Safe transition of the coronavirus dashboard

Julian Flowers

27/01/2021

Table of Contents

[Background 1](#_Toc63021668)

[Delivery 2](#_Toc63021669)

[Users 3](#_Toc63021670)

[How it works 3](#_Toc63021671)

[Daily process 3](#_Toc63021672)

[Data delays 5](#_Toc63021673)

[Dashboard team 7](#_Toc63021674)

[Current staffing and roles 8](#_Toc63021675)

[Adding new metrics 9](#_Toc63021676)

[Risks and resilience - transition 10](#_Toc63021677)

[Strengthening rotas 10](#_Toc63021678)

[Recommendations 10](#_Toc63021679)

# Background

PHE has published daily COVID pandemic indicators since early March. The coronavirus dashboard (<https://coronavirus.data.gov.uk>) is the public facing site for daily publication of official UK figures on the COVID pandemic. It has undergone 3 major redesigns since April to accommodate increasing volumes of data, reflect the changes in the pandemic, and improve transparency, navigation, and clarity in response to user research findings.

Daily release and development is now a huge team effort requiring collaboration across several teams in PHE, NHS England, DHSC (Test and Trace), Cabinet Office, SofS Office, and the DAs.

The bulk of the development and daily effort is undertaken in PHE by a team of data scientists and analysts, digital staff, and developers. Most of the analytical resource is sourced from Health Improvement teams.

It relies heavily on NHS England infrastructure - data is stored and processed in the NHS Foundry system (provided by Palantir) and the major infrastructure is hosted on the NHE England Azure tenant.

### Delivery

To deliver the dashboard we:

* collate data from over 20 sources
* process over 500m rows of data daily to generate 100m data points which populate the website front end with > 200 metrics for 6791 MSOAs and 700 higher geographical areas[[1]](#footnote-1)
* publish data on tests, cases, deaths, healthcare metrics and vaccinations.
* present information as a series of summaries, charts (epidemic curves) and tables
* make small area data available via an interactive map which includes a slider to view trends over time.
* update the R number and growth rate weekly.
* provide localisation services via postcode searching, and map zooming.
* provide easy read reports daily for every area for which we have data
* provide extensive access to all the data via downloads, API calls, chart images

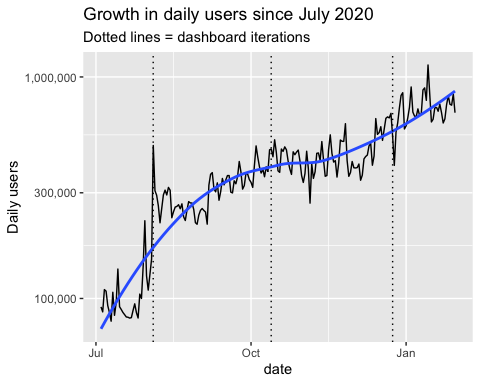
We also produce a range of data and outputs for comms, management and internal use. These include:

* Daily data summaries for FES, IDs, Epicell and other
* LSOA data (daily and weekly) for the PowerBI Portal (shared with DsPH)
* Pre-release data via email which includes daily headline metrics. This is generated automatically by the deployment process and circulated to a range of recipients including No 10, Cabinet Office, Patrick Vallance, SofS.
* Custom data feeds for the No 10 Press Briefing team
* Ad hoc analysis for internal queries, media requests, PQs and select committee report.

### Users

95% of our users are members of the general public but the service is widely used in central government including PHE, T&T and JBC; local government; the NHS; national and local media; academics, and international reporting agencies.

The site now averages over 800,000 users a day - 8 x the users in July and 80 x the users in April. 20,000 easy read documents are downloaded daily. There are over 60 million API calls (hits) every day.



## How it works

### Daily process

There are a series of daily steps.

