**Background**

**The feeds**

Network Rail (NR) provides the following operational data feeds[[1]](#footnote-1):

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Acronym** | **Description** | **Frequency** |
|  | **BPLAN** | Train planning data, including locations and sectional running times. | Twice a year |
| Codes for Operations, Retail, & Planning - a Unified Solution | **CORPUS** | Location reference data. | Monthly |
|  | **MOVEMENT** | Train positioning and movement event data. | Real-time |
| Real-Time Public Performance Measure | **RTPPM** | Performance of trains against the timetable, measured as the percentage of trains arriving at their destination on-time. | One message per minute |
|  | **SCHEDULE** | Daily extracts and updates of train schedules from the Integrated Train Planning System (ITPS). | Overnight each night |
|  | **SMART** | Train describer berth offset data used for train reporting. | Monthly |
| Train Describer | **TD** | Train positioning data at a signalling berth level. | Real-time |
| Temporary Speed Restrictions | **TSR** | Details of temporary reductions in permissible speed across the rail network. | Once a week on a Friday morning |
| Very Short Term Plan | **VSTP** | Train schedules created via the VSTP process | Real-time |

Three NR feeds are used in this dissertation: CORPUS, MOVEMENT, and SCHEDULE.

Darwin provides schedule information when a train is activated, or when a new schedule, or changes to an existing schedule are received. Schedules cover the complete journey of a single train are always sent in full.

**Choo-choo: TD, TRUST, and Darwin**

The TD feed provides information about the position of trains through a network of *berths*. A berth usually represents a signal. TD was discarded as too low-level for this project.

Train Running Under System TOPS (TRUST) is a NR system used for monitoring the progress of trains and tracking delays in the UK.

Darwin uses both TRUST and TD for real-time data, and also incorporates Darwin workstations, Customer Information Systems (CIS), and internal messaging systems.

There are three train movements. TD feeds into TRUST, and TRUST into DARWIN.

The MOVEMENT feed contains data from TRUST.

# Perhaps this should just be a cold overview. I’m figure it out later. I can always re-integrate the content. I’d rather have a background phase here.

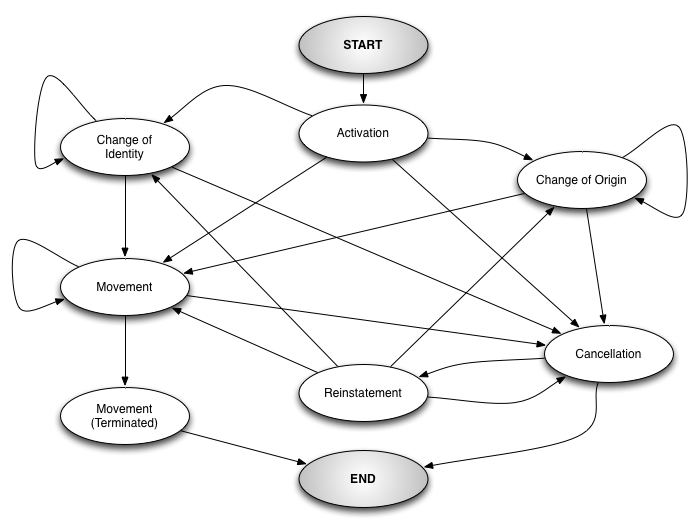


Figure 1. A finite state machine showing the order in which TRUST messages may be received.

|  |  |  |
| --- | --- | --- |
| **Type** | **Name** | **Description** |
| 001 | Train activation |  |
| 002 | Train cancellation |  |
| 003 | Train movement |  |
| 004 | N/A |  |
| 005 | Train reinstatement |  |
| 006 | Change of origin |  |
| 007 | Change of identity |  |
| 008 | Change of location |  |

**Geolocating trains**: **CORPUS (and NaPTAN)**

The Rail Delivery Group (RDG) is in the process of developing a Locations Proof of Concept[[2]](#footnote-2) which unifies multiple RDG and NR location services. However, at time of writing, this is not operational.

