DG DIGIT
Unit D1

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

Data analytics for Member States and Citizens

Date: 22/07/2020
Doc. Version: Final

THE REPORT HAS BEEN PRODUCED FOR THE EUROPEAN COMMISSION BY:

The research presented in the report has been carried out within the scope of the study Data Analytics for Member States and Citizens (Framework Contract DI/07624 - ABC IV Lot 3) commissioned by the European Commission, Directorate-General for Informatics, to Deloitte and the Lisbon Council for Economic Competitiveness and Social Renewal. The project has been carried out within the scope of the ISA² Action 2016.03 – Big Data for Public Administrations. More information is available at https://ec.europa.eu/isa2/sites/isa/files/library/documents/isa2-work-programme-2016-detailed-action-descriptions_en.pdf.

Contact information
DIGIT-DATA-SERVICES@ec.europa.eu

DISCLAIMER

TABLE OF CONTENTS

1. CASE STUDY: REPRODUCIBLE ANALYTICAL PIPELINES (RAP) - UNITED KINGDOM

1.1. Introduction

Reproducible Analytical Pipelines (RAP) is a methodology for the production of statistical publications, that was developed during a collaboration between the Government Digital Service (GDS) and the Department for Digital, Culture, Media & Sport (DCMS) in 2016. The project aimed to improve the production of a statistical bulletin by introducing techniques from software engineering, data science, and academia. The use of open source software was critical to the success of the project which reduced production time of the statistical bulletin by an estimated 75%.

The outputs from the project were published openly and widely, and the methodology which came to be known as Reproducible Analytical Pipelines (RAP) has been adopted widely across local and national government in the UK, now totalling some 30 projects. A community comprising mostly statisticians, analysts, and data scientists, has grown to provide support across the public sector in the implementation and development of RAP techniques, supported by the Office for National Statistics (ONS) and GDS.

RAP recognises three things:

- Workflows with a large number of manual steps are time consuming, potentially error prone, difficult to audit, and tedious.
- Analysts increasingly have programming skills, and are keen to use statistical programming tools in their daily work.
- A number of tools developed in the fields of software engineering, DevOps, and academia can easily be adapted for use in government to tackle the challenges associated with producing official statistics.

RAP comprises a set of methodologies, open source tools, and examples, with capabilities including data gathering and integration, analysis tools, data visualization and access to expertise, including data analysts and data scientists.

1.2. Development of the work

Following a successful proof of concept between GDS and DCMS to automate the production of a statistical bulletin, the project was widened in mid 2017 to involve the Department for Education (DfE) and Ministry of Justice (MoJ). With the assistance of the ONS Good Practice team, a RAP community was set up, which continues to iterate and develop the work with regular meetings, and the creation of a website.

Furthermore, it supports and is part of multiple cross-government frameworks and guidelines, including the UK Government Data Ethic Framework (see section 'Reproducibility'), the Office for National Statistics (ONS) guidance on Quality Statistics in Government (see section 'Reproducible Analytical Pipelines'), the National Data Strategy and the Aqua Book: guidance on producing quality analysis for government.

RAP is now used for at least 30 projects across the UK public sector, in projects ranging from the production of Official Statistics for the School Census, monthly publication of economic forecasts, local authority-specific health "at a glance" documents, Annual Employment Allowance take-up, calculating the cost of reoffending and the production of index of price changes for rail fares.

1.2.1. Example: Economic Estimates for DCMS Sectors Statistical First Release (SFR)

The proof of concept for RAP was the production of the Economic Estimates for DCMS Sectors Statistical First Release (SFR). In 2016, the publication was produced with a mixture of manual and semi-manual processes. The proof of concept was designed to test if the production of the SFR could be sped up, while maintaining the high standard of the publication.

The outcomes of the proof of concept were eesectors an open source R package which converted into code all the steps to reproduce part of the Economic Estimates for DCMS Sectors Statistical First Release publication, and eesectorsmarkdown a rendering of the publication in RMarkdown to be used alongside eesectors. Both these outputs were published openly on Github so that the code could be examined and re-used by other departments.

The project demonstrated that using a RAP methodology and developing analysis as code could reduce by 75% the time taken to produce the same statistical analysis the following year, in comparison to the manual/semi-manual process.

Although DCMS no longer uses the original proof of concept, the department continues to use the RAP methodology and has published a new open source package named eeqva which also uses the R language, in addition to implementing more of the data preparation pipeline in Python.

1.3. Delivery model

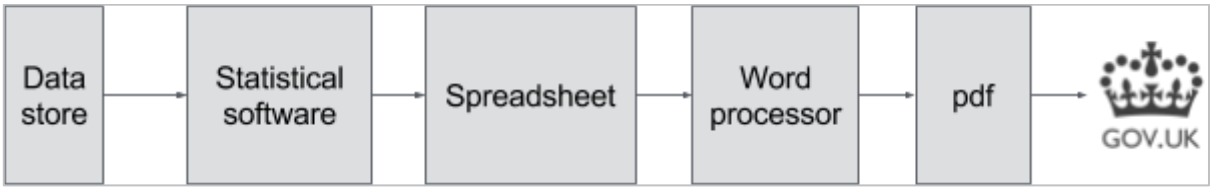
This project was initially set up through a small collaborative team, comprised of two data scientists, one each from DCMS and GDS. The second iteration involved a wider collaboration with statisticians and data scientists at the DfE and MoJ, with additional data science support provided by GDS.

It is now supported by the Government Statistical Service Good Practice Team (GPT) and the Government Digital Service (GDS), who have brought together a new champions group from 31 departments to support and promote the use of RAP. The network includes around 80 champions from 31 government departments, including departments in Scotland, Wales, and Northern Ireland.

These champions are statisticians and data scientists who have the software development skills necessary to deliver RAP data products, and can share their expertise with others by mentoring and providing advice.

1.4. What a Reproducible Analytical Pipeline does

The usual process of statistics production in Government varies widely across departments, and individual teams within departments, but processes often have some or all of the following steps:



The current statistics production process

This workflow is characterised by:

- **Wide use of proprietary software**. Which software applications are used depends on the purchasing decisions of an individual department, and the expertise of individual analysts within a team.
- **Multiple software applications**. Often multiple software applications are used for very specific tasks. For example, a proprietary database client may be

used to access the data from a proprietary database; proprietary statistical software package might be used for making the calculations relevant to the bulletin; proprietary spreadsheet software may be used to organise the resulting statistics intro an appropriate form; whilst proprietary word processing software is almost universally used to format the final publication.

- **Multiple 'copy and paste' operations**. Since multiple software applications tend to be used, data is often manually 'copied and pasted' from one application to the next.
- **Quality assurance**. Statistical publications almost always undergo some form quality assurance (QA). Since the process of producing a bulletin is often highly manual, quality assurance is also highly manual. Frequently QA involves a member of the team 'eyeballing' the final bulletin to ensure there are no obvious errors, or actively re-creating the same statistics used in the bulletin to ensure that the same values are arrived at. The process of QA can be time consuming, and tedious for the analysts involved.

The RAP workflow aims to be much simpler, with the most time consuming and error-prone steps replaced by bespoke software owned and managed by the team that produces the publication:

Data product production using Open Source programming languages such as R

The RAP workflow is characterised by:

- **Analysis as Code.** The key idea behind the Reproducible Analytical Pipeline is just that, to create a pipeline from the datastore to the final analysis that is easily reproducible. This is done by enshrining all the processes behind a piece of analysis or publication in code - a practice that can be described as Analysis as Code (AAC).
- **Language agnostic.** In principle the production of analysis as code should not be dependent on a single programming language, a RAP can be produced in any language available to the analyst. In practice however, RAPs have tended to be written in just two languages: Python, and more predominantly, R. This is because these languages (particularly R) tend to be available to analysts on government computing infrastructure, and both languages (particularly R) have a mature ecosystem of modules that easily facilitate the production of statistical bulletins.
- **Open source, not proprietary.** One of the great strengths of the RAP ecosystem and community is that programming languages that tend to be used to create analysis as code (R and Python) are open source. This has meant that from the very beginning it was possible to share the source code of the first prototypes freely across government departments and beyond. If a team develops all its analysis through manual processes, it is much more challenging to share its processes between individuals in the team, let alone with other analysts in other departments, or indeed the general public.
- **Auditability.** Since analysis is developed as code, it becomes possible to make use of a Version Control System (VCS) just as a software developer would. Such systems allow for every change to the code to be recorded and documented in a clear and easily auditable fashion. This is something that is

practically impossible with a manual process, but essential for analysis produced by a government department which may be used to inform policy.

- **Automated Testing.** Since the analysis is enshrined in code it is possible to implement automated testing. This means that the process of QA can move from another, difficult to audit, manual process into a largely automated process that can be performed instantly and repeatedly. Publications will always need a manual check for consistency, but by adopting the same kinds of testing used by software developers, analysts can greatly reduce the propensity for errors in analysis. Analysts can also make use of 'DevOps' (Developer Operations) practices, which when applied to the data realm are sometimes called 'DataOps'. These tools include the use of third-party services which can trigger the compilation of the publication and automated testing when a new change is made, further reducing the burden of testing on the individual analyst.
- **Everything, all in one place.** Tools such as Rmarkdown can be used which allow the analyst to combine the logic (code) of the analysis with the final publication itself, in one single document. These documents essentially become templates that can be automatically updated with the latest data, and 'compiled' to produce an updated version of the statistical bulletin. It also becomes possible to automatically provide context around a given statistic, for example 'this measure <increased or decreased> by X% compared to the previous year.' without human intervention. Since government departments often have a high turnover of analytical staff, another benefit of combining the logic with the formatting is that it ameliorates the issue of institutional knowledge transfer.

1.5. Identifying a user need

It is useful to consider two sets of users when talking about statistical bulletins:

- End users of the statistical publications. This could be ministers, members of the public, businesses, or public servants within the same or other departments.
- Public servants responsible for producing the publications. This could be the analysts responsible for the actual hands-on work, or managers responsible for the delivery of the publications.

In the former case, a clear indication that a new approach is needed would be a requirement for timely and accurate statistics delivered at a frequency difficult or impossible using existing means. Such situations can occur at times of crisis, for example during widespread public disorder in August 2011 which became known as the London riots. Ministers needed statistics about the unrest much more frequently than normal, which put pressure on the teams responsible for producing them.

Statistics for which there is a statutory requirement to be delivered in a limited period of time also add pressure to statistical teams. So-called ad-hoc statistics which are often produced at the behest of members of parliament (MPs), and freedom of information requests (FOIs) from members of the public, both of which are unplanned and unpredictable, are good examples.

A grassroots user need arising from the public servants responsible for the publications is simply to have interesting and rewarding work. The traditional process of statistical bulletin production is often time-consuming and repetitive, and detracts from more interesting and in-depth analytical tasks better suited to the expertise of the analyst.

Managers may find themselves stuck between these two imperatives: a need to provide accurate and timely statistics to (often high profile) clients, and to provide interesting and rewarding work to analysts, and to upskill and induct new team members as others leave, all in the context of shrinking public sector budgets.

Most (if not all) publication teams are likely to have some combination of these user needs.