* **Metric creation and data collection**
  + Morning process
  + Afternoon process
* **Running the pipeline**
* **Data deployment**
* Data is ingested from a wide range of data sources daily including PHE, NSHE, Test and Trace, Scottish, Welsh and Irish govt and public health agencies into the NHS Foundry System
* Data is a mix of large record datasets, spreadsheets and API calls
* Each data set undergoes a set of transformations (cleaning, aggregating) and calculations to extract the aggregate data from inputs or create metrics from the raw data. About x transformations are performed on the data
* This generated a set of outputs which are then transferred to blob storage on the NHS England Azure tenant. From there the data are uploaded to the dashboard database and made available
* Data is analysed to provide outputs for sitreps, Epicell, local authorities, DHSC and others.
* Data is released at 4 pm daily in a single deployment.
* Daily deaths (28 days)
  + The SGSS team send case data to NHS Digital batch tracing service for over night processing - this adds the fact of death. Tracing is done in batches of 250k records
  + The traced files are shared with Epicell first thing in the morning who then link additional data to generate a line list of deaths which is sent to the Foundry system - usually by 1pm
* SGSS
  + The dashboard takes several daily SGSS feeds for cases, positivity and tests during the course of the day.
  + SGSS is updated by overnight feeds from labs

Deployment is the final step completing the process of transforming data in the pipeline, creating and transferring the datafiles to the dashboard, the data processing for populating the dashboard databases, APIs (the *etl*), the map and releasing the data to the public domain.

The final step of deployment is manual requiring the entry of a token into a webpage by the G6 analyst on duty and the tech lead monitoring the environment. Data is (almost always) released in one go. This ensures that everyone sees the data at the same time and the deployment process has security to controls to prevent premature date release.

We aim to release the data at 4pm every day. The time of release relates to when there were daily press conferences which started at 4pm. At this time DHSC provided the daily data, and tweeted the daily figures and dashboard daily was released once DHSC had tweeted. This was erratic and could be delayed until well into the evening. There were only two or three people involved in the daily release process at this point which meant long hours waiting for a small team.

Once the press conferences stopped in July and the daily publication transferred from the CO and DHSC to the dashboard there was public expectation of seeing the figures at the start of the PC around 4pm. Since then we have attempted to keep to this time despite huge growth in the volume of data, volume of users, and complexity of the process.

### Data delays

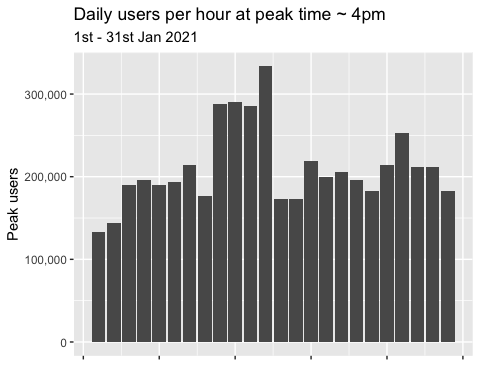
Delays in data release exemplify the challenges of daily publication and the complexity of the process, and highlight where we need to build resilience. We have been collecting data on deployment times since the beginning of August (see figure).

#### Shifting publication time

We know from user research data that delays impact public trust in the data; delays have sometimes become media stories in their own right; and we are tied back into the schedule of press conferences because we provide a custom data for feed to the No 10 press conference team. We have tried to get more leeway in the target time but senior stakeholders are very reluctant to agree.

As a consequence we constantly review the process to see where we can be more efficient and keep time. To achieve 4pm releases we need all the date to have arrived in the Foundry system by about 2:30. Data from DAs tend to be published between 2 and 2:30pm so are the last data to be entered along with the daily vaccination data. Any data delays can delay the publication process.

Because of the pattern of traffic on the site with very large spikes about 4 pm every day (see below), if we release the data at peak traffic, despite having massive resilience in the site (for example we can be running 200 servers at peak), it can crash the service so we are very reluctant to cut things fine. This also cuts the time for final QA to minutes. The G6s get perhaps 5 minutes to review the data on the development server before going live.



