Task 3

Study on data tools and technologies used in the public sector to gather, store, manage, process, get insights and share data

Matthew Upson

Data scientist and machine learning specialist

Public Digital affiliate

public digital

Based on four case studies

1. the Reproducible Analytical Pipelines in the United Kingdom;
2. the New Zealand government's Integrated Data Infrastructure and Social Investment Analytical Layer;
3. Findata, a Finnish agency to enable the secondary use of social and healthcare data in the research, public, and private sectors;
4. and KOKE, an analytics solution for fraud detection in use by the Estonian Tax and Customs Board.

Based on four case studies

1. the **Reproducible Analytical Pipelines in the United Kingdom;**
2. the **New Zealand government's Integrated Data Infrastructure** and **Social Investment Analytical Layer;**
3. Findata, a Finnish agency to enable the secondary use of social and healthcare data in the research, public, and private sectors;
4. and KOKE, an analytics solution for fraud detection in use by the Estonian Tax and Customs Board.

UK Reproducible Analytics Pipeline (RAP)

What it is: a methodology for improving the production of statistical publications

Developed by GDS and the DCMS in 2016.

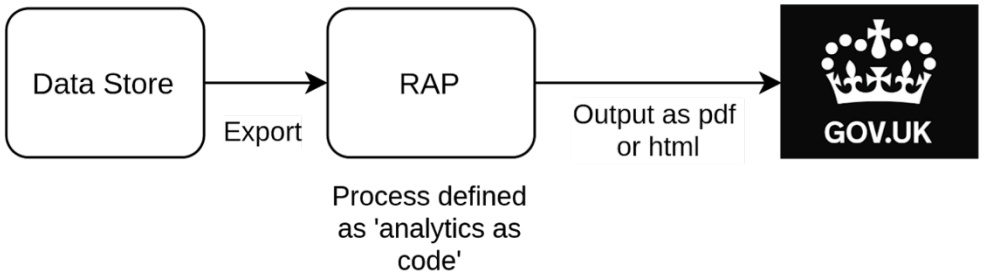
As technology, it used:

- open source software
- R and Python languages
- Software version control (Git / Github)

What was achieved: More than 30 projects, some achieving a 75% time saving over traditional methods of statistics production. Wide, and growing adoption across UK government departments.

Typical manual process for producing a statistical publication

Producing a statistical publication using a Reproducible Analytical Pipeline



Lessons learned

- **Open source** software can greatly improve data analytics in government
- RAP **only solves part of the problem**
- Approaches like RAP can take **time to implement**
- There can be **skills shortages** for the adoption of approaches like RAP
- RAP is as much **about people** as it is technology

