A model for implementing RAPs that doesn’t cost half a billion pounds

The tender finally clarifies what NHS England mean by their “Federated Data Platform”, because dictionary definitions weren’t going to help very much, and didn’t.[1]

One approach to the FDP is for it to suck – to suck all data into one place and rely on goodwill and good luck for governance.

An alternate approach, that does not suck, is to send analyses to where the large pools of data already are, with a standard query layer on top, which is the practice underpinning the recommendations of the Goldacre Review. But this tender leaves best practice as optional.

The model doesn’t need much change:

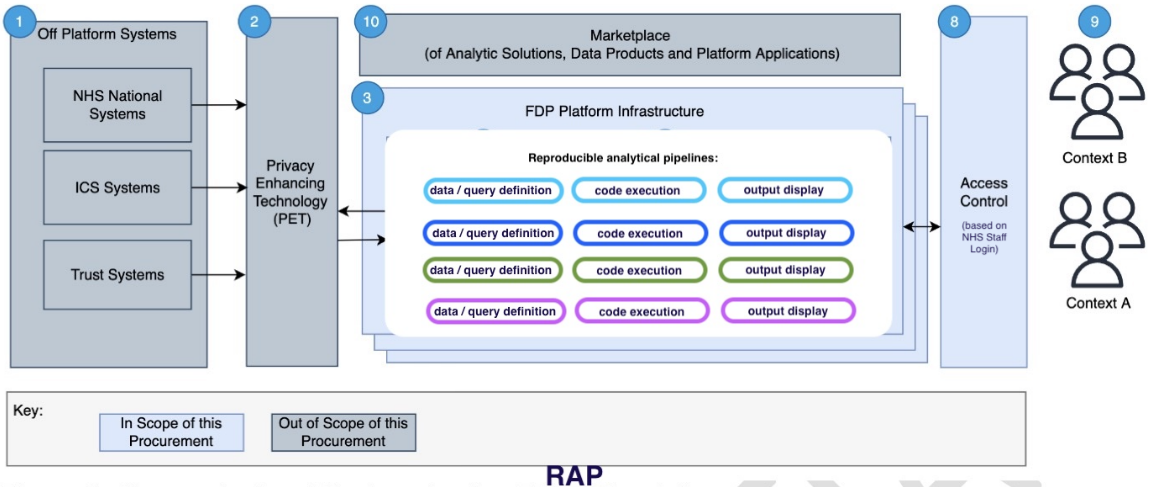

Figure 2: Conceptual architecture for the ~~FDP-AS~~ RAP solution

The NHS has available and freely reusable data infrastructure – the only thing it doesn’t offer is secrecy to analysts (features HDRUK and NHSE seem willing to pour money into).

NHS England had the choice between confidentiality for patients and secrecy for itself, and picked itself over patients, as bureaucracies usually do.

Principles

1. All analyses should be transparent and auditable.[2]

2. There must be no sucking of data into one system.

3. There should be no copies of sensitive personal data when it can already be analysed in-situ in an existing location – where data can be analysed in place, it should be.

[1] We’ll cover that here https://medconfidential.org/2023/the-palantir-procurement-part-one/

[2] Insufficient: https://questions-statements.parliament.uk/written-questions/detail/2022-07-08/HL1602/
Awaited: https://questions-statements.parliament.uk/written-questions/detail/2023-01-12/hl4714

Analyses

4. Ad–hoc analyses in the environment have outputs checked as for the research TRE (the development of a RAP is an ad-hoc analysis).

5. Any Reproducible Analytical Pipeline that has been run once can be requested (and approved) to run repeatedly on a desired schedule with no output checking.

6. Patients should be able to see which projects have used data about them, and what the benefits of those projects are, and how various IG/dissent was (dis)applied.

Key aspects of Interoperable Architecture, Operation & Governance

A. A standards-based interop layer to ensure an analysis has the right IG to run on the right data, and to facilitate any output checking necessary for the results of that data/analysis (ie, this is the data-outwards view of the system).
   - This layer is what those with pools of data, GP, HES, etc, will implement.
   - This is OpenSAFELY for GP, OpenSAFELY or equivalent for hospital data (NHSD has it working in their internal environments, but E is blocking the use of it for anything other than COVID).

B. An analysis interface built upon the to which analyses are submitted and outputs returned to the user (ie, this is the user-inwards view of the system)
   - There should be many different ways to build an analysis, and many different treatments of returned analysis.

Just as ONS has multiple data products with different interfaces for different users of the same data , the same principle applies here.

This should cost no more than £48m per year. Individual sectors should be capped at £4.8m per sector.

Existing suppliers.

The NHSD TRE is compatible with this model, and OpenSAFELY's software and standards is an example of part 3, and their use of Stata/R/etc would be part 4.

NHSD's databricks can operate separately at layers 3 and 4, but it is unclear whether NHSD has configured it to do so.

Palantir's visualisation tools rely on a data model of sucking everything data inside their databases; whether Palantir would be willing to connect their visualisation tools on top of NHS controlled databases is unclear (even with the capacity to create the same tables/schemas/etc as in Palantir's own databases), but it can do so if Palantir are willing to do development work (that all providers should be expected to do for a contract of this size).

medConfidential