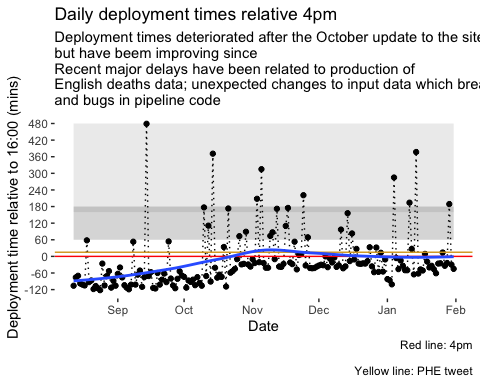
#### SGSS and daily deaths

We are very dependent of PHE colleagues in the SGSS team and Epicell who provide the various feeds of SGSS data we use, and the daily deaths data respectively.

If case data or the deaths data don’t flow properly the dashboard output can be delayed. The process of outputting SGSS and death data have been impacted by growing volumes of data. The SGSS system has experienced issues in collating overnight flows from labs such as running out of disk space, failed software updates which if not detected until the morning affect multiple daily processes.

The production of deaths data is nearing technical limits. It relies on batch tracing of case data through the DBS service provided by NHS Digital. This service looks up cases in the Personal Demographic Service to establish the fact of death in those who have died and the death has been notified. THe DBS has a batch limit of 250,000 records and takes around 90 minutes per batch. Even though PHE limits the lookups to the last 100 days to reduce the number of cases, this still means processing 2599465 every night (in 13 files). Often files are not fully processed and have to be sent back and this eats in the time for running the linkage processes to generate the deaths data. Coupled with a data processing routine based on Stata, the volume of data is creating increased risks of delays in the death data and there have been recent significant delays.

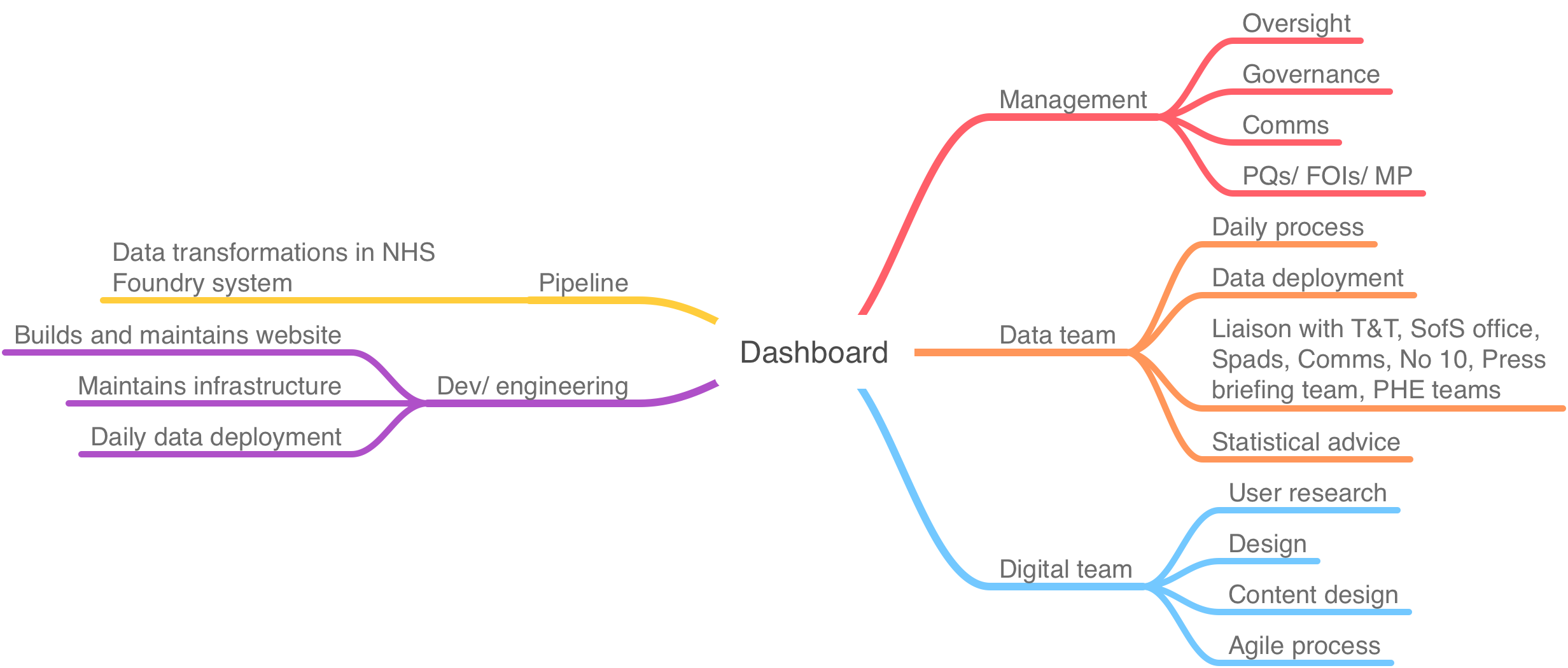
Once deaths data arrives in Foundry it takes about an hour to process through the data pipeline.



