintained as continuous service with clear ownership and routine updates, particularly where information is spread across teams and systems.

Recommended approach

We recommend treating the catalogue as a simple “front door” that starts small and scales over time:

- **Reduce discovery friction:** Publish a central catalogue for an initial set of high-use datasets and their access routes, with scope expanding over time to include relevant ALB and DfT-funded service data.

- **Standardise metadata:** This could include use a short, maintainable entry template that include the responsible owner/team and support contact, dataset contents, access route, licensing, coverage, refresh frequency, known issues, and links to documentation.

- **Make datasets easier to combine:** Keep common IDs consisted over time, publish mapping tables where multiple ID systems exist, and document how datasets relate to each other.

- **Match access to the use case:** Provide APIs for filtered or frequent access, while retaining low-friction bulk downloads for large or static datasets and research workflows, using common formats such as CSV and, where appropriate, Parquet.

Practical next steps

To make this sustainable, DfT could start by focusing on a small initial set, setting clear ownership for upkeep, and expanding coverage once the Minimum Viable Product (MVP) is working in practice:

1. Agree the initial scope and dataset list with data owners (DfT-first, then expand where feasible)

2. Publish an MVP catalogue using a standard entry template and named owners for updates.

3. Prioritise a small number of joinability fixes that unlock clear value (e.g., identifiers/ mapping tables).

4. Expand coverage and strengthen standards over time based on demand, readiness, and user feedback.

3. User enablement and support are key to API adoption

Even when data is available, uptake can be low if users do not know how to get started, who to contact, or how to raise issues. This is especially true for users with limited technical capability. Once a dataset is reliable and easy to find, light-touch support can help users adopt API access in practice.

Recommended approach

We recommend a pragmatic enablement offer to help users adopt APIs in practice:

- **Make it easy to get started:** Provide simple guidance (e.g., FAQs, worked examples, and clear “how to access” instructions) and test it with users. Where APIs return developer-oriented formats (e.g., JSON), include simple examples showing how to convert outputs into analysis-ready tables

- **Provide a clear route for questions and issues:** Use a single support or feedback channel so users know where to go. Keep it manageable with a lightweight triage process so queries reach the right data owner/team and recurring issues feed into documentation and “known issues” updates.

- **Support practical reuse:** Publish a small amount of “getting started” code (e.g., a script or notebook) for selected high-use datasets or APIs. Maintain bulk download options for national datasets where full extracts and historical time series matter (e.g., STATS19).

- **Build on existing community tools:** Where researchers already use third-party tools to work with DfT data, keep formats and field names consistent so those tools keep working (e.g., the STATS19 R package supporting reproducible access to STATS19). Validate what support is most useful through focused sessions (e.g., office hours, workshop, hackathon, training day).

Practical next steps

To keep delivery manageable, DfT could start by:

1. Publishing a minimal support pack for the initial set of datasets/APIs (e.g., FAQ, “how to access” and worked examples).

2. Setting up one support or feedback channel and define a simple triage route.

3. Updating guidance and examples based on recurring questions and common friction points.

4. Prioritisation and service decisions guide delivery and charging

Pricing and licensing can influence uptake, particularly for smaller organisations and local authorities. If charging is considered in future, it should reflect where value is created and to avoid creating barriers to use.

More broadly, where API access ais not yet in place, DfT needs to decide which datasets to progress next, what access routes to provide (e.g., API and/or bulk downloads), and what service expectations are realistic to sustain, including any charging.

Recommended approach

We recommend setting a clear decision process for what to do next: which datasets to prioritise, what access to offer (API and/or bulk downloads), and what level of service to commit to. If charging is considered, it should support sustainability without creating barriers to use:

- **Prioritise the next datasets using evidence signals:** Build on the Prioritisation Framework to identity an initial set of candidates for API adoption, drawing on demand signals, data readiness and intended outcomes and beneficiaries.

- **Test a targeted approach before scaling:** Preserve low-friction baseline access for public-interest, research, and low-volume users, and focusing any charges on sustained high-frequency or commercial use, typically via intermediaries that capture downstream value.

- **Differentiate tiers by service quality:** If tiered charging model is considered, base them on predictable change control/versioning, enhanced support, reliability expectations rather than volume alone.

- **Pilot and reinvest transparently:** Pilot with a small number of datasets and reinvest any revenues into data quality and governance.

Practical next steps

DfT could start by:

1. Agree governance principles for prioritisation (e.g., evidence signals, readiness checks, and decision points on API vs bulk).

2. Use the Prioritisation Framework to identify an initial set of candidates for API adoption and a short pipeline, documenting assumptions and constraints.

3. If charging is in scope, agree charging principles and decision criteria, using the Hypothetical Charging Model as a starting point.

4. Define a small pilot set with clear success measures (e.g., uptake, administrative burden, user outcomes), and review before any expansion.

This report aims to establish: a decision-ready evidence base of DfT's data landscape and API patterns that work elsewhere; a grounded understanding of what users need and where value arises; and a structured approach to prioritisation, benefit valuation, and (if required) marginal cost recovery. Taken together, these elements support a practical, incremental route to sustainable API delivery that improves access while protecting baseline public value.

Section 1. Introduction

1.1 Context

The Department for Transport (DfT) is taking steps to improve how transport data is discovered, accessed, and reused across the sector, including exploring where APIs could reduce friction and enable more timely, automated use of data. This work sits within the strategic context of DfT's Data Action Plan and Transport Data Strategy, which emphasise improving data accessibility and enabling more consistent, reusable dissemination. In practice, however, API readiness remains uneven, and many organisations still rely on manual data-sharing processes.

To improve the use of and access to data, DfT commissioned PUBLIC to undertake research into the market for APIs for DfT data and the implications of releasing more information in this format. This research had two components:

1. **User research** examined how the transport sector uses APIs, and the benefits and barriers associated with their use; and

2. **Economic analysis** estimated the potential value of releasing more DfT data via APIs.

Together, this work informs DfT's decisions on future data dissemination and the case for investment in APIs.

1.2 Project objectives and phases

This project set out to:

- Understand how API access to DfT data could benefit the transport sector and the wider economy;

- Estimate the economic value of these benefits;

- Prioritise datasets with the greatest potential value if released via APIs;

- Assess willingness to pay for APIs of DfT data;

- Develop a hypothetical charging model; and
- Identify barriers to API use and the support required by organisations with lower levels of API readiness.

To deliver these objectives, the project was structured into five phases between November 2025 and February 2026:

1. **Phase 1:** Supply Analysis - DfT Data and API Discovery;
2. **Phase 2:** Demand Analysis - User Research;
3. **Phase 3:** Prioritisation and Interim Report ;
4. **Phase 4:** Benefit Framework, Quantification and Charging Model;
5. **Phase 5:** Final Delivery.

1.3 Project scope

This Final Report synthesises the full programme of work and sets out findings and recommendations for DfT.

Specifically, this report presents:

- **Findings from the supply-side analysis**, including mapping and structured assessment of 85 DfT datasets (including some derived datasets), covering attributes such as completeness, quality, and metadata, alongside a rapid review of comparable public-sector API benchmarks (e.g., the Department for Education, Transport for London) to understand API availability, readiness, and good practice (Section 3).
- **Insights from the demand-side user research**, drawing on 55 interviews (including 1 written response) with a representative mix of transport data stakeholders. The report outlines 5 user personas, and synthesises how users currently use DfT data, their priority needs, the barriers they face, the benefits they associate with improved API access, and early indications of willingness to pay (Section 4).
- **Dataset prioritisation framework**, combining both supply and demand insights with feasibility considerations to identify datasets with the highest potential value if released via APIs (Section 5).
- **Benefits framework and economic analysis,** translating user research evidence into 5 priority benefits and a structured assessment of how value arises across personas, supported by transparent proxy measures and valuation mechanisms to inform economic appraisal and charging considerations (Section 6).

- **Hypothetical charging model,** setting out a benchmark-informed framework for marginal cost recovery for API access, with indicative charging triggers and an application to prioritised datasets (Section 7).

- **Practical recommendations,** setting out next steps and options to improve dissemination and enable API adoption over time (Recommended approach).

Section 2. Methodology

This section summarises the project methodology at a high level. The work was delivered through five interlinked workstreams, combining supply-side analysis, benchmarking, and demand-side research to assess the case for expanding API access to DfT data, prioritise candidate datasets, and explore potential benefits and sustainability options. Further methodological detail (including sampling, assessment criteria, benchmark descriptors, and limitations) is provided in Appendix 1.

Table 1: Summary of workstreams

Workstream	Description	Evidence base
DfT data mapping	Mapped and assessed DfT datasets to establish a baseline view of discoverability, access routes, metadata quality, and indicators of API readiness.	Structured review of 85 datasets listed on Find Transport Data and Transport Statistics Finder, including related/derived variants of core datasets.
API benchmarking review	Reviewed comparator public-sector APIs to identify transferable good practice (e.g., documentation, access models, licensing, and sustainability mechanisms).	Desk-based review of comparator APIs and published documentation.
User research	Conducted interviews to understand current use of DfT data, priority needs, barriers to use, perceived benefits of API access, and early willingness-to-pay signals.	55 interviews (including one written response) with transport data stakeholders (24 Nov-11 Dec 2025).
Benefit framework	Synthesised user research into a structured set of priority benefits and value pathways, supported by a transparent approach to valuation.	Coded interview findings, triangulated with supply-side assessment and prioritisation outputs; supported by published sources used for proxy valuation measures.
Hypothetical charging model	Developed a benchmark-informed approach to marginal cost recovery for API access, including indicative charging triggers and application to prioritised datasets.	Benchmark evidence on charging/licensing approaches, supplemented by interview insights on willingness to pay and affordability constraints.

Limitations

Findings reflect what is publicly discoverable and what stakeholders reported in interviews. Some modes and stakeholder types were under-represented during recruitment (notably aviation and, to a lesser extent, maritime), so findings should be interpreted accordingly.

Section 3. Supply Analysis - DfT Data and API Discovery

Supply analysis, comprising data mapping and API benchmarking, aimed to establish a clear picture of the current data and API landscape relevant to DfT. The data mapping developed a structured inventory of DfT-published datasets, including formats, update cycles, metadata quality, and access routes, recognising that in some domains (particularly aviation) widely used public datasets are made available via sector regulatory bodies rather than DfT. The API benchmarking review looked outward rather than inward. It examined how other UK public-sector bodies design, document, and deliver APIs, identifying common patterns, good practices, and maturity models to inform what is feasible and effective for DfT.

3.1 DfT data mapping

The data mapping exercise provides a structured view of DfT's dataset estate. It allows the Supply Analysis to move beyond a simple inventory and assess value, readiness, and gaps.

Datasets were grouped into ten user-oriented categories, such as operations, behaviour, assets, and safety. This shows how current provision aligns with how stakeholders work with transport data. Assigning each dataset a primary category and subcategory enables both high-level pattern analysis and detailed filtering.

The analysis reveals variation in availability, format, openness, geography, timeliness, and access routes. It shows not only what data exists, but how usable and consistent it is. Clear differences emerge between bulk-download datasets and those supported by APIs. There are also inconsistencies in metadata quality.

Together, this evidence supports gap identification, API readiness assessment, and subsequent prioritisation and valuation.

Framework

Datasets cluster into ten clear categories. These categories are adapted from the portal topic lists but refined to better reflect how users interpret and navigate transport data.

Together, they provide a consistent structure for grouping datasets and reveal the dominant themes within DfT's data estate. The ten categories, along with examples of their most common sub-categories, are set out in Table 2.

Table 2: Data Categorisation

Category	Descriptor Tags
Travel Demand & Behaviour	travel behaviours, attitudes, surveys
Traffic Operational Data	traffic volumes, journey times, congestion, flow counts
Public Transport	timetables, stops/access points, open bus data, fares
Road Freight & Logistics	HGV/van statistics, international freight flows
Infrastructure & Assets	EV charging points, roadworks, asset condition
Licensing & Regulation	vehicle licensing, blue badge, taxis, speed compliance
Safety, Accidents & Responses	collision records, casualties, road safety tables
Non-land Transport	aviation and maritime statistics
Environment & Sustainability	emissions, RTFO, sustainable aviation fuel stats
Others / Cross-cutting	multi-domain or miscellaneous datasets

This structure was used as it better adheres to typical user needs such as, behaviour, operations, assets, safety, etc while still incorporating DfT portal topics as to not stray from their original framework. This also makes the process reproducible as our chosen categories encase any possible transport data topic. Each dataset was allocated by using the given topic and description notes. To better specify our categorisation, we assigned a primary category plus a subcategory to allow users to filter both ways. Subcategories were used to capture more granular specifications (e.g., Traffic → Journey times, Traffic → Flow counts), which allowed high-level filtering without the addition of an unnecessary amount of categories.

Metadata Summary

A structured metadata model was applied to every dataset to ensure comparability across the two portals and to enable a systematic assessment of dataset usability. These fields capture what users need to understand a dataset's scope, format, openness, and update patterns. Granular detail is available in Appendix 2.

Headline insights

1. **Content concentration:** There is a strong concentration of datasets related to road and multi-modal traffic, including road traffic estimates, journey-time measures. A second large cluster of datasets also focuses on roads, but with a stronger consumer perspective rather than logistics. This includes travel demand and road safety statistics, which represent a major focus for DfT data

collection and primarily support everyday use and monitoring rather than long-term planning. Public transport represents a clear third concentration, with a particular emphasis on bus data such as bus timetables and stop access nodes (e.g., NaPTAN). Freight and logistics and licensing/regulation form further dense clusters. Non-land transport and environment/sustainability are smaller but specifically distinct groups. Overall, these patterns reinforce DfT's road-focused and operational approach to data collection.

2. **Data access and licensing:** Most datasets are published as downloadable files, primarily containing spreadsheets (e.g., CSV, ODS, XLS/XLSX) or ZIP archives. JSON is also used, mainly for datasets that provide API access. The majority of datasets are open and published under the UK Open Government Licence (OGL), although licensing information is not always applied consistently across sources.

3. **Update frequency:** Update frequencies range from yearly and quarterly to monthly and live. Operational datasets are typically updated daily or in real time, while statistical series are usually updated quarterly or annually. Several datasets describe their time coverage as "to the present" or "ongoing", while others are limited to a specific year or survey period.

4. **APIs and real-time feeds:** APIs and real-time feeds represent a small proportion of the dataset landscape but are strategically important. Most datasets are static tables or spreadsheets, while a small group provides automated access through APIs such as the Transport Statistics API, the Roadworks Service API, and Bus Open Data feeds for real-time vehicle locations and timetables. These datasets that provide API access focus on operational and real-time transport use cases.

5. **Datasets structure and metadata:** Most datasets follow a consistent structure with clear geographic and temporal fields, such as region, or date. However, metadata quality varies significantly across datasets and sources. File sizes, formats, and licensing information are not always listed or consistent, and other metadata fields may be missing. The collection therefore spans highly structured, large-scale datasets as well as smaller, more fragmented ones. While this creates significant opportunity, it also highlights the need for metadata cleaning and validation with DfT before the data can be used reliably as a reference source.

3.2 Comparative review of public sector APIs

Headline insights

1. **Standardisation is the primary adoption lever.** The most effective APIs reduce cognitive and engineering overhead through predictable response structures (commonly JSON), consistent naming and error handling, and documentation that moves beyond endpoint lists into "how to use this in practice" guidance (e.g., examples, common workflows, and clear definitions). This lowers time-to-first-success for developers and reduces the volume of clarifications and support requests the API owner must handle.

2. **Openness is usually paired with lightweight governance rather than hard gating.** Benchmark examples show that it is entirely possible to keep public data broadly accessible while still protecting performance. Registration or API key models are frequently used not to restrict access, but to enable rate limits, usage monitoring, and fair use. In effect, “open” is treated as a default stance, while API keys are treated as an operational control that keeps the service stable at scale.

3. **Charging is not a default model; where it exists, it is rules-based and narrow.** The benchmark suggests an established norm of free baseline access for public-value data. Where fees appear, they tend to be tied to a defined transaction type or exceptional usage pattern rather than general access to the data itself. The design implication for DfT is that cost recovery, if required, is most defensible when it is tightly scoped and transparently explained - because broad paywalls create friction, suppress adoption, and reduce the public benefit of open data.

4. **Data type determines what “good” should optimise for.** Operational transport APIs are judged primarily on reliability, stability under load, and clarity of entitlement/usage rules. Statistical APIs are judged primarily on queryability, metadata richness, and reproducibility (i.e., enabling users to extract the right slice of data without repeatedly handling very large files). Reference APIs sit between: their value comes from being authoritative, stable, and easy to integrate.

A practical categorisation: three API product types

The benchmarking points away from a single uniform DfT API pattern and towards a portfolio approach with three distinct product types. These categories provide a simple way to understand the design priorities behind the case studies below.

5. **Operational APIs (real-time, disruptions, high-frequency):** These support live operational and passenger-facing use cases where timeliness matters. They optimise for reliability, stability under load, clear usage rules, monitoring, and predictable change control.

1. **Reference APIs (authoritative lookups, identifiers, network components):** These provide stable building blocks that many other services depend on, such as identifiers and canonical lists. They optimise for stability, schema consistency, versioning, and ease of integration.

2. **Statistical APIs (tables, indicators, time series):** These support analytical use cases where users need to repeatedly access subsets of large datasets. They optimise for queryability, metadata richness, and reproducible extraction workflows, reducing reliance on file downloads.

Case study 1: TfL as an example of an Operational API

TfL's experience points to a clear takeaway for DfT: when data is used to run live services, the policy challenge is not just "publish it", but "run it as a dependable public service". For information that changes very frequently (e.g., data captured every minute, or multiple times per hour), users build tools and operational processes that rely on the feed being consistently available, especially during peaks and major disruption. In this context, the key design choices become governance choices: how DfT ensures continuity, manages spikes in demand, and avoids sudden changes that break downstream services.

What this looks like in practice is an approach where baseline access remains open, but DfT uses light-touch access management (e.g., issuing access credentials) to protect reliability and build an evidence base on usage. This is less about creating barriers and more about having a practical way to: (i) manage demand fairly across users, (ii) maintain service performance, and (iii) understand which datasets are delivering the most value in real-world use. The broader lesson is that the public value of operational transport data is highest when it is paired with the operational disciplines that keep it usable at scale (clear expectations for use, visibility on performance, and predictable update practices) so that openness translates into sustained adoption rather than a brittle service that performs well only under light use.

Case study 2: DVLA as an example of a Reference service model with explicit operating rules

DVLA's model supports a second key takeaway for DfT: some data services deliver the most value not because they are "live", but because they are trusted, stable building blocks that other organisations can safely embed into important processes. In these cases, the policy question shifts from "how fast can users get updates?" to "how confidently can users rely on the information staying consistent over time?"

Where a dataset functions as an authoritative reference point, such as a definitive list, identifier, or verification-style lookup, the biggest source of user cost is uncertainty: changing definitions, inconsistent fields, or updates that arrive without warning. DVLA's approach illustrates why clarity and stability matter: it pairs a clearly defined service with explicit operating rules so that users know what to expect and can build services and workflows that do not break unexpectedly. This is also where managed access becomes relevant as a governance tool. Requiring users to obtain access credentials can be a proportionate way to protect service integrity, ensure fair use, and maintain a clear line of sight over who depends on the service, particularly where demand is high or the service is operationally costly to run.

