# Data requirements

## Data Source

At the start of the project Ordina pointed us to a train information archive to give us a headstart. Briefly looking into it, the archive seems to be entirely maintained by train hobbyists working together in an online group. The data is registered as open source and the only language for the data documentation or discussions thereof is Dutch.

The data in the archive is, according to the group themselves, acquired directly from the [NS API](https://www.ns.nl/reisinformatie/ns-api) and another open source archive called [NDOV Loket](https://ndovloket.nl/)

## Ideal Situation

In an ideal situation we would have enough time to collect this data ourselves, from the [NS API](https://www.ns.nl/reisinformatie/ns-api) for example. If we did this we could find additional data that was dropped by the first group for reasons unknown to us.

### Ideal Data

To properly fulfill our goal of predicting chance of delay and delay duration for Ordina employees, we should aim to make our models as accurate as possible. We should use data that is as representative of reality as possible.

To keep our data representative we should take into account all different kinds of things that can cause delays or extend delay durations. From our domain researching we know delays are caused, among other factors, by the following phenomena:

- Any kind of railroad malfunction that either prevents train travel or safety thereof

- Inclement weather such as snow or heavy storms that threaten the safety of train travel

- Large objects or people occupying the railway a train needs to go through

- Delayed trains occupying their station for too long, causing the next trains to also be delayed in a kind of avalanche effect

- By fault of the Train conductor(s)

- A lengthy process of boarding/unloading passengers from the train, for example during rush hours

To optimally predict delays our data should take the cause of delay into account for predictions