Data Standards & Governance

Rail Data Marketplace

Hayden Sutherland

July 2022

DRAFT

Rail Delivery Group

National Rail

Data Governance – high level approach

DRAFT

RDM
Rail Data Marketplace

Types of data

- Guidance on the types of data which will be appropriate for publication on RDM

Data standards

- Guidance on suitable standards for data which is published.

Data management

- Guidance on how data should be managed through its life.

Data quality

- An approach to the description of data quality

Licensing & monetisation

- Application of data licenses and guidance on when and how the commercialisation of data may be achieved.

Enforcement of minimum governance standards

- The circumstances under which RDM will take action to address a failure to meet minimum standards

Data Security

- Guidance on how the platform and data should be protected

Background

DRAFT

RDM

Rail Data Marketplace

Statements already made about RDM data standards & quality in previous project documentation (PID):

- Provide data which is transparent, and where possible presents a ‘single source of truth’, meets minimum standard of quality, and avoids a high barrier to entry for users.
- See cleaner and more useful data that conforms to agreed standards and that can be relied upon
- Build upon the work already undertaken to present a data catalogue and set standards for data format, quality and cleansing as well as an approach to orchestration as needed
- RDM will set and enforce minimum data standards and provide a feedback mechanism for consumers which promotes improved data quality. Minimum data standards will include the requirement upon data providers to ensure that datasets do not include personally identifiable information and RDM will be responsible for removing from the service any datasets which it becomes aware is breach this condition.
- Develop mechanism for open data scoring, and guidance on open data standards.
- Data should be transparent, provide single source of truth, meet minimum standard of quality, and avoid a high barrier to entry for users.
- Develop approvals guidance and framework for new datasets to ensure appropriate and high-quality data.