For DfT, the implication is that any "quietly critical" datasets that underpin multiple downstream uses should be treated as national digital infrastructure: prioritise stability, clear definitions, and predictable update practices before speed. If DfT chooses to introduce charging in any areas, DVLA's model also reinforces a useful principle: cost recovery is most defensible when it is tightly scoped and rules-based (e.g., linked to defined transactions or exceptional usage), rather than applied broadly in ways that could reduce uptake and dilute public value.

Case study 3: DfE Explore Education Statistics as an example of a Statistical API

DfE's Explore Education Statistics (EES) offers a third key takeaway for DfT: for large statistical datasets, the main barrier to value is rarely "access", but the effort users spend finding, understanding, and repeatedly reshaping data after downloading it. The most effective approach is therefore not simply to provide another download route, but to make it easy to pull out the exact slice of data a user needs, in a way they can repeat reliably over time.

In practice, this matters when datasets are large and multi-dimensional, covering multiple geographies, time periods, and measures, where users typically only need a small subset for a specific question (e.g., one region over a ten-year period, or a particular indicator across local areas). EES shows that the highest-value design choice is to support structured filtering and extraction alongside clear contextual information: definitions, coverage, caveats, and update schedules. This shifts the user experience from "download everything and tidy it up" to "select what you need and extract it consistently", which reduces duplication, manual processing, and error risk. It also strengthens transparency and reproducibility: users can rerun the same extraction later and understand what has changed.

For DfT, the implication is that many of its statistical publications would unlock more value if they were treated as a service for analysts and decision-makers, not just a repository of files. That means prioritising features that reduce analyst burden - targeted extraction, strong contextual information, and predictable publishing practices - so that open data translates into more timely insight, easier evaluation, and wider reuse across government, academia, and industry.

Implications for DfT

Across the case studies, the key implication is that DfT should treat APIs as a portfolio of service types, not a single "publish data" decision. The right approach depends on how the data is used and what problem it solves.

- **For fast-changing operational data** (e.g., information captured every minute or multiple times per hour and relied on during disruption), DfT should prioritise reliability and predictability. The policy challenge is to run these releases as dependable public services: managing peaks in demand, setting clear expectations for use, and introducing changes in ways that do not break downstream tools. Light-touch access management (e.g., issuing access credentials) can be appropriate here as a governance tool to protect service performance and build evidence on demand.

- **For authoritative "building block" datasets** (such as identifiers, stable lists, or network components), the value comes from confidence and consistency over time. DfT should treat these assets as digital infrastructure: clear definitions, stable structures, and predictable update practices matter more than speed.

- **For large statistical datasets**, the barrier to value is often the effort required after download. DfT can unlock more reuse by enabling targeted extraction (pulling the subset needed) and pairing it with strong context, including definitions, coverage, caveats, and update schedules, so analysis is easier to repeat and less error-prone.

This portfolio view should directly guide prioritisation and the economic case: DfT would be likely to realise the most value where APIs materially reduces user burden or enables time-critical use, provided they are paired with proportionate access arrangements, predictable change practices, and clear supporting information.

Section 4. Demand Analysis - User Research Findings

This section presents findings from user interviews, structured around key themes and observations drawn from all interviews, alongside a summary of persona-specific findings from five user personas. Detailed insights for each persona are provided in Appendix 3.

4.1 Cross-cutting themes and observations

Interviews highlighted seven common themes across user personas, showing where API-enabled access is most valuable and what conditions stakeholders consider important for it to work.

1. **A clear, long-term mission on seamless multimodal travel in the key.** Interviewees highlighted that interoperability must be designed from the start, rather than added later. While a single multimodal access point was not seen as an immediate need, DfT could play a stronger coordinating role across the transport data ecosystem. This includes acting as a system integrator and standard-setter to reduce fragmentation among DfT's ALBs, such as DVLA, DVSA, and National Highways, improving integration across modes, and addressing data gaps in freight, maritime, and aviation. Progress also depends on broader sector-wide factors, including senior-level technical literacy and clearer alignment around shared outcomes.

2. **Data quality is critical for the effective use of APIs, particularly with high-frequency data.** Interviewees noted that poor-quality high-frequency data can undermine trust, regardless of access methods. Recurring issues include accuracy, completeness, consistency, provenance, and documentation, which are often more pressing than the absence of new access layers. NaPTAN was repeatedly mentioned as an example, where stop locations once accurate 20-30 years ago are now materially misaligned for modern journey planning. Problems such as inconsistent, misspelt, or uncoded terms further complicate reliable integration.

3. **Limited awareness of the data held or funded by DfT emphasises the importance of improving data discoverability.** Many users are unaware of

DfT's open data platforms, such as the Transport Statistics Finder. Even those who know where to look face two main challenges. First, it is difficult to get a full picture of the transport data assets, particularly those held by ALBs or delivered through DfT-funded services. Second, users struggle to locate specific datasets and key information relevant to their needs. In many cases, users must download large files before assessing their suitability.

Interviewees called for a single, authoritative DfT data catalogue detailing what data exists, ownership, coverage, known quality issues, access routes, and update schedules. Improved visibility of existing data would support more meaningful discussions about priorities, helping to ground future data needs in a shared understanding of the current data landscape.

4. **API access should not be treated as a one-size-fits-all solution, instead, it should match the use case.** Large, static datasets for processing, aggregation, and research, are often better delivered via scheduled data drops, as opposed to APIs, due to volume, rate limits, and reliability issues.

   From the user interviews, APIs, however, are well suited for two use cases: The first, is for filtered data retrieval, such as querying limited time ranges, specific fields or subsets. The second is for where higher-frequency or near-real-time access is required, such as powering real-time user interfaces. In these cases, APIs provide efficient access without unnecessary data transfer.

5. **Pricing and licensing models are important determinants of API adoption, particularly for smaller organisations.** SMEs are particularly sensitive to high upfront or blanket fees and prefer charge-per-use models where charges apply, as these allow costs to scale with realised value and encourage experimentation and innovation. Local authorities also discussed the costs of accessing high-value commercial datasets, noting that DfT could help reduce duplication and fragmentation in procurement across the system. Mobile network data was frequently cited as a priority, with coordinated approaches seen as a way to improve value for money.

6. **Interviewees emphasised the importance of clear, outcome-led prioritisation, supported by evidence of what works in practice.** They want DfT to be explicit about the behaviours it aims to support (e.g., safety, modal shift, accessibility, fair markets) and the intended beneficiaries (e.g., local authorities, operators, startups, researchers, citizens). These priorities should then shape which datasets are developed and maintained, and how, with clarity on where standards apply and how responsibilities are shared between DfT, the market, and local authorities. This helps avoid data being made available without clear ownership or purpose.

   Interviewees also suggested using recurring demand signals to monitor and refine priorities over time. Patterns in Freedom of Information (FOI) and Environmental Information Regulations (EIR) requests indicate where the same datasets are repeatedly requested or demand persists despite access barriers. Usage and discovery analytics (including searches, page views, downloads, API calls, access requests, and failed queries) were also highlighted as useful for understanding which assets attract sustained interest, where users

encounter dead ends, and which datasets may appear low-value due to poor visibility rather than lack of demand. Systematically applying these signals would allow DfT to identify latent demand (high interest, low access) from genuinely low-value assets, supporting more iterative and evidence-based prioritisation.

7. **Users expressed strong preference for incremental delivery over highly bespoke solutions.** Interviewees warned against letting the pursuit of perfection delay progress and recommended establishing minimum viable standards that can be iterated based on real-world use of datasets and APIs. Early design decisions should consider the wider ecosystem to ensure that standards and interfaces do not limit future interoperability or scalability. Overly bespoke APIs were seen as counterproductive, as the additional complexity - imposing a burden on transport operators and local authorities responsible for data collection - often outweighs the benefit of extra data points.

   Rather than addressing all modes simultaneously, interviewees also suggested starting with a small number of high-use, foundational datasets, such as BODS, Street Manager, traffic data, the NTS, and STATS19, and expanding deliberately to lower-demand assets over time.

4.2 Summary of user personas

User research shows that users interact with DfT data in different ways. To capture these differences in a way that is useful for data strategy and API design, we developed a set of user personas that summarise users' goals, needs, pain points, and benefits of better API access.

From our user research, we developed five user personas based on:

- The types of users, including direct users (i.e., who use DfT data directly for operational and research purposes), data intermediaries (i.e., who use DfT data as an input their own products or services), and providers of public goods and services (i.e., public sector organisations that use DfT data to inform policy, planning, or public services).
- Their objectives for accessing DfT data, and
- The potential benefits they seek from improved API access.

A summary of the five user personas, and how each aligns with stakeholder groups and user types, is provided in Table 3. It is important to note that a single stakeholder may take on different personas depending on their objectives and how they use DfT data.

Table 3: User persona overview

User Persona	User Type	Relevant Stakeholder Groups
Local Transport Authority Analyst	Provider of public goods and services	Local councils, combined authorities, and regional transport authorities
Transport researcher	Direct user	Researchers, transport authorities

Operational Controller	Direct user	Transport operators, journey planning developers, and Mobility as a Service (MaaS) providers
Data Engineer & Integrator	Data intermediary	Transport consultancies
Developer & Innovator	Data intermediary	Journey planning developers, MaaS providers, ticketing and payment platforms, and transport consultancies

1. Local Transport Authority Analyst (Provider of public goods / services)

This persona uses DfT data alongside local-level datasets to support planning, monitoring, and policy decisions. They tackle practical, often time-sensitive questions for local planning and investment cases, operational monitoring, and reporting to elected members. Typical stakeholders include analysts working in local councils, combined authorities, and regional transport authorities.

2. Transport Researcher (Direct user)

This persona uses DfT data directly to produce analysis and evidence for research and policy. They typically work with national datasets and research-grade formats to understand travel behaviour, model demand, evaluate interventions, and answer strategic or policy questions. Typical stakeholders include researchers and analysts from academic institutions, industry associations, and transport authorities.

3. Operational Controller (Direct user)

This persona maintains real-time situational awareness across modes and must respond quickly to changes on the network. They use DfT's real-time or near-real-time data as a shared reference layer, combined with their own or third-party operational data, to understand the planned network (e.g., timetables, routes, fares) and keep live passenger-facing information consistent across systems. Typical stakeholders include transport operators, journey planning developers, and MaaS providers involved in operational control and service management.

4. Data Engineer & Integrator (Data intermediary)

This persona focuses on cleaning, transforming, and combining DfT data, often alongside other public or private sources, to create analysis-ready datasets and data pipelines. Downstream users, including developers, consultants, and operators, depend on the outputs of this work rather than on raw DfT datasets. Typical stakeholders include technical specialists in transport consultancies and similar service providers.

5. Developer & Innovator (Data intermediary)

This persona uses DfT data to build products, tools and user-facing applications, with a focus on service quality rather than intensive data preparation. Typical

stakeholders include journey planning developers, MaaS providers, ticketing and payment platforms, and transport consultancies.

Together, the five personas summarise how stakeholders engage with DfT data, their recurring pain points, priority needs, expected benefits from improved API access, and early indications of willingness to pay (WTP).

Table 4 provides a high-level comparison across five areas: DfT data usage, key barriers, priority needs, expected API benefits, and early WTP signals. Detailed persona-specific analysis is set out in Appendix 3, with cross-cutting themes differentiated by how they manifest across personas rather than repeated.

Table 4: Summary of persona-specific findings

Persona Features	1. Local Transport Authority Analyst	2. Transport Researcher	3.Operational Controller	4. Data Engineer & Integrator	5. Developer & Innovator
Key DfT data usage	STATS 19 BODS Street Manager Traffic count data	National Travel Survey STATS19 BODS NaPTAN Vehicle licensing statistics	BODS Disruptions and incidents data Street Manager NaPTAN	BODS NaPTAN	BODS NaPTAN Street Manager
Key barriers	Infrastructure and IT constraints make it costly and complex for local teams to ingest, store, and analyse high-volume DfT datasets. Limited in-house technical capabilities constrain automation, API use, and effective reuse of DfT data. Publication lags and unpredictable refresh cycles make some national datasets poorly suited to time-sensitive local planning and monitoring. Weak coordination and feedback loops lead to duplicated effort, misaligned solutions, and slower delivery for local policy and analytical use cases. Data quality, structure, and coverage limitations reduce the usefulness of national datasets for local analysis and decision-making. Fragmented discovery and uneven signposting make it time-consuming to identify, access, and assemble the data needed for local analysis.	Fragmented discovery and signposting across GOV.UK make it difficult for researchers to identify what DfT data exists, where to find it, and which sources are authoritative. Publication delays, limited historical snapshots, and gaps in near-real-time data constrain timely policy evaluation and longitudinal research. Inconsistent data quality and unclear methodology increase uncertainty and add significant overhead to preparing DfT datasets for analysis and reuse. APIs often create friction for research use, particularly around documentation, tooling, and reliability. Gaps in data coverage and licensing constraints limit the scope of research and slow collaboration.	Real-time DfT feeds are insufficiently timely or complete for operational control, reducing confidence in live decision-making. Data quality issues in real-time operational feeds reduce trust in DfT data as a reliable source for live decision-making. Weak and inconsistent standards across operators and modes create significant integration overhead and reduce the usability of operational data. Gaps in dataset coverage and granularity limit end-to-end operational visibility and consistent passenger information across modes. Governance, commercial, and contractual constraints limit consistent operational data sharing and improvement across the ecosystem.	Fragmentation and weak standardisation of datasets limits engineers' ability to write reusable, stable processing rules. Existing data quality is reduced by inconsistent or erroneous fields and incomplete coverage, increasing integration effort and the unreliability of data pipelines. Poor API documentation makes it difficult for data engineers to interpret fields and coverage correctly. Emerging transport data is not being systematically captured or exposed, increasing reliance on third party sources.	Data quality and inconsistent, and competing "sources of truth" were described as the dominant constraint for building reliable products. Limitations in refresh rates, consistency, and real-time usability constrain the delivery of reliable, user-facing experiences. Fragmented discovery and unclear signposting increase onboarding time and integration effort for product teams. Misaligned standards and identifiers increase integration effort and limit seamless multimodal product development. Gaps in dataset coverage limit product scope and the quality of user experiences. Fragmented reporting and bespoke data-sharing processes create delivery friction and slow product development.
Priority needs from DfT	Sustained investment in shared infrastructure and skills to reduce duplication and support local analytical capacity. Timely and predictable data releases aligned to local policy,	A central, inspectable DfT data catalogue to act as a single source of truth for reproducible and longitudinal research across all modes and ALBs. A minimum research metadata standard embedded in both the data catalogue and API	Reliable, nationally consistent real-time data that accurately reflects the live network and can be queried flexibly for operational decision-making. Clear, nationally defined standards are needed to enable interoperability and consistent	Enforce consistent standards and identifiers across suppliers, modes and datasets. Predictable API versioning and change communication. Support machine-to-machine workflows as a baseline expectation by assuming the consumer	Data that meets common national standards for structure, depth, quality, and coverage to support scalable product development. Clear signals on which datasets are priorities, which are intended to become authoritative,

	planning, and funding cycles. APIs and access methods designed for local authority and "average technically interested" users, not only large vendors or advanced technical teams. Centralised analysis to produce structured, derived insights alongside raw data delivery.	documentation to support reproducibility and methodological transparency. APIs and data structures designed around research and analytical workflows, rather than solely for operational or commercial integration.	operational data supply across modes.	of DfT data is a server, not a person.	and where third-party innovation is encouraged. Access to datasets with clear end-user experience value and higher potential for economic impact.
Benefits from improved API access	Enable an integrated, coherent multimodal view of local travel patterns. Strengthen evidence-based policymaking and forecasting. Enable richer, more timely road safety and compliance analysis.	Strengthen evidence-based policymaking and forecasting. An integrated, coherent multimodal view of travel patterns.	Deliver more reliable real-time operational data. Accurate, real-time journey planning and reliable passenger information.	An integrated, coherent multimodal view of travel patterns. More reliable real-time operational data. Accurate, real-time journey planning and reliable passenger information.	Unlock cross-sector innovation, research, and new products. Accurate, real-time journey planning and reliable passenger information.
Early WTP signals	Very low	Very low	Low	Medium	Medium

A Note on Maritime Data

In two maritime stakeholder interviews, DfT's ports and maritime statistics were described as a key input for analysis and reporting. One interviewee characterised the quarterly ports freight statistics as high quality, consistent, and sufficiently granular for their needs, and noted that DfT's accompanying analysis provides useful context. They also described DfT as approachable and responsive to occasional questions.

Barriers were relatively specific, focusing on timeliness, definitions, and usability. One interviewee highlighted that detailed annual data is published with a long lag, and quarterly releases can also be delayed, limiting usefulness for understanding current market trends. In terms of usability, one interviewee described the access route they typically use as time-consuming, requiring users to download a zip file and manually open and save around 30-40 spreadsheets, sometimes needing format conversion, which can be difficult for less specialist users. While other access routes are available, this was the route the interviewee reported using. Two niche issues were raised: an outdated "major ports" definition, and port labelling that combines multiple terminals under a single port entry (e.g., Port of London), which can affect perceived rankings.

Priority needs therefore focused on earlier availability of detailed data and more queryable access. One interviewee suggested a Eurostat-style interface for interrogating specific cuts online, reducing repeated bulk downloads and improving navigation. Regarding APIs, one interviewee indicated that an API-backed query interface could add value if it materially accelerates access to up-to-date data, while another saw no added value from APIs for quarterly data in their current workflow.

On charging, willingness to pay was linked mainly to whether improved access materially accelerates availability (e.g., closer to month-end). One interviewee noted potential resistance in a publicly funded context and sensitivities around confidential port data, while another indicated that charging was not relevant given the quarterly cadence and spreadsheet-based workflow.

Section 5. Prioritisation of DfT Datasets

5.1 Prioritisation framework

DfT's dataset landscape is broad, and not all datasets are equally suitable - or equally valuable - to convert into APIs. Prioritisation is therefore necessary to focus effort on the datasets where an API would deliver the greatest user value and can be delivered and operated reliably.

To support transparent, repeatable decision-making, we propose a simple scoring-based prioritisation framework that ranks candidate datasets for API development using five criteria: demand evidence, timeliness criticality, GDS alignment, dataset size, and data quality. The framework is designed to balance user value (building APIs that solve real problems) with delivery feasibility (focusing effort where DfT can deliver a reliable, maintainable service).

The framework complements the insights from user research by translating them into a structured decision tool. It also supports governance: scores can be refreshed periodically, assumptions can be evidenced, and decisions can be explained consistently across stakeholders. Full scoring definitions, scales, and weighting rationale are set out in Appendix 4.

5.2 Emerging priority datasets and domains

This section summarises the results of the dataset prioritisation scoring framework. Each dataset is assigned a composite score out of 100 using a weighted sum of: Demand evidence (40%), Timeliness criticality (20%), Data quality (20%), and GDS alignment (20%). In this iteration, data size is excluded from the weighted score due to incomplete coverage in the mapping database (see Table A.2)

To support decision-making and avoid false precision, results are presented in priority bands:

- **Band 1: 80+ (inclusive)**
- **Band 2: 70-79**
- **Band 3: 60-69**

- **Band 4: 50-59**
- **Band 5: 40-39**

The full dataset prioritisation register is set out in Appendix 6, and detailed scoring definitions and methodology are provided in Appendix 4.