## Dashboard team

The team of people working on the dashboard has grown rapidly and there are now five teams involved. There are a few people in the team who have worked continually on the dashboard since the summer, and two or three since March. The lech lead/ lead developer has worked on the project since March *without* a day off.

#### Five teams



## Current staffing and roles

The current staffing is x wte and is a mixture of fixed-term appointees, volunteers/ loanees and contractors.

* **Dev team** - 3 wte (1 FTE tech lead/ lead developer, 1 FTE contractor, 1 FTE apprentice)
  + Front and back end development - API/ web pages/ databases
  + Liaison with Microsoft team who manage Azure infrastructure
    - Data stores
    - Cache
    - CDN
    - APi Management services
    - Resilience
    - Scaling
* **Data team** - 7 wte (3 WTE G6, 2 WTE G7, 1.2 wte SEO)
  + Daily process
  + Daily deployment (G6s)
  + Layouts
  + Data design
  + Negotiation on data additions
  + Data for daily press conferences
  + Pre-release data
  + Ad hoc analysis
* **Pipeline team** - 2 wte
  + Data flows and data management
  + Data onboarding
  + Working with Palantir
* **Digital team** - 4 wte (Senior Product Manager, Delivery manager, 1 wte user researcher, content designer, 0.5 wte interaction designer…)
  + User interaction design
  + Content design
  + Agile management
* **Management team** (1 wte)
  + Work closely with Senior Product Manager and heads of analysis to determine feasibility and priority of requests
  + Establish governance as arrangements
  + Work with senior stakeholders
  + Procure resources
  + Set priorities and direction
  + Respond to users and requests

## Adding new metrics

Many people ask us to add metrics.

We follow a number of principles in determining priority of adding metrics

1. Will it help transparency e.g. in showing data underpinning Govt decision making
2. Will it help the public understand the current state of the pandemic
3. Is there a data flow
4. Are the metrics for the data defined and agreed across the system
5. Can the DAs supply the data
6. Can the metric be visualised
7. Can it be generated daily

## Risks and resilience - transition

**The key principle has to be safe transition.**

Transitioning elements of

Work has been undertaken to look at the feasibility of moving the pipeline to EDGE - this was partly because of uncertainty about Foundry but the future

### Strengthening rotas

* Unless there is a change in reporting policy the dashboard will remain a seven day per week process
* This means that the analytical, development and pipeline teams will need to continue to provide a 7-day service and operate a rota
* This means each team needs at least 4 people to allow a 1 in 3 weekend rota
  + 3 - 4 devs who operate the technical process and deployment and fix issues. Often development has to take place at the weekend
  + 3 - 4 pipeline developers who be available to fix issues with the pipeline
  + 4 - 5 analysts who can run the daily analytical process, oversee the pipeline

## 

## Recommendations

1. Govern the dashboard as part of PHE transition and ensure that senior stakeholders are aware
2. Increase the resilience of the dashboard team required for the 7-day a week service to ensure rotas and on-call arrangements. This means at least 4-5 people working in each of the data, pipeline and dev teams to enable a 1 in 3 weekend rota at the most onerous.
3. Review the backlog and existing priorities
4. Increase the resilience of the production process for daily deaths

### 

1. Nations, regions, upper and lower tier local authorities, NHS trusts [↑](#footnote-ref-1)