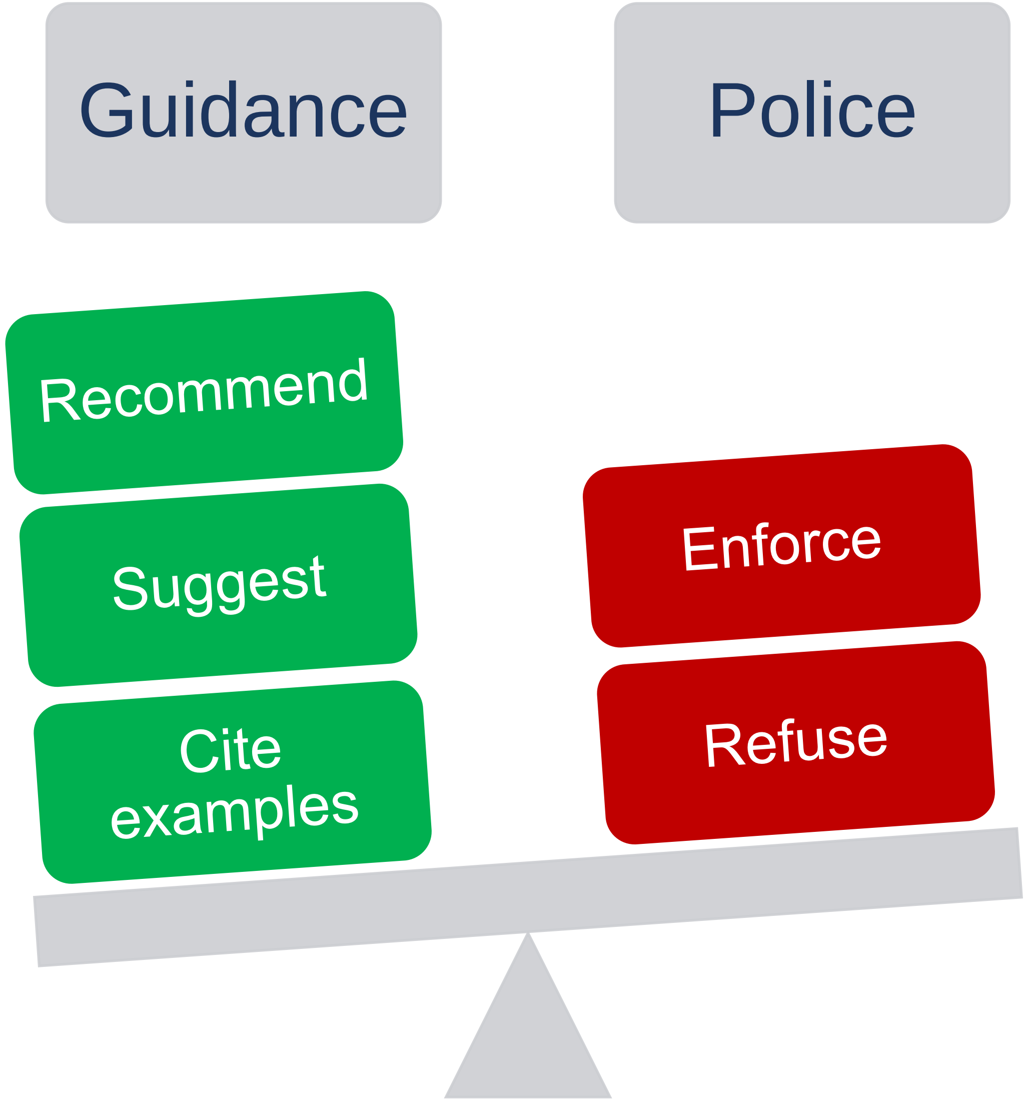
RDM's Role in Governance & Standards

DRAFT

RDM
Rail Data Marketplace

The RDM exists to bring together and make available fragmented sources of rail data into an easy-to-use digital service

RDM will provide guidance to publishers on what data sources they can publish and how they can make them **findable, accessible, interoperable and reusable** and will only enforce governance in exceptional circumstances.



Data Governance – Types of data

DRAFT

RDM
Rail Data Marketplace

- Anyone can publish data on the RDM after creating a Publisher account and agreeing to the Terms & Conditions
- The RDM service will not be prescriptive about the data published, as long as it:
  - Is their own data to publish or they have the owner's permission
  - Is data about the rail service: operations, trains, passengers, etc.
    Or has some rail sector value: events, weather, other modes, etc.
  - Must not contain personally identifiable information (PII)
- Publishers will be asked to tag their data using a publicised taxonomy for consistency.
- Publishers will be asked to provide information relating to the data source (metadata) covering such things as the type, format and ownership of data.

Note: Although RDM aims to be a 'single source of truth' – multiple data sources can be published about the same type of data (e.g. more than one organisation can publish Train Delay data, as this can be obtained in multiple ways and with different granularity / quality)