To interpret the banding, we triangulated results against the explicit interview question on API priorities (Table 5). For this comparison, each dataset was mapped to the interview "dataset type" categories based on its dataset title and scope as recorded in the prioritisation register.

Table 5: Dataset types prioritised for API access by Interviewees

Type of data sets	Times mentioned
Buses and Public Transport Operations (incl. NaPTAN, BODS, operational data)	12
Road Events, Restrictions and Disruption (inc. Street Manager, DRO / TROs, disruption data)	6
Road Traffic Demand (incl. car usage / road traffic, traffic counts, congestion data, journey time data, road geometry/network)	5
Travel Behaviour and Survey Statistics (incl. National Travel Survey, Statistics API)	3
Safety and compliance (incl. Road Safety Data (STATS19))	3
Vehicles licensing and Electric vehicles and charging (incl. Vehicle licensing EV data)	1
Active travel (incl. micro mobility)	1
Fares, ticketing, and payments (incl. Central ticketing / ticketing data)	1
Advanced products (i.e., Digital twin / synthetic data)	1
Cross-cutting / others	1
Road Infrastructure	0
Freight and logistics	0
Accessibility and disability	0
Maritime and ports	0
Aviation	0
Environment and fuels	0

We use these insights to validate our prioritisation which is summarised in Table 6.

Table 6: Summary of Prioritisation Bands

	Band 1	Band 2	Band 3	Band 4	Band 5
Score	80+ (inclusive)	70-79	60-69	50-59	30-49
Selected Datasets [Dataset score] Full lists in Appendix 5	Roadworks service API (Street Manager) [92] (already an API) Bus Open Data - Published bus timetables [80] (already an API) Daily domestic transport use by mode [80] Large Goods Vehicle (LGV) vocational driving tests [80]	Bus Open Data - Published bus locations [76] (already an API) National Coach Dataset [76] Traffic Flows by Borough [76] (already an API) Transport Statistics API [76] (already an API) National public transport access nodes (NaPTAN) (76) (already an API) National Travel Survey: 2024 (72)	Road Safety Data (Collisions) [68] Road Safety Data (Vehicles) [68] Travel time measures for the Strategic Road Network and local 'A' roads [68] Taxi and private hire vehicle statistics, England [68]	Road traffic estimates in Great Britain: 2024 [56] (partially available via API) Disabled parking badges [56] Road Casualties by Severity [56] Casualties involved in reported road accidents (RAS30) [56]	Concessionary travel statistics: year ending March 2024 [48] National Travel Attitudes Study (Waves 1-10) [48] International road freight statistics [48] Linking police and fire road collision data [44] Public attitudes towards transport [44]
Alignment with Interviews	Band 1 has few datasets but aligns with top priorities: 50% cover public transport operations and road disruptions. Live APIs (Street Manager, BODS) validate both value and feasibility.	Band 2 clusters around priorities: 22% PT operations (mostly already APIs) and 39% within the top three areas, with the remainder reflecting demand and safety priorities.	Band 3 mainly covers lower-urgency statistical and safety datasets, with few outliers driven by readiness rather than demand. Adds value but best delivered later rather than in the first API tranche.	Band 4 is largely non-operational statistics, with 41% in non-prioritised areas (accessibility, maritime, fuels). Likely publishable but low API priority without policy or user demand.	Band 5 covers low-frequency reports and studies; 35% sit outside stated priorities. Indicates structurally publishable data but low immediate case for APIs.
Interview Insights	Buses & Public Transport: Core feed but constrained by coverage, refresh, and quality; users need faster updates and richer performance/history data. Road events & disruption: Current data is fragmented and often scraped, with gaps in local closures and inconsistent feeds. Users want standardised, queryable APIs rather than infrastructure-heavy push access.	Road traffic demand: Priority is making existing traffic data queryable and timely, with demand for road-level granularity. Travel behaviour: The National Travel Survey is useful but static; users want API access enabling flexible querying while preserving dataset integrity.	Safety & compliance: STATS19 suited to API access for targeted, record-level queries; sensitive fields likely require tiered access.	-	-
Further Details	Band 1	Band 2	Band 3	Band 4	Band 5

Section 6. Benefit Framework

6.1 Purpose and positioning of the benefit framework

This section sets out the benefit framework used to assess how different user groups derive value from enhanced access to DfT datasets via APIs, and how this value is distributed across use cases.

The framework serves two related purposes:

- to provide a structured and evidence-led account of the main benefits reported by users; and
- to support charging considerations by anticipating where benefits are sufficiently material, repeatable, and scalable to justify paid API access.

Rather than seeking to establish overall value-for-money, the framework focuses on identifying where improved access generates sustained operational, commercial, or user-facing value that may support cost recovery or pricing in specific contexts.

The analysis builds on the prioritised benefits identified through user research and translates them into comparable, transparent proxy measures and valuation mechanisms.

6.2 Overview of priority benefits

The five priority benefits reflect common value drivers across multiple user types.

Table 7 explains how each priority benefit creates value and their alignment with DfT's data action plan.

Table 7: Priority benefits and value mechanisms

Benefit	Core value mechanism	Data Action Plan Alignment
Faster time-to-insight	Reduced manual processing → increased analytical capacity	Embracing technology & AI
Automation & integration	Lower marginal cost of reuse → scalable delivery	Embracing technology & AI
Timeliness & responsiveness	Faster response → improved service continuity	Informing individual choices
Cost savings / reduced dependency	Avoided procurement and duplication	Better transport system
Consumer-facing UX	Reduced uncertainty → higher user trust and uptake	Better transport system

Faster time-to-insight

Many users reported that a significant proportion of their analytical effort is currently absorbed by downloading, cleaning, reconciling, and reformatting DfT datasets before any substantive analysis can begin. This manual overhead is repeated across teams, organisations, and time periods, particularly where datasets are refreshed regularly or used for recurring monitoring and reporting. Improved API access reduces this friction by enabling more direct, structured, and repeatable access to data, allowing analysts to focus more consistently on interpretation, insight generation, and decision support rather than data preparation.

Automation and system integration

For more technically mature users, the value of improved access lies less in one-off analysis and more in the ability to integrate DfT data into automated pipelines, dashboards, and operational systems. Interviewees highlighted that current access routes often prevent stable automation, creating reliance on manual intervention or single points of failure within teams. API-based access enables programmatic ingestion and reuse, reducing repeated manual processing and allowing systems to scale with frequency of use rather than analyst capacity. This benefit becomes increasingly material where datasets are high-frequency or embedded in ongoing operational workflows.

Improved timeliness and responsiveness

Several users emphasised that the usefulness of DfT data is highly sensitive to how current it is. Publication lags, unpredictable refresh cycles, and manual update processes limit the ability to use data for time-sensitive monitoring, operational control, or rapid analysis. Improved access via APIs can reduce delays between data generation and use, supporting more timely decisions and responses. This benefit is particularly relevant for operational and near-real-time use cases, where even modest improvements in freshness can materially improve service continuity and effectiveness.

Cost savings and reduced dependency

Across both public and private sector users, improved access was linked to reduced reliance on external data sources, bespoke processing solutions, and consultancy support. Interviewees described situations where organisations duplicate effort locally to clean or reconstruct datasets or procure alternative sources to compensate for access limitations. More reliable, well-structured API access can reduce these costs by lowering the marginal effort required to reuse DfT data and by providing a clearer, more authoritative source of information. The resulting savings are typically realised through avoided spend and reduced duplication rather than through headline efficiency gains.

Better consumer-facing services and user experience

For users building passenger-facing products and services, improved access to DfT data supports more consistent, reliable, and timely information for end users. Interviewees highlighted that fragmented access, inconsistent definitions, and variable data quality currently create uncertainty for passengers, who often cross-check multiple sources or experience conflicting information. Enhanced API access can improve the completeness and consistency of journey information, supporting clearer communication and greater user trust. Within this benefit, a key mechanism is the reduction of uncertainty experienced by end users, which improves perceived reliability and overall service experience.

6.3 Analytical approach

For each benefit, we set out:

- how the benefit arises in practice;
- the user personas most affected;
- the mechanism through which value is created;
- the proxy used for quantification;
- the indicative calculation; and
- the main assumptions and boundary conditions.

This approach enables consistent comparison across datasets and user groups, while remaining proportionate to available evidence.

Table 8: Overview of Benefit Framework

Benefit characteristics	Faster time-to-insight	Automation & system integration	Improved timeliness & responsiveness	Cost savings & reduced dependency	Better consumer-facing services & user experience
What this benefit enables in practice	Analysts spend less time downloading, cleaning, and reconciling datasets and more time interpreting results and informing decisions.	Data can be ingested programmatically into pipelines, dashboards, and systems, reducing repeated manual handling and single-analyst bottlenecks.	Users can act on fresher data, enabling quicker operational responses and more up-to-date analysis.	Organisations rely less on bespoke processing, proprietary datasets, and external consultants to access transport data.	Passengers and end-users receive more reliable, consistent, and timely information, improving confidence and trust in transport services.
How it aligns with DfT's Data Action Plan	Embracing technology & AI Programmatic access reduces analyst overhead and enables reuseable tooling	Embracing technology & AI Supports system-to-system workflows, continuous models and operational automation.	Informing individual choices Fresher data improves passenger information and near-real-time decision	Creating a better transport system Lowers duplication, procurement spend and supports efficient service planning.	Creating a better transport system More consistent passenger information and trust in services; also feeds broader system performance.
Primary beneficiaries (personas)	Local Transport Authority Analyst; Transport Researcher; Data Engineer & Integrator	Data Engineer & Integrator; Developer & Innovator; Operational Controller	Operational Controller; Local Transport Authority Analyst; Developer & Innovator	Local Transport Authority Analyst; Data Engineer & Integrator; Developer & Innovator	Developer & Innovator; Operational Controller; Local Transport Authority Analyst
Typical use cases	Policy analysis, evaluation, reporting, ad-hoc analysis, repeated monitoring tasks.	Automated reporting, live dashboards, system-to-system data feeds, analytics pipelines.	Incident response, service monitoring, operational planning, near-real-time analytics.	Replacing paid data sources, reducing consultancy spend, avoiding duplicated local processing.	Journey planning, passenger information, real-time service updates, customer-facing products.
How API access changes scale or reliability	Reduces repeated preparation effort across teams and time periods, making routine analysis more repeatable and scalable.	Enables stable, reusable integrations that scale with frequency of use rather than analyst capacity.	Reduces publication lag and manual refresh delays, increasing the operational usefulness of the data.	Lowers marginal cost of reuse and reduces lock-in to alternative data sources or bespoke solutions.	Improves consistency, completeness, and timeliness of information, reducing the need for users to cross-check multiple sources.
Nature of value created	Analytical and operational efficiency gains.	Operational efficiency and scalability.	Operational effectiveness and service continuity.	Direct cost savings and avoided expenditure.	User experience improvements and reduced uncertainty for end-users.
How this benefit is quantified (high-level)	Estimated using analyst time saved on data preparation tasks, multiplied by the number of analysts and frequency of use.	Estimated through the volume of manual tasks replaced and the frequency of automated runs, translating reduced effort into avoided staff time.	Estimated using reductions in data latency and the value of faster decisions for operational or analytical users.	Estimated through avoided spend on external data, consultancy, or duplicated internal processing.	Estimated using reductions in passenger waiting time or uncertainty per trip, valued using standard value-of-time assumptions.
Key assumptions and limitations	Assumes time saved is redeployed productively; integration overhead is modest.	Assumes organisations are able to implement and maintain automation; benefits increase with repeat use.	Assumes timeliness is operationally binding and users can act on fresher data.	Assumes displaced spend is genuinely avoidable rather than partially substitutive.	Assumes improved information is trusted and used by passengers; uncertainty reduction is the primary defensible pathway.

6.4 Uncertainty reduction as a core value mechanism

Among the five benefits identified, “Better consumer-facing services and user experience” captures improvements in reliability, clarity, and trust for passengers. Within this benefit, one mechanism stands out: the reduction of uncertainty.

Uncertainty reduction is not only a recurring theme in user interviews. It is also the most extensively examined mechanism in the academic literature on the value of data. For this reason, we examine it separately below. This section does not introduce a new benefit. Rather, it deepens the analysis of a core mechanism that underpins improvements in consumer-facing services.

Table 9: Overview of key references

Author(s) / Source	Context	Core Concept	Relevance to Uncertainty Reduction	Implication for API-enabled Data
The Economics of Urban Transportation Small K.A. & Verhoef, E.T. (2007)	Transport economics (theoretical foundation)	Value of reliability and variability in travel time	Establishes that variability (uncertainty) in travel time imposes measurable welfare costs beyond mean travel time	Reducing uncertainty in travel conditions (via better information access) generates welfare gains
The valuation of reliability for personal travel Bates, J., Polak, J., Jones, P. & Cook, A. (2001)	Empirical modelling of travel time reliability	Reliability as distinct component of generalised cost	Demonstrates that travellers value reductions in uncertainty separately from average time savings	Information that reduces uncertainty has monetisable value
The value of reliability Fosgerau, M. & Karlström, A. (2010)	Behavioural modelling of route choice	Scheduling model under uncertain travel times	Quantifies economic value of reduced variance in arrival time	Supports modelling benefits through reduced variance enabled by better data
Dynamic network models and driver information systems Ben-Akiva, M., de Palma, A. & Kaysi, I. (1991)	Route choice under information conditions	Value of traffic information systems	Shows that information reduces uncertainty and changes user behaviour, generating measurable welfare gains	APIs that improve timeliness and consistency increase effective information value
Traveler response to information Chorus, C., Arentze, T. & Timmermans, H. (2006)	Behavioural transport research	Information precision and decision-making	Demonstrates that improved information quality reduces uncertainty and improves route choice outcomes	Higher-quality API dissemination amplifies decision value of data
Howard, R. (1966); Raiffa, H. (1968)	Decision theory (foundational)	Expected Value of Information (EVI)	Formalises economic value of uncertainty reduction in decision-making	Provides theoretical basis for Value of Information modelling of API-enabled data

In economics and decision theory, the value of data is typically formalised as the value of information: the expected improvement in outcomes that arises when uncertainty is reduced. This concept is embedded in established frameworks such as Expected Value of Information (EVI) and Bayesian decision theory, and is widely applied in health economics, environmental policy, transport appraisal, and risk analysis. Across these domains, data is valuable to the extent that it reduces uncertainty about future states of the world and improves decision quality. Table 9 provides an overview of the academic literature linking transport decision-making and uncertainty.

Unlike time savings or cost avoidance, which are context-specific and often proxy-based, uncertainty reduction has a well-developed theoretical and empirical foundation. It provides the most rigorous conceptual explanation for **why data generates economic value.**

The relationship between data value and APIs

This report concerns the value of disseminating DfT datasets via APIs. APIs do not generate new data; they reduce the friction associated with accessing and using existing data. Friction arises where users must manually download, reconcile, clean, reformat, or cross-check datasets before they can inform decisions.

APIs reduce this friction by enabling:

- Structured and repeatable access,
- Timelier data retrieval,
- Programmatic integration into systems,
- Reduced reconciliation and duplication effort.

The economic implication is as follows: APIs derive their value from the value of the underlying data. If the data has no decision-relevant value, reducing friction of access will not create value. Conversely, where data materially reduces uncertainty in decision-making, lowering the barriers to accessing and integrating that data amplifies its realised value.

In this sense, APIs are an enabling mechanism. Their contribution is to increase the effective availability and usability of information that has intrinsic decision value.

Implications for monetisation and modelling

Because the most theoretically robust foundation for data value lies in uncertainty reduction, the most rigorous approach to monetisation is likely to involve a **Value of Information (VoI)** framework.

A VoI approach would:

- Define the decision context (e.g. operational control, scheme appraisal, passenger choice);
- Characterise the level of uncertainty under current access arrangements;
- Estimate how improved API-enabled access reduces that uncertainty; and
- Quantify the expected improvement in outcomes resulting from better-informed decisions.

Under this framework, APIs are treated as an intervention that increases effective information availability by lowering access friction. The resulting benefit is measured not only in time saved, but in improved decision quality, reduced variance, and enhanced welfare.

Accordingly, while multiple benefit channels are identified in this report, uncertainty reduction provides the strongest academic foundation for robust economic estimation. Exploring a Value of Information model in a future phase would therefore provide the most defensible basis for monetising the long-term value of data dissemination.

However, developing such a model would require dedicated analytical resource, including structured modelling work and sensitivity analysis to establish credible probability distribution. It would also necessitate sustained stakeholder engagement to define decision contexts, validate assumptions, and secure agreement on how uncertainty and outcomes should be parameterised.

Section 7. Hypothetical Charging Model

7.1 Purpose and positioning

This section sets out an indicative framework for marginal cost recovery associated with API-based access to selected DfT datasets.

The purpose of this section is not to propose a tariff, nor to recommend the introduction of charging. Rather, it sets out the conditions under which charging could be considered proportionate and defensible, drawing on:

- Benchmark evidence from comparable UK public-sector data services;
- User research findings, including early willingness-to-pay signals; and
- The benefit framework developed in Section 6.

The analysis assumes that baseline public-value access remains open and low-friction. Any charging would be narrow in scope and activated only where sustained or high-intensity use generates measurable marginal costs for DfT (e.g., infrastructure scaling, monitoring, or enhanced service support).

Charging is therefore positioned as a sustainability mechanism, not a revenue-generation tool.

7.2 Principles underpinning marginal cost recovery

The hypothetical framework is guided by five principles.

Proportionality to technical load

Charging should only apply where usage materially increases marginal infrastructure, governance, or service costs.

Protection of baseline public value

Access for public-interest, research, and low-volume users should remain free or low-cost. Charging should not restrict legitimate analytical or civic uses.

Narrow and rules-based triggers

Any charging mechanism should be clearly linked to defined usage thresholds or service characteristics, rather than broad user categories.

Transparency and predictability

Usage thresholds, enforcement mechanisms, and access pathways should be explicit and stable.

Reinvestment in data quality and governance

Any revenues generated should be transparently reinvested in improving data completeness, metadata, documentation, and reliability, reinforcing the conditions that underpin user benefits.

These principles are consistent with benchmark practice across UK public-sector APIs, where free access is the default.

7.3 Evidence from benchmarking

The benchmarking exercise identified recurring charging and control archetypes across public-sector comparators (e.g., Met Office, Companies House, OS Data Hub, DVLA, Darwin Data Feeds):

1. Free baseline access with published rate limits;
2. API key registration to enable monitoring and fair use;
3. Freemium models with metered usage beyond free quotas;
4. Per-transaction charging for defined, restricted services; and
5. High-volume usage thresholds triggering invoicing or bespoke agreements.

A consistent pattern across benchmarks is that charging is not applied to general access to public-value datasets. Instead, fees are activated where:

- Usage is sustained and high-volume;
- Services are embedded in commercial or operational systems at scale; or

- Premium service characteristics (e.g. SLAs, guaranteed uptime, advanced filtering, or dedicated support) are required.

This suggests that, if DfT were to consider marginal cost recovery, the most defensible model would be a **free baseline + defined high-intensity trigger structure**, rather than a blanket subscription model.

7.4 Alignment with this report's API taxonomy

Our supply analysis (Section 3) distinguished between three API product types:

- Operational APIs (real-time, high-frequency, disruption-sensitive)
- Reference APIs (authoritative identifiers and stable lookups)
- Statistical APIs (large, multi-dimensional datasets requiring queryability)

Marginal cost exposure differs across these categories.

