Response to ORR's Approach to Transparency

Peter Hicks
peter.hicks@poggs.co.uk

Background

I am the author of OpenTrainTimes, a website which seeks to improve engagement between the travelling public and the rail industry. It operates predominantly on data released by Network Rail and ATOC, showcasing positive uses of Open Data and opening the way for further datasets to be opened up and released.

The project started in early 2011 to produce an open version of National Rail Enquiries' Darwin system, following their refusal to allow what most would call fair access to developers wishing to use the data. I secured access to full timetable data from Network Rail in January 2011, and was the first person outside the industry to be allowed access to their real-time data feeds some months later.

Through my experience in IT systems, I have helped to bridge the gap between the rail domain and the needs of developers. To spread knowledge, I set up and currently curate a wiki at http://wiki.openraildata.info/ to help others understand and work with the data that Network Rail have released.

Throughout the project, I have built up a network of contacts within and outside the industry. I have presented the project and the background at several events – at OpenTech 2011, to the Mayor's Digital Advisory Board in December 2011, and more recently at the Open Knowledge Festival in Helsinki and at Loco2's 'Off The Rails' hack-day.

I am happy to discuss any of the points in this document in greater detail, as well as assist – where I can – in solving some of the technical hurdles to opening up the rail industry's data.

Q1: We would like to hear consultees' views on the content and functionality of the NRT Portal

The content within the portal is good. It appears to be aimed at the journalist or researcher looking for figures rather than a developer looking for raw data.

The ability to view the data online, as well as being able to download as a spreadsheet, is an excellent balance between raw data and making the data instantly understandable.

Q2: We are interested to hear views on what other areas of our work consultees believe should be published and why

Not answered.

Q3: We would be interested to hear consultees' views on our proposals around the publication of the results from our safety inspections and reports on the comparative performance of duty holders from our audit and inspection activities.

Not answered.

Q4: We would also be interested to hear views on the benefits and otherwise of duty holders reporting on best practice by the publication of specific KPIs

Not answered.

Q5: We would be interested to hear consultees' views as to the potential use that could be made of Network Rail historic performance data. In particular the extent to which this data provides a means by which the market, via third party developers, could meet consumer demand for real time train information products and services and/or information about performance at even more disaggregation than the current route sector publication

Historic performance data is not likely to be of much interest to the majority of developers – much of the focus is on real-time, live data. It will likely be of most interest to those wishing to perform trend analysis and monitor performance.

Network Rail could look at using long-term historical performance data to prove the value of any improvements they make to the network. Even if the immediate figures show no great improvement, the visibility of the data to the general public can be used as a focal point for improvement.

Q6: In what areas of its business could Network Rail, in your view, become more open, and what information or data would you like to see made available as a result?

Network Rail is already in a good position for furthering transparency in their business:

- The appointment of Mark Farrow as Head of Transparency, and David Higgins' internal statement that he is firmly behind openness and transparency has already created a cross-functional platform for encouraging and developing transparency and openness.

- Their decision to respond to Freedom Of Information (FOI) requests from the public via mySociety's 'What Do They Know?' website, despite not being covered under FOI legislation, is a bold and positive step.

I have already engaged directly with Network Rail on three occasions and asked them to make further specific datasets available – data they already hold in electronic format and could release without significant technical effort.

Given I am an outsider to the industry and have never worked within Network Rail, ATOC or any TOC, there are likely to be numerous other datasets which may be beneficial to release. Based purely on feedback from the developer community in recent weeks, there is demand for the following datasets:

Geospatial

- High-level centre-line data on the railway network is already available[1], however it appears to be in bitmap format and unsuitable for re-use. Developers are using data from OpenStreetMap at present

- Low-level track layout and signal positioning data, useful for detailed analysis of train movements. Network Rail already have a Signalling Systems Collaboration extranet[2], run by Mike Christelow in their IM function, which distributes data on new signaling schemes and alterations to the railway network. Extending the scope of this extranet to include current signaling reference data will be a big step to helping developers work with, and importantly, innovate with the existing data feeds

Real-time operational data

- Train formation data, such as number of carriages and class of train. Whilst not 100% reliable, it is held within systems under Network Rail's control or influence

- Data on freight train movements - currently filtered out from the Train Describer and Movements feeds on the grounds of commercial sensitivity. If not released, there is still value in performing analysis on the justification for keeping the data confidential to ensure the reasons are legitimate and robust

- The inclusion of more data in to the Train Describer feed, which is currently limited to Class 1, 2 and 9 trains. Empty train movements (Class 0, 3 and 5) are excluded. The community would like to see Class 4, 6, 7 and 8 movements specifically filtered out, with all other data allowed – the reverse of the present situation, but still filtering out freight movements

[1] http://data.gov.uk/dataset/railway-network-inspire
[2] http://td-updates.webexone.com/

- On-board energy metering data, either in real-time through the existing datafeeds platform, or aggregated on a frequent basis such as daily.

Other operational data

- Electronic versions of the Weekly and Periodical Operating Notices (WON/PON) along with Signalling Notices

- To build upon the data in the Engineering Access Statement[3], the Confirmed Period Possession Plan (CPPP). This dataset is useful for the analysis of engineering work taking place, and augments the data in the WON and PON

ORR should seek to ensure that Network Rail's ORBIS programme includes scope for identifying datasets not outlined above that may be useful and/or valuable to release. One area that would be of great interest is data from the New Measurement Train.

