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

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. INTRODUCTION

1.1. Purpose of the Report and Relation with other project work

The data explosion is affecting all aspects of the society and the economy – and public administration is no exception. Data is a fundamental resource for carrying out all government activities, from regulation to service provision. And governments everywhere and at all levels are looking into the opportunities of data driven innovation, and in many cases experimenting with it. IDC estimates that central government is the fifth largest industry of the big data analytics market, covering about 7% of the expenditure, and growing fast. A recent study by Deloitte (2016) identified 103 cases of big data analytics in government. In that regard, the Communication on "Data, Information and Knowledge management" calls for a more strategic use of data, information and knowledge. In this context, a data strategy (DataStrategy@EC) and a related Action Plan have been set-up in 2018, with the objective of transforming the EC in a data-driven organisation. The eight actions of the Action Plan are centred around 5 different dimensions: data, people, technology, organisation, policy. The data strategy highlights indeed that these dimensions need to mature and evolve harmonically to deliver a real transformation on how data is used in the decision-making processes. In 2019, an operational governance framework has been set up to closely follow-up the implementation and the evolution of the Action Plan. The 2016-2020 ISA² (Interoperability solutions for public administrations, citizens and businesses) programme funded with a budget of 131 million euro, aims to support the development of digital solutions that enable public administrations, businesses and citizens in Europe to benefit from interoperable cross-border and cross-sector public services.

All these initiatives foster data-centric public administration. But where do we stand? To understand that the European Commission has commissioned the study **Data Analytics for Member States and Citizens,** which provides policy Directorate Generals of the European Commission and Member States public administrations with a knowledge base and guidance on the adoption of public sector data strategies, policy modelling and simulation tools and methodologies, and data technologies fostering a data-centric public administration.

Specifically, the study covers three domains in relation to data analytics in government:

1. **Data strategies, policies and governance**: initiatives in the public sector both at the strategic level, such as data strategies, data strategies, data governances and data, management plans; and at organisational level, aimed to create units or departments, and to elaborate new processes and role.
2. **Policy modelling and simulation**: initiatives to improve policy analysis through new data sources, robust and reliable models to perform "what-if" scenarios, predictive analytics and hypothesis testing, and tools allowing policy makers to carry out scenario analysis through intuitive interfaces.
3. **Data technologies**: new architectures, frameworks, tools and technologies to be used by public administrations to gather, store, manage, process, get insights and share data. This domain includes the study of how data are governed as well as data collaboratives, and in particular stresses the joint analysis of governance and technologies.

1.2. Structure of the report

The report is structured as follows:

- About the methodology utilized, comprising the choice of cases, the methodology of analysis and the list of cases;

- A cross-analysis of cases, focussing in recommendations and identifying key success factors;
- The four case studies, with detailed information about the delivery model, implementation processes and main technology choices, but also their impact and lessons learned.

2. METHODOLOGY

2.1. Choice of the cases

The study team carried out an extensive desk research leveraging on events, previous knowledge and expertise, and a literature review. In this way, the study team collected a long list of 35 cases, which have been selected based on the information available and the evidence of real application. The study team has complemented the list of cases also through two scoping interviews: one with Petra Simperl, Director of the Southampton Data Science Academy, and another one with Ben Welby, Policy Analyst in Digital Government and Open Data at OECD. Out of the list of 35 cases, the study team and the European Commission agreed to choose five for in depth analysis based on domain balance, data availability, maturity, and other four criteria:

- Extent to which the project or service allow public administrations to process and analyse data
- Extent to which the project or service creates a reusable infrastructure to cater for several use cases and needs at a horizontal way (e.g. across departments, ministries, etc)
- Extent to which the project or service has national level scale or significant use across local administrations
- Extent to which the infrastructure is actively used (i.e. has impact, not academic or theoretical)
- Extent to which the project or service uses a modern approach to technology, therefore exemplifying best practice.

2.2. Methodology of analysis

The analysis of the cases has been carried out by means of the following methodologies:

- Desk research, which consists in the analysis of documents and reports, scientific articles, and websites;
- Interviews to informants, which are experts that directly worked at the development of the model and that make regular use of it. For the interviews, the study team uses a template mirroring the structure of the cases.

The general structure of the template for the case studies is the following:

- Introduction
- Development of the work
- Delivery model
- User needs
- Resource considerations
- Implementation process
- Technology choices
- Lessons learned

The desk research is complemented by questions on the following topics:

- Key outcomes achieved
- Contribution to a national digital plan
- Service funding

- Delivery model
- Design principles and reasons behind technology choices
- Success factors, bottlenecks and lessons learnt
- Advice to prospective adopters.