Operational APIs

Operational APIs supporting real-time or near-real-time use cases are most likely to generate infrastructure scaling costs under sustained high-frequency usage. Marginal cost recovery, if required, would most plausibly be triggered by:

- Sustained high call volumes;
- High-frequency polling beyond published limits; or
- Requests for enhanced service-level guarantees.
- Baseline access for low- to moderate-frequency use would remain free.

Reference APIs

Reference services are typically lower-frequency but widely embedded. Marginal cost recovery in this context would be defensible only where:

- Very high-volume automated lookups occur; or
- Defined transaction-based services impose measurable processing or governance costs.
- This aligns with benchmark examples where per-transaction charging applies to specific restricted services rather than general reference access.

Statistical APIs

For large statistical datasets, the main cost drivers are likely to relate to:

- High-volume, repeated extraction of large slices of data;
- Advanced filtering or compute-intensive queries; or
- Enhanced support or service features beyond baseline.

In many cases, the marginal cost of typical analytical use may be modest. Charging would therefore require careful evidence of sustained infrastructure load before activation.

7.5 Illustrative charging triggers

Given current information constraints, detailed tiers and price thresholds cannot be specified at this stage. We do not currently have systematic, consistent access to:

- Observed usage volumes at scale;
- Robust marginal infrastructure cost estimates; or
- Evidence on behavioural responses to different threshold levels.

Accordingly, this section identifies categories of potential charging triggers, rather than defined tariffs:

- Sustained high-volume API calls beyond published rate limits;
- Persistent high-frequency polling of real-time feeds;
- Premium service characteristics (e.g., SLAs, priority support, version guarantees);
- Dedicated or bespoke integration arrangements;

These triggers would need to be linked explicitly to measurable incremental costs before charging is introduced.

7.6 Interaction with the benefit framework

Section 6 identifies five priority benefit mechanisms, including:

- Faster time-to-insight;
- Automation and system integration;
- Improved timeliness and responsiveness;

- Cost savings and reduced dependency; and
- Improved consumer-facing services and uncertainty reduction.

Importantly, the charging framework does not seek to capture the full value of these benefits. Instead, it recognises that some benefit pathways - particularly automation, embedded integration, and high-frequency operational use - are associated with sustained, repeatable API usage that may increase technical load.

In these contexts, marginal cost recovery may be proportionate because:

- Integration is ongoing rather than episodic;
- API calls are automated and frequent; and
- Service reliability and monitoring requirements increase.

Charging would therefore be aligned with usage intensity and infrastructure impact, not with the scale of downstream economic value realised by users.

7.7 Indicative application of the charging model

To illustrate how the hypothetical charging model could be applied in practice, we present an indicative application to datasets within Band 1.

As detailed usage metadata was not available for this analysis, the application models potential usage intensity scenarios rather than defining specific charging thresholds.

Table 10 sets out this overview and the associated rationale.

	Street Manager Roadworks API	Traffic Regulation Orders / Road Restrictions	BODS – Published Bus Timetables	BODS – Vehicle Locations	NaPTAN	Traffic Flows by Borough	Road Traffic Statistics (TRA)
API Product Type	Operational	Operational	Operational	Operational	Reference	Statistical	Statistical
Value Concentration	Diffuse	Diffuse	Mixed	Mixed	Mixed	Diffuse	Diffuse
Indicative WTP Signal	Low	Low	Low–Medium	Low–Medium	Medium	Low	Low
Proposed Charging Model	Free baseline + rate limits; charge only for exceptional usage	Free baseline access	Free baseline; tiered usage for high-volume commercial reuse	Free baseline; tiered usage for high-volume commercial reuse	Free baseline; enterprise licence for SLA guarantees	Free, open API	Free, open API
Rationale for	System-critical operational dependency; charging only	Public coordination function with limited	Strong public value but identifiable	Operational dependency with commercial	Acts as national digital infrastructure; charging only	Limited private surplus; high	Benefits accrue primarily to public

Charging Model	defensible for exceptional or abusive usage	appropriable private surplus	commercial reuse at scale	downstream products	justified for enhanced guarantees	local public value	planning and analysis
Suggested pricing basis	Free baseline + paid tier	Free baseline + paid tier	Free baseline + paid tier	Free baseline + paid tier	Free baseline + paid tier	£0	£0
Suggested charge metric	Annual high-volume tier	Annual high-volume tier	Annual high-volume tier	Annual high-volume tier	Annual high-volume tier	N/A	N/A
Suggested charge amount	1500	1500	1500	1500	1500	0	0
Suggested charge unit	£/year (ex VAT) indicative	£/year (ex VAT) indicative	£/year (ex VAT) indicative	£/year (ex VAT) indicative	£/year (ex VAT) indicative	Free (OGL)	Free (OGL)
Benchmark triangulation anchor	Met Office DataPoint (~£1,500/year paid plan) + TfL quota-tiering pattern	Met Office DataPoint (~£1,500/year paid plan) + TfL quota-tiering pattern	Met Office DataPoint (~£1,500/year paid plan) + TfL quota-tiering pattern	Met Office DataPoint (~£1,500/year paid plan) + TfL quota-tiering pattern	Met Office DataPoint (~£1,500/year paid plan) + TfL quota-tiering pattern	Benchmark norm: free baseline for public-value/statistical APIs; charging only narrow/exceptional	Benchmark norm: free baseline for public-value/statistical APIs; charging only narrow/exceptional
Charge justification (triangulated)	Primary benefits include automation/timeliness (B7/B10) and UX (B16). Paid tier aimed only at high-volume commercial reuse to cover cost-to-serve; aligns with benchmarked fair-use + tier uplift models.	Primary benefits include automation/timeliness (B7/B10) and UX (B16). Paid tier aimed only at high-volume commercial reuse to cover cost-to-serve; aligns with benchmarked fair-use + tier uplift models.	Primary benefits include automation/timeliness (B7/B10) and UX (B16). Paid tier aimed only at high-volume commercial reuse to cover cost-to-serve; aligns with benchmarked fair-use + tier uplift models.	Primary benefits include automation/timeliness (B7/B10) and UX (B16). Paid tier aimed only at high-volume commercial reuse to cover cost-to-serve; aligns with benchmarked fair-use + tier uplift models.	Primary benefits include automation/timeliness (B7/B10) and UX (B16). Paid tier aimed only at high-volume commercial reuse to cover cost-to-serve; aligns with benchmarked fair-use + tier uplift models.	Benefits are largely diffuse public-good (B6/B15). Charging would suppress reuse; align with benchmark norm of free baseline access.	Benefits are largely diffuse public-good (B6/B15). Charging would suppress reuse; align with benchmark norm of free baseline access.

7.8 Empirical case study

To illustrate how usage metadata can inform the hypothetical charging model, we examine the **Road Traffic data microsite** as a worked example as we have metadata for this dataset.

Observed usage profile

Usage data from the Road Traffic microsite indicates:

- Approximately **470,000 views** from Feb 2025 – Feb 2026
- Approximately **60,000 active users** from Feb 2025 – Feb 2026
- An average of **7.85 views per active user** from Feb 2025 – Feb 2026

- A long tail of page paths (over 33,000 unique paths), with usage dispersed across many local authority and manual count point pages

Two characteristics are notable:

1. **Broad user base**: high number of unique users.

2. **Low average intensity**: relatively small number of views per user.

This suggests the dataset is widely accessed but not heavily concentrated among a small number of repeat users.

Intensity and concentration analysis

The hypothetical charging model is predicated on charging only where usage intensity materially increases marginal infrastructure costs. In this case:

- Average user engagement is low.

- Usage appears distributed across many local authority and count-point pages.

- There is no clear evidence of extreme concentration among a small cohort of high-frequency users.

Absent a heavy-tail intensity pattern (e.g., top 5-10% users accounting for the majority of traffic), the empirical basis for intensity-triggered charging is weak.

In other words, Road Traffic data appears to exhibit **broad, light-touch access** rather than automated, high-volume extraction behaviour.

Economic interpretation

APIs generate value by reducing friction of access. However, reduced friction does not automatically imply that charging is appropriate.

For charging to be justified under a marginal cost recovery principle:

- High-intensity usage must materially increase infrastructure or monitoring costs; and

- That usage must be concentrated among a limited segment of users.

The Road Traffic usage profile suggests:

- Value is diffuse across many users.

- Access is likely analytical or exploratory in nature.

- Infrastructure burden per user is low.

Under these conditions, introducing a usage-based charging threshold would:

- Generate limited revenue;
- Risk discouraging beneficial use; and
- Potentially impose administrative costs disproportionate to recovered revenue.

Conclusion

In the case of Road Traffic data, the empirical usage profile does not support the introduction of intensity-based charging. Accordingly, the dataset has been provisionally classified as non-chargeable under the hypothetical model, pending future API-specific usage evidence.

Appendices

Appendix 1. Methodology

1.1 DfT data mapping

The data mapping workstream establishes a baseline by cataloguing datasets available via Find Transport Data or Transport Statistics Finder. The mapping links each dataset to its full time series where possible and records formats, update cycles, metadata quality, and API availability.

This creates a clear view of the data supply landscape: identifying priority datasets for user research, exposing gaps and inconsistencies, and providing the evidence needed to match user needs to specific datasets in Phase 2. It also enabled a structured assessment of API suitability in Phase 3 and supported Phase 4 by grouping datasets into categories aligned to appropriate benefit-valuation approaches.

Approach

We created a comprehensive inventory by examining every dataset listed on the Find Transport Data and Transport Statistics Finder portals, recognising that some entries are derivatives of a core dataset.

Where a dataset appeared on both portals, it was logged once. For datasets listed as a single-year release (e.g., “Dataset X - 2022”), we traced the landing page for the full multi-year series and used that as the canonical entry so the time-series integrity was preserved. Datasets were identified using four methods:

- browsing by topic and organisation,
- keyword searches within each portal,
- reviewing dataset landing pages, and

- extracting information from dataset metadata.

Inclusion criteria

A dataset was included if it:

- is held by DfT,
- appears on either of the two selected portals,
- contains structured transport-relevant data (CSV, ODS, spreadsheets, JSON feeds, APIs, ZIPs), and
- is publicly available and accessible through those portals.

Exclusion criteria

Datasets clearly owned or maintained by DfT arm's-length bodies (e.g., DVLA, National Highways) were excluded. If a dataset was presented through multiple outlets (e.g., a report plus a spreadsheet), all related assets were linked within a single dataset row.

Limitations

The mapping only reflects datasets made publicly available via the two portals, so internal or unpublished datasets were not captured. Metadata quality varies across datasets: gaps in our database reflect cases where information was unavailable or inconsistent and requires later clarification from DfT.

1.2 API benchmarking review

We adopted a case study approach to benchmark DfT's data APIs against other established public-sector APIs, to identify transferable design and operating patterns (e.g., access and documentation practices, licensing, and sustainability controls) that DfT could adapt.

Identification and sampling

We first compiled the API longlist from api.gov.uk's API Catalogue and transport-relevant entries identified via Find Transport Data. All APIs identified were included in the benchmarking dataset, without filtering for perceived quality or fit.

To draw practical and transferable lessons for DfT, we then prioritised non-DfT transport APIs. We also included a cross-sector comparator (i.e., the Department for Education's Explore Education Statistics API) to compare whether charging and

delivery patterns differed between operational, real-time transport APIs and statistical API models.

Data extraction and mapping process

For each API, we extracted information from official documentation (including developer portals and guidance pages) and mapped it against a consistent set of descriptors:

- access model and openness,
- data type (operational, reference, statistical),
- response format and standardisation signals (e.g., JSON-first approaches),
- licensing and terms of use, and observable sustainability features (e.g., rate limits, authentication requirements, monitoring mechanisms, and any charging policy where stated).

Where information could not be verified from documentation, fields were left blank rather than inferred. As a result, gaps in the benchmarking sheet reflect either (i) inconsistent disclosure across providers, or (ii) cases where documentation links were no longer functional at the time of review.

Analytical approach: case studies plus cross-cutting synthesis

Analysis was undertaken in two stages. First, we developed case studies for a small number of representative APIs (e.g., Driver and Vehicle Licensing Agency, Transport for London, and Department for Education) to illustrate how different public sector bodies implement openness, standardisation, and sustainability and how this varies by API purpose and data type. Second, we synthesised cross-cutting findings across the full mapped set to identify common patterns and their implications for DfT's prioritisation and design choices.

Limitations

This benchmarking is bounded in three key ways. First, the longlist included only APIs discoverable via api.gov.uk and Find Transport Data and is not an exhaustive census. Second, the analysis relied on publicly available documentation, limiting comparability where information was incomplete or outdated. Third, APIs vary widely in maturity and capability, so findings indicated broad patterns rather than definitive performance assessments. More detailed comparisons, such as uptime, latency, user volumes, or operating costs, would require direct provider engagement and/or technical testing.

1.3 User research

Our user research approach comprised stakeholder mapping and sampling, user recruitment and interviews, and analysis.

Stakeholder identification and recruitment

We collaborated with the DfT to develop a stakeholder longlist using multiple sources, including web searches, DfT contacts, PUBLIC's network. A purposive sampling strategy was employed to ensure representation across organisation types, sizes, transport modes, geographies, and levels of digital maturity. Relevant organisation types included:

- DfT's ALBs
- Local councils, combined authorities and regional bodies
- Regional transport authorities
- Other government departments
- Transport operators
- Journey planning developers
- Ticketing and payment platforms
- Mobility as a Service (MaaS) providers
- Transport consultancies
- Academic institutions and researchers
- Others (e.g., charity, industry associations)

To broaden engagement, we also used snowball sampling and a DfT-approved public call for participation, allowing organisations to self-select and book interviews.

In total, 159 invitations for interview were issued, 62 responses were received, and input was collected from 54 organisations during the three-week research window between 24 November and 11 December 2025.

A summary of stakeholder representation with a full list of participating organisations is provided in Table A.1 below.

Table A.1: Stakeholder engagement

Sector	Category	Target sample size	Number orgs engaged	Organisations interviewed*
Public Sector	DfT's ALBs	2	1	Active Travel England
	Local councils, combined authorities and regional bodies	5	8	Middlesbrough Council Birmingham City Council Gateshead Council Hertfordshire County Council Barnet Council Teignbridge District Council Greater London Authority (Data for London Library) Local Government Association
	Regional transport authorities	3	5	England's Economic Heartland Transport for Greater Manchester Transport for the South East Midlands Connect Solent Transport
	Other government departments	2	2	Ministry of Housing, Communities and Local Government Department for Science, Innovation and Technology / Government Digital Service
Public / Private Sector	Transport operators	3	3	Greater Anglia First Bus East Midlands Railway
Private Sector	Journey planning developers	5	3	QRoutes Google Maps Transit App
	Ticketing and payment platforms	5	2	Passenger Fairtiq
	Mobility as a Service (MaaS) providers	5	4	Co-Wheels Beryl Lime Via Transportation
	Transport consultancies	5	18	Vivacity NayaOne City Swift One Auto API INRIX KL Systems Arcadis Mobility Ways Transport API Valerann Data Wharf RASIC Consultant (SWR + former Arup) Port Centric Logistics Partners Journeo

				Alchera Technologies Agilysis Vianova
Third Sector and Associations	Academic institutions and researchers	3	5	University of Glasgow University of Bristol University of Leeds University of Exeter University College London
	Others (e.g., charity, industry associations)	2	3	RAC Foundation (two teams) Rail Delivery Group British Ports Association
Total	54			

*One organisation provided written responses instead of an online interview due to availability constraints during the user research window. Another organisation participated in two separate interviews with different teams. In total, 54 interviews were conducted, along with one written response, across 54 organisations.

Interview design and execution

An Interview Guide with sector-specific scripts (public, private, and third sector) was developed and shared with DfT for validation. Interviews explored:

- Current use of DfT and other transport data.
- How organisations access and process data.
- The potential value of API-based access to DfT data.
- Barriers to API use, especially for organisations constrained by legacy systems, limited digital skills, or restricted resources, and the types of support required from DfT.
- Perceived economic benefits and early indications of willingness to pay.
- Suggested prioritisation of DfT data for API development.

All participants received a Participant Information Sheet before the interview, outlining the research purpose and how their data would be handled. All interviews were transcribed for notetaking and analysis purposes only. For one organisation unable to attend, interview questions were shared by email for written responses.

Analysis approach

Each interview transcript was coded in a structured way using the following fields:

- Digital/API maturity

- Current use of DfT datasets
- Relevant dataset categories and sub-categories
- Current access routes
- Preferred data/API characteristics
- Perceived benefits and value themes
- Barriers and API readiness
- Willingness-to-pay stance and rationale
- Priority use cases and datasets
- Support needs

Coded insights were then used for thematic analysis using affinity mapping to identify recurring patterns across users. This formed the basis for user personas that captured users' barriers, needs, and behaviours.

Limitations

Our sampling strategy achieved broad representation across stakeholder groups, organisation sizes, and regions. However, several gaps should be considered when interpreting findings.

There was limited participation from maritime stakeholders (two organisations) and no aviation stakeholders took part despite invitations. As a result, insights relating to these modes should be interpreted with caution. Participation from journey planners, MaaS providers, and ticketing/payment platforms was also lower than expected, as many access DfT data indirectly via intermediaries such as transport consultancies. The research is based on self-reported behaviours, which may introduce bias. Participants may overstate usage, understate willingness to pay, or describe workflows differently from actual practice.

Overall, the research provided strong insight into user needs, pain points, and the potential benefits of improved access to DfT data, though these limitations should be kept in mind when interpreting the results.

1.4 Benefit framework

To develop the benefit framework, we translated qualitative interview evidence into a structured and auditable assessment of the value users associate with improved API access to DfT datasets. The approach followed four stages:

Structured coding of interview evidence

Interview transcripts and notes were systematically coded in an analysis spreadsheet, capturing datasets referenced, user workflows, pain points, and associated benefit themes during the user interview analysis.

Consolidation into a benefit taxonomy

Individual benefit statements were grouped into a single benefit taxonomy (20 categories). This ensured consistent classification of similar concepts (e.g., reduced manual handling, improved reliability), while retaining traceability to the underlying qualitative evidence. This taxonomy is available in full in Appendix 7.

Prioritisation through frequency analysis

We established salience by counting the number of interviews in which each benefit was mentioned. This provided a transparent basis for narrowing the framework to the five most frequently cited benefits, representing the strongest cross-cutting demand signals across user types.

Quantification and monetisation

For these five prioritised benefits, we developed benefit models to translate qualitative value statements into quantifiable metrics and, where feasible, monetised estimates. The modelling approach and assumptions are set out in the Benefit Framework section.

Limitations

The benefits have been identified using insights from user interviews to ensure the framework remains grounded in lived user experience. While this strengthens its practical relevance, it also limits the framework to the scope and composition of our sample, which, although broadly representative, remains subject to sampling limitations and potential bias.

1.5 Hypothetical Charging Model

Charging models benchmarking

A focused case-study sample was benchmarked to understand how UK public bodies manage access, charging and/or usage controls for data services, and how “charging triggers” operate in practice. The purpose was not to catalogue every model, but to identify transferable patterns relevant to high-volume, frequent usage.