Q7: We are interested in hearing views on the scope of our and industry activities; whether the sector is moving in the right direction; whether the pace is right; and whether there are other areas that consultees believe would benefit from greater transparency and why

As per Q6, Network Rail is clearly working in the right direction. They are a large organisation where change does not happen quickly, but they have made several quick wins over the past year. The pace must continue, and I believe they understand that transparency is an iterative, constant process.

Trailing somewhere behind Network Rail is ATOC, who have quite some distance to go.

For a number of years, they had a publically accessible Application Programming Interface (API) in to the Live Departure Boards system, which allowed developers to write their own interfaces to existing train prediction and disruption data. After detail of the API – used by several developers for personal projects – became widely known, 'tokens' became required to use the service. Although the terms for being granted a licence to use the service are seen by ATOC as acceptable, this is not necessarily a view shared by developers[4] [5].

[3] http://www.networkrail.co.uk/aspx/3741.aspx

[4] http://www.theregister.co.uk/2010/11/03/ldb_api_atoc_alex_hewson_mystery/

[5] http://placr.co.uk/blog/2011/05/why-train-departure-information-is-not-currently-open-data/

Raw, real-time information on train departures from ATOC's Darwin system should not be locked away from the hands of those in a position to innovate. As Network Rail have proved, it is possible to distribute real-time information in a cost-effective manner without impact on existing operational systems.

One of the driving forces behind OpenTrainTimes was to launch a technical response to the lack of open real-time train information and ask questions on how to fill in the gaps.

Aside from real-time data, ATOC have other data[6] available, but priced such that it is out of the reach of the average developer. According to their document on licence fees and data charges[7] are as follows:

- An annual licence charge of £5,005 for daily or weekly distribution of data

- An annual datafeed charge of £5,525 for fares, timetable and routeing guide data

A fee in excess of £10,000 per annum prohibits any but the existing players with in the rail industry from innovating with this data.

To respond to this, ATOC should be encouraged by whatever means to reduce the charges – ideally to zero (or a figure which appropriately reflects the costs in supplying the data) – that they impose on people who want to use their data.

Aside from ATOC, individual TOCs need encouragement and a framework in which to become more transparent. It could be a condition of future franchise agreements that TOCs adhere to a minimum standards level for openness and transparency.

Even something as simple as publishing statistics on train loading – busiest trains and quietest parts of trains for example – can have a direct impact on how passengers travel by allowing them to self-manage and even out loading on peak trains. Publishing this raw data in electronic format allows it to be integrated in to other systems.

Q8: We are interested in consultees' views on the use of our statutory powers and how they believe they could be applied in the context of transparency

[6] http://www.atoc.org/about-atoc/rail-settlement-plan/data-feeds/types-of-data

[7] http://www.atoc.org/clientfiles/File/RSP%20Licence%20Fee%20&%20Datafeeds%20Charge%20FY2011-13%20v04-00.pdf

ORR is in a powerful position, and can easily help lead the way within the industry.

To encourage transparency now, it can begin or continue to talk with industry players about the importance of being open where relevant and start a culture shift to 'open by default' for new datasets. There are many experienced people in government and the SME sector who can lend their skills to help.

To encourage transparency in the future, it can ensure that appropriate and flexible mechanisms exist – such as clauses in future franchise agreements – to incentivise the industry to remain open.

Where there is resistance, ORR can intervene and engage with senior management to force change. The path to transparency is not always easy.

Q9: Presentation of the data or information is key and we would like to hear views as to the likely risks and pitfalls and how best to address them

The biggest risk to presenting data is to make it difficult to process.

Broadly speaking, there are at least three types of rail industry data – static, dynamic API and dynamic raw.

Static datasets should be offered in both raw and interpreted formats. For example, if a PDF of statistics is provided, an Excel spreadsheet – or ideally a neutral format such as CSV – should also be provided with the raw data. This helps prevent against consumers of the data having to re-key it between formats, ultimately increasing the value of the dataset.

Dynamic APIs should, wherever possible, also include the capability to download the entire dataset. Innovation happens at a faster rate, and with fewer barriers, when the entire picture can be seen. This does not devalue the API – for the majority of use cases, an API is perfect.

Dynamic raw data should always be presented in a standard and consistent manner using an open protocol such as Sparql, HTTP or Stomp, sometimes with a light level of adjustment. For example – Network Rail's Open Data platform interfaces a heavyweight enterprise messaging system with a lightweight, open platform that developers can consume.

The messages that leave Network Rail's messaging system are in a very raw format, and a light level of reprocessing takes place to transform the messages in to a format is more easily interpreted. It is then delivered via an open and easy to implement protocol to end users.

Q10: We would be interested to hear of any other initiatives in the sector or elsewhere where transparency

Not answered.

Q11: We are also interested in hearing about the risks and any unintended consequences

Not answered

Q12: Consultees views are sought on how we should go about evaluating the risks and benefits of more transparency and what factors we should take into account, including how we should measure whether our objectives for transparency are being achieved.

Looking solely at data that would be useful for developers to analyse and use, the most important metric is that the needs of the majority of those consuming the data are fulfilled.