SIRI data summary

Received data

The SIRI-VR data that we are currently receiving essentially consists of a series of `<VehicleActivity>` blocks like this:

```
<VehicleActivity>
  <RecordedAtTime>2017-09-03T23:59:51+01:00</RecordedAtTime>
  <ValidUntilTime>2017-09-03T23:59:51+01:00</ValidUntilTime>
  <VehicleMonitoringRef>SCCM-54307</VehicleMonitoringRef>
  <MonitoredVehicleJourney>
    <LineRef>X5</LineRef>
    <DirectionRef>OUTBOUND</DirectionRef>
    <FramedVehicleJourneyRef>
      <DataFrameRef>1</DataFrameRef>
      <DatedVehicleJourneyRef>467</DatedVehicleJourneyRef>
    </FramedVehicleJourneyRef>
    <PublishedLineName>X5</PublishedLineName>
    <OperatorRef>SCCM</OperatorRef>
    <VehicleFeatureRef>lowFloor</VehicleFeatureRef>
    <OriginRef>0500CCITY476</OriginRef>
    <OriginName>Parkside Bay 16</OriginName>
    <DestinationRef>0500HSTNS064</DestinationRef>
    <DestinationName>Market Sq Stop D</DestinationName>
    <OriginAimedDepartureTime>2017-09-03T23:30:00+01:00</OriginAimedDepartureTime>
    <Monitored>true</Monitored>
    <InPanic>0</InPanic>
    <VehicleLocation>
      <Longitude>-0.2354520</Longitude>
      <Latitude>52.2262192</Latitude>
    </VehicleLocation>
    <Bearing>288</Bearing>
    <Delay>PT35S</Delay>
    <VehicleRef>SCCM-54307</VehicleRef>
  </MonitoredVehicleJourney>
</VehicleActivity>
```

Individual blocks are occasionally repeated, frequently once in adjacent SIRI messages, occasionally up to 8 times. There are also occasional examples of blocks that differ *only* in reported position (in particular with the same RecordedAtTime). This represents a bus in two places at the same time, and could cause a naive speed calculation to attempt to divide by zero.

From a review of actual data on three weekdays 2017-04-26, 2017-08-30 and 2017-09-04 the following appear to be true:

RecordedAtTime

A plausible timestamp for the event. Generally a few seconds in the past relative to time of receipt, very occasionally up to 75 minutes in the past or up to 60 seconds in the future.

ValidUntilTime

Always the same as `RecordedAtTime`.

VehicleMonitoringRef

Looks plausibly to be a vehicle identifier qualified by it's operator. Always the same as `VehicleRef`. Appears to match a field on a Whippet ticket issued on the Universal. About 600 distinct values.

LineRef

Looks to be an identifier for the *Line* (e.g. timetable) to which this journey relates. Probably needs to be qualified by `OperatorRef` for uniqueness. Always the same as `PublishedLineName`. About 130 distinct values (170 whan qualified by `OperatorRef`).

DirectionRef

Always 'INBOUND" or 'OUTBOUND'.

DataFrameRef

Always '1'.

The SIRI standard says (part 2, page 117):

> Unique identifier of data frame within participant service. Used to ensure that the DatedVehicleJourneyRef is unique within the data horizon of the producer. Often the OperationalDayType is used for this purpose.

DatedVehicleJourneyRef

Integers from 1 to about 10,000, occasionally appearing with one or more leading zeros -- unclear if they should be interpreted as numbers or strings. These seem to indicate vehicle journey in some sense, and increase throughout the day resetting to 1 at midnight.

The SIRI standard says (part 2, page 117):

> A reference to the DATED VEHICLE JOURNEY that the VEHCLE is making.

[There's some suggestion that DATED VEHICLE JOURNEY and VEHCLE may be terms from the 'NeTEx data modle', itself derived from Transmodel (SIRI part 2, page 114)]

However there are multiple examples of reports for buses that are clearly on the same journey (as indicated for example by `OriginRef`, `Destinationref` and `OriginAimedDepartureTime`) but which have the same `DatedVehicleJourneyRef`.

PublishedLineName

See `LineRef`.

Defined by SIRI (part 2, page 121) as

> Name of Number by which the LINE is known to the public.

OperatorRef

One of

ATS CBLE FECS GP SCCM SCNH WP ZSIN

VehicleFeatureRef

Present in only 16% of records. If present, only ever 'lowFloor'.

OriginRef, OriginName, DestinationRef, DestinationName

One of about 420 stops.

The Refs seem to correspond to NAPTAN atcocode; the names do not necessary correspond to any of the names in NAPTAN (e.g. 02900040 which has Name 'Luton Stn Stand 10' but appears in NAPTAN as "commonname = 'Luton Station Interchange'; shortcommonname ="; landmark = 'Luton Railway Station'; street = 'Station Road'; indicator = 'Stand 10'; localityname = 'Luton').