Identification and sampling

Eight benchmark cases were identified:

- Met Office DataPoint (Met Office)
- HM Land Registry (HM Land Registry)
- SAIL Databank (Swansea University, Wales)
- TfL Unified API (Transport for London)
- OS Data Hub (Ordnance Survey)
- Companies House (Companies House)
- Access to Driver Data (ADD) API (DVLA)
- Darwin Data Feeds (National Rail)

The sample was compiled from three inputs:

- **Earlier benchmarking work**, focusing on APIs with documented non-free tiers or explicit access controls.
- **Interview recommendations**, which highlighted SAIL, Met Office, and HM Land Registry as useful comparators.
- **Desk research**, identifying additional public-sector APIs with published charging or quota-based models.

The shortlist was selected to cover a range of approaches:

- free access within published limits (Met Office; Companies House hard rate limits),
- quota-based access via keys (TfL),
- freemium metered usage (OS),
- licence-based re-use rights (HM Land Registry),
- per-transaction charging for restricted data access (DVLA),
- project-based cost recovery (SAIL), and a documented high-volume threshold with caps and invoicing (Darwin).

Data extraction and mapping

For each benchmark, information was extracted from official sources such as developer portals, terms and conditions, charging policies, and any published high-volume usage guidance. Each case was mapped into a standardised spreadsheet covering:

- Example / organisation
- Model archetype
- Charging trigger
- What is monetised or controlled
- Free access offer and published free limits
- Whether a paid path exists and the basis for charging
- Paid limits / thresholds / triggers
- Overage and enforcement mechanisms
- Access pathway to higher tiers
- Relevance

Analytical approach

The completed mapping was used to compare benchmarks and group them into recurring charging/control patterns. This synthesis enabled comparison of the triggers used to manage heavy usage and recover costs.

Limitations

The sample is purposive rather than exhaustive. Findings reflect what organisations publish publicly; detail varies significantly between providers. Benchmarked services differ in purpose (e.g., open operational feeds vs restricted data access), meaning results indicate transferable patterns rather than direct price comparators.

Appendix 2. Supply Analysis

Table A.2, below, sets out the metadata fields included their purpose within the mapping of DfT datasets, and the percentage of datasets where each was available. This makes visible the strengths and gaps in metadata quality across the portals and informs where further validation or standardisation may be required.

Table A.2: Overview of Metadata in database

Metadata	Purpose	Availability
DfT Dataset	Identifies the dataset by its official title.	100.00%
Organisation	Indicates which department or public body publishes or maintains the dataset.	100.00%
Source	Shows which portal or platform the dataset was obtained from	100.00%
Category	Provides a high-level classification used for grouping datasets	100.00%
Subcategory	Adds a more specific thematic tag under the broader category.	100.00%
Mode	Indicates the transport mode covered (e.g., road, rail, bus).	100.00%
Description	Summarises the content and purpose of the dataset.	100.00%
Link to Information Page	Provides access to the main landing page where dataset details are published.	100.00%
Topic	Displays the topic label assigned by portal.	98.82%
API?	Indicates whether the dataset has an API programmatic access route.	98.82%
Date Added	Shows when the dataset was first listed on the portal.	98.82%
Form of Information	Describes the type of output (e.g., report, spreadsheet, survey)	96.47%
Area	Specifies the geographic coverage of the dataset.	95.29%
Open / Closed Data	Indicates whether the dataset is open under OGL or restricted.	90.59%
Data Format	Lists the technical file format(s) available.	82.35%
Last Time when Dataset Framework was Adjusted	Records the most recent update or structural change to the methodology behind the dataset.	74.12%
Cost / License	States licensing conditions or any associated costs.	35.29%
Update Frequency	Indicates how often the dataset is updated.	30.59%
Time Period	Shows the temporal coverage of the dataset.	29.41%
Mode (as on the DfT website)	Shows the transport mode covered (as displayed on the DfT's data portals)	21.18%
Data Size	Provides information on dataset volume or file size.	5.88%

Appendix 3. Persona-based interim findings

I. Local Transport Authority Analyst

Local Transport Authority Analyst

DfT data usage

Local Transport Authority Analysts use DfT data, alongside local datasets, to support traffic management, road safety, and local transport analysis. Their work focuses on answering practical, often time-sensitive questions for planning, monitoring, and reporting to elected members.

Key DfT datasets used:

The analyst typically requires regionally relevant data in accessible formats for local analysis, rather than large-scale, national datasets. DfT assets include:

- STATS 19 CSV downloads.
- Real-time BODS and Street Manager feeds, where local systems have the capacity to ingest and process them.
- Traffic count data.
- Static downloads (mainly CSV or Excel) from the DfT website for regionally relevant data, which is then typically processed through Power BI dashboards.

Non-DfT dataset used:

External data sources the Local Transport Authority analyst interacts with include:

- APIs or tools provided by consultancies (e.g., Arup).
- INRIX bulk raw files.
- Google API.
- EV charging data through private organisations like ZapMap.

For more complex analysis, more digitally mature authorities have developed internal query layers. Authorities with more limited data capability often rely on external consultants to interpret and apply the data to specific transport projects, which can increase delivery costs.

Key Barriers

Infrastructure and IT constraints make it costly and complex for local teams to ingest, store, and analyse high-volume DfT datasets.	• Resource-constrained public-sector IT environments limit tooling choices (often requiring use of Microsoft-only stacks) and restrict flexible data processing. • High-volume datasets and slow downloads were described as operationally burdensome, requiring significant local storage and compute; some teams reported needing to delete historical data to manage capacity (e.g., with INRIX bulk files).

	• Push-only interface designs (e.g., Street Manager) require teams to ingest and store full data feeds locally and build their own querying capability, rather than running simple, ad-hoc queries.
Limited in-house technical capabilities constrain automation, API use, and effective reuse of DfT data.	• Technical skills for automation and API use were described as limited and sometimes concentrated in a single individual, creating fragility and bottlenecks. • A lack of worked examples and “getting started” guidance increases reliance on manual processes. • Modelling and advanced analysis are often outsourced, with dependence on third parties, adding complexity and cost.
Publication lags and unpredictable refresh cycles make some national datasets poorly suited to time-sensitive local planning and monitoring.	• Interviewees noted that key DfT and national statistics are often published with long delays, meaning outputs used in local transport plans can be “practically two years out of date” by the time decisions are taken. • Similar concerns were raised about census and ONS datasets, where long publication cycles were described as making the data effectively obsolete for current planning. • Unpredictable publication timing and refresh cadence were described as limiting automation and regular monitoring workflows. • Delays in accessing national datasets were also said to slow project delivery, including extended data discovery phases at project start or when responding to data requests from bodies such as National Highways.
Weak coordination and feedback loops lead to duplicated effort, misaligned solutions, and slower delivery for local policy and analytical use cases.	• Underrepresentation of local authority research and policy users in earlier DfT API and standards design was said to have produced solutions poorly suited to local analytical workflows. • Multiple authorities were reported as independently repeating the same basic data preparation work (ingesting, cleaning, reshaping, and storing data locally for tools such as Power BI), creating duplication and inefficiency. • A lack of higher-level, derived outputs (e.g., pre-processed metrics or ready-to-use summaries) was described as increasing duplication across regions. • Fragmentation across operators, modes, and local data standards was said to limit integration and make it harder to set and enforce minimum data or API standards.

	• Uncertainty around licensing, onward sharing, and who pays for data was described as causing delays, with a preference for OGL to enable smoother sharing between local authority partners.
Data quality, structure, and coverage limitations reduce the usefulness of national datasets for local analysis and decision-making.	• Insufficient granularity and small sample sizes were described as constraining local-level analysis, with some data not available at local authority level. • Data structure and consistency were reported as poorly optimised for reuse, complicating integration across systems and formats (e.g., quality issues in EV and NTS outputs, and difficulty extracting usable data from complex or publication-style tables). • Methodological opacity was flagged, including unclear definitions (e.g., urban/rural classifications) and limited access to underlying methodologies. • Gaps in coverage were highlighted, including missing modes (e.g., pedestrian data) and some required datasets not being published. • Limited nationally accessible disruption data, particularly for locally managed unplanned road closures, was said to require workarounds such as scraping operator websites or negotiating access locally.
Fragmented discovery and uneven signposting make it time-consuming to identify, access, and assemble the data needed for local analysis.	• Interviewees reported difficulty locating the right datasets on DfT websites, including finding the correct CSVs and navigating outdated or incomplete documentation. • Analysis often requires long time series across multiple authorities (e.g., 10 years), making manual downloads slow and hard to maintain compared to programmatic access. • Uneven signposting across operators was noted, with some offering clear catalogues of available feeds while others require repeated bilateral conversations to establish what data exists.
Budget constraints limit data acquisition, analysis, and tooling investment.	• Tight budgets were repeatedly cited as constraining the ability to acquire data (e.g., installing cameras), commission analysis from consultants, and invest in infrastructure or vendor tools.
Priority Needs from DfT	
Sustained investment in shared infrastructure and skills to reduce duplication and support	• Build and operate national platforms where economies of scale exist, including centralised procurement and provision of high-value commercial datasets that are currently inconsistently purchased across local authorities (with mobile network data most frequently cited).

local analytical capacity.	• Provide long-term resourcing to maintain, clean, and improve core datasets, including funding to strengthen upstream data collection and quality. For example, ring-fence funding for monitoring and evaluation within local authority programmes, incentivising better local data collection and analysis that can justify schemes and feed back into DfT evidence. • Address skills gaps through coordinated support as well as tooling, including plain-English "getting started with APIs" guidance, worked examples (e.g., R, Python, Power Query1), clear schema and parameter documentation, and access to knowledgeable user support. • Improved resourcing and access routes were seen as reducing reliance on intermediaries (e.g., consultants or larger authorities) for basic analytical capabilities.
Timely and predictable data releases aligned to local policy, planning, and funding cycles.	• More frequent updates for operationally critical datasets (e.g., traffic, walking and cycling), favouring usable refresh cadences (e.g., ~ 15-minute intervals) over infrequent historic releases. • Key planning datasets published on a fixed annual timetable that aligns with common funding deadlines, with a clear commitment that the full year's data is available no later than the end of February. • Clear, reliable publication schedules for major datasets to support automation, routine monitoring, and planning workflows. • Predictable, timely releases allow councils to base decisions on current evidence, reduce reliance on proxy data, and avoid using outdated information.
APIs and access methods designed for local authority and "average technically interested" users, not only large vendors or advanced technical teams.	• Provide simplified, queryable APIs over foundational datasets (e.g., Street Manager, BODS, disruptions), with a longer-term ambition to support natural-language querying for non-technical users. • Offer lighter-weight standards and formats suitable for small teams alongside richer profiles for advanced users. • Recognise that for some datasets (e.g., road traffic and collision statistics), existing CSV publishing remains sufficient unless APIs enable higher frequency, greater detail, or materially new use cases.

[1] An example of this is the STATS19 R package developed independently of DfT which offers an alternative way to access this data for those familiar with the R language. See *Lovelace, R., Morgan, M., Hama, L., Padgham, M. stats19 A package for working with open road crash data, 2019* for an example of community-led work to add-value to datasets by making them more readily available.

Centralised analysis to produce structured, derived insights alongside raw data delivery.	• Centralise analytical processing where appropriate to share "crunching" at the DfT level (e.g., aggregations, joins, derived metrics) to avoid each local authority replicating ETL pipelines independently. • Provide derived outputs such as national dashboards and visual products for high-use datasets. • Offer policy- and place-relevant views, including local-authority-specific portals, clear boundary filtering, and borough-level granularity. • Expose processed datasets through Power BI-friendly access routes, with simple URLs and keys, standard filters, stable schemas, and predictable refresh cycles to support routine local reporting and monitoring.
Benefits from improved API access	
Enable an integrated, coherent multimodal view of local travel patterns.	• Queryable access to DfT datasets was seen as enabling a more integrated, cross-modal view of travel by allowing councils to combine national and local data. • Ability to pull locally relevant slices of national datasets and combine them in tools such as Power BI would support dashboards and day-to-day decision-making between major model updates. • More complete modal coverage, including currently missing modes, would strengthen local baselines for scheme assessment and modelling. • Simpler standards and APIs would make it easier to combine national and local datasets and share local inputs back upstream to support national modelling and reuse.
Strengthen evidence-based policymaking and forecasting.	• Better API access was linked to more timely, decision-relevant evidence for investment cases, policy monitoring, and reporting to elected members. • Finer-grained, query-ready data with predictable update schedules was seen as supporting automation and routine monitoring. • Easier access to large datasets and provision of structured or derived indicators were seen as reducing local processing effort, duplication across authorities, and reliance on paid-for camera data, consultants, or intermediary platforms.

Enable richer, more timely road safety and compliance analysis.	• More frequent, link-level traffic volumes with breakdowns by road user type would help fill gaps in local count networks and improve road safety metrics. • Improved access would also support better monitoring of congestion and compliance, and clearer, more interpretable outputs (e.g., graphs and dashboards) for internal decision-making and reporting.
Early WTP signals	
Willingness to pay is very low. Local Transport Authority Analysts reported limited capacity to pay for data due to stretched local authority budgets, with additional charges for access, licensing, or API usage seen as unrealistic. WTP arises primarily where authorities already purchase high-cost commercial datasets (e.g., mobile network or freight data), in which cases interviewees saw clear value in DfT centralising procurement to secure national licences, reduce costs through economies of scale, and ensure consistent access across authorities.	
"If the last data is from a year and a half ago, it's hard to see whether policy has had any effect." A Regional Transport Authority	

II. Transport Researcher

Transport Researcher	
DfT data usage	
Transport Researchers use DfT data directly to produce analysis and evidence for research and policy. They work with national, research-grade datasets to understand travel behaviour, model demand, evaluate interventions, and support policy development through historic analysis and demand forecasting.	
Key DfT datasets used: **NTS**, accessed through the UK Data Service (both standard and secure versions). **STATS19** (road collision data) for road safety research and evaluation.	Non-DfT dataset used: Data from other providers linked to the DfT ecosystem: (i) National Highways Motorway Incident Detection and Automatic Signalling (MIDAS) data; and (ii) Local authority number plate data for air quality monitoring.

BODS API (Bus Open Data Service). **NaPTAN** (often exported as raw files). **Vehicle licensing statistics** (e.g VEH0120 and 0124 tables) via multiple CSV files. **Data from DfT website and other Gov.uk resources** is sometimes scraped for non-API assets, like Excel files.	Third party intermediaries who process DfT data (like Agilis's Crash Maps).

Key Barriers	
Fragmented discovery and signposting across GOV.UK make it difficult for researchers to identify what DfT data exists, where to find it, and which sources are authoritative.	• Lack of a central, well-signposted repository for DfT data, with discovery fragmented across GOV.UK pages and legacy documentation. • Overlapping or duplicate data listings, creating uncertainty over which sources should be treated as authoritative. • GOV.UK navigation itself described as a barrier to discovery, characterised by one participant as a "firewall" to finding information. • Difficulty establishing the existence, location, and status of niche but research-critical datasets. • Specific uncertainty around assets such as DVLA-DVSA source tables, availability of historic BODS trajectories, and the current state of Street Manager datasets.
Publication delays, limited historical snapshots, and gaps in near-real-time data constrain timely policy evaluation and longitudinal research.	• Many core datasets (e.g., NTS, traffic and vehicle statistics) are published annually or quarterly, often with 12-18 month delays, limiting assessment of recent policy interventions and alignment with live policy debates. • Limited or inconsistent availability of historical snapshots constrains longitudinal analysis, sometimes requiring researchers to collect and store data themselves over multiple years. • Concerns were raised about the refresh and retention of operational and near-real-time data, particularly within BODS, including refresh intervals typically cited as 20-30 seconds where some use cases require closer to 5 second updates, alongside limited availability of timetable and trajectory snapshots and gaps in real-time coverage. • The absence of real-time train location data was highlighted by some participants as a significant gap for research use.

Inconsistent data quality and unclear methodology increase uncertainty and add significant overhead to preparing DfT datasets for analysis and reuse.	• Inconsistent naming conventions, categorisation, anomalies across datasets, and weak or incomplete metadata reduce confidence and limit reuse. • Limited transparency on methodology, data cleaning, reconciliation, and treatment of gaps or exclusions, particularly for datasets derived from upstream sources (e.g., DVLA-DfT pipelines). • Key attributes split across multiple tables with different update frequencies (e.g., VEH0124 annual vs VEH0120 quarterly), with unclear guidance on how tables should be reconciled, especially for niche vehicle categories. • Separate United Kingdom and Great Britain tables with differing totals, often without clear methodological explanation. • Format and usability issues, including large file sizes, publication-style CSVs, embedded notes or headers, inconsistent column structures, and other non-analysis-friendly formats that require significant manual restructuring. • Use of non-standard geospatial formats, such as DRO fields published in OSGR rather than WGS84, creating additional processing overhead before integration with other datasets.
APIs often create friction for research use, particularly around documentation, tooling, and reliability.	• Barriers were framed less around whether APIs exist and more around whether they support common research workflows. • Documentation was described as focusing on endpoint guidance, with limited coverage of data meaning, caveats, provenance, and methodological context needed to interpret outputs. • Developer-oriented formats (e.g., JSON) often require reshaping and conversion that is time and skills-intensive for researchers. • Tight rate limits, API key constraints (including limits per legal entity), and fragility of long-running or continuous calls were cited as operational constraints. • Registration, credential management, and usage rules were described as adding administrative burden. • Reported difficulties using APIs with common research tools (e.g., R), sometimes requiring a switch to unfamiliar languages (e.g., Python). Variable response times made rate limits hard to manage in practice, and network interruptions during long-running calls were described as leading to gaps in collected data.

	• Discontinued or unreliable APIs were described as undermining long-running pipelines, with some participants noting that poor-quality data delivered via APIs can be worse than well-maintained legacy formats.
Gaps in data coverage and licensing constraints limit the scope of research and slow collaboration.	• Limited or absent coverage in some areas, particularly freight and maritime, constrains research in these domains. • Incomplete fares data and gaps in bus datasets were linked to inconsistent operator submissions or uneven GPS tracking. • Licensing conditions can restrict data sharing between organisations (e.g., with local authority partners), slowing collaborative research.
Priority Needs from DfT	
A central, inspectable DfT data catalogue to act as a single source of truth for reproducible and longitudinal research across all modes and ALBs.	• To be useful for research, stakeholders noted that such a catalogue would need to clearly set out: • Dataset existence and status (live, historic, deprecated, experimental). • Coverage (temporal, spatial, modal, population). • Granularity (e.g., street-level vs regional, per-vehicle vs aggregate). • Update frequency and publication lag. • Access route (API, bulk download, secure access, licensed). • Owning organisation and authoritative version. • Reproducible “getting started” code or reference implementations to support consistent access and analysis, including community-maintained tooling where relevant.
A minimum research metadata standard embedded in both the data catalogue and API documentation to support reproducibility and methodological transparency.	• To be useful for research, stakeholders noted that such a standard would need to clearly set out: • Data generation and cleaning steps. • Reconciliation rules across sources (e.g., DVLA - DfT pipelines). • Reasoning for exclusions and breaks in series. • Versioning and revision policy (what changes retrospectively, what doesn’t).