1.6. Identifying publications most likely to benefit from RAP

RAP is most suited to statistical bulletins which are produced periodically, and follow the same rough pattern each time they are published, but with updated data. Many statistical publications in the UK government fit this pattern; these stand to gain the most from the implementation of a RAP.

Very short publications, or those that require innovative or constantly changing analysis or formatting are less well suited as there is an overhead associated with editing the source code behind an analytical pipeline. It should also be stated that not every statistical publication or piece of analytical work would benefit from the full RAP process implemented in the original proof of concept, for which a conscious decision was made to try to automate as much as possible. However, almost all analysis would benefit in some way from some of the methods included under the RAP moniker.

National Health Service (NHS) Scotland have created a useful checklist[1] for teams considering taking a RAP approach, and categorised levels of automation[2] for teams to aim at. This is a useful guide, as publication teams may not have the requisite skills to transition their publication into a complete RAP, without further training or support from the community.

1.7. Resource considerations

Reproducible Analytical Pipelines rely on Free and Open Source Software (FOSS), so there is usually little or no capital investment implicated by a decision to convert existing analysis into a RAP. The main constraints are human, principally:

- **Capability**. Whilst many analysts have some familiarity with the tools used in building a RAP, they will often require additional training or mentoring to develop the full range of skills required to implement a complete RAP. Most statistical teams in the UK government are not multi-disciplinary, so it is a more common pattern to upskill existing analysts rather than recruit software developers who would already be familiar with many of the skills required to implement a RAP.
- **Time**. Whilst developing the capability of analysts requires a commitment of time, converting from a manual/semi-manual publication also can take time, and requires some understanding from senior managers that the investment of time is justified.
- **Access to the requisite tools**. Some analysts have reported that it is difficult to get access to the tools required to build RAPs within their departments, with security concerns cited as a frequent reason by IT departments for unwillingness to allow analysts access.

1.8. Implementing the RAP

A broad outline to implementing the RAP following the methodology used in the proof of concept (equivalent to a Level 7 on NHS Scotland's scale of maturity[2]) follows:

1. **Ingest and standardise the data**. A given statistical publication may make use of data from a range of sources, which may include international statistical bodies, other UK government departments (particularly the ONS), or third-party companies with contracts to provide the data. Data rarely arrives in a format that is suitable for developing a RAP, so the first step is to standardise the data into a format that can easily be reused throughout the publication.

1 https://www.isdscotland.org/About-ISD/Methodologies/_docs/Checklist-for-TPP-and-RAP-processes_v0-3.pdf

2 https://www.isdscotland.org/About-ISD/Methodologies/_docs/Reproducible_Analytical_Pipelines_paper_v1.4.pdf

Often this means that the data is converted to long or 'tidy' format[3], and code is developed to convert the incoming data (which may arrive in a mix of proprietary and non-proprietary formats). The code is built in such a way that it can be reused for the subsequent publication using new data, for example by creating a 'class' for each dataset (if using an object orientated programming (OOP) language)[4]. As with all the following steps, all the code should be managed with a version control system.

2. **Develop code to produce each of the elements of the publication**. Every publication comprises various elements derived from data, for example: tables, figures, and individual statistics. In the original RAP proof of concept, a software function was developed which corresponded to each of these elements, with each function accepting as an input the dataset described above. In an OOP language like R and Python, the code to produce these elements can be written as methods of the data class defined in 1.[5]
3. **Recreate the publication narrative**. The narrative of the publication (which often remains relatively unchanged from year to year) can be reproduced in a literate programming tool, for example Rmarkdown[6] for R, or Jupyter Notebooks[7] for Python, allowing it to be merged with the source code developed in 1. and 2.[8]

1.9. Re-usability

The more mature RAPs are built as modular pieces of software, which intrinsically support re-use elsewhere. However, it is unlikely that a RAP may be re-used in its entirety: necessarily the start and end of each pipeline is unique. Most government departments have their own distinct data infrastructure, whilst each publication usually requires a unique mix of data from different sources. Equally, no two statistical publications are entirely alike, and differences in style, formatting, etc. may exist at the inter and intra-departmental level. For these reasons, the components most likely to be reused widely are those that 'produce each of the elements of the publication' as described above. The following table summarises the propensity for components to be re-used by other teams.

Description	**Re-use by the same team in the future**	**Re-use by other teams in the same department**	**Re-use by teams in other departments**
Ingest and standardise data	All the components should be reusable. Only if the data sources change would these components	Some components may be reusable if the other team also works with the same data.	Some components may be reusable by other government departments (OGDs), although perhaps only when using open

3 https://vita.had.co.nz/papers/tidy-data.pdf0

4 At this point tests can be written to ensure the integrity of the data as it is imported. These may take the form of very simple checks that the right number of values have been imported, or could be more complex statistical checks to look for outliers in the incoming data based on data from previous years.

5 Together with classes written for each data source (1.), these functions can be combined into a software module using the packaging systems native to the language it is written in. Within these packages can be embedded further tests and documentation.

6 https://rmarkdown.rstudio.com/

7 https://jupyter.org/

8 Now that the whole publication has been enshrined in code, further steps can be taken to ensure its reproducibility, for example creating a dedicated environment in which to run publication code (for example creating a docker container), and implementing automated testing and compilation (Continuous Integration) using a third party (often free) service, for example Travis.

	need to be adjusted[9].		data.[10]
Produce elements of the publication	All the components should be reusable. Unless new elements are required, it is likely that these components could remain unchanged from publication to publication.	Some components are likely to be reusable that reflect common approaches taken across the department.	Some generic components unspecific to the individual publication and department may be shared between departments. For example, govstyle[11], a library for creating figures in a gov.uk style.
Recreate the publication narrative	All the components should be reusable. Whilst small changes will almost always be required to the final document, the expectation is that most of the logic produced to recreate a publication will remain unchanged.	Generic components relating to formatting, etc. are likely to be reusable.	Generic components relating to formatting, etc. may be reusable.

Note that a component will only get reused if there is knowledge about that component and a willingness to share the source code. The community supporting RAP plays a key role here, as typically analysts from different government departments do not share their work with each other on a regular basis, unless there is an explicit reason to do it. Some departments, like DCMS have opted for an entirely open approach to the development of their RAPs, which are published on Github, greatly facilitating reuse. Others, such as MoJ have taken a more conservative approach, and do not share all their source code externally, although sharing does take place internally, facilitated by the department's Analytical Platform (see example). Note that prior to RAP, there were few possibilities for analysts from one department to share their source code with another department.

9 Since many publications consume data from various sources, it is possible for the incoming data sources or format to change. This can be outside of the control of a publication team.

10 Since data sharing between government departments is the exception rather than the norm, re-use of these components is relatively unlikely to be possible between departments, and is most likely to occur when using open data sources.

11 https://github.com/ukgovdatascience/govstyle

1.9.1. Example: Using RAP on the MoJ Analytical Platform

The MoJ has prioritised the building of a state of the art Analytical Platform (AP) for all its analysts, based on principles from the UK government Data Ethics Framework, and the Technology Code of Practice. The AP is a cloud platform deployed on Amazon Web Services (AWS) which allows users to sign into an analytical environment from their browser giving them access to all the tools and data that they require.

Reproducibility in the context of government analytics can be defined loosely as the ability to *run the same analysis on the same data, in the same environment*. Whilst RAP is concerned with the first two (and predominantly just the first), the AP effectively solves the third issue by allowing environments to be defined in code, leveraging the power of container technology. Just as the code in a RAP can be source controlled, so too can all the prerequisites to recreate the computing environment for the AP, including all the software versions and prerequisites that are required, and that might have a significant impact on the replicability of a particular piece of analysis.

A number of RAP projects have been deployed on the AP. One example is the automation of the HMPPS workforce statistics quarterly publication. The publication consists of a bulletin and an accompanying Excel workbook: together these provide statistics on members of staff in HMPPS, including numbers of staff in post and leaving rates.

The code for this RAP was developed in an integrated development environment (IDE) made available on the AP, which is linked to GitHub in to manage version control. GitHub facilitates collaborative working, and provides code review tools that are used by analysts working to peer review each other's code.

As part of the AP, the MoJ have also undertaken a lot of work to improve its flows of data. Incoming data is processed, and made available to users through AWS Simple Storage Service (S3). This data can then be used as the single source of truth which is referenced in individual RAPs ensuring reproducibility.

The raw data required for this publication are stored on AWS Simple Storage Service (S3). The RAP accesses the raw data, manipulates them, generates the publication, and writes the manipulated data and publication back to s3. Working in this way ensures that all analysts have access to the same data and the latest draft of the publication.

The AP enables the code to be written in a managed environment, which is defined in a file that is stored together with the code on GitHub. This means that any analyst can quickly ensure that they have the correct environment necessary to run the code.

Together, the managed coding environment and the use of cloud services to store the code and data have allowed this RAP project to be easily handed over to other analysts for updates and maintenance. The HMPPS Workforce Statistics publication has been automated since June 2019, and whereas the old version would take the best part of two weeks (at one full-time equivalent) to put together, plus a further week for quality assurance, the code that produces the new version takes less than 10 minutes to run, and the quality assurance can now be completed in less than a day. This has freed up time for additional and more in-depth analyses.

1.9.1.1.

1.10. Technology

1.10.1. Design principles

The basis of RAP is taking software development principles and using them to produce periodic statistical publications. The tools and techniques of software and DevOps: version control, continuous integration and deployment (CI/CD), automated testing, and test coverage are combined with analysis written in code (usually in R or Python) to create a pipeline of steps that can easily be reproduced semi-automatically. These techniques when applied to data are often described as DataOps[12]. The underlying

12 https://www.dataopsmanifesto.org/

principle is to reduce time consuming and potentially error prone manual intervention where it is unnecessary.

1.10.2. Technology choices

RAPs are produced in open source programming languages, typically R or Python. This is in line with GDS guidance to 'Choose analytical tools that keep pace with user needs'[13], since these languages are commonly used in industry and academia for cutting edge analysis and data science. Both are supported by large communities of users who develop open source code incorporating the latest analytical methods. Most RAPs have been built using the R language and the associated Rmarkdown which allows the narrative of statistical publications to be combined with the code which produces the statistics, tables, and figures). Some departments, notably the Department for Transport (DfT) and parts of the ONS have also experimented with developing RAPs using Python.

R is an excellent choice for the initiative for three reasons:

- Firstly, statisticians and data scientists leaving university in recent years will almost certainly have been introduced to R, so most new recruits to government departments are already somewhat familiar with the language.
- Second, the language has sophisticated support for so-called 'literate programming' where the narrative of a publication and the code to produce it are combined in a single location - indeed this is a common use for the language in academia.
- Finally, the language is typically quite well supported within government departments, where due to security restrictions it can be difficult to get access to more fully-fledged programming languages (for example Python).