Data Governance – Data standards & specifications

DRAFT

RDM

Rail Data Marketplace

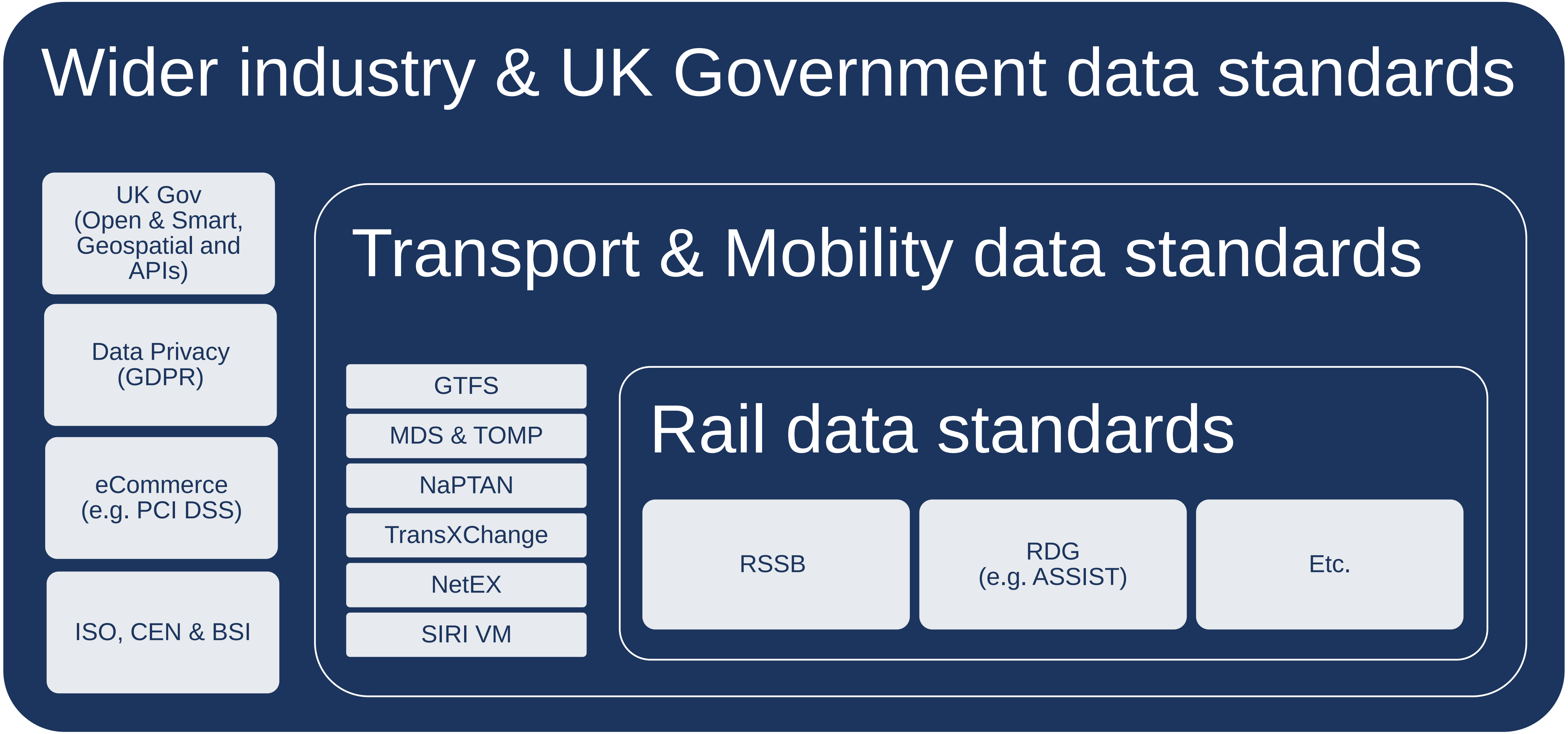
- The RDM service will NOT be prescriptive about the format or standard of the data that is published.
- However…
  - We suggest that data is published as a REST API (Unless is it either an a-synchronous stream of data or a very large payload better suited to a flat file download)
  - We recommend that common and Open Standards are used to define the data sources E.g. Open API Specification (OAS)
  - We advise that publishers review what other data sources are available on RDM & in the wider data ecosystem, and align to the standards use by those where possible (e.g. UK Gov CSV guidance in using RFC4180)
- Data Publishers will be asked to consider whether their data source should conform to a published standard relating to rail, the transport & mobility industry or Government and broader industry.
  - https://transparencee.org/analysis/data-standards-what-are-they-and-why-do-they-matter/

The Data Standards Landscape

DRAFT

RDM

Rail Data Marketplace



Wider industry & UK Government data standards

UK Gov (Open & Smart, Geospatial and APIs)

Data Privacy (GDPR)

eCommerce (e.g. PCI DSS)

ISO, CEN & BSI

Transport & Mobility data standards

GTFS

MDS & TOMP

NaPTAN

TransXChange

NetEX

SIRI VM

Rail data standards

RSSB

RDG (e.g. ASSIST)

Etc.