	• Allow users to inspect samples and metadata without downloading large files.
APIs and data structures designed around research and analytical workflows, rather than solely for operational or commercial integration.	• To be useful for research, stakeholders noted that this would need to include: • Column- and filter-level queries over full datasets. • Support for common research tools and practices (including R and Python parity, reproducible pulls and historic access where appropriate). • Raw data, alongside API release, particularly Parquet format, for making bulk datasets available. This can provide many of the benefits of API's, including the ability to download only the datasets needed by researchers using queries, without the costs of setting-up backend infrastructure.
Benefits from improved API access	
Strengthen evidence-based policymaking and forecasting.	• More timely and granular programmatic access to core datasets supports forecasting, policy evaluation, and assessment of future legislation. • APIs could materially reduce routine data handling by automating tasks such as merging large split files and allowing users to retrieve only the fields they need, reducing time spent managing large tables and enabling greater focus on interpretation and modelling. • Clearer documentation on data cleaning and stronger consistency across related tables improves confidence in official statistics and reduces time spent resolving discrepancies. • Well-supported APIs, including worked examples and “getting started” guidance, lower barriers for less technical researchers and students, supporting broader uptake and more consistent, reproducible analysis across teams.
An integrated, coherent multimodal view of travel patterns.	• Easier discovery and combination of datasets across organisations and modes, ideally through a DfT-coordinated “one-stop shop,” supports more integrated and timely analysis of travel behaviour. • Improved integration enables faster insight into shifts in travel patterns, including responses to policy changes or external shocks. • Better access to under-documented areas (e.g., freight, maritime, aviation) expands what can be studied and compared across modes.

Early WTP signals
Willingness to pay is very low among Transport Researchers. Researchers rely on free, OGL-licensed data, and charging was seen as likely to create inequalities between well-funded and financially constrained organisations, particularly universities and research centres, and to discourage use, especially for research and evidence-generation activities. They also pointed to the public value of transport research, arguing that its role in supporting policy development and public benefit weakens the case for charging.
“If the last data is from a year and a half ago, it’s hard to see whether policy has had any effect.” A Regional Transport Authority

III. Operational Controller

Operational Controller

DfT data usage

Operational Controllers maintain real-time situational awareness across transport networks and respond quickly to disruptions. They use DfT's real-time or near-real-time data primarily as a shared reference layer, alongside their own or third-party operational systems, to understand the planned network and keep live passenger-facing information consistent.

Key DfT datasets used:	Non-DfT dataset used:
They interact with a limited set of DfT and DfT-associated datasets that support coordination and information consistency: • BODS • Disruptions and incidents data • Street Manager, and • NaPTAN. Some operators noted that data flows are often predominantly upstream, with operational data supplied to DfT feeding national systems and APIs, rather than DfT data acting as a primary operational input. Some operators are beginning to explore DfT and DfT-associated RDM datasets internally.	Operational Controllers rely heavily on non-DfT data for day-to-day control, including: • Operator internal systems for live operations and decision-making. • Industry rail data (e.g., LENNON national rail data) accessed via internal data platforms (e.g., Snowflake SQL Queries). • Geographic and mapping data from external and industry sources.

Key Barriers

Real-time DfT feeds are insufficiently timely or complete for operational control, reducing confidence in live decision-making.	• Some real-time feeds were described as having latency or refresh rates that are "not fast enough" for operational use. • Operational events (e.g., emergency roadworks) may only appear in DfT systems after activity has begun, creating gaps between the live network and downstream data. • Misalignment between operator systems, local authority systems, and DfT feeds requires manual reconciliation during incidents.
Data quality issues in real-time operational feeds reduce trust in	• Inaccuracies, missing data, and limited provenance were described as undermining confidence in operational use.

DfT data as a reliable source for live decision-making.	• BODS-related issues were highlighted, including unexplained gaps, missing records, and uneven maintenance across local authorities leading to outdated or inaccurate data. • One operator noted that the ABODS real-time flow is “not working 100% correctly yet,” with discrepancies versus operator systems reducing confidence in it as a live source of truth. • Deriving higher-value operational signals (e.g., trip updates) from raw vehicle positions was described as costly and complex for operators.
Weak and inconsistent standards across operators and modes create significant integration overhead and reduce the usability of operational data.	• Operator submissions were described as highly inconsistent, with no clear precedent for cross-operator alignment or agreed aggregation definitions. • Lack of consistent identifiers and standards across rail and bus was reported as a barrier to integration. • Practical GTFS issues were highlighted, including poor shapes, inaccurate stops, and missing fields (e.g., bus bay information present in source data but not carried through). • Translation between formats (e.g., NetX to GTFS) was described as error-prone and resource-intensive, adding operational complexity.
Gaps in dataset coverage and granularity limit end-to-end operational visibility and consistent passenger information across modes.	• Limited multimodal integration was highlighted, for example rail data supporting entry/exit points but not full end-to-end journeys. • Some datasets were described as overly aggregated, preventing finer operational segmentation. • Absence of accessible, high-resolution “source” train location data constrains real-time operational use. • Interviewees noted missing protocols or feeds for certain modes (e.g., car share). • Uncertainty about the existence or availability of specific operational datasets (e.g., parking pressure or curbside data) was reported.
Governance, commercial, and contractual constraints limit consistent operational data sharing and improvement across the ecosystem.	• Tension between DfT expectations and operator concerns about commercial sensitivity constrains what data can be shared. • Legacy systems, suppliers, and contracts often lack clear provisions on data-sharing rights and ownership. • Improving and aligning operational data was described as requiring dedicated headcount or subsidy, which is not always available.

	• Fragmented local authority approaches and bespoke, case-by-case arrangements drive inconsistent reporting. • These factors increase coordination effort to reconcile, maintain, and operate multiple parallel feeds and processes.
Priority Needs from DfT	
Reliable, nationally consistent real-time data that accurately reflects the live network and can be queried flexibly for operational decision-making.	• Set clear national baselines for timeliness, refresh rates, latency, and rate limits for operationally critical datasets (e.g., BODS, ABOD, roadworks, disruptions), prioritising consistency over perfect standardisation. • Provide time- and location-specific APIs that support filtering by route, asset, and time window to enable live operational use. • Improve train location data, acknowledging that signalling-based inference is often ambiguous and currently requires operators to calculate positions themselves. • Move BODS toward a consistent national update baseline (e.g., 5-10 seconds) rather than uneven ~30-second refresh rates across operators. • Address known gaps in BODS coverage, including missing assets, incomplete operator participation (notably smaller operators), and modal omissions (e.g., some ferries). • Establish a nationally coordinated highways disruptions dataset to provide early, authoritative visibility of roadworks and closures, including emergency works that currently surface late. • Publish clear, machine-readable refresh and availability schedules so operators can automate ingestion, reconciliation, and monitoring. • Be explicit about DfT data mandates and interests, clearly stating what operators must publish, in what form, and which additional datasets DfT is actively seeking to access. • Continue addressing commercial-sensitivity claims from operators which slow access to passenger count and load data.
Clear, nationally defined standards are needed to enable interoperability and consistent operational data supply across modes.	• Define minimum common standards for identifiers, aggregation rules, core fields, and metadata across rail, bus, and highways. • Ensure aligned formats and identifiers to support smooth cross-mode integration and reduce downstream reconciliation.

	• Provide clear, statutory-level guidance on standards, naming conventions, and aggregation rules, set centrally by DfT and agreed cross-industry. • Standardise data at source rather than relying on downstream translation or voluntary compliance.
Benefits from improved API access	
Deliver more reliable real-time operational data.	• Standardised, well-documented APIs and identifiers make operational data easier to use consistently across organisations and modes. • API-based delivery improves provenance and traceability, reducing inconsistencies associated with ad hoc, file-based exchanges (e.g., spreadsheets and opaque processing). • More granular passenger loading data supports better capacity planning and more accurate real-time passenger information.
Accurate, real-time journey planning and reliable passenger information.	• Improved APIs, data quality, and standardisation support more accurate trip planning that combines schedules, real-time updates, disruptions, and fares. • Better access reduces downstream “patching” needed to keep passenger-facing outputs consistent across systems. • Reliable crowding indicators improve trust by ensuring apps and displays reflect actual network conditions. • Fresher real-time data improves passenger experience and can support more sustainable travel choices.
Early WTP signals	
Willingness to pay is low. Operators indicated they would only pay for data where it delivers clear commercial and operational value, stressing that APIs should not be designed or priced until that value is evident. They emphasised that value comes from improving journey quality rather than marginal gains in journey time or cost, which are largely fixed.	
“We have a sizable team just scrubbing and sanitising the data before it’s usable, because if the data is wrong, users lose trust immediately.” A Mobility as a Service (MaaS) stakeholder	

IV. Data Engineer & Integrators

<table>
<tr><th colspan="2">Data Engineer & Integrators</th></tr>
<tr><td colspan="2">DfT data usage</td></tr>
<tr><td colspan="2">The Data Engineer & Integrator cleans, transforms and combines DfT data, often alongside other public or private sources, to create analysis-ready datasets and pipelines. Their work underpins downstream user-facing applications and operational dashboards.</td></tr>
<tr><td>Key DfT datasets used:

• BODS feeds covering disruptions, schedules, bus geometry, and vehicle movements, used to power real-time mapping, performance analytics, and journey context.

• NaPTAN bus stop data, used as a core reference layer for stop locations, naming, and interchange logic across modes.</td><td>Non-DfT dataset used:

• Rail Data Marketplace data on disruption data, station calling patterns, actual train movements, and vehicle-to-service associations. Accessed via API or flat files/CSVs as a fallback through OGL3 and other open licenses.

• DVLA and DVSA APIs

• Bulk licence providers (Experian, UK Vehicle Data)

• Integrations directly from operators.

• Other commercial aggregators.</td></tr>
<tr><td colspan="2">Key Barriers</td></tr>
<tr><td>Fragmentation and weak standardisation of datasets limits the engineers’ ability to write reusable, stable processing rules.</td><td>• Inconsistent naming conventions, identifiers, and formats and measurement standards between operators.

• Events represented differently across geographical regions.

• NaPTAN and bus-stop names are not consistent with corresponding train station names. Misalignment prevents matching engines recognising valid interchange points which breaks “through-journey” integration across modes.

• Inconsistency across operational datasets, like disruption reporting.

• APIs are not centrally discoverable, with TfL referenced as a positive contrast.</td></tr>
</table>

Existing data quality is reduced by inconsistent or erroneous fields and incomplete coverage, increasing integration effort and the unreliability of data pipelines.	• Vehicle-level integration is constrained by anonymisation that removes asset-level identifiers. Operationally and analytically linked to DfT, DVLA's Vehicle Registration Mark (VRM)-based approaches can create synchronisation issues, as registration marks can move between vehicles; Vehicle Identification Number (VIN)-based querying more reliably ties data to a specific asset. • Incomplete geographical coverage, with some councils' data missing. • BODS fares data is "not consumer facing quality yet", requiring additional processing. Disruption data is patchy and inconsistently implemented, limiting reliable cross-referencing of disruption events against stops, services, or journeys. • Instances of non-compliance, where operators have limited incentive to provide data they are formally obligated to supply.
Poor API documentation makes it difficult for data engineers to interpret fields and coverage correctly.	• Technical access guidance exists, but the underlying data content, structure, and meaning are insufficiently explained. • Very large or "monolithic" outputs that bundle too much data together are difficult to code and maintain. • API schema changes pose commercial risk of breaking downstream services if they are not clearly communicated in advance.
Emerging transport data is not being systematically captured or exposed, increasing reliance on third party sources.	• Limited visibility of what DfT holds and when new data becomes available. • Lack of publicly available feeds in areas of high interest, investment, and policy change such as micro-mobility (scooter/bike docking locations) and active travel (missing or non-API exposure of walk links and cycle infrastructure). • Limitations in real-time coverage. • Cultural reluctance to share data, weak commercial incentives for some operators, and ill-defined contractual data ownership reduce data availability.
Priority Needs from DfT	
Enforce consistent standards and identifiers across suppliers, modes and datasets.	• Clear minimum standards for temporal granularity (timestamps, refresh cadence), spatial consistency (location precision, naming conventions), and update frequency. • Standardised identifiers that reliably join datasets end-to-end such as NaPTAN and BODS station naming alignment.

	• Actively enforce existing obligations by monitoring compliance with current standards, such as timeliness and completeness of real-time submissions and intervene where variation undermines usability.
Predictable API versioning and change communication.	• APIs with explicit, stable versioning, treating each version as a contract that downstream systems can depend on. • Minor, non-breaking changes (e.g., adding optional fields) should be allowed within a version but breaking changes (renamed fields, identifiers, semantics or changed data types) must only occur in a new major version. • Communication of schema and data-model changes early and through predictable channels. • Clearly defined and published deprecation policy when fields, endpoints, or datasets are to be deprecated, including a minimum notice period, a fixed end-of-life date, and explicit guidance on replacement fields or APIs.
Support machine-to-machine workflows as a baseline expectation by assuming the consumer of DfT data is a server, not a person.	• Replacing “email reminder” access patterns with persistent credentials (service accounts) and stable endpoints or storage locations. • Provide automation-friendly access patterns like APIs or, where more appropriate, event-driven mechanisms (webhooks, Pub/Sub topics, queues). At minimum, a reliable “new data available” signal is enough. • Publishing in clear, regular update schedules so ingestion pipelines can be built without polling or human intervention.
Benefits from improved API access	
An integrated, coherent multimodal view of travel patterns.	• Lower effort required to clean and reconcile data. • Easier to conduct analysis-ready outputs for downstream users and services, rather than relying on ad hoc downloads, inconsistent formats, or scraping. • For vehicle-linked use cases, API access becomes significantly more valuable when records can be queried and joined using stable identifiers. • More consistent, well-documented interfaces with improved metadata.
More reliable real-time operational data.	• Consumption of current, service-based feeds rather than repeated bulk downloads and re-hosting.

	• High-frequency refresh and stronger disruption coverage make fast-changing feeds operationally usable. • Reduced downstream maintenance by removing the need to rebuild bulk ingestion workflows when data changes.
Accurate, real-time journey planning and reliable passenger information.	• Simpler passenger-facing outputs from complex public transport data, such as timetables, locations, disruptions, and stop/interchange information for improving routing and inclusivity. • More reliable journey modelling. • More consistent information across data sources reducing downstream cleaning.

Early WTP signals

Willingness to pay is medium. Suggested pricing models include per-API call or volume-based pricing, but only if the data is reliable, accurate and available, with pricing reflecting the actual cost of processing. Integrators would not pay for premium access tiers, higher rate limits or faster refresh APIs. However, they might consider incurring a cost for data they cannot get elsewhere or data where DfT has purchased exclusive access on behalf of all local authorities and the consultant integrators they work with.

> "You've got loading data for different train operating companies, all in different formats, all with different measurement standards."
>
> A Third Sector stakeholder

> "The ability to access the data faster, make more calls on the API, is not really [something we'd pay for]. We will figure out an engineering way around having to do that."
>
> A Transport Consultant stakeholder

V. Product Developer & Innovator

Product Developer & Innovator
DfT data usage

This persona uses DfT and wider transport data to build and improve products, tools, and applications, with usage varying by mode and product focus.

Key DfT datasets used:	Non-DfT dataset used:
Depending on the sector(s) they are working across, usage might include: • **BODS**, including timetable/route data and real-time vehicle movement feeds where relevant. • **NaPTAN** (often via file formats). • **Street Manager** for roadworks and road closure information. • Some stakeholders reported using **formal request routes** (e.g., FOI/EIR) where required data is not otherwise accessible.	• Operator-provided data to supplement or validate public datasets where quality or coverage is a concern. • Other public-sector sources (e.g., ONS Census), local authority datasets via data-sharing agreements, and rail data services/marketplaces where applicable. • Third-party tooling and vendors that normalise feeds into standard developer formats (e.g., GTFS). • Commercial providers, including mapping platforms (e.g., OpenStreetMap, Google) and mobility datasets such as mobile network data and road restriction/traffic regulation data obtained via partnerships or aggregators (e.g., Ito).

Key Barriers

Data quality and inconsistent and competing “sources of truth” were described as the dominant constraint for building reliable products.	• Quality concerns spanned accuracy, latency, weak provenance, and uneven maintenance, creating downstream rework and product risk. • Competing “sources of truth” across BODS, local authorities, and operators created uncertainty over which data should be treated as authoritative. • BODS-specific issues included inconsistent coverage across operators and regions (including mismatches between schedules and real-time), missing data requiring fallbacks to other sources, duplicate data and overlapping calendars, intermediate uploads rather than finalised data, and services disappearing from the dataset. • Direct operator or local authority feeds were sometimes described as higher quality in practice, as operational dependence allows issues to be identified and corrected more quickly. • NaPTAN issues included generic stop naming and outdated records due to inconsistent local authority updates, alongside a need for richer stop attributes (e.g., shelters, seating) to support planning and accessibility features.

	• Fares data was described as complex and, in some cases, not available through expected access routes, including perceived limitations in rail-related data provision.
Limitations in refresh rates, consistency, and real-time usability constrain the delivery of reliable, user-facing experiences.	• Update frequency was raised as a concern, with live bus data commonly refreshing with less frequency as needed. • Refresh cadence was described as inconsistent across operators and regions, making it difficult to deliver uniform real-time experiences at scale. • Where only raw vehicle position data is available, developers reported needing to derive higher-value signals (e.g., trip updates) themselves, adding cost and engineering complexity. • High-volume real-time feeds were described as difficult to manage without effective filtering and query mechanisms.
Fragmented discovery and unclear signposting increase onboarding time and integration effort for product teams.	• Developers reported difficulty understanding what DfT datasets exist, how they relate, and which should be treated as authoritative, complicating integration planning and product roadmaps. • One interviewee cited DfT's former Transport Direct service as a useful reference for a unified multimodal data source, suggesting a similar approach could simplify developer onboarding.
Misaligned standards and identifiers increase integration effort and limit seamless multimodal product development.	• Mismatches between standards such as NaPTAN and GTFS (e.g., bus bay information present in NaPTAN but not in GTFS). • Poor shapes and inaccurate stop representations in GTFS requiring additional correction before product use. • Lack of consistent identifiers across modes, particularly between rail and bus. • Rail data not available in commonly used developer formats, leading some developers to pay third parties to translate it.
Gaps in dataset coverage limit product scope and the quality of user experiences.	• Lack of usable, high-resolution train location source data, with signalling-based location described as imprecise. • Historical weather data described as siloed or hard to access. • Uncertainty about the existence of parking pressure data. • Need for more granular congestion and traffic-flow data (e.g., street-level or specific-road views, broken down by time of day).

Fragmented reporting and bespoke data-sharing processes create delivery friction and slow product development.	• Fragmented local authority reporting requirements were described as driving bespoke dashboards and outputs, increasing development and maintenance effort. • Manual or bespoke data-sharing processes were cited as a source of frustration, including safety incident reporting and resource-intensive arrangements for scooter trials.
Priority Needs from DfT	
Data that meets common national standards for structure, depth, quality, and coverage to support scalable product development.	• To be useful for product developers, stakeholders noted that this would need to include: • Minimum national data standards set and enforced where the policy objective is national growth and market scale, particularly for UK SMEs. • Alignment with established international standards where interoperability matters, with UK-specific deviations clearly documented. • General Transit Feed Specification (GTFS) and GTFS Real-Time (GTFS-RT) data for all transport modes. • Integration of fares data using GTFS Fares v2. • Reporting frameworks like Mobility Data Specification (MDS). • Embedding standards into contracts, guidance, validation pipelines, and data publication processes rather than relying on voluntary compliance. • Stakeholders noted that consistent national standards would allow innovators to scale services across more areas without extensive local data fixing.
Clear signals on which datasets are priorities, which are intended to become authoritative, and where third-party innovation is encouraged.	• To be useful for product developers, stakeholders noted that this would need to include: • An authoritative cross-modal data directory spanning DfT, operators, and local authorities, setting out priority datasets and their intended future role (e.g., system of record, enabling dataset, experimental) so developers know where to invest and where to build complementary services. • Better alignment of DfT funding levers (e.g., Innovate UK's Small Business Research Initiative) into joined-up programmes that reinforce priority datasets, standards, and outcomes rather than fragmented one-off calls.