The source code behind RAPs is then typically packaged into a software module that can be shared and installed easily. In addition to using R or Python, RAPs typically make use of a number of open source software packages. Many RAPs make use of Git (https://git-scm.com/) for source control, which is a fundamental basis to producing reproducible work. Git is the industry standard for version control among software developers, works across multiple platforms, and can typically be installed without the need for special permissions on many systems. Git provides an unequivocal and unbroken record of all the changes that have been made to a statistical publication, without the need for tedious manual record keeping. Alongside Git, many practitioners use Github (https://github.com/ or similar source control repositories like GitLab: https://about.gitlab.com/ and Bitbucket: https://bitbucket.org/product) to store and share development of RAPs (for example https://github.com/moj-analytical-services/mojrap). Using Github allows departments to share their work openly (if possible) providing unparalleled levels of transparency, and allowing other departments to make use of, and learn from their work. Using GitHub also allows statisticians to make use of a range of tools that improve the quality of the code that underlies the statistical publications. These include https://travis-ci.org/ and https://www.appveyor.com/ which allow for tests of the RAP to be run each time a change is made, ensuring consistency. Some RAPs also make use of https://codecov.io/, a tool which identifies weaknesses in the current suite of tests covering the source code. Both these types of tool provide a quantum leap in quality assurance, which traditionally is a manual process.

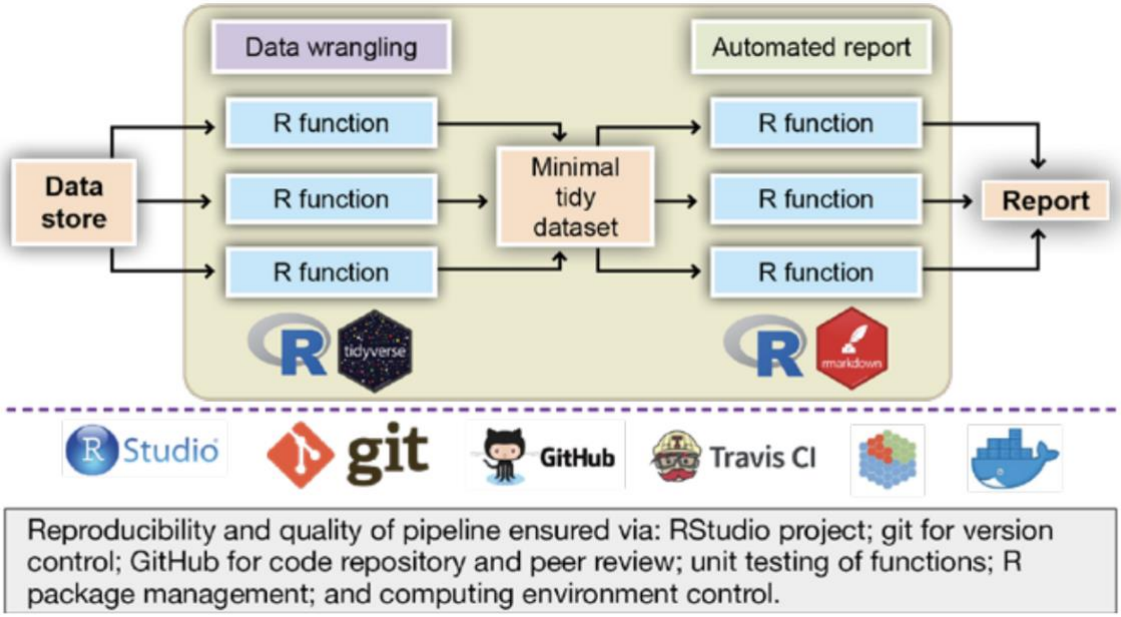
1.10.3. Using open source

RAPs are built almost exclusively using the R and Python languages. Multiple open source packages (too many to list) are therefore used in their development. Because of the highly specific nature of many of the data preparation processes, it would be

13 https://www.gov.uk/guidance/choose-tools-and-infrastructure-to-make-better-use-of-your-data#case-study---using-data-science-with-the-ministry-of-justice-analytical-platform

difficult, if not impossible, to automate all of the data preparation steps behind each publication using proprietary software. In addition, since a key part of the RAP project has been to develop a supportive community across government, using proprietary software would have hampered this process, as it would have limited potential contributors to only those whose departments had invested in the proprietary tools. RAPs typically make use of the version control software git, which is essential to providing an auditable history for each publication. Git is also the industry standard for version control in software development, so is the obvious choice for version control. One advantage of using these tools is that we can reduce the number of steps where the data needs to be moved from one program (or format) into another. This is in line with the principle of reproducibility given in guidance on producing quality analysis for government (the AQUA book), as the entire process can be represented as a single step in code, greatly reducing the likelihood of manual transcription errors. Moving away from proprietary software, towards Open Source, may also have the additional benefit of being more compatible with tried and tested software development tools and techniques (as well as the obvious no longer needing to pay for it).

1.10.4. Architecture



Source:

https://www.isdscotland.org/About-ISD/Methodologies/_docs/Reproducible_Analytical_Pipelines_paper_v1.4.pdf

1.10.5. Hosting

Since RAP is a methodology rather than a service per se, it has typically been implemented by individual statisticians and data scientists working on their local environments. Canonical versions of each RAP are typically stored in a version control repository hosted on Github (or similar). Some departments have also opted to develop cloud based analytical environments (see MoJ Analytical Platform example) which further improve the reproducibility of individual publications by ensuring that a computing environment can easily be reproduced. These cloud environments leverage containerisation technology (for example docker: https://www.docker.com/), and should be considered the gold standard. The UK Government adopted a 'Cloud First' policy in 2013 for all technology decisions, which requires public sector organisations to consider cloud solutions before alternatives. By Cloud First, the policy means the public cloud rather than a community, hybrid or private deployment model.

1.11. Lessons learned

1.11.1. Open source technologies can greatly improve data analytics in government

RAP has proven to be a great success in improving reproducibility and reducing production times of statistical publication. Whilst it is hard to quantify these improvements, the RAP proof of concept was reported to have reduced the time taken to produce the publication in future years by 75%, whilst the MoJ recently reported a time reduction from two weeks down to just over a day for one publication. Despite the difficulty in quantifying the benefits, the utility of RAP has been widely recognised across the UK government, and it continues to grow in popularity as an approach.

1.11.2. Analytical teams often lack the necessary skills to do RAP

It became apparent following the proof of concept that there are a limited number of analysts in government with the skills to support the work going forward. Indeed; so, in demand have the skills needed to build RAPs become, that they have begun to feature in some analyst job adverts[14]. In order for RAP to be really successful, it needs to be supported by an approach to upskilling existing analytical staff and recruiting more highly skilled analysts. This is a key feature of NHS Scotland's approach to RAP which offers training courses for analysts wanting to work on RAPs. The ONS and GDS also continue to play a critical role in supporting government in developing RAPs by promoting the community, and producing a wide range of training materials that include a website[15], an ebook[16], and a free open online course[17]. The situation is somewhat hampered by the structure of analytical teams in the UK government which are traditionally composed only of analytical professionals. Analytical teams could be more like multidisciplinary digital teams, which combine a number of different professions in the same team - indeed this is supported by point 6 of the GDS Service Standard[18].

Adopting such an approach and recruiting software and DevOps engineers to handle the back end infrastructure of RAPs would mean that there is less burden on individual analysts to learn all the required skills, some of which are not typically within the analytical domain. In general, taking a more service-oriented approach to the production of statistical publications would bring many of the advantages of the way digital teams work to the area of government statistics.

1.11.3. RAP can take time to implement

Whilst RAP holds great promise in reducing production times in future publications, the initial process of converting publications is time consuming, not least because analysts may need to upskill in order to do it. Senior buy-in is critical to getting RAP projects off the ground, as they may be relatively slow to show returns on the time invested.

14 E.g. an assistant statistician job at the Department for International Development (February 2020): https://www.civilservicejobs.service.gov.uk/csr/jobs.cgi?jcode=1664806 and an analyst job at the Cabinet Office (May 2019): https://cabinetofficejobs.tal.net/vx/mobile-0/appcentre-1/brand-2/candidate/so/pm/1/pl/16/opp/3824-3824-Civil-Service-Workforce-and-Pay-Analyst-Analysis-and-Insight/en-GB

15 https://ukgovdatascience.github.io/rap-website/index.html

16 https://ukgovdatascience.github.io/rap_companion/

17 https://www.udemy.com/course/reproducible-analytical-pipelines/

18 https://www.gov.uk/service-manual/service-standard/point-6-have-a-multidisciplinary-team

1.11.4. RAP only solves part of the problem

As noted earlier, a loose definition of reproducibility is to be able to *run the same analysis, on the same data, in the same environment*. Whilst RAP does provide good solutions to *analysis*, it does not address the other two issues. Departments like MoJ have solved the *environment* issue by standardising this through an Infrastructure as Code (IAC) approach, but this is not widespread. The *data* issue is much more pernicious because ensuring that the raw data the department receives is always well formatted and of high quality is often outside their control. RAP should be part of a more complete data strategy that encompasses all areas of government data use.

1.11.5. RAP is as much about people as it is technology

Whilst the idea and technology behind RAP are clearly compelling, RAP would not have been as successful in the UK government had it not been for a community of motivated individuals pushing for its adoption. RAP has been one of the most compelling and widespread applications for the skills brought by data scientists, which the UK government began recruiting in 2014/2015.

1.11.5.1.

1.12. ANNEX - Data collection activities

1.12.1. List of consulted stakeholders

- Dr Matthew Upson. Data Scientist. Formerly UK Government Digital Service
- Dr Matthew Gregory. Lead Data Scientist. UK Government Digital Service
- Dr Matthew Dray. Data Scientist. UK Cabinet Office
- Duncan Garmondsway. Data Scientist. UK Government Digital Service
- Chloe Pugh. Data Analyst. UK Ministry of Justice
- David Read. Technical Architect. UK Ministry of Justice

1.12.2. References

- The UK Government Data Ethic Framework (see section 'Reproducibility') https://www.gov.uk/guidance/5-use-robust-practices-and-work-within-your-skillset#reproducibility.
- The Office for National Statistics (ONS) guidance on Quality Statistics in Government (see section 'Reproducible Analytical Pipelines') https://gss.civilservice.gov.uk/policy-store/quality-statistics-in-government/#accuracy-and-reliability.
- National Data Strategy: https://www.gov.uk/guidance/national-data-strategy
- Aqua Book: guidance on producing quality analysis for government: https://gss.civilservice.gov.uk/policy-store/quality-statistics-in-government/#accuracy-and-reliability.
- UK Government Cloud First policy: https://www.gov.uk/guidance/government-cloud-first-policy
- Choose tools and infrastructure to make better use of your data (GDS Guidance) https://www.gov.uk/guidance/choose-tools-and-infrastructure-to-make-better-use-of-your-data#case-study---using-data-science-with-the-ministry-of-justice-analytical-platform

2. Case study: The Integrated Data Infrastructure and Social Investment Analytical Layer - New Zealand

2.1. Introduction

New Zealand's Integrated Data Infrastructure (IDI) is a large research database holding anonymised data from across the public sector about citizens, linked to data about life events such as education, income, migration, justice and health. The IDI is longitudinal, meaning that it tracks anonymised individuals and households throughout their lives, and as such is exceptionally useful for answering questions about groups of people or businesses with similar characteristics over time. It is updated on a quarterly basis. The IDI has been described internationally as a success for New Zealand, and an exemplar for other countries to learn from in terms of getting the most from harnessing public sector data, and facilitating evidence based policy making.