2.3. List of cases

For task 3 the team so far collected a long list of 100 cases, 17 of which are acceptable for an in-depth analysis. Out of this 17, five have been chosen for the in-depth analysis, consisting in the elaboration of a case study based on desk research and interaction with two informants.

The document features four case studies:

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%.

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.

Findata, a Finnish agency to enable the secondary use of social and healthcare data in the research, public, and private sectors. 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.

KOKE, an analytics solution for fraud detection in use by the Estonian Tax and Customs Board. 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.

3. CROSS ANALYSIS OF THE CASE STUDIES

3.1. Introduction

The 21st century has seen unparalleled changes in the way that organisations manage and use data. Citizens are accustomed to using services provided by technology companies that are able to gather, process, and derive insights from data with unprecedented speed and utility. Public sector organisations face a number of technological and cultural challenges in the way they deal with data if they wish to meet the expectations of increasingly data savvy populations.

This analysis considers the following four case studies:

- the New Zealand (NZ) government's Integrated Data Infrastructure (IDI) and associated tools;
- the Reproducible Analytical Pipelines (RAP) methodology used by administrations in the United Kingdom (UK);
- Findata, a Finnish agency to enable the secondary use of social and healthcare data in the research, public, and private sectors;
- and KOKE, an analytics solution for fraud detection in use by the Estonian Tax and Customs Board.

These four case studies provide a relatively wide ranging cross section of data analytics practices across the public sector.

The objective of this document is to carry out a cross-analysis of the case studies to identify these key factors in the delivery model, implementation processes and main technology choices. From them arise six overarching recommendations that can help the European Commission and the Member States improve their own data analytics strategy.

The overarching recommendations are:

1. **Put user needs before organisational needs**
2. **Work in the open and foster reusability**
3. **Adapt to data readiness**
4. **Use open source**
5. **Invest in data capability at all levels**
6. **Break down silos**

3.2. Recommendations

Put user needs before organisational needs

The European Commission should aim to meet the needs of both consumers of public sector data products, and the needs of the analyst users that produce it. Clearly the needs of consumers (be they individual citizens, businesses, public bodies, or decision makers) to have access to timely and accurate information is critical to any data infrastructure and analysis strategy. However, it is also important to recognise analysts as a user group with distinct and often varying needs, and often the capability to meet their own needs if given sufficient flexibility. The case studies examined in this analysis demonstrate the ability of analysts to build the tools they need to do their work better, and by working openly, to share those tools with the wider community and enable their reuse.

Work in the open and foster reusability

The European Commission should embrace open ways of working and embed the same approach to Member States. In two of the case studies that we examine in this analysis, working in the open has been a major contributor to success. The decision in NZ to work openly on the SIAL and SIDF has led to significant cost savings among other public sector bodies who do not, as a result, need to repeat the same work. Similarly, working openly in the production of RAPs has fostered the creation of a community that spans all the devolved administrations in the UK, and some regional public sector bodies: a grassroots movement for modernisation of tooling and practices that originates from the analysts themselves.

Adapt to data readiness

The Commission should recognise that different public sector bodies have different needs and capabilities and a 'one size fits all' approach to data analysis tools and infrastructure is unlikely to be appropriate. It is also important that tools and infrastructure are interoperable, support common standards (for example data formats), and should be able to scale to support future needs. The implementation of RAP, for instance, varies significantly between organisations depending on requirements and capability. NHS Scotland defines seven levels of maturity that an agency can adopt, all based on the principles of RAP, and all built using open source technologies that can be easily adapted and developed as required.

Use Open Source

The organization and the Member States should start prioritising the use of open source technologies in future developments.

Advances in statistical techniques, the availability of large amounts of data, and the availability of cheap computing power have led to rapid changes in the field of data analysis. Software companies and researchers routinely publish their research and tools freely under open source licenses. These tools are almost uniformly written in open source languages. Allowing analysts to use the same open source tools ensures that they can keep up to date with developments in the field. This is critically important as public sector bodies increasingly adopt machine learning and artificial intelligence: the field moves so quickly that what was once considered to be cutting edge can be obsolete in a matter of months.

Furthermore, open source languages act as a 'programmatic glue' that can combine disparate data sources, varied analysis, and multiple outputs with minimal effort. This is why the R language has been an indispensable part of the RAP project: it offers flexibility rarely seen in proprietary tools. Moreover, public sector bodies often differ in their choices of proprietary software for all manner of budgetary and political reasons. Adopting common open source tools like Python and R removes these barriers to sharing, enabling reuse.