	• A multimodal central transport data marketplace distinguishing between DfT-owned data, DfT-endorsed datasets, and market-provided services. • Stakeholders noted that clearer signalling would reduce uncertainty, support earlier feasibility assessment, direct private investment toward priority areas, avoid duplication of effort, and increase confidence about where DfT intends to lead versus enable.
Access to datasets with clear end-user experience value and higher potential for economic impact.	• To be useful for product developers, stakeholders noted the importance of engaging with them to identify which data unlocks viable products rather than speculative use cases. Market interest includes but is not limited to: • High-granularity, frequently updated operational data (e.g., real-time feeds such as BODS, where some use cases require 5-10s refresh rather than ~30s), including street-level and time-of-day variation data, validated incidents and disruptions, and weather context where it affects journey reliability and planning. • Baseline capacity data across modes to support performance measurement, demand management, and system improvement. • Richer asset and service attributes to support accessibility, inclusion, and service design, including occupancy indicators, accessibility features (e.g., disabled spaces, bike storage), and stop-level attributes (e.g., shelters, seating). • Improved visibility of datasets supporting micromobility and active travel. • Fares data standardised across modes.
Benefits from improved API access	
Unlock cross-sector innovation, research, and new products.	• Central, standardised, and well-documented APIs reduce aggregation effort by removing the need to harmonise fragmented data sources. • More consistent access lowers the cost and time required to build, test, and scale products on top of DfT data. • Improved access was also linked to easier evaluation and benchmarking of interventions (e.g., comparing treated areas against controls), as well as lower-friction access models for SMEs (e.g., selective “pick-and-mix” access rather than large bulk dumps).
Accurate, real-time journey planning and	• Higher-quality, more consistent real-time and operational data (including disruptions and fares) supports more reliable journey planning and a better passenger experience, reducing the need for

reliable passenger information.	users to cross-check multiple apps or reconcile conflicting information. • A clearer national “source of truth” for live public transport data was seen as necessary for maintaining user trust in real-time journeys and reducing reputational risk for developer-built services. • Improved completeness and consistency reduces downstream “fixing” work (e.g., repairing shapes, correcting stops, deriving trip updates, compensating for missing fields), allowing teams to focus more on user experience and faster rollout of coverage. • Some participants noted that high-quality, open operational data can level the playing field for smaller developers by reducing reliance on costly proprietary datasets and narrowing the advantage of large platforms.

Early WTP signals

Willingness to pay is medium. Product Developers and Innovators generally expect public-sector data to be free. High fixed licensing costs were described as largely inaccessible for SMEs, particularly where similar insights can be achieved using open alternatives. Charging for data was therefore often viewed as potentially anti-competitive, reinforcing advantages held by large global firms.

WTP was expressed primarily for data that is exclusive or scarce (e.g., connected-vehicle feeds) or where uniform, high-quality coverage is available across geographies (e.g., comparable to Google or TomTom pricing at around $10k per country per month). Several participants noted that any charges introduced by DfT would likely be passed on to end customers, many of whom are publicly funded, creating circular costs with limited public value and a risk of reduced uptake.

> “The key question is whether the data quality is good enough to justify using it instead of sourcing data directly ourselves. BODS is useful to the extent that it provides around 70-80% of what we need, but the rest still has to be located or fixed.”
>
> A Mobility as a Service (MaaS) Stakeholder

Appendix 4. Prioritisation framework and scoring methodology

Background on Framework

The framework gives DfT a consistent way to rank candidate datasets for API development using a small set of criteria that reflect what users said matters most, while still keeping delivery considerations visible. This helps ensure API development effort is targeted at datasets where an API will improve usability and unlock value, rather than treating API enablement as an end in itself.

The scoring balances three core signals of user value and need:

- **Demand evidence** - whether there is clear, demonstrated demand for API access to the dataset,
- **Timeliness criticality** - whether users' use cases depend on freshness or high-frequency updates that make file-based publishing insufficient.
- **Dataset size** - reflecting consistent feedback that large datasets become substantially more manageable when they can be accessed programmatically through filtering, pagination, incremental pulls, and integration into existing pipelines, rather than requiring repeated bulk downloads and manual processing.

Alongside these value signals, the framework incorporates two criteria that protect deliverability and adoption:

- **GDS alignment** - captures whether an API can realistically be delivered in line with government expectations and for usable, secure, well-documented, and operable services, drawing on the GDS API technical and data standards and API Hub alpha lessons learned, reducing the risk of releasing endpoints that are hard to integrate or maintain.
- **Data quality** - reflects whether the underlying data is consistent, complete, and trustworthy enough for users to embed in products and operational processes. Where a dataset is high value, but quality is not yet sufficient, the framework helps flag it for foundation fixes before API development proceeds.

Finally, the approach strengthens governance by making prioritisation decisions transparent and auditable. Each score is accompanied by a short justification and linked evidence, such as user research insights, usage analytics, or a data readiness assessment, enabling DfT to refresh scores over time and explain clearly why particular datasets are sequenced as build now, foundation fixes then build, or next wave as priorities, constraints, and user needs evolve.

1) Demand evidence

Demand evidence was used as the primary input to the dataset prioritisation framework, reflecting the principle that API development should be led by demonstrated user need rather than inferred technical opportunity alone.

Demand was measured using qualitative interview frequency analysis drawn from 55 stakeholder interviews conducted as part of this study. During these interviews, respondents were asked about the datasets they currently use, struggle to access, or would find most valuable if made more accessible or available via an API. Mentions of datasets were coded and aggregated at the dataset sub-category level to ensure consistency and avoid over-weighting individual named datasets.

Each dataset sub-category was then assigned a demand score on a five-point ordinal scale, based on the number of interviews in which it was referenced:

- A score of **5** was assigned where a dataset sub-category was mentioned in **15 or more interviews**, indicating very strong and recurrent demand across user types.
- A score of **4** was assigned where a dataset sub-category was mentioned in **10-14 interviews**, indicating strong but slightly less universal demand.
- A score of **3** was assigned where a dataset sub-category was mentioned in **5-9 interviews**, indicating moderate and context-specific demand.
- A score of **2** was assigned where a dataset sub-category was mentioned in **1-4 interviews**, indicating limited or niche demand.
- A score of **1** was assigned where a dataset sub-category was **not mentioned** in any interview, indicating no evidenced demand within the interview sample.
- No dataset sub-category was assigned a score of **0**. A minimum score of **1** was applied to all datasets to **hedge against potential blind spots in the interview sample**, recognising that some datasets may serve emerging, specialist, or under-represented user groups whose needs were not fully captured within the 55 interviews.

2) Timeliness criticality

Timeliness criticality was included as a core scoring dimension to reflect interview evidence that the value of an API increases significantly as update frequency increases. Across stakeholder interviews, users consistently emphasised that high-frequency datasets are materially more useful when accessed via APIs, as they support automation, near-real-time decision-making, and integration into operational systems. By contrast, low-frequency datasets can often be accessed effectively through periodic downloads with limited loss of value.

Timeliness criticality therefore captures the extent to which API delivery meaningfully improves usability relative to alternative publication methods, based on how frequently the underlying data is updated.

Each dataset sub-category was assigned a timeliness score on a five-point ordinal scale, based on its typical update frequency:

- A score of **5** was assigned to **live or near-real-time** datasets.
- A score of **4** was assigned to datasets updated **daily**.
- A score of **3** was assigned to datasets updated **weekly**.
- A score of **2** was assigned to datasets updated **more frequently than annually** (e.g., monthly, quarterly, or biannually).
- A score of **1** was assigned to datasets updated **annually**.

This scoring reflects the principle that **API provision delivers the greatest marginal benefit where timeliness is critical**, and progressively less additional value as update frequency decreases. As with demand evidence, timeliness criticality was not used in isolation but combined with other scoring dimensions to inform overall prioritisation.

3) GDS alignment

GDS alignment was included as a scoring dimension to assess the extent to which datasets are already structured in a way that aligns with established public sector API standards, and therefore how readily they can be exposed via an API with minimal additional transformation. GDS guidance emphasises the importance of following existing API structures and open standards to support consistency, reuse, and ease of integration.

To maintain consistency across the framework, GDS alignment was scored on the same 1-5 ordinal scale used for other criteria. Each dataset sub-category was assessed against four core characteristics commonly shared by well-established public sector APIs:

- Openness - whether the dataset is published as open rather than closed or restricted
- Form of information - whether the dataset is structured (e.g., spreadsheet or table) rather than unstructured (e.g., narrative documents or PDFs)
- Licensing - whether the dataset is published under the Open Government Licence (OGL)
- Data format - whether the dataset is available in machine-readable formats typically used in APIs (e.g., JSON, XML, CSV)

Scores were assigned as follows:

- A score of **5** was assigned where the dataset **aligned with all four** characteristics, indicating strong GDS alignment and high readiness for API delivery.
- A score of **4** was assigned where the dataset **diverged** from the benchmark on **one** characteristic.
- A score of **3** was assigned where the dataset **diverged** on **two** characteristics.
- A score of **2** was assigned where the dataset **diverged** on **three** characteristics.
- A score of **1** was assigned where the dataset **diverged on all four** characteristics, indicating low alignment with established API standards.

Lower scores indicate that additional standardisation, restructuring, or policy alignment would be required before API development could be efficiently undertaken. As such, GDS alignment functions as a feasibility and sequencing indicator within the prioritisation framework, rather than a measure of user demand or intrinsic dataset value.

4) Dataset size

Data size was included as a scoring dimension to capture the extent to which API delivery reduces user burden relative to file-based access. Interview evidence and benchmarking both indicate that very large datasets, particularly those that are high-frequency, granular, or longitudinal, are significantly harder for users to access, store, and process when published as static files. In these cases, APIs provide substantial additional value by enabling incremental access, automation, and selective retrieval.

By contrast, very small datasets (for example, a single-page PDF or a small static table) generally do not benefit materially from API delivery, as the cost of integration can outweigh the usability gains.

Data size therefore acts as a **proxy for handling complexity**, rather than a precise measure of file size alone, and was scored based on the typical volume, granularity, and update structure of the dataset.

Each dataset sub-category was assigned a data size score on a **five-point ordinal scale**, as follows:

- A score of **5** was assigned to **very large datasets**, characterised by extensive, high-frequency, or highly granular data (e.g., continuous feeds, event-level records, or large time-series where API access materially improves usability). In practice, “very large” denotes datasets that are impractical to distribute or work with as bulk downloads (e.g., streaming/near-real-time or routinely refreshed at event level), making API access effectively necessary.

- A score of **4** was assigned to **large datasets** with substantial volume or longitudinal depth, where APIs enable filtering, pagination, or incremental updates that significantly reduce user handling effort. Here, “large” denotes datasets that remain feasible as bulk files but typically exceed standard download-and-open workflows (e.g., multi-GB files), so API based querying materially improves usability.

- A score of **3** was assigned to **moderate-sized datasets**, typically structured tables that are manageable as files but would benefit from API access for automation or repeat use.

- A score of **2** was assigned to **small datasets**, such as limited tables or infrequently updated files, where API access provides only marginal additional benefit.

- A score of **1** was assigned to **very small datasets**, such as static documents or single-page PDFs, where API delivery offers little practical advantage.

This scoring explicitly prioritises **very large datasets**, reflecting the principle that API conversion should focus first on cases where it delivers the greatest reduction in user effort and the greatest improvement in accessibility and scalability.

(Dataset size is not included in this iteration of our scoring, as we discuss below).

5) Data quality

Data quality was included as a scoring dimension to reflect the finding from interviews that poor data quality can materially limit the value of API access, regardless of demand or timeliness. In several cases, stakeholders indicated that issues such as incompleteness, inconsistency, or lack of trust in the data reduced their willingness to integrate datasets into automated workflows. As a result, data quality was treated as a gating factor rather than a direct driver of prioritisation.

To maintain consistency with the wider framework, data quality was scored on the same 1-5 ordinal scale used for other criteria. However, unlike demand or timeliness, higher scores indicate fewer reported quality concerns, and therefore greater readiness for API delivery.

Data quality scores were derived from interview evidence by counting the number of times quality issues were raised in relation to each dataset sub-category:

- A score of **1** was assigned where **data quality issues were mentioned five times or more**, indicating persistent or widely recognised quality concerns.

- A score of **3** was assigned where **data quality issues were mentioned fewer than five times,** indicating some concerns but not to a degree that clearly dominates user experience.

- A score of **5** was assigned where **data quality issues were not mentioned at all** in interviews, indicating no evidenced quality barriers to use within the sample.

This scoring approach ensures consistency across criteria while clearly signalling that **lower data quality scores reduce the suitability of a dataset for immediate API prioritisation.** Where a dataset scores poorly on data quality, the recommended priority is to address underlying quality issues first, with API development treated as a **secondary or follow-on intervention** rather than the primary response.

Scoring approach

To combine the individual scoring dimensions into a single prioritisation metric, the framework uses a weighted sum approach. This enables consistent comparison across dataset sub-categories while reflecting the relative importance of different considerations identified through interview evidence, benchmarking, and policy guidance.

Each dataset sub-category receives a score on a 1-5 ordinal scale for each included dimension. These scores are multiplied by their respective weights and summed. The resulting total is then normalised and presented as a score out of 100, providing an intuitive and easily interpretable prioritisation metric.

Weighting rationale

Weights were assigned to reflect the relative contribution of each dimension to the likely value and feasibility of API development:

- **Demand evidence (40%)**
  Demand carries the highest weight, reflecting the principle that API investment should be led by demonstrated user need. Datasets frequently referenced by stakeholders are more likely to achieve uptake and deliver sustained value once exposed via an API.

- **Timeliness criticality (20%)**
  Timeliness captures the degree to which API delivery materially improves usability compared to file-based access. Interview evidence consistently highlighted high-frequency datasets as priority candidates for APIs, justifying a substantial but secondary weighting.

- **Data quality (20%)**
  Data quality functions as a readiness and sequencing signal. Persistent quality issues can undermine the value of API access, regardless of demand. Assigning a material weight to this dimension ensures that prioritisation reflects practical deliverability.

- **GDS alignment (20%)**
  GDS alignment reflects structural and policy readiness for API delivery.

Datasets already aligned with open standards and established API patterns can typically be implemented more efficiently and with lower delivery risk.

Exclusion of data size in the current iteration

At this stage, we have excluded Data Size from our weighted scoring. This is due to incomplete metadata coverage (see Table A.2). Including data size at this stage would risk skewing results, as size information is currently available for only a small proportion of datasets.

Data size remains an important contextual consideration, particularly for very large or high-frequency datasets, and is intended to be reintroduced into the scoring framework in future iterations as data coverage improves.

Interpretation of scores

The final prioritisation score, expressed out of 100, should be interpreted as an indicative signal of relative priority rather than a definitive investment decision. The scoring framework is designed to surface high-value candidates for API development, highlight datasets requiring preparatory work (such as data quality improvement), and support transparent, evidence-led decision-making.

Current Limitations

While the scoring framework provides a structured and evidence-led basis for prioritising datasets for API development, several limitations should be considered when interpreting the results. These limitations reflect both data availability constraints and the methodological choices made to ensure transparency and consistency.

Incomplete coverage of data size information

The most significant limitation relates to the data size scoring component. As shown in Table A.2, data size information was available for only 5.88% of datasets in the mapped database. This reflects gaps in publicly documented metadata rather than a judgement about the intrinsic importance of dataset size. The practical implication is that the current prioritisation results, particularly those presented in the Appendix, should be treated as provisional. As data size metadata is improved and coverage increases, scores for this dimension may change, which in turn may alter the relative ranking of some datasets. The framework is therefore designed to be iterative, with the expectation that prioritisation will be revisited as data gaps are filled.

Reliance on qualitative interview evidence for demand and quality

Demand evidence and data quality scores are derived from interview mention frequency, which captures expressed user needs and concerns within the interview sample rather than observed usage or revealed demand. While this approach

provides a transparent and defensible proxy for relative importance, it may under-represent datasets that serve emerging, specialist, or less vocal user groups, or datasets whose value becomes apparent only when combined with other data sources.

To mitigate this risk, the framework avoids zero-scoring for demand and uses ordinal bands rather than precise thresholds. Nonetheless, results should be interpreted as indicative of relative priority, not definitive measures of future API traffic or economic value.

Data quality treated as a readiness signal rather than an outcome

Data quality is intentionally framed as a readiness and sequencing indicator, not as a direct driver of priority. While this reflects interview evidence that poor quality undermines API value, it also means that some high-demand datasets may appear lower in the prioritisation despite their strategic importance. In these cases, a low data quality score should be interpreted as a signal that quality improvement should precede or run alongside API development, rather than as a recommendation to deprioritise the dataset entirely.

Appendix 5. Further details on prioritisation bands

Band 1 (80+): highest priority datasets

Datasets scoring 80+ (inclusive):

- Roadworks service API (Street Manager) (92) (already an API)
- Bus Open Data - Published bus timetables (80) (already an API)
- Daily domestic transport use by mode (80)
- Large Goods Vehicle (LGV) vocational driving tests (80)

How this band aligns with interviewees' priorities:

Band 1 contains a small number of datasets, but it strongly reflects the top interview priorities: 50% (2 of 4) fall under the two most frequently mentioned priority types -

Buses/Public Transport Operations (12 mentions) and Road Events/Restrictions/Disruption (6 mentions).

The presence of already-live APIs (Street Manager, BODS timetables) provides an additional validation signal that the scoring is surfacing dataset types that are both feasible and valued.

Interview insight

Buses and Public Transport Operations:
BODS was consistently described as a core national feed for bus operations because it provides the basics many services depend on: timetables and real-time vehicle locations. However, interviewees emphasised that its practical value is often constrained by coverage gaps, refresh performance, and underlying data quality, with some teams needing significant “scrubbing” and, in some cases, blending BODS with alternative sources for schedules. Several also pointed to “next level” needs: faster location refresh, plus bus performance and historical trajectory data, noting that these are either not available via the API or are more restricted, despite being important for evaluation and operational insights.

Road Events, Restrictions and Disruption:
Interviewees prioritised API development for road events, restrictions, and disruption because current provision is fragmented and hard to use operationally. They described reliance on scraping operator websites, gaps in locally managed unplanned road-closure data, and disruption feeds that are patchy and inconsistently implemented. They also highlighted Street Manager as an example where the current “push” approach can require users to build and maintain their own infrastructure before they can query the data, reinforcing demand for standardised, queryable APIs that reduce integration effort and improve consistency downstream.