Access to the IDI is stringently controlled, and only possible at one of three secure physical locations called Data Labs. In addition, Stats New Zealand, the agency responsible for the IDI, assesses applications for access against a number of criteria (Figure 1.), and adheres to the 'five safes'[19] framework to ensure that data is used in a responsible and safe manner.

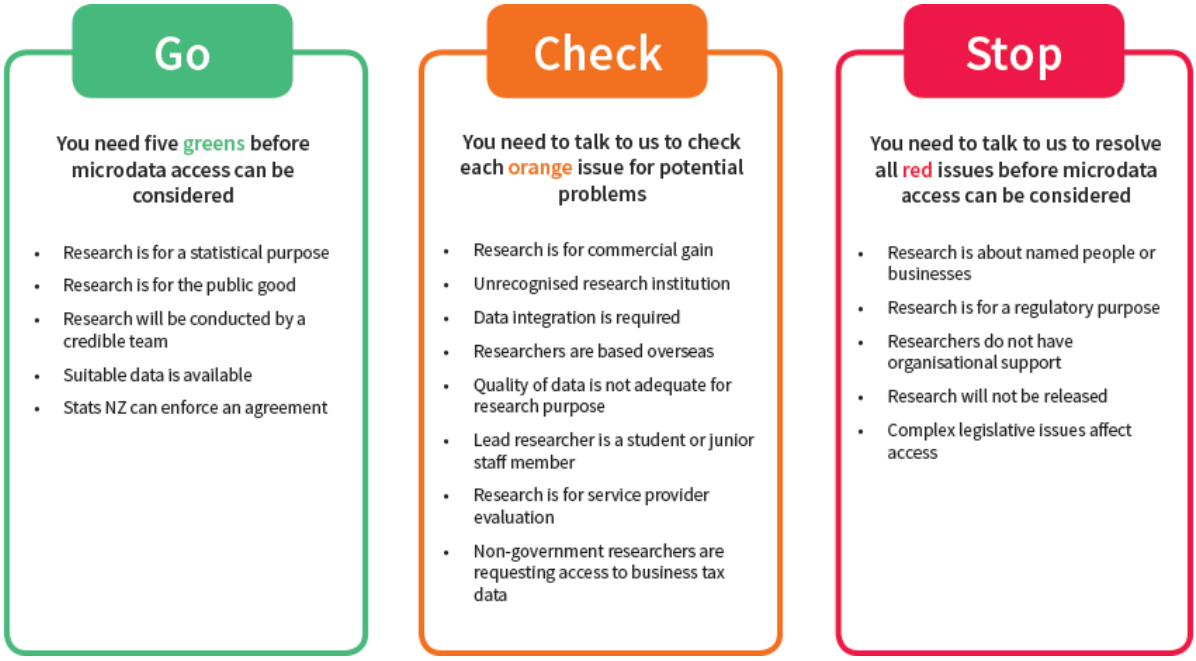


Figure 1. Potential barriers to a successful application for IDI access[20]

Several additional tools have been built to facilitate use of the IDI. The Social Investment Analytical Layer (SIAL) helps agencies understand the potential Return On Investment (ROI) before investing in a new service. It is a collection of tables derived from the IDI that can be recreated using open source code. The SIAL tables provide the user with a uniform data structure, whereas the data structure in the IDI reflects the multiple organisation from where the data come. These tables are 'event based' meaning that they record events in an individual's life through their interactions with government bodies, for example income records recorded by Inland Revenue, health interventions recorded by the Ministry of Health, and interactions with the courts or police service recorded by the Ministry of Justice and Police Force respectively.

19 https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure#five-safes

20 https://sia.govt.nz/assets/Documents/Beginners-Guide-To-The-IDI-December-2017.pdf

The Social Investment Data Foundation (SIDF) builds on the IDI and the SIAL, and allows public servants and researchers to answer more in-depth questions about individuals in the IDI, and to generate service metrics that summarise an individual's interactions with government over a given time period, for example: time spent receiving a benefit, total money received while on a benefit, or number of benefit spells over a given time period.

2.2. Development of the work

The authority for Statistics New Zealand (the primary statistical agency) was granted by the NZ Cabinet in 1997, and after some data integration projects for specific purposes, the Cabinet agreed in 2011 to consolidate previously separate integration projects into the IDI prototype. Further Cabinet agreement in 2013 led to the expansion of the IDI to create a cross-government data integration service[21]. Since then the IDI has continued to be updated on a quarterly basis, and the number of new datasets available in the IDI and demand for the service continues to increase.

In 2017, six data scientists at the Social Wellbeing Agency SWA (then the Social Investment Unit) created the Social Investment Analytical Layer (SIAL) to make accessing data from the IDI easier during the completion of a the Social Housing Test Case[22]: a study on the ROI of social housing. Creation of the SIAL cost an estimated $144k NZD. As part of this work the SWA made the SIAL available as open source software available on Github giving "a head start for future analysis needing to allocate 'events' and their costs to individuals"[23]. To date, six organisations (in addition to the SIA) have used the code to create their own instances of SIAL tables within the IDI infrastructure saving an estimated $1m NZD that would have been spent on developing code with similar functionality in other departments. Further savings are likely to accumulate each time the SIAL is reused[24].

After initial creation, the SWA sought feedback from other social sector agencies, to review the code. The Ministry of Justice contributed new code to more accurately calculate the cost of court cases. They also created a more effective way to calculate the cost of education events – as recommended by the Ministry of Education. Both code modifications have been integrated into the SIAL.

2.3. User needs

The IDI, SIAL, and SIDF address at least four principle user needs.

In order to make evidence-based decisions, policy makers need to be able to measure the impact of interventions, ideally in monetary terms, and to understand what would have happened if the intervention had not been made. This usually requires both longitudinal data (i.e. spanning a period of time long enough to be able to measure impact), and joined data that combines information about individual citizens across a number of datasets, often provided by a number of different agencies (see Figure 2).

Public bodies also have a responsibility (often statutory) to produce timely statistics for publication to be consumed by other public bodies, businesses, and individuals. These statistics often involve combining data collected by several public bodies, and like the above need, may need to be historical (longitudinal).

In order to produce accurate research relating to public issues in New Zealand, researchers need to have access to the most up-to-date data in a way that protects the privacy of citizens and is subject to the necessary scrutiny.

21 http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-how-it-works.aspx

22 https://sia.govt.nz/how-we-can-help/measuring-outcomes/social-housing-test-case-2/

23 https://sia.govt.nz/assets/Uploads/sh-technical-report.pdf

24 https://data.govt.nz/use-data/showcase/sia-open-source/

Finally, users of the IDI need to be able to access data in a way that minimises the amount of time taken to prepare the data for their analysis, so that they can concentrate on extracting insights from the data.

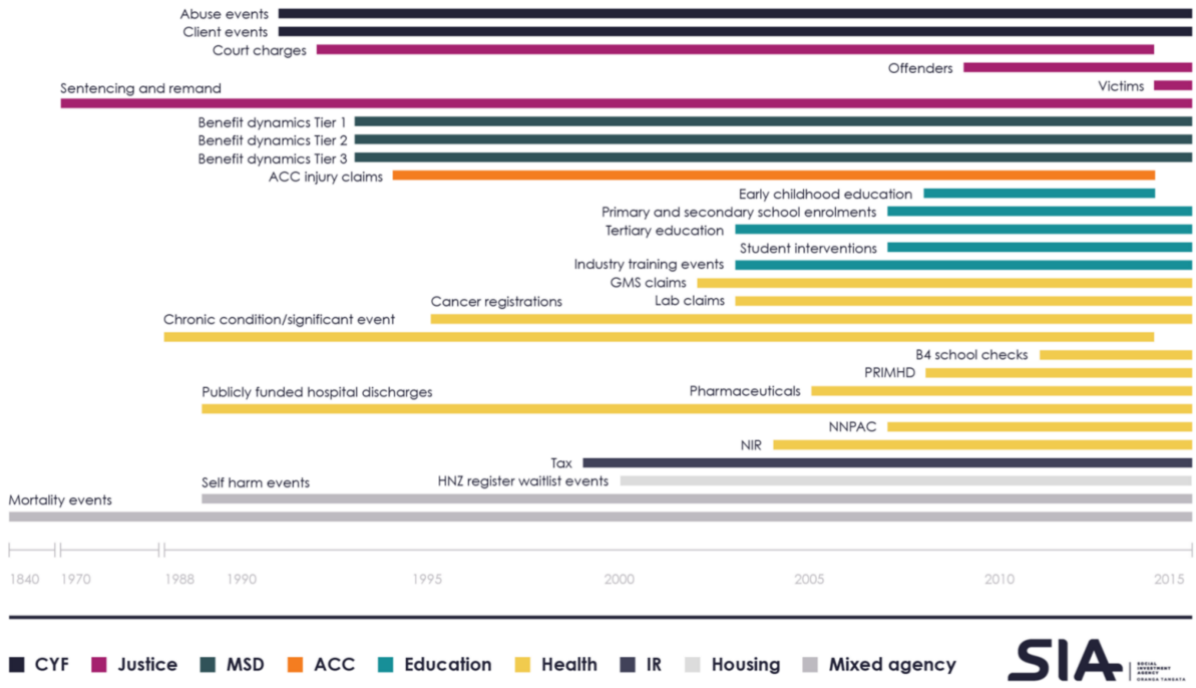


Figure 2. Schematic of data included in the IDI, from *SIA's Beginner's Guide to the IDI*[25].

2.4. Why is the IDI valuable?

The IDI successfully satisfies the first three user needs described above. Individual citizens can be tracked anonymously through around 550 datasets from 14 organisations, which include data from as early as 1840. This database is made available to all government departments and researchers who successfully fulfil the access criteria. Furthermore, by using the SIAL and SIDF, government spend can be calculated for each individual interaction an individual has with the state, for example: hospital procedures, sick pay, changes in employment and other changes that may occur following an accident.

The IDI additionally allows researchers and analysts to ask '*What if…*' questions through techniques such as Propensity Score Matching (PSM). PSM facilitates comparison between groups of individuals who have received an intervention and similar groups of individuals who have not, without necessitating a careful *apriori* experimental design to remove experimental bias. Such techniques have been used by the Ministry of Justice in studies of recidivism, and by the SIA in a study of the return on investment of social housing investment[26].

Few other countries have managed to build infrastructures like the IDI; the UK makes an interesting point of comparison. Like New Zealand, the UK does not have a single compulsory ID card or number, which creates a hurdle for creating integrated datasets. In both countries, the various other identification numbers must be used to try to match individuals, for example National Health Service number, driving license number, passport number, social security number, etc. In both countries it is likely

25 https://sia.govt.nz/assets/Documents/Beginners-Guide-To-The-IDI-December-2017.pdf

26 https://sia.govt.nz/how-we-can-help/measuring-outcomes/social-housing-test-case-2/#New_and_reusable_tools

that some individuals cannot be matched with these ID numbers alone, and so a process of probabilistic matching is performed which results in a 'best guess' match.