CORPUS is location reference data. It is used in conjunction with the National Public Transport Access Node (NaPTAN) database, a nationwide system for uniquely identifying all points of access (nodes) to public transport in the UK. Only rail stations are, naturally, of concern here. A frankly alarming number of codes are used to refer to locations in the UK rail network; they are described below, along with which datasets support them.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CORPUS** | **NaPTAN** | **Description** |
| **STANOX (Station Number)** | Y |  | First two digits are the geographic area. Can refer to non-station locations such as sidings and junctions. Numbers run broadly north-to-south. |
| **UIC** | Y |  |  |
| **3ALPHA / CRS / NRS / TLC** | Y | Y | 3-character limit. Used primarily to identify stations and on seat reservation labels. |
| **TIPLOC (Timing Point Location)** | Y | Y | Relates to points used in deriving train schedules. 7-character limit. A station often has multiple TIPLOCs if it consists of multiple groups of platforms on different lines. |
| **NLC (National Location Code)** | Y |  | 6-character limit. Identifies locations on the railway. Used for retailing and accounting purposes. |
| **NLCDESC** | Y |  | A description of the NLC. No limit. |
| **NLCDESC16** | Y |  | A description of the NLC. 16-character limit. |
| **ATCO** |  | Y |  |
| **EASTING, NORTHING** |  | Y | Geographic Cartesian coordinates for a point. Uses EPSG:27700[[3]](#footnote-3), the British National Grid / Ordnance Survey system. |
| **NAME** |  | Y | The name of the location. |

Each MOVEMENT message, depending on the system, has either a STANOX or TIPLOC code included. Both CORPUS and NaPTAN are therefore necessary to ensure a message can be geolocated.

**The weather**

**Historic data**

The primary source of weather data is the Met Office Integrated Data Archive System (MIDAS). MIDAS is a database of land and marine surface observations, collected from 1853 to the present day, by the Met Office station network. MIDAS offers several datasets. The most comprehensive is the Hourly Weather Observation Data, which contains meteorological values measured on an hourly time scale. These observations include 104 fields[[4]](#footnote-4), though many are for quality control, or too specific to necessitate inclusion.

Station data is available from the Centre for Environmental Data Analysis (CEDA). There are 507 stations, but several are located overseas. Each station is geolocated by latitude and longitude.

**Forecast data**

One of the objectives of this dissertation was the application of a trained model to unseen data. Unseen train data are simply train schedules. Unseen weather data are *forecasts*. Forecasts are available from Met Office Datapoint, a service allow accessing to freely available Met Office data feeds. They may be obtained as 3-hourly site-specific forecasts up to 5 days’ in advance. There are over 5,000 UK forecast sites. There are 10 available fields for forecasts, of which only 7 could be meaningfully mapped to equivalent MIDAS fields. For several, this was a simple unit conversion. For weather type and visibility, codes were used[[5]](#footnote-5).

|  |  |  |
| --- | --- | --- |
| **MIDAS** | **Datapoint** | **Notes** |
| *prst\_wx\_id* | Weather type | MIDAS code definition depends on *src\_opr\_type.* Use mapping between relevant table and Datapoint code definition |
| *rltv\_hum* (%) | Screen relative humidity (%) | No conversion necessary |
| *visibility* (decametre) | Visibility | Used code definition |
| *air­­\_temperature* (°C) | Screen temperature (°C) | No conversion necessary |
| *wind\_direction* (degrees) | Wind direction (16-point compass) | Map degrees to compass directions, with 0 as N. |
| *wind\_speed* (knots or ms-1) | Wind speed (mph) | MIDAS unit depends on *wind\_speed\_unit\_id* |
| *q10mnt\_mxgst\_spd* (knots or ms-1) | Wind gust (mph) | MIDAS unit depends on *wind\_speed\_unit\_id* |
|  | Feels-like temperature (°C) | No equivalent |
|  | Precipitation probability (%) | Clearly nonsensical |
|  | UV Index | No equivalent |

1. https://www.networkrail.co.uk/who-we-are/transparency-and-ethics/transparency/open-data-feeds/ [↑](#footnote-ref-1)
2. https://wiki.openraildata.com/index.php?title=Locations\_PoC [↑](#footnote-ref-2)
3. https://epsg.io/27700 [↑](#footnote-ref-3)
4. https://artefacts.ceda.ac.uk/badc\_datadocs/ukmo-midas/WH\_Table.html [↑](#footnote-ref-4)
5. https://www.metoffice.gov.uk/services/data/datapoint/code-definitions [↑](#footnote-ref-5)