Band 2 (70-79): high priority datasets

Datasets scoring 70-79:

- Bus Open Data - Published bus locations (76) (already an API)
- National Coach Dataset (76)
- Traffic Flows by Borough (already an API) (76)
- Transport Statistics API (76) (already an API)
- Vehicle licensing statistics: July to September 2021 (76)
- National public transport access nodes (NaPTAN) (76) (already an API)
- Road traffic statistics (TRA) (76) (partially available via API)

- Bus Open Data - Bus Fares datasets (72) (already an API)
- Electric vehicle charging infrastructure statistics (72)
- Electric vehicle public charging devices January 2022 (72)
- Licensed Vehicles - Type, Borough (72)
- Walking and Cycling by Borough (72)
- Quarterly traffic estimates (TRA25) (72)
- Vehicles involved in reported road accidents (RAS20) (72)
- Reported road accidents, vehicles and casualties tables for Great Britain (72)
- Reported road accidents (RAS10) (72)
- National Travel Survey: 2024 (72)
- Transport Statistics Great Britain (72)

How this band aligns with interviewees' priorities:

- 22% (4 of 18) of Band 2 datasets fall under the most frequently mentioned interview priority type - Buses and Public Transport Operations (12 mentions). These are: BODS bus locations; National Coach Dataset; NaPTAN; BODS bus fares - and three of these four are already APIs, reinforcing the "follow proven patterns" narrative.
- 39% (7 of 18) of Band 2 datasets fall under the top three interview priority types combined (Buses/Public Transport Ops; Road Events/Disruption; Road Traffic Demand, noting that some road traffic data is already available via API). This reflects the concentration of higher-scoring datasets in operational and high-utility areas that stakeholders explicitly prioritised.
- More broadly, Band 2 also includes substantial representation of categories that interviewees prioritised next - Road Traffic Demand (5 mentions) and Safety/Travel Statistics (3 mentions each) - which is consistent with the framework's weighting toward demand and timeliness.

Interview Insight

Road Traffic Demand:
Interviewees flagged road traffic demand (e.g., car usage, traffic levels) as an area of recurring demand for API-enabled access. With most road traffic data already available via API, they framed the priority less as creating new datasets and more as making existing road datasets easier to query, refresh, and combine with other sources, reducing reliance on manual, time-limited downloads and enabling more

timely policy and operational analysis. One interviewee noted that high-granularity traffic flow data (e.g., at the level of specific roads) would be more actionable than general national or borough averages.

Travel Behaviour:
Interviewees also pointed to the National Travel Survey (NTS) as a strong candidate for API-enabled access. One stakeholder noted that, while the existing DfT NTS outputs are useful, an API that preserves the integrity of the dataset while allowing users to interrogate it in more flexible ways could unlock additional insights.

Band 3 (60-69): medium-high priority datasets

Band 3 contains datasets that remain credible candidates for API delivery, but where the marginal benefit of "API-first" delivery is typically **less immediate** than Bands 1-2. In practice, this band is dominated by **analytical and monitoring datasets** (notably safety and transport statistics) and a smaller number of operationally relevant measures (e.g., travel time measures), where value is often driven by consistency, queryability, and ease of reuse rather than true real-time integration.

Datasets scoring 60-69:

- Road Safety Data (Collisions) (68)
- Road Safety Data (Vehicles) (68)
- Travel time measures for the Strategic Road Network and local 'A' roads (68)
- Taxi and private hire vehicle statistics, England (68)
- Developing faster indicators of transport activity (64)
- Journey time statistics (Up to 2019) (64)
- Light rail and tram statistics, England: year ending March 2024 (64)
- Bus statistics data tables (64)
- Walking and cycling statistics (64)
- Vehicles statistics (collection) (64)
- Quarterly vehicle speed compliance tables (SPE25) (64)
- Road freight: domestic and international statistics (64)
- Road Safety Data (Casualties) (60)
- Rail factsheet: 2024 (60)

- Road lengths in Great Britain: 2024 (60)
- Rail passenger numbers and crowding on weekdays in major cities in England and Wales (60)
- Haulier coronavirus testing statistics (60)

How this band aligns with interviewees' priorities:

The composition of Band 3 aligns with the second-tier priorities identified in interviews. Stakeholders explicitly prioritised Safety and compliance and travel behaviour and survey statistics (both 3 mentions), which is reflected in the prominence of road safety and broader transport statistics within this band. The presence of travel time and journey time datasets also aligns with the wider Road Traffic Demand theme (5 mentions), albeit these items are typically less "live" than the operational disruption and public transport feeds that dominate Bands 1-2.

Band 3 is still largely populated by categories that interviewees did prioritise. Only 18% (3 of 17) of datasets in this band fall into categories that received 0 mentions in the explicit prioritisation question (Table 6). These are:

- **Freight and logistics (0 mentions)**: Road freight: domestic and international statistics (64); Haulier coronavirus testing statistics (60). However, this may partly reflect limited freight representation in the interview sample (as noted in the methodology).
- **Road infrastructure (0 mentions)**: Road lengths in Great Britain: 2024 (60)

These items are noteworthy **outliers**: they score relatively strongly on the composite metric, despite their dataset types not being prioritised explicitly for APIs in interviews. This suggests that their scoring is being supported by other dimensions (e.g., structural readiness or fewer reported quality concerns), but that the case for API delivery may require **additional validation** (targeted user discovery, confirmation of timeliness needs, and clarity on intended API use cases) before sequencing them ahead of higher-salience categories.

Overall, Band 3 represents a set of datasets where API delivery is likely to add value, particularly by improving **consistency, discoverability, and machine-readable access**, but where the immediate operational case is generally weaker than Bands 1-2. As such, Band 3 is best suited to **sequenced delivery** (or incorporation into broader programme work on standardisation and statistical query access), rather than being treated as the first tranche of API investment.

Interview insight

Some interviewees pointed to road safety data (including STATS19) as a strong candidate for API-enabled access because many use cases depend on pulling targeted, record-level subsets (e.g., accidents in a given geography and time window) rather than downloading the full dataset. One stakeholder said an API "would be great" for this kind of dynamic search, while noting that sensitive "police cut" fields would likely require a tiered/privileged access route rather than being fully open.

Band 4 (50-59): lower-medium priority datasets

Band 4 contains a large number of datasets that typically score **lower on timeliness criticality** (often annual/periodic publications) and/or have more limited evidence of immediate API-specific value. This band includes several clusters in dataset types that interviewees did **not** prioritise for APIs, alongside a smaller set of datasets that sit within prioritised themes but are less time-critical or more publication-style outputs.

Datasets scoring 50-59:

- Road traffic estimates in Great Britain: 2024 (56) (partially available via API)
- Disabled parking badges (56)
- Road Casualties by Severity (56)
- Casualties involved in reported road accidents (RAS30) (56)
- Disabled parking (Blue Badge) scheme statistics (56)
- Vehicle excise duty evasion statistics (52)
- Crime Survey for England and Wales: self reported road offences (52)
- Reported road casualties Great Britain: e-Scooter factsheet (52)
- Reported road casualties Great Britain: fatal casualties (52)
- Reported road casualties Great Britain: Injury-based factsheet (52)
- Reported road casualties Great Britain: motorcyclist factsheet (52)
- Reported road casualties Great Britain: older adults factsheet (52)
- Reported road casualties Great Britain: pedal cyclists factsheet (52)
- Reported road casualties Great Britain: pedestrians factsheet (52)

- Road safety factors initial analysis (52)
- Seatbelt and mobile phone use surveys (52)
- Reported road casualties Great Britain: casualty-based factsheet (52)
- Sea passenger statistics, all routes: 2024 (52)
- Port freight quarterly statistics: January to March 2024 (52)
- Seafarers in the UK Shipping Industry: 2024 (52)
- Shipping fleet statistics: 2024 (52)
- Domestic road freight statistics (52)
- Road goods vehicles travelling to Europe (52)
- Van Statistics (52)
- British social attitudes survey: 2017 (52)
- Disability, accessibility and Blue Badge statistics (DPTAC) (52)
- Renewable Transport Fuel Obligation (RTFO) statistics (52)
- Port freight annual statistics: 2024 (52)
- Sustainable Aviation Fuel (SAF) Mandate 2025: consultation response (52)

Justification using interviewees' priorities:

- Band 4 still includes some datasets linked to categories that interviewees prioritised (e.g., **Road Traffic Demand** and **Safety and compliance**), but these are largely periodic statistical publications rather than operational feeds - consistent with lower timeliness scores.
- A substantial portion of Band 4 is made up of dataset types that were not prioritised explicitly for APIs by interviewees (0 mentions), particularly **maritime/ports, accessibility/disability, freight/logistics, and environment/fuels**.
- 41% (12 of 29) datasets in Band 4 fall into dataset types that received 0 mentions in the explicit prioritisation question.
- Notable outlier clusters (relative to Table 6) include:

- **Accessibility/Blue Badge** datasets (multiple entries scoring 52-56) despite Accessibility and disability receiving 0 mentions.
- **Maritime/ports** datasets (multiple entries at 52) despite Maritime and ports receiving 0 mentions.
- **Environment/fuels policy outputs** (RTFO; SAF consultation response) despite Environment and fuels receiving 0 mentions.
- These clusters suggest areas that may be structurally "publishable" but where the case for API investment is currently weak without additional user validation or a strategic policy driver.

Band 5 (30-49): low priority datasets

Band 5 represents the lowest-prioritised datasets in the current iteration. They are characterised by weaker evidence of API-specific value under the framework, typically reflecting lower timeliness criticality, narrower expressed demand, and/or alignment with dataset types that stakeholders did not identify as priorities when asked directly (Table 6). These bands therefore form the clearest set of candidates for longer-term roadmap consideration or targeted further discovery, rather than near-term API investment.

Datasets scoring 30-49:

- Concessionary travel statistics: year ending March 2024 (48)
- National Travel Attitudes Study (Waves 1-10) (48)
- International road freight statistics (48)
- Linking police and fire road collision data (44)
- Public attitudes towards transport (44)
- Public attitudes towards train services: 2018 (44)
- Air passenger experience of security screening at UK airports (44)
- Road conditions in England to March 2024 (44)
- Vehicle speed compliance statistics for Great Britain: 2024 (40)
- Reported road casualties Great Britain: road user risk (40)
- Reported road casualties Great Britain, annual report (40)
- Reported road casualties Great Britain: involvement rates (40)

- Reported road casualties in Great Britain, involvement by road user (40)
- Serious e-scooter casualties: comparing police and hospital data (40)
- Transport and environment statistics (40)
- Search and rescue helicopter statistics (40)
- Road conditions (36)

Justification using interviewees' priorities:

- A number of datasets in this combined band relate to categories that were mentioned as API priorities (e.g., Safety and compliance and Travel behaviour/survey statistics, each 3 mentions). However, within Bands 5-6 these tend to be lower-frequency publications (annual reports, attitudes studies, or analytical outputs) where the incremental benefit of APIs is less immediate than for operational or high-frequency datasets.
- Several datasets map to categories that received 0 mentions in the explicit prioritisation question, particularly aviation, environment and fuels, freight and logistics, and road infrastructure, which is consistent with their placement in the lowest scoring bands.

Share of "0-mention" categories and outliers:

- **35% (6 of 17)** datasets in Band 5 fall into dataset types that received 0 mentions in the explicit prioritisation question (Table 6). These introduce clear outliers relative to stakeholder-stated priorities:
- **Freight and logistics (0 mentions)**: International road freight statistics (48)
- **Aviation (0 mentions)**: Air passenger experience of security screening (44); Search and rescue helicopter statistics (40)
- **Environment and fuels (0 mentions)**: Transport and environment statistics (40)
- **Road infrastructure (0 mentions)**: Road conditions in England to March 2024 (44); Road conditions (36)

These outliers are informative rather than contradictory: they indicate datasets that may be relatively "publishable" in structural terms, but where there is currently limited evidence from the 55 interviews that API delivery is the most urgent intervention. This should not be read as an indication that these datasets are of no value; rather, they appear lower-salience for API delivery within the current evidence base.

Appendix 6. DfT data prioritisation

Table A.3: Dataset prioritisation

Index	DfT Dataset	Score	Currently an API
1	Roadworks service API (Street Manager)	92	Yes
2	Bus Open Data - Published bus timetables	80	Yes
3	Daily domestic transport use by mode	80	No
4	Large Goods Vehicle (LGV) vocational driving tests	80	No
5	Bus Open Data - Published bus locations	76	Yes
6	Road traffic statistics (TRA)	76	Partial
7	National Coach Dataset	76	No
8	Traffic Flows by Borough	76	Yes
9	Transport Statistics API	76	Yes
10	Vehicle licensing statistics: July to September 2021	76	No
11	National public transport access nodes (NaPTAN)	76	Yes
12	Bus Open Data - Bus Fares datasets	72	Yes
13	Electric vehicle charging infrastructure statistics	72	No
14	Electric vehicle public charging devices January 2022	72	No
15	Licensed Vehicles - Type, Borough	72	No
16	Walking and Cycling by Borough	72	No
17	Transport Statistics Great Britain	72	No
18	Quarterly traffic estimates (TRA25)	72	No
19	Vehicles involved in reported road accidents (RAS20)	72	No
20	Reported road accidents, vehicles and casualties tables for Great Britain	72	No
21	Reported road accidents (RAS10)	72	No
22	National Travel Survey: 2024	72	No
23	Road Safety Data (Collisions)	68	No
24	Road Safety Data (Vehicles)	68	No
25	Travel time measures for the Strategic Road Network and local ‘A’ roads: July 2024 to June 2025	68	No
26	Taxi and private hire vehicle statistics, England	68	No
27	Developing faster indicators of transport activity	64	No
28	Journey time statistics (Up to 2019)	64	No
29	Light rail and tram statistics, England: year ending March 2025	64	No
30	Bus statistics data tables	64	No

31	Walking and cycling statistics	64	No
32	Vehicles statistics (collection)	64	No
33	Quarterly vehicle speed compliance tables (SPE25)	64	No
34	Road freight: domestic and international statistics	64	No
35	Road Safety Data (Casualties)	60	No
36	Rail factsheet: 2024	60	No
37	Road lengths in Great Britain: 2024	60	No
38	Rail passenger numbers and crowding on weekdays in major cities in England and Wales: 2024	60	No
39	Haulier coronavirus testing statistics	60	No
40	Road traffic estimates in Great Britain: 2024	56	Partial
41	Disabled parking badges	56	No
42	Road Casualties by Severity	56	No
43	Casualties involved in reported road accidents (RAS30)	56	No
44	Disabled parking (Blue Badge) scheme statistics	56	No
45	Vehicle excise duty evasion statistics	52	No
46	Crime Survey for England and Wales: self reported driver behaviour	52	No
47	Reported road casualties Great Britain: e-Scooter factsheet	52	No
48	Reported road casualties Great Britain: fatal 4 factsheet	52	No
49	Reported road casualties Great Britain: Injury severity within injury-based reporting systems	52	No
50	Reported road casualties Great Britain: motorcyclist factsheet	52	No
51	Reported road casualties Great Britain: older and younger driver factsheets	52	No
52	Reported road casualties Great Britain: pedal cyclist factsheet	52	No
53	Reported road casualties Great Britain: pedestrian factsheet	52	No
54	Reported road casualties Great Britain: casualties and deprivation factsheet	52	No
55	Road safety factors initial analysis	52	No
56	Seatbelt and mobile phone use surveys	52	No
57	Port freight annual statistics: 2024	52	No
58	Sea passenger statistics, all routes: 2024	52	No
59	Port freight quarterly statistics: January to March 2025	52	No
60	Seafarers in the UK Shipping Industry: 2024	52	No
61	Shipping fleet statistics: 2024	52	No
62	Domestic road freight statistics	52	No
63	Road goods vehicles travelling to Europe	52	No
64	Van Statistics	52	No

65	British social attitudes survey: 2017	52	No
66	Disability, accessibility and Blue Badge statistics	52	No
67	Renewable Transport Fuel Obligation (RTFO) statistics	52	No
68	Sustainable Aviation Fuel (SAF) Mandate 2025: First provisional	52	No
69	Concessionary travel statistics: year ending March 2024 (revised)	48	No
70	National Travel Attitudes Study (Waves 1-10)	48	No
71	International road freight statistics	48	No
72	Linking police and fire road collision data	44	No
73	Public attitudes towards transport	44	No
74	Public attitudes towards train services: 2018	44	No
75	Air passenger experience of security screening: 2024	44	No
76	Road conditions in England to March 2024	44	No
77	Vehicle speed compliance statistics for Great Britain	40	No
78	Reported road casualties Great Britain: road user risk	40	No
79	Reported road casualties Great Britain, annual report	40	No
80	Reported road casualties Great Britain: involving driving for work	40	No
81	Reported road casualties in Great Britain, involving illegal alcohol levels	40	No
82	Serious e-scooter casualties: comparing police and hospital data	40	No
83	Transport and environment statistics	40	No
84	Search and rescue helicopter statistics	40	No
85	Road conditions	36	No

Appendix 7. Benefit taxonomy

Table A.4: Benefit taxonomy developed from User Interviews

Code	Title	Description
B1	Better policy evaluation & impact attribution	Using data/APIs to evaluate interventions, funded schemes, or policy changes (incl. before/after comparisons, benchmarking).
B2	Improved road safety insights & outcomes	Better identification of risk, collisions, casualties, and the safety impact of interventions (e.g., STATS19-type use cases).
B3	Stronger strategic planning & scenario modelling	Better forecasting, scenario testing, simulation/synthetic data, demand modelling, long-term network planning.
B4	Better operational decision-making	Using data for day-to-day operations (service control, incident response, resource deployment, operational KPIs).
B5	Stronger business cases & investment justification	Evidence for bids, business cases, scheme appraisal, prioritisation decisions (often tied to clearer visuals/communication too).
B6	Faster time-to-insight (less manual wrangling)	Reduces time spent downloading, stitching files, cleaning, reformatting, and repeated extraction.

B7	Automation & integration into systems	Enables programmatic ingestion into pipelines, dashboards, analytics tools; fewer “single analyst” bottlenecks.
B8	Scalable access to large / high-frequency datasets	Makes very large datasets and high-frequency feeds usable without brittle manual workflows.
B9	More useful granularity & richer detail	More detailed spatial/temporal/asset-level detail (e.g., 15-minute splits, finer geography, asset identifiers).
B10	Improved timeliness & responsiveness	More frequent updates, reduced publication lag, real-time / near-real-time availability where it matters.
B11	Better data quality, accuracy & trust	Higher reliability/cleanliness/completeness; fewer errors; better confidence in decisions derived from the data.
B12	Standardisation & “single source of truth”	Consistent definitions, authoritative reference, schema alignment; reduces disputes and fragmentation across sources.
B13	Better documentation, metadata & discoverability	Clearer guidance, metadata, examples, schemas, and “findability” so users can self-serve confidently.
B14	Improved interoperability & data linkage	Easier joining across datasets (common IDs, geography alignment, consistent structures), enabling richer insights.
B15	Remembered cost savings / reduced dependency	Lower spend on consultants, duplicate procurement, proprietary alternatives, repeated local rework.
B16	Better consumer-facing services & user experience	Enables passenger/consumer products: consistent journey info, fares, live service quality, reduced confusion.
B17	Accessibility & inclusion improvements	Better accessibility/capacity information and flows that support inclusive travel planning and assistance.
B18	Transparency, scrutiny & accountability	Supports open analysis, reproducibility, public trust, external scrutiny, and clearer accountability.
B19	Innovation, new products & economic growth	Enables new commercial offerings, market activity, and broader economic/public value through reuse.
B20	Research & education enablement	Lower barriers for academic/research use, replicable methods, training, open-source tooling, and skills development.