Whilst in New Zealand this work to integrate government data began in 2011, in the UK a widespread effort to integrate datasets from multiple departments has not been attempted. Some smaller efforts which span a handful of departments and datasets, for example the Longitudinal Education Outcomes (LEO) data[27] do however exist. Part of the difficulty in the UK has been overcoming legal barriers to data sharing between departments which is dependent on having a valid 'legal gateway'. This situation improved with the enactment of the Digital Economy Act in 2017 which included seven provisions[28] for data sharing, but these provisions are quite specific and do not cover a general purpose integrated data infrastructure like the IDI. Whilst the notion of an integrated data infrastructure may form part of the new Government Data Strategy[29] due to report in 2020, the UK lags New Zealand by at least a decade.

2.5. What the Social Investment Analytical Layer (SIAL) does

Like many large databases, the IDI can be confusing for new users, in part due to the variety of sources from which data are collected. Different agencies tend to have their own way of representing data, and representations from one department are not necessarily compatible with those of another. This is the problem that the SIAL tries to solve: it arranges the data from a subset of the c. 550 datasets in the IDI into a format that can more easily be worked with. This means that IDI users do not need to be an expert in the format of datasets from multiple agencies: they just need to understand how the SIAL is structured. This is work that many users would need to do prior to working with the IDI, which is why the SIAL has been used by several departments across the government.

Specifically, the SIAL does the following:

- Arranges data into a consistent format, making it easier and faster for authorised IDI users (researchers and analysts) to use and understand.
- Allows authorised IDI users (researchers and analysts) to quickly and easily undertake cross-sector analysis.
- Provides government spend in NZ$ for every individual in the SIAL, linked to events such as an individual claiming benefit, or accessing medical services.

2.6. What the Social Investment Data Foundation (SIDF) does

The SIDF is a layer built on top of the IDI and SIAL. Like the SIAL, it facilitates easier analysis of the IDI. Taking tables from the IDI and the SIAL as an input the SIDF calculates summaries of counts, costs, and duration for each person and government service described in the SIAL, for example:

- Time spent receiving a benefit
- Total money received while on a benefit
- Number of benefit spells over a given time period

27 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/790223/Main_text.pdf

28 https://www.registers.service.gov.uk/registers/information-sharing-agreement-powers-and-objectives-0001

29 https://www.gov.uk/guidance/national-data-strategy#what-happens-next

2.7. Technology

2.7.1. Technology choices

The IDI is built on existing technology in use by government departments in NZ. Along with the related Longitudinal Business Database (LBD), the IDI is a large relational database hosted in Microsoft Structured Query Language (MS SQL) Server. For this task, where the data are well structured, and not enormous, a SQL database is an obvious choice, and a tried and tested technology used by many thousands of businesses worldwide. Since the data that are stored in the IDI come from many different government agencies, SQL is also a good choice, as many organisations are likely to either already be using SQL, or data storage (for example spreadsheets) that can easily be ingested by a SQL database.

SQL is a ubiquitous technology that can be accessed by a variety of programs. The secure IDI environments allow access to proprietary statistical software such as SPSS, SAS, and Stata, in addition to open source software such as R, and in the future, Python.

The SIAL is a series of SQL scripts which create new tables by combining data from the tables which exist in the IDI. These scripts are executed through SAS, a proprietary enterprise analysis software often used by government departments for simple analysis. SAS offers a range of connectors for accessing data stored in various forms, in this case a MS SQL database.

The SIDF is a built mostly using SAS scripts which access and manipulate the data stored in the tables in the IDI and those created by the SIAL.

2.7.2. Using open source

In July 2016 The NZGOAL Software Extension was published. The New Zealand Government Open Access and Licensing (NZGOAL) framework guides government agencies on how to apply Creative Commons Licences to enable publicly funded data, information and other content to be legally reused. The Software Extension is guidance on how to apply open source licences to enable publicly funded software development to be reused.

Guided by the using the NZGOAL Software Extension[30] the SIA licensed content about their projects as CC-BY-SA (Creative Commons Attribution Share-alike) and the code itself with a GNU GPLv3 (General Public) Licence, both licences enabling people to share and modify the content and code.

In principle however, whilst the code for the SIAL and SIDF have been published openly, they are only useful to other users of the IDI, they could not be re-used by anyone without access to the IDI infrastructure. Even if this were not the case, whilst the code is open source, the programs (MS SQL and SAS) required to run the code are both proprietary, although the open source language R is also available in the IDI secure environments.

2.7.3. Architecture

The IDI is a relational database implemented in MS SQL Server. The database itself is large (totalling over 1TB) but not sufficiently large to require specialised big data technologies such as Spark, or Hadoop. This is because typically a given analysis would only make use of around 10% of the 550 available data sources.

The SIAL and SIDF are created using open source code within a user's own workspace within the IDI environment.

30 https://www.ict.govt.nz/guidance-and-resources/open-government/new-zealand-government-open-access-and-licensing-nzgoal-framework/nzgoal-se/

2.7.4. Hosting

The IDI is cloud hosted in two data centres in NZ and one in Sydney, Australia. Access to the database is provided at three secure locations in New Zealand known as Data Labs.

2.8. Lessons learned

2.8.1. For the IDI

2.8.1.1. The service will start small, be tested regularly, and iterated to meet users' needs

The IDI started as a prototype and was iterated from there, until 2011 when it moved from one-off data integration to providing a whole data integration service. The LBD originally started at the beginning of 2006 as a two-year feasibility project.

Part of their success in opening up data is down to working with users to understand their data needs and connect them with relevant government contacts. At the end of June 2019, the percentage of positive responses by customers who were asked if they could find what they were looking for was 40%. The main barriers New Zealand have experienced are having good visibility of what data is already available, lack of clarity about what rights people have to data and inconsistent processes for requesting data.

2.8.1.2. Good security practices are essential when managing personal data

Data in the IDI is de-identified, with information like names, dates of birth, and addresses removed. Numbers that can be used to identify people, like tax references, are encrypted and Stats NZ check research results before they're released to make sure individuals can't be identified.

After integrated data has had identifying information removed, only vetted and approved researchers can access selected datasets for their specific project. Research must be for the public good and data can only be accessed in Stats NZ's secure research data facilities. Before any new data is added to any Stats NZ service, a privacy impact assessment is carried out to consider any risks.

2.8.1.3. Standards for public trust and transparency

They apply ethical, statistical, and security best-practice standards to the data being collected, and people who use the data must apply the same standards. Stats NZ staff and researchers who use data have to sign a statutory declaration of secrecy which is a lifetime agreement.

2.8.1.4. Cloud and onshore environments to ensure sustainability

All research services are operated out of cloud environments, with two onshore data centres and one in Sydney. Given the volatile nature of New Zealand's physical environment, this is essential for resilience. To ensure these centres are always available and are resilient, they have production and disaster recovery servers operating out of geographically separated Infrastructure as a Service (IaaS) data centres.

2.8.1.5. Keep the project accountable and updated

Stats NZ regularly publish progress against their strategy and report to the Minister of Statistics and New Zealand's parliament on progress against their objectives, impact statements, and financial performance.

Also, Stats NZ publishes regular updates to its release calendar and both research services are growing in terms of numbers of datasets.

2.8.2. For the SIAL and SIDF

2.8.2.1. Open source saves time and resources

The code behind the SIAL and SIDF is published under a permissive license allowing it to be re-used by government and academic users of the IDI. These two pieces of infrastructural code were produced in the course of two projects, and might not have been published if it were not for a commitment to publish openly. This commitment has saved at least six other departments from the need to produce their own data manipulation code to coerce the IDI into an easier to use format.

2.9. ANNEX - Data collection activities

2.9.1. List of consulted stakeholders

- Data Scientist (x2). NZ Social Investment Agency

2.9.2. Sources

- https://sia.govt.nz/assets/Documents/Social-Investment-Analytical-Layer-Code-User-Guide-December-2017.pdf
- https://github.com/nz-social-investment-agency/social_investment_analytical_layer
- http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-how-it-works.aspx
- http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-data.aspx
- https://sia.govt.nz/assets/Documents/Beginners-Guide-To-The-IDI-December-2017.pdf
- https://sia.govt.nz/assets/Uploads/Measuring-the-wellbeing-impacts-of-public-policy-social-housing.pdf
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181260
- https://statsnz.contentdm.oclc.org/digital/collection/p20045coll17
- https://thehub.sia.govt.nz/assets/Uploads/Growing-up-schools.pdf

3. CASE STUDY: FINDATA - FINLAND

3.1. Introduction

Finland has a long history of collecting extensive data in registers but making use of the data has been complicated and inefficient, due to several causes, including the privacy restrictions and the supporting technologies. In 2019 a new Act on Secondary Use of Health and Social Data[31] (hereinafter "Act" or "new Act") entered in force in Finland. With the new enabling legislation, Finland has become the first country in the world to successfully enact a law on the secondary use of well-being data that meets the requirements of the European General Data Protection Regulation (GDPR).

The new legislation enables and expands the use of social and healthcare data from the traditional areas of scientific research and statistics to those of management/control of social welfare and healthcare, development and innovations, knowledge management, high-level education, authorities' planning and forecasting

[31] *Act of Secondary Use of Health and Social Data;* https://stm.fi/en/secondary-use-of-health-and-social-data

tasks, and steering and supervision of work. The new Act facilitates the establishment of a new one-stop-shop operator and central data permit authority in Finland, known as Findata, that collects and co-ordinate well-being data on the Finnish population for use in several areas. The Findata organisation has been established and operates within the National Institute for Health and Welfare, but as a separate entity and with dedicated technologies.

Findata is responsible for streamlining and securing the secondary use of social and health data. It guarantees a flourishing ecosystem (both organisational and technological) around the secondary use of social and health data streamlining the processes for the issuing of research permits and data collection and ensuring that data are being used in secure environments, thereby maintaining the trust that the general public have in authorities and the public sector.

Three key technological systems have been prepared to support Findata operations: a permit and information portal (that is already available), a data description system (metadata) and data management system for the data collection and processing. It also includes a remote desktop for data pseudonymisation and anonymisation services. The secure remote environment with associated tools are used by data scientists to experiment with data.

This case study reports how the work on Findata has been developed, and it describes the delivery model focusing on the services provided to the users and the supporting systems. The general part includes a brief introduction to the implementation approach.

This study introduces the design principles underpinning the solution such as agility, good security practices, the variety of tools offered to data scientists and sustainability.

In the technology section, the virtualised environment, available for data scientists to experiment with data using pre-installed analytics tools, is analysed. The study also focuses on the reusability of the solution. An overview of the Findata ICT environment is also provided.

The final section details relevant lessons learnt and the challenges and benefits of Findata.

3.2. Development of the work

From a legislative perspective, Finland has been developing its legislation to support electronic healthcare and social welfare services in the Act. The working Committee prepared the new Act from April 2015 to December 2017. The first official proposal was submitted for the first hearing in August 2016, and in October 2017 the Act was proposed by the Government. Parliamentary work, expert hearing and database, and social affairs and health committee suggested amendments in April 2018. An adjusted proposal was returned to the Committee in October 2018. Finally, the new Act was approved by Parliament in March 2019 and entered into force on the 1st of May 2019.