UK Gov : API & Geospatial Standards

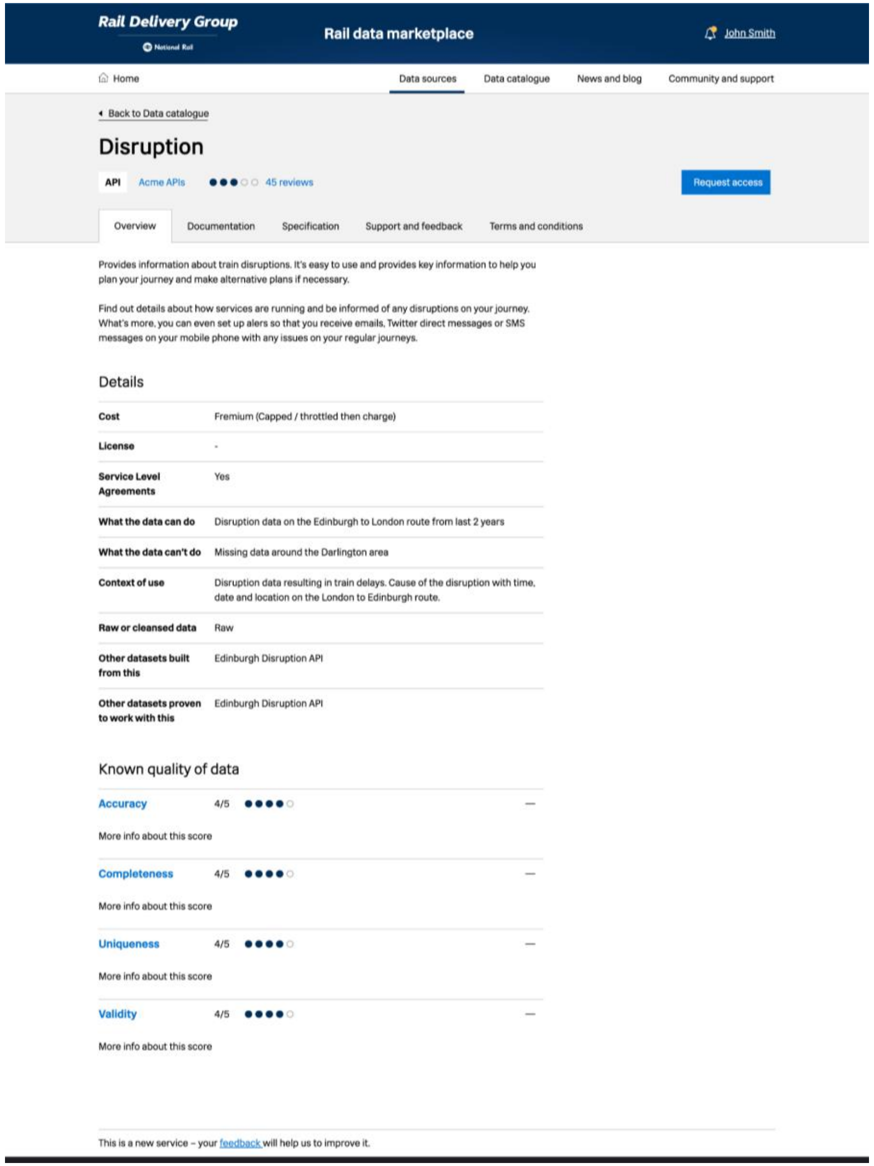
- To encourage interoperability, Data Publishers are encouraged to apply the Gov.uk web-based application programming interface (API) standards guidance, which were created to help public sector organisations deliver the best possible services to users including:
  - Using REST APIs with a JSON payload and HTTPS to secure connections
  - Using the ISO 8601 standard to represent date & time
  - Using the Unicode (UTF-8) standard for encoding text
  - API technical and data standards (v2 - 2019) - GOV.UK (www.gov.uk)
- The Open Data Institute also publish some very useful guidance into Open standards and open APIs
  - https://theodi.org/topic/open-standards-and-open-apis/
- Where appropriate, Data Publishers are also asked to consider use of the Geospatial data standards register for location-based information
  - https://www.gov.uk/government/publications/uk-geospatial-data-standards-register/national-geospatial-data-standards-register