NZ Integrated Data Infrastructure (IDI) and SIAL

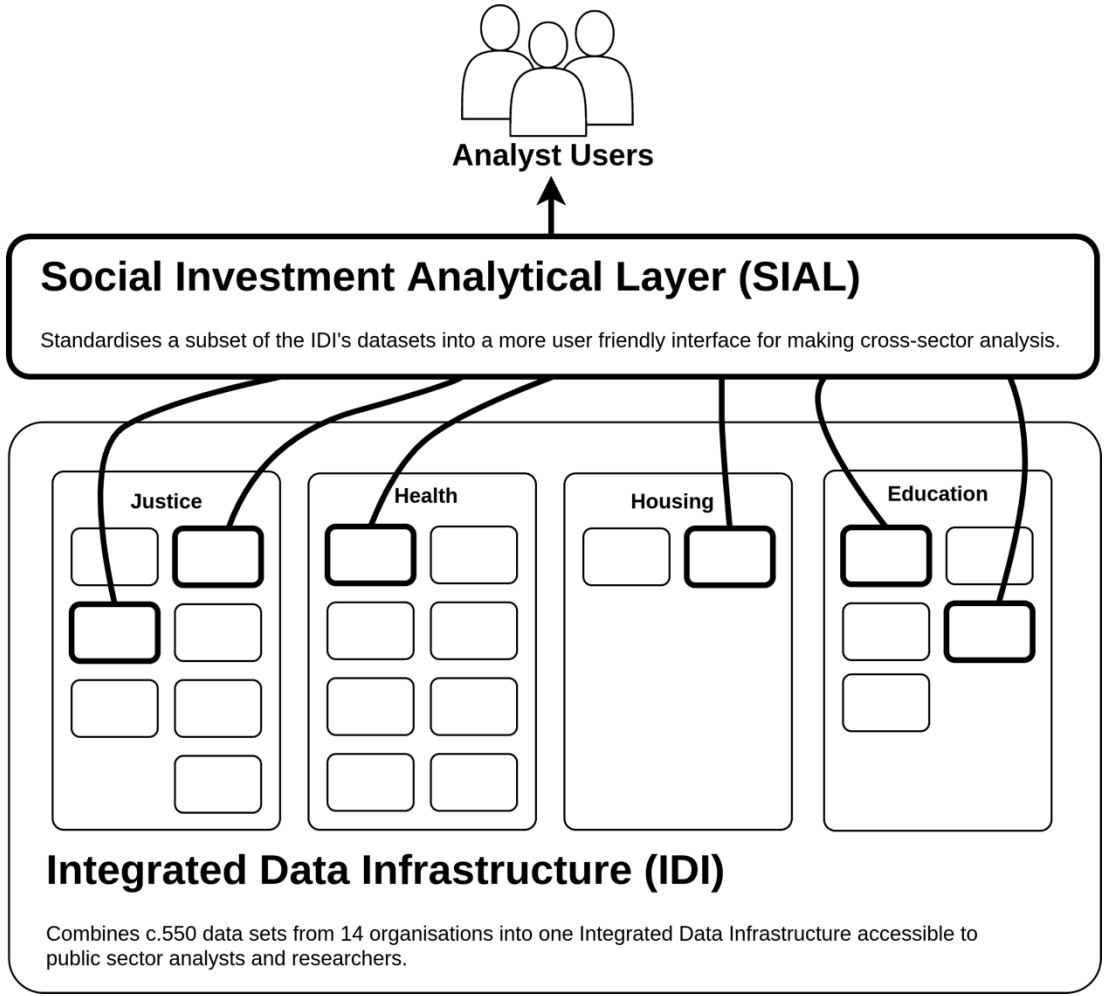
What it is: The IDI is a research database that holds anonymised data from across the public sector. The Social Investment Analytical Layer (SIAL) is a tool to improve the usability of the IDI by processing the raw data into an easier to use format.

Developed by Statistics New Zealand in 2017.

As technology, it used:

- MS SQL
- SAS, SPSS, R
- Software version control (Git / Github)

What was achieved: Access to public sector data for public sector agencies, and researchers subject to approval. The SIAL generated approximately **$1m NZD** in time savings since it was shared openly and has been re-used multiple times.


Analyst Users
Social Investment Analytical Layer (SIAL)
Standardises a subset of the IDI's datasets into a more user friendly interface for making cross-sector analysis.
Justice
Health
Housing
Education
Integrated Data Infrastructure (IDI)
Combines c.550 data sets from 14 organisations into one Integrated Data Infrastructure accessible to public sector analysts and researchers.

Lessons learned

- Start small, test regularly, and iterate **to meet users' needs**
- **Good security practices** are essential when managing personal data
- Standards for **public trust and transparency**
- Adopting and sharing **open source software** saves time and resources

To sum up, Task 3 case studies highlight the need to:

- Start with user needs (and recognise analysts as users)
- Work in the open and foster reusability
- Adapt to data readiness
- Use open source
- Invest in data capability at all levels
- Break down silos

Angie Kenny
angie@public.digital
Website https://public.digital/
Twitter @publicdigitalHQ

public digital