The data permit authority (Findata), was established by the new Act, to ensure the ethically sustainable use of data, and it started operating at the beginning of 2020. The granting of data permits for health and social data was centralised with the new authority, allowing data from numerous different data controllers to be gathered from one source. A centralised data permit system for the processing of permits and secure user environments, where data can be worked on, is under implementation and will be ready soon. With the new Act, there are clear legal grounds for using register data in innovation and development activities. Companies can get collated, aggregated statistical data for these purposes more quickly and comprehensively.

From a technological perspective, several (pilot) projects[32] played an essential role in building an innovation ecosystem. The project's main accomplishments were the

[32] *Pre-production projects; https://www.sitra.fi/en/articles/one-stop-shop-well-data-isaacus-laid-foundations-future/*

creation of a prototype for the one-stop-shop service model, the building of new technical infrastructure and greater expertise in the use of new technologies and multi-stakeholder collaboration..

A project steered by the National Institute for Health and Welfare (THL), Statistics Finland and the Finnish Social Science Data Archive developed a solution for creating and managing common metadata descriptions, integrated into a public web service where they can be used by other stakeholders (e.g. researchers, for example).

The projects by the Hospital District of the City of Kuopio, the Hospital District of Helsinki and Uusimaa (HUS) and the Hospital District of Southwest Finland have created a data lake solution to collect well-being data that has previously been dispersed into separate places in operational systems. This approach enables the faster and more comprehensive use of data than before.

In addition, a project by the Hospital District of Southwest Finland has refined data collected from the Data lake and created views for different types of users at hospitals, while a project by the City of Kuopio is piloting the use of data collected from the data lake. The aim is to predict the optimal service process for youth work, family social work and home service for the elderly.

A data-secure environment for the use of well-being data has been designed in a project carried out by Statistics Finland, THL, the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI.fi) and the Institute for Molecular Medicine Finland (FIMM).

A project led by BBMRI.fi has created a common process and tools for all biobanks for the formulation of research data. The pilot combines digitised samples collected by the biobanks from breast cancer patients and the data related to the samples, and information from pharmaceutical reimbursement statistics and the Finnish Cancer Registry with clinical treatment information.

A project led by the National Archives of Finland (link in Finnish) created a Finnish permit service operating on a one-stop-shop basis, in which researchers can apply for authorisation to use the data and material in the social welfare and healthcare register. In addition, the project implemented an information and support portal offering information about the access limitations of social welfare materials, the prerequisites for using data, and the various intended uses of the data.

3.3. Delivery model

The pilot projects described above aimed at preparing the delivery model of Findata. It provides well-being data and open data from different information sources and registers[33] on a one-stop-shop basis. The permit and guidance services for the users of this data were made available in one place. When data is processed, special attention must be paid to privacy protection, data security, and the right of individuals to decide who the information concerning them is used by and for what.

Before describing the Findata delivery model, it is important to clarify that, there are two different levels of data and different ways to access related datasets:

1. **Individual level data:** this data can be used for scientific research, statistics, education, authorities' steering, supervision, planning and forecasting. This data is available in a remote access environment for a set period. The data has been anonymised or pseudonymised. A data utilisation plan is required for access to data sets.
2. **Statistical level data:** The data of this level can be used for the aforementioned purposes and, in addition, for development and innovation and knowledge management. This kind of data is directly delivered to customers.

[33] For example the cancer, infectious disease, hospital discharge, national vaccination registers, among others.

Findata provides and will provide soon the following main services supported by related technologies/systems:

1. **Help desk for data users** (already available): it came online on the 1st of November 2019 in the form of email and phone.
2. **Application for permits and data requests** (already available): from the 1st of January Findata issues and grants research permits, including ethical evaluation for anonymised statistical data. From the 1st of April 2020 it also accepts applications for individual-level data. Users can apply through the permits and information portal.
3. **Data service** (coming soon): Findata will collect relevant data from different registers and edit, combine and anonymise the data before distributing it to users. This is possible through the data management system. Additionally a data description system is available as a centralised place for saving the metadata of available materials. The solution includes, among other things, a metadata editor for editing and updating description data.
4. **Secure remote use environment** (coming soon): as future features, the Findata infrastructure will be able to provide a secure remote user environment with associated tools for customers. The remote environment is further described and analysed in the "Technology choices" section.

Starting from 2021, Findata will also be permitted to use data for secondary use from Kanta services. Kanta is a place for digital patient records and client data: information about a patient's health status, medical history and prescriptions is available to all treatment providers, both public and private, in one safe and secure location[34].

3.4. Implementation of Findata

The systems and technologies that enable the implementation of the Act and the operations of Findata have been developed following an agile implementation methodology by a consortium led by Sitra, the Finnish Innovation Fund, an independent public foundation which operates directly under the supervision of the Finnish Parliament.

Sitra, together with a set of institutional and technological partners, has been preparing the launch of Findata by launching several pre-production projects. The first projects were launched in summer 2016, and the experiences from these projects were collected and integrated into a plan of action to enable the operational launch of Findata 2020 (initial plan was 2018).

Sitra's projects prepared the launch of Findata by piloting, testing and developing service models and processes, metadata descriptions, data lakes and co-operation models with authorities and stakeholders.

This implementation approach was chosen mainly due to the need to deal with a very tight time scheduling and the need to create a plan of action based on concrete outcomes of the projects.

3.5. Re-usability

The concept of reusability in the Findata ecosystem is severalfold. Firstly, Findata has reused the existing technology platform of a cloud provider (nominally Microsoft Azure). Secondly, the platform makes use of existing open source technology, for example in the backend data storage an open source technology Apache Hadoop has been used. In the analytical environment R, Python, Jupyter and other open source tools are made available for the data analysis. Furthermore, various tools built in one of Sitra's pilot projects: Isaacus, have been made available as open source software[35]

[34] *Kanta -The Place for Digital Patient Records and Client Data; https://www.slideshare.net/THLfi/kanta-the-place-for-digital-patient-records-and-client-data*

[35] https://github.com/Sitra-Isaacus

and formed the basis of the technology used in Findata. Finally, the analytical platform can be scaled up into multiple simultaneous instances provided to different groups of users. Similarly, the data storage solution is modular, and can be scaled as the data storage requirements grow.

In addition, the data platform could also be potentially used by other institutional organisations since it is possible to transfer the material to a non-Findata secure environment. However, this can only be done for necessary reasons and the rest of the operating environment must meet the security requirements set by Findata.

3.6. Technology

3.6.1. Design principles

The design principles underpinning the solution can be summarised as follows:

1) **Agility** - To deliver to a very tight schedule, an iterative and agile implementation approach has been preferred, based on pilot projects carried out by different stakeholders coordinated by a single entity (Sitra).
2) **Good security practices** - For the construction of a data-secure environment, the following practices are guaranteed:
    - anonymisation: data sets are made available without individual direct identifiers. The data permit authority Findata sees to anonymisation centrally;
    - data encryption;
    - strong authentication;
    - a secure remote access environment for the individual level of data sets;
    - security in the Findata operating environment will be controlled by an information security inspection body and National Supervisory Authority for Welfare and Health (Valvira).
3) **Sustainability** - using a virtualised cloud infrastructure[36].
4) **Variety of tools and technologies offered to data scientists:** the data scientists can ask to set-up in the cloud environment the software they prefer.

3.6.2. Technology choices

3.6.2.1. Data storage

Data is stored by Findata in a data lake. A data lake is a sensible solution for the services that Findata offers because it allows for data from multiple disparate sources to be stored together easily. The specific implementation used by Findata is based on Apache Hadoop[37], an open source technology that was trialled in three precursor projects[38]. Hadoop has the benefit of being 'schema on read', meaning that data does not need to have a strictly defined format when it is added to the data lake, it is instead applied when the data is accessed. This approach is unlike many other data storage technologies, which may require all data standardisation to be done as the data is ingested; Hadoop thus offers great flexibility to deal with disparate data sources in a wide variety of ways.

Hadoop is a 'big data' technology which can be scaled easily to accommodate very large quantities of data, and is used by large tech firms such as Uber, Airbnb, Netflix, and Twitter[39]. In fact, there are probably grounds to argue that Hadoop is

[36] Virtualised cloud environments ensure sustainability by obviating the need to procure and manage physical computing infrastructure, thereby outsourcing those tasks to the cloud service provider.
[37] http://hadoop.apache.org/
[38] https://www.sitra.fi/en/projects/isaacus-pre-production-projects/#results
[39] https://stackshare.io/hadoop

overspecified for the task at hand, as it is unlikely that Findata's data storage requirements will be in the same order of magnitude (usually petabyte scale) as the tech giants who have popularised this technology. That said, the scaleable nature of the technology means that it can be scaled down as well as up. Indeed this was recognised in the Isaacus pre-cursor project: it was noted that the compute requirements were not likely to be constantly heavy, and additional resources could be brought online to accommodate the peaks in demand as required.

There are also some downsides to the technology; handling smaller quantities of data may actually be slower and more difficult to perform using Hadoop than other more traditional technologies, for instance a relational database management system (RDBMs) like Postgresql or MySQL. In the Isaacus precursor project, this seems to have been overcome by using the data lake to store just the rawest form of the data, whilst exposing cleaner, anonymised data to users via more traditional RDBMs.

Another downside of Hadoop, which was highlighted in the evaluation of the precursor projects is overcoming a technical skills shortage. A data lake like Hadoop requires a level of ongoing technical support that will almost certainly need to be sourced from outside the public sector, whereas many public bodies retain the technical capabilities of maintaining more traditional data storage systems like RDBMs. In the case of Findata, this expertise is being provided by Tieto, a Finnish digital services and software company.

3.6.2.2. Analytics Environment

Findata provides a secure virtualised environment (ePouta virtual private cloud), for processing pseudonymized individual-level data.

Users receive usernames for the remote environment when:

1. Findata approves permit application
2. Findata collects material from data controllers
3. Users complete the remote environment order form
4. Users define an agreement with Findata to use the remote access environment.

Within the environment, users are granted access to a range of statistical software required for data analysis: SPSS, Stata and R software are included as standard in all machine packages. The software range will also be expanded. Users can request other software that fits their needs. The usage fee for the remote operating environment is based on the selected machine package available for the remote environment:

- S (small):
    - 8 GB RAM
    - 4 cores
- M (medium):
    - 16 GB RAM
    - 6 cores
- L (large):
    - 32 GB RAM
    - 8 cores
- XL (MaxPower):
    - 90 GB RAM
    - 20 cores

It is possible to get an additional order and for an additional fee SAS software.

It is also possible to install other programs on the remote environment, either open access or licensed programs. There is an hourly rate. In the case of a program that requires a license, the applicant must have that license itself.

The basis for determining prices is laid down in the payment decree of the Ministry of Social Affairs and Health.

A task force is also investigating the potential use of Artificial Intelligence (AI).

3.6.3. Using open source

As noted, the data lake where the various data sets and registers are stored is based on the open source Apache Hadoop and related ecosystem of open source tools. A requirement highlighted in the assessment of the data lake pilot projects was that the system should be portable, meaning that the system could be moved and expanded as required, without being locked-in to one particular provider; using open source tools is the best way to achieve this.