Data Governance – Data management

- RDM will provide content and guidance on areas of data management best practice such as:
    - Ongoing management & support of data sources
    - Data accuracy, timeliness and error correction
    - Data testing and validity
    - Data security
    - Legal & regulatory compliance (e.g. user anonymisation)
    - Taxonomy & metadata classification
    - Documentation and specifications
- ONS provide some useful guidance on data structures & format
    - https://style.ons.gov.uk/category/data-visualisation/datasets/
- Data Publishers are expected to take appropriate steps to ensure that their data is managed on an ongoing basis.
- The RDM community functionality can also be used to provide peer advice & cite examples

Data Governance - Data Quality

- RDM will allow both Data Publishers and Data Consumers to independently rate the quality of data sources aligned to the UK Government Dimensions of Data Quality
    - Accuracy
    - Completeness
    - Uniqueness
    - Validity
- Data Publishers will be asked to provide information to allow Consumers to understand the likely quality of data at the point of publication and to maintain that data to ensure it remains an accurate representation of the data quality.



Disruption

Provides information about train disruptions. It's easy to use and provides key information to help you plan your journey and make alternative plans if necessary.

Find out details about how services are running and be informed of any disruptions on your journey. What's more, you can even set up alers so that you receive emails, Twitter direct messages or SMS messages on your mobile phone with any issues on your regular journeys.

Details

Cost	Fremium (Capped / throttled then charge)
License	-
Service Level Agreements	Yes
What the data can do	Disruption data on the Edinburgh to London route from last 2 years
What the data can't do	Missing data around the Darlington area
Context of use	Disruption data resulting in train delays. Cause of the disruption with time, date and location on the London to Edinburgh route.
Raw or cleansed data	Raw
Other datasets built from this	Edinburgh Disruption API
Other datasets proven to work with this	Edinburgh Disruption API

Known quality of data

Accuracy	4/5	●●●●○	—
More info about this score			
Completeness	4/5	●●●●○	—
More info about this score			
Uniqueness	4/5	●●●●○	—
More info about this score			
Validity	4/5	●●●●○	—
More info about this score			

This is a new service – your feedback will help us to improve it.

Data Governance - Data Security

- Data is an asset
  - To RDG : The User accounts are the foundation of RDM
  - To Data Publishers : Data sources are valuable (not just monetised ones)
- Data must be protected from unauthorised use & disclosure
  - In transit, at rest, and at end of life
- The UK Gov National Cyber Security Centre provide comprehensive guidance on this topic
  - 10 Steps to Cyber Security : Data security https://www.ncsc.gov.uk/collection/10-steps/data-security
- The platform we have chosen will be secure upon delivery.
  - It will not be vulnerable to any of the OWASP Top 10 and OWASP API Top 10 vulnerabilities or any other known or latent security vulnerabilities

Data Governance – Licensing & Monetisation

DRAFT

- Data Publishers will be required to state any conditions of use of the data and ensure that an appropriate data license is available to Consumers. In the majority of cases, RDM will be able to suggest a suitable template licence agreement.
- The RDM service will facilitate charging / billing for data sources that publishers wish to commercialise.
- RDM may charge a commission on chargeable data sources or for data publication in line with the Terms and Conditions.
- Data Publishers may not make a charge for data which they have a statutory, regulatory, or contractual obligation to make freely and publicly available unless they can clearly demonstrate how they have added value to that data.
- RDM will NOT set specific rates for monetised data sources although suggestions on charging & licensing can be provided. It is the publisher who will set the rates.
- Consumers who pay for data are likely to have higher expectations on standards, quality and service levels

Data Governance – Enforcement of minimum governance standards

- RDM reserves the right to remove data sources where:
  - It breaches RDM Terms & Conditions
  - OR Where an appropriate level of data governance is manifestly absent,
  - OR Where it consistently fails to meet user expectations
  - AND the Data Publisher has been made aware of these shortcomings and has failed to address them in a timely manner.
- RDM reserves the right to immediately remove any data sources which contains personally identifiable data or breaches GDPR regulations.
- RDM reserves the right to remove any comment, post, document or rating which is untrue, defamatory or offends common decency.
- RDM reserves the right to deny access of a Data Consumer to a data source where there is evidence that the Consumer is breaching the terms and conditions of use.