OriginAimedDepartureTime

Defined by SIRI (part 2, page 120) as

> Timetabled departure time of VEHICLE from Origin.

This, along with `OriginRef`, `OriginName`, `DestinationRef`, `DestinationName`, is part of VehicleJourneyInfo group which seems to contain information intended for human consumption rather than

for automatic processing.

Monitored

Always 'true'

InPanic

Always '0'

Longitude

-0.755235 to 0.63038 (Milton Kenes/Kettering/Oakam to somewhere between Newmarket and Bury St Edmunds)

Latitude

52.0085564 to 52.8346291 (Milton Kenes/Safron Walden to Spalding)

Bearing

0 to 354 in steps of 6. Most values appear with similar frequency, except 0 which occurs 4 time more often than other values. SIRI standard (part 2, page 116) defines this to be in compass degrees (0-360) with north equal to 0 degrees and east to 90.

Delay

A time delta in ISO format. Positive and negative. SIRI standard (part 2, page 116) defines this to be a limited version of xsd:duration that may only contain Month, Day, Hour, Minuite, Second and Millisecond terms. Further (pary 2, page 125) to be

> Delay to a precision of seconds. Early times are shown as negative values.

VehicleRef

See `VehicleMonitoringRef`.

Extracting trips

The raw data gives bus positions. Partitioning the data by some combination of `VehicleMonitoringRef`, `LineRef`, `DirectionDef`, `DatedVehicleJourneyRef`, `OperatorRef`, `OriginRef`, `DestinationDef`, `OriginalAimedDepartureTime`, and date might be expected to result in the paths corresponding to individual timetabled 'trips' (`DataFrameRef` is

irrelavent, given that it is always '1').

An exhaustive search across actual data on three weekdays 2017-04-26, 2017-08-30 and 2017-09-04 suggests that partitioning by anything other than

```
vehiclemonitoringref, datedvehiclejourneyref, originaimeddepartureti
vehiclemonitoringref, originref, originaimeddepartureti
vehiclemonitoringref, destinationref, originaimeddepartureti
vehiclemonitoringref, lineref, datedvehiclejourneyref, originref, d
departureti)
vehiclemonitoringref, lineref, datedvehiclejourneyref, destinationr
aimeddepartureti)
```

selects unrelated trips.

Leaving out `VehicleMonitoringRef`, and selecting on (`OriginRef`, `DestinationRef`, `OriginalIntendedDepartureTime`) produces a moderate number of duplicates which all look to be errors. Two particular versions:

- Different vehicles travelling clearly separate routes but both claiming to be on the same one (e.g. two buses claiming to be running the same Cambridge Science Park --> Addenbrooke's Hospital Bus Station journey while one of them was in Peterborough).
- Different vehicles each apparently travelling on separate parts of the route making up the trip (e.g. a Bedford to Luton trip where one vehicle appears to run into Bedford bus station, a different one runs from Bedford to Wilstead, and a third from Wilstaed to Silsoe where the trip gets lost).

Once extracted, the resulting paths are by and large consistent with expected timetable trips, with the following occasional mis-feastures:

- Position dropouts during trips. These may be more common in some places than others, perhaps reflecting poor GPS and/or mobile coverage. They may also be more common on fast route sections (busway, A14, A428, A1307), though even random dropouts will also be more obvious at speed.
- [Often related to dropouts] Very-occasional widely off-route points (e.g. a point north of Ely for a bus clearly going from Fulbourn to the City Centre)
- Paths starting before the origin stop or ending after the destination. In most cases these look to be vehicles travelling to the origin or onward from the destination. In many cases these additional trips start from or end at recognisable bus depots or known layover locations.
- Paths starting after the origin stop or ending before the destination. These could be caused by position dropouts, or by the driver being late or early updating the bus's route information. For example it seems common for Universal buses to switch from 'To Addenbrokes' to 'To Edington' on approach to the hospital even though the route officially terminates/starts outside outpatients. This does mean that

trips frequently appear not to reach their destinations.

- Paths that lappearto be partial concatenations of more than one trip.

Reporting Frequency

Having identified trips based on (VehicleMonitoringRef, OriginRef, OriginalAimedDepartureTime), it's possible to analyse the frequency with which buses on a trip report their position. This appears to be predominantly every 20 seconds (95% of reports are within 21 sec of a previous report, 96% within 30 sec, 99% within 60 sec, and 99.9% within 240 seconds. The maximum observed delta is about 2.5 hours, but in all investigated cases deltas of over an hour are caused by otherwise bogus data (typically a bus advertising that it is on a journey that it can't possibly be completing - typically one it completed in the past or will be completing in the future).

The distribution shows additional small peaks at 40, 60, 80, 100 sec, etc. presumably caused by failed or lost transmission. There are also peaks at 10 and 30 sec. (but not at 50 or above) suggesting some initial transitions on 10 second intervals.