Among the tools available for users in the virtualised ICT environment, a number of open source tools are made available, including R which is widely used among academic researchers, public sector and industry data professionals.

3.6.4. Architecture

The virtualised ICT environment of the Findata technological architecture is represented in Figure 4 below:

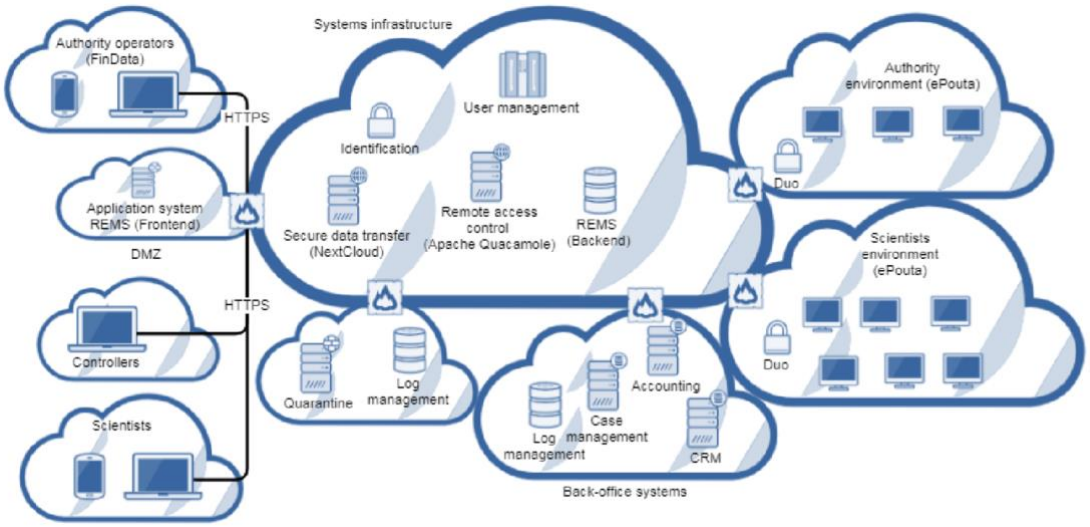


Figure 4 – Findata ICT enviroment

The data scientists and controllers use the application system REMS (front-end) to communicate with Findata. In particular, through the front-end, a data scientist communicates with Findata operators to request access to the virtual environment as well as software and tools use permits.

Findata harmonises all data between the infrastructure's different systems, allowing or denying permit applications for the use of the virtual environment. It is furthermore responsible for the anonymisation of data and for data registry.

In addition to the identification system and the user management system, NextCloud and Apache Quacamole are used to manage secure data transfer and remote access control, respectively. These two system increase security, in particular:

- Thanks to NextCloud, Findata provides a safe and fast transmission of the data to the users.

- Remote access control ensures that access to data is possible in visualisation mode only. The system prohibits the usage of removable media, allowing only peripherals such as a mouse, keyboard and monitor. Users are thus unable to transfer data in or out of the system infrastructure.

Additionally, the data received in the system are temporarily stored in Quarantine. There, their consistency is checked, and they are checked for viruses before being decrypted and made available in the environment.

The Back-office Systems make sure that all the archiving operations are done correctly. In particular, the Log management stores any information regarding users and their behaviour in the system (e.g. how much time they are online).

The Scientists' Environment (ePouta) provides data scientists with their own virtualised environment. Findata populates these environments with data and tools needed by the users. Once a data scientist completes its work, Findata deletes its virtual environment.

3.6.5. Hosting

The secure environment for data processing used by data scientists and authorities is hosted in ePouta Virtual Private Cloud (VPC) (an IaaS Cloud Computing service). It allows users to provision virtual machines and storage resources within the VPC. The virtualised infrastructure consists of, but is not necessarily limited to, these resources:

- Virtual machines (instances);
- Block devices that can be attached to virtual machines (volumes);
- Virtual networks that can be used to connect virtual machines.

The ePouta service is designed for processing sensitive data. It is suitable for all fields of science, and for government and research-sector organisations. The cloud service combines virtual computational resources with the users' own resources using a Virtual Private Network solution. The service is easily scalable to users' requirements.

Users can manage their resources using a web interface accessible through a web browser and through a set of APIs which allow programmatic management of resources. In order to access and use the service, the user must have a CSC user account.

3.7. Lessons learnt, benefits, challenges

Finland has succeeded in creating a new ecosystem built around the use of health data through a national project culminating in the ground-breaking new legislation. But this result was not achieved without important lessons learnt, useful also for other MSs and for the European Institutions.

First of all, the Digital Skills gap: emerging technologies enabling big data innovations and integration challenges between different data sources and data types require a good mix of skills that can enable the use of the new technology and methods, analytics skills and achieve a good understanding of research practices. The majority of pilot projects, focusing on Digital Health HUB development, were highly technical. The projects tested the integration of data from existing systems with the aid of new technological solutions, developed existing systems to meet future requirements and developed entirely new systems. Technical problems were caused either by new technology, such as the support for and functionality of early-stage data-lake software versions or by new requirements.

Additionally, from a legislative perspective, is also to be taken into account the rapid development of technology that makes it difficult for legislation to keep up and for legislators to understand development. Technological solutions change so fast that they should not be recorded in legislation. Instead, the law should include the requirements to be met by technological solutions.

From a technological perspective, one of the main benefits of Findata ecosystem is related to its capability to enable effective, secure and safe processing and access to data. Thanks to its technological architecture and overall governance processes, retrieving combined health and social data from different sources is easier and faster and possible with just one permit application, removing the need to approach each authority and data source separately. Indeed, in the past, these processes were very time-consuming.

Another benefit regards the data protection that is constantly increasing. Using Findata, personal data are not delivered to a personal computer or removable media, but they can only be accessed in the remote access system, where they can be used in a controlled and secure way.

Currently, Findata is a system only available in Finland. The system could be extended to other European and non-European countries, but giving access to non-Finnish users is difficult at the moment since the identification of external applicants is not possible. Globally acceptable solutions for strong authentication are still under investigation.

The initiative faces several challenges. The first big challenge is represented by the tight time schedule. In addition, some operating steps are not ready yet. They will be ready during this or the next year. The second big challenge regards metadata. Indeed, some data cannot be linked because they are not well described or structured. The plan is to achieve a standard data structure for all data in a couple of years.

3.8. ANNEX - Data collection activities

3.8.1. List of consulted stakeholders

- Mr Heikki Lanu
- Ms Johanna Seppänen
- Mr Jaana Sinipuro

3.8.2. Interviews

- Mr Heikki Lanu (Head of ICT) from National Institute for Health and Welfare
- Ms Johanna Seppänen (Director) from National Institute for Health and Welfare

3.8.3. Presentation

- How the legislation for the secondary use of social and health care data and implementation was prepared the Isaacus project.
- Implementation of the national Social and Health Data permit authority Findata. Johanna Seppänen, PhD, Director.
- SECONDARY USE OF HEALTH AND SOCIAL DATA IN FINLAND. 22.11.2019. Joni Komulainen, Ministrial adviser, Master of Laws.
- A FINNISH MODEL FOR THE SECURE AND EFFECTIVE USE OF DATA.
- Interview for Findata case study 23.3.2020. Mr Heikki Lanu Head of ICT.

3.8.4. List of consulted documents

- Secondary use of health and social data; https://stm.fi/en/secondary-use-of-health-and-social-data?p_p_id=56_INSTANCE_7SjjYVdYeJHp&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=3&_56_INSTANCE_%C2%AD7SjjYVdYeJHp_%C2%ADlanguageId=en_US
- Act of Secondary Use of Health and Social Data; https://stm.fi/en/secondary-use-of-health-and-social-data
- Findata; https://www.findata.fi/en/

- Johanna Seppänen, Secondary use of Finnish Social and Health Data - a new Act and Data Permit Authority; https://www.himsseuropeconference.eu/sites/himsseuropeconference/files/sponsors/hims s-eu-19/presentations/seppanen_secondary_use_of_finnish_social_and_health_data_-_a_new_act_and_data_permit_authority.pdf
- Jaana Sinipuro et al, A finnish model for the secure and effective use of data, Sitra, 2019; https://www.sitra.fi/en/publications/a-finnish-model-for-the-secure-and-effective-use-of-data/
- Joni Komulainen, Secondary use of health and social data in Finland- How to securely fully utilise the health and social data for research, development and innovation activities, education and knowledge management duties, 2019;
- Hannu Hämäläinen, How the legislation for the secondary use of social and health care data and implementation was prepared – the Isaacus project, eHealth Network, 2019.
- Anna Hammais, Juha Varjonen, and Arho Virkki (Editors), The Clinical Data Refinery, Management and Administration of the Analytics Environment, Kliininen Tietopalvelu Centre for Clinical Informatics. Available at: https://github.com/Sitra-Isaacus/VSSHP-tietoallas-dap/blob/master/book/cci_book.pdf, 2018.
- Richard Darst, Mikko Hakala, and Kimmo Kaski, Evaluation of the Isaacus project's data lake solutions in research use, Department of Computer Science Aalto University School of Science. Available at: https://media.sitra.fi/2017/02/07110947/datalakeevaluation30042017final.pdf
- https://github.com/Sitra-Isaacus/HUS-tietoallas-main/blob/master/HUS%20Suunnittelu%20v1.1.pdf, 2017
- Liisa Jansson, Challenges and solutions for building social and health information pools [translated], MSc Thesis. Available at: https://lutpub.lut.fi/bitstream/handle/10024/160273/diplomityo_jansson_liisa.pdf?sequence=1

4. CASE STUDY: KOKE SYSTEM AND AUTOMATED RISK MODELS - ESTONIA

4.1. Introduction

EMTA (Eesti Maksu- ja tolliametile) is the Estonian Tax and Customs Board[40]. The Estonian Tax and Customs Board's activities include administration of state revenues, implementation of national taxation and customs policies and protection of the society and legal and economic activities. The journey towards data analytics started in 2004 upon its Management request. Starting with a pilot, now they have optimised their daily work leveraging on data analytics and they also achieved significant results in terms of cost reduction. The use of data analytics within EMTA is now also driven by the Estonian Digital Government strategy[41] on this matter, which promotes further and further big data take up across administrations.

EMTA uses big data and data analytics technology for fraud detection and evaluation purposes. Through data analytics, they redefined their strategy towards the identification of cases to verify. They moved from an "unstructured approach" to this "case selection towards data-driven methods" based on an algorithm identified risk coefficient for each case, with the overall objective of increasing tax compliance and preventing fraud. For this purpose, EMTA analyses a large amount of structured data coming from government sources, mainly such as business registers and tax declarations.

Nowadays, EMTA management is also taking decisions about the organisation's activities based on big data mainly. In this context, EMTA identified and developed different risk models that follow different algorithms to identify different tax risks. The main is the risk model that analyses VAT declarations. All the risk models are developed in-house using SAS, a proprietary analytics software.

EMTA has harmonised several systems in order to meet risk management expectations. The current study is focused on the system used by the tax audit unit (KOKE – Control Environment) that receives the list of high-risk cases automatically calculated by the VAT risk model.

4.2. Development of the work

EMTA Tax audit unit used KOKE as the everyday working platform for all auditors. The work on the KOKE system started in 2008, with an initial analysis of auditors business needs and how to satisfy them with an efficient IT system. In 2010, the European Regional Development Fund funded the initiative. In 2011, the system was developed by Cybernetica[42] following a waterfall approach.

The cost of the system has been 189.204,04 euros, entirely funded by the European Regional Development Fund (ERDF). Maintenance costs are around 30.000 euro/year, and are of course strongly linked to the changes implemented in the system.

Although KOKE system data fields were initially defined by EMTA with external developers, EMTA can change data fields, according to the needs, autonomously and so without further expense. Adding modules or connections to other databases however has to be done by contracting external developers.

4.3. Delivery model

Taxpayers submit tax declarations to have a tax return on the e-MTA portal[43]. In Estonia, every taxpayer has an ID card that automatically logs the citizens in the system for presenting the declaration. Each application is automatically evaluated by

[40] Estonian Tax and Customs Board; https://www.emta.ee/eng
[41] Estonian digital government strategy; https://e-estonia.com/
[42] Cybernetica; https://cyber.ee/
[43] E-MTA portal; https://www.emta.ee/eng

risk models in order to identify risks. The cases with higher risks are stored and elaborated in KOKE. Once the cases are collected in the system, auditors can review their tax declarations and integrate them with more information (there is a space on KOKE where the auditors can add information manually). The process for the selection of the audit cases can be summarised in three main steps, as shown in the figure below:

1. The taxpayer submits a VAT declaration and the appendix to the value-added tax declaration in the KMD form (value-added tax return) on the e-MTA portal;
2. The automated risk models (the most important is the VAT risk model) calculate the high-risk cases and send them to the auditors (KOKE).
3. KOKE receives the list of the cases with a higher risk of fraud, and the auditors can start to analyse them.

Figure 1 - Process for the selection of the high-risk cases[44]

The users of the KOKE are auditors (responsible for preparing the work and elaborating all the data), service department consultants and managers (responsible for approving the work).

EMTA is allowed by the regulation on Statutes of the register of taxable persons[45] to collect and analyse personal data in the respect of the General Data Protection Regulation.

4.4. Re-usability

KOKE is a scalable solution, and it can be used by other units (e.g. the taxation department).

4.5. Technology

4.5.1. Design principles

The design principles underpinning the solution can be summarised as follows:

1. **The system started small and iterated to meet users' needs:** the initial version of the VAT risk model analysed only VAT payments, later on, the model evolved to address different needs.
2. **Good security practices:** the EMTA approach is to ensure that security issues are appropriately addressed through the following key business decisions:
    a. policy based on ISKE[46] (three-step reference system for information systems), in the current version 8.00. This includes audit at least every three years with the external auditor reviewing both the organisation of information security and its implementation;
    b. individual and shared responsibilities for information security;

[44] "KOKE4 – audit desk program", Herje Vahemäe, Republic of Estonia Tax and Custom Board.

[45] Statutes of the register of taxable persons https://www.riigiteataja.ee/akt/112032019012?leiaKehtiv

[46] ISKE manuel; https://www.ria.ee/sites/default/files/content-editors/ISKE/iske-implementation-manual.pdf

c. preventive technical and organisational measures for the protection of private life and tax secrecy;

d. on-premise solutions (as opposed to cloud);

e. identification of a Data Embassy, to ensure the digital continuity of the country, located in Luxembourg.

3. **Flexibility:** KOKE is a flexible system since it has the possibility to access different risk models. This is really helpful for fraud detection.
4. **Independence from the vendor** – The methodologies are mostly automated, and the systems are managed in house.

4.5.2. Technology choices

All modules of KOKE run on an Oracle database.The aggregated data coming from the database are elaborated and also visualised by the top management using SAS tools[47], in particular by the intelligence department and the internal control for monitoring and reporting purposes.

4.5.3. Using open source

No open-source software is used. EMTA uses only proprietary solutions. The code is not public.

4.5.4. Architecture (Data sources and Data flow)

As shown in the picture below, the data are collected from the KMD form filled by the taxpayer and submitted in the portal. The data are stored in the KMD database (Oracle). The database is divided into two types of data: raw data and aggregated data. Indeed, the database has an aggregated mid-layer responsible for aggregating data as follow:

- total of VAT declaration annex form where sales invoices are declared (INF A);
- VAT declaration annex form where sales invoices are declared (INF A) per partner;
- total of VAT declaration annex form where purchase invoices are declared (INF B);
- VAT declaration annex form where purchase invoices are declared (INF B) per partner.

The aggregated data are analysed by the automated risk models, and in particular by the VAT risk model called VATSUM that is model linked to the KOKE and VAT listing (where taxpayers must provide all invoices that they have exchanged with transaction partners) that elaborates dataset of information concerning taxpayer´s VAT declaration and the appendix to the value added tax declaration. The VAT risk model also takes into account salary declaration data, data from Vehicle registry, data from Business Registry and risk information that is collected in EMTA (including customs risks).

From the analysis of the risks models, the cases with higher risk to KOKE are identified and analysed by the KOKE system.

The KOKE system is composed of different modules that can be updated with different regularity. The modules can be updated on a daily basis by different data sources such as VAT overpayments, VAT registration application and payroll tax declaration. Based on the needs, the modules can also be updated on a weekly and/or monthly basis.

[47] SAS tools; https://www.sas.com/fi_fi/customers/estonian-tax-and-customs-board.html

The aggregated data are also sent to other SAS tools for data elaboration and visualisation used by senior management (intelligence department and internal control).

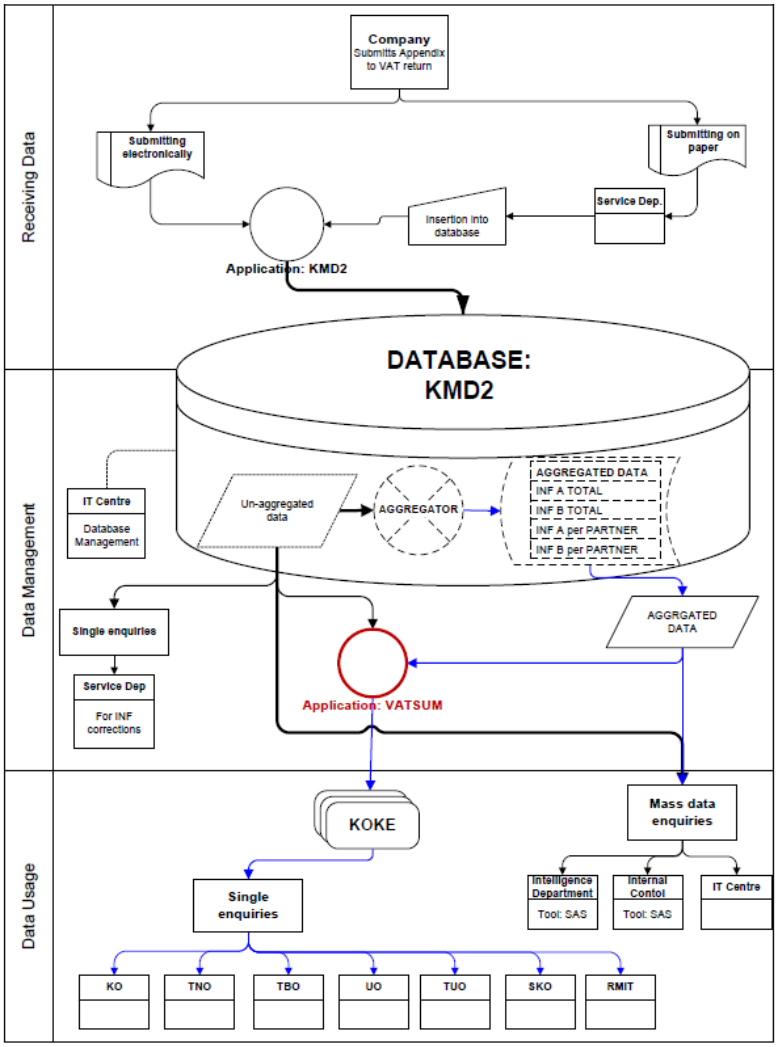


Figure 2 - Data flow[48]

4.5.5.

4.5.6. Hosting

EMTA opted only for on-premise infrastructure, mostly due to privacy and security constraints.

4.6. Lessons learnt, benefits, challenges

4.6.1. Increment in fraud detection

The system helped EMTA in fraud detection. Since the system is very flexible, it is possible to assess different risk models easily and quickly, addressing different types of frauds.

[48] Source: "KOKE4 – audit desk program", Herje Vahemäe, Republic of Estonia Tax and Custom Board.

4.6.2. Change management issues

The process of change management is expensive in terms of cost and time. Indeed every change in the system must be outsourced, and it is time-consuming because of the time needed to add new functionalities.

4.6.3. Improvements on KOKA system

EMTA is evaluating two possibilities to answer new requirements (in particular, the issue related to flexibility): create new modules in KOKA or create a new environment. Currently KOKA is mainly used for auditing and reused by other service departments, but creating a new environment can provide to EMTA the possibility to increase the audience including also additional services. In this way the new system can be used by all the Estonian tax and support services. The main idea is to have a more flexible frame that provides the possibility to change content and logics autonomously. Today the change management process is not efficient, it is really expensive in terms of costs and time.

4.6.4. Increase the efficiency

Thanks to the use of analytics solutions for fraud detection, the Estonian Tax & Customs authority has reduced the number of people in the organisation substantially as they were able to optimise their working processes by involving innovative technology solutions.

4.6.5. Technical people with a strong math background but without experience in fraud

The Estonian customs tried to involve people with very strong math and statistical background but without any experience in the domain of fraud and discovered that they find it difficult to imagine how fraudsters think. For fraud identification, EMTA needs data-savvy subject matter experts to understand who is breaking the rules. That is why EMTA figured out that it is easier to teach fraud experts to work with big data than the other way around.

4.7. ANNEX - Data collection activities

4.7.1. List of consulted stakeholders and interviews

- Mr Ailo Jõgi from Estonian Tax and Custom Board
- Ms Herje Vahemäe from Estonian Tax and Custom Board
- Mr Jaanus Timusk from Estonian Tax and Custom Board
- Ms Natalja Samsonova from Estonian Tax and Custom Board

4.7.2. Presentations

- "Automated notification system IRIS", Republic of Estonia Tax and Custom Board.
- "KOKE4 – audit desk program", Herje Vahemäe, Republic of Estonia Tax and Custom Board.

4.7.3. List of consulted documents

- Estonian Tax and Customs Board; https://www.emta.ee/eng
- Estonian digital government strategy; https://e-estonia.com/
- Cybernetica; https://cyber.ee/

- SAS tools; https://www.sas.com/fi_fi/customers/estonian-tax-and-customs-board.html
- Big data analytics for policy making; A study prepared for the European Commission DG INFORMATICS (DG DIGIT); https://joinup.ec.europa.eu/sites/default/files/document/2016-07/dg_digit_study_big_data_analytics_for_policy_making.pdf