Invest in data capability at all levels

The data landscape is changing rapidly, and the pace of that change is increasing. Member states should recognise the need to invest in the capabilities of their personnel in order to keep pace with these changes. The RAP project provides a good example of this. Because the project relied predominantly on open source software, it did not imply a big new capital investment, but did require capability building both among the analysts who would use and develop the tools, and among the managers responsible for them. As public sector organisations become increasingly sophisticated in their exploitation of data, these organisations must ensure that the whole organisation develops data literacy as a core

skill, and that the benefits that data can bring are not siloed among small groups of highly data literate specialists.

Break down silos

The commission should work to break down the siloing of data within public sector organisations, and encourage Member states to do the same, whilst prioritising proportionate measures for data security and protection that ensure that the public trust that their data are being well managed.

One of the biggest data problems that the public sector faces is that data are often siloed in different organisations, in different formats, and on different infrastructure. Both the IDI and Findata develop legislative and infrastructural solutions to these problems, whilst some of the issues that are solved by RAP exist only because of inconsistencies in the way data is stored and managed by UK Government departments.

However, member states should be aware that citizens may be concerned about the collation of data sets within government servers, and the release of this data to organisations outside of the public sector. Both the IDI and Findata have strong approval processes in place to ensure that this is done appropriately, and technical solutions in place to safeguard citizens' privacy.

3.3. Critical success factors

Meeting user needs

Consumers as users

In all of the case studies that are considered in this analysis, a common user need is that of consumers of public sector data to have access to timely and accurate information to inform decision making. These users may be individual citizens, businesses, researchers, public bodies, or decision makers. Clearly this is a key group of users, and many initiatives in the public sector data space will target the outcomes experienced by these users.

One element of this that the IDI and Findata both address is the provision to users of a single point of contact and process for requesting and accessing data. Findata aims to provide a 'one stop shop' where those who want access to Finnish social and healthcare data can go, whilst the IDI is wider ranging, and stores many datasets from across NZ Government departments. Both projects simplify the situation for would-be users by reducing duplication in the application process for data access, ensuring consistent standards, and levels of data protection and security.

The analyst as user

Analysts should be recognised as a user group in their own right. In the RAP and IDI case studies we have also identified the needs of the individual public sector analysts or researchers who need to interact with the data on a daily basis (hereafter 'analyst users'). Often, legacy processes for working with public sector data can be repetitive, time consuming, and may not best utilise the skills of the analyst. Part of the success of these two case studies is that they both addressed this user need: the IDI with the creation of the SIAL and SIDF, and RAP with its aim of automating repetitive and labour intensive tasks. Meeting this analyst user need is consistent with the primary need of meeting consumer's expectations - if repetitive tasks are automated, there may be more room to conduct more valuable analysis, and the resulting data products may be more timely and of better quality.

Analyst users should be able to exercise sufficient autonomy over the tools that they use. Not all analyst users are alike, and whilst some will be comfortable using

modern analytical tools like R and Python, many (probably most) analyst users will be more comfortable working with spreadsheets like Microsoft Excel or Google sheets. Best practice accommodates all types of analytical users and allows them to access data in the way they find most comfortable. The precursor data lakes which form the basis of Findata's data storage were designed to cater for the needs of in-house business intelligence (BI) staff, doctors and medical thesis workers, and computational researchers[1] - use cases that span from the ubiquitous spreadsheet, to artificial intelligence research using cutting edge open source tools.

Failing to provide analysts with sufficient autonomy can be costly. Research from the UK Government Digital Service (GDS) suggests that spreadsheets are so prevalent that it would be fair to say they are the default model for government data[2]. Whilst it is recognised that spreadsheets lead to many errors when relied on for business processes[3][4][5], attempts to replace them frequently fail when they are supplanted by tools that the analysts cannot adapt so easily[6]. Indeed, while successful, the future of the current implementation of Estonia's KOKE system is under review for this very reason: lack of autonomy, and the need to outsource work to further adapt the system. By contrast the success of RAP and the SIAL and SIDF tools is that given enough autonomy, skilled analyst users in NZ and the UK were able to develop their own tools internally to solve problems they encounter, obviating the need to outsource. When this autonomy is coupled with open source software allowing analysts to share their work with other teams, departments, or even governments, the benefits are multiplied enormously.

Reusability and Open Source

Using and writing open source software fosters reusability

There are two ways in which open source software helps with reusability. Firstly, if analysts use open source tools for their analysis, or the technical infrastructure on which analytical environments are built is based on open source tools, it allows analysts and data engineers to make use of innumerable onlines resources. Github, for instance, the platform where many RAPs, and the SIAL and SIDF are published openly, is used by more than 40 million users, from around 2.9 million organisations worldwide[7]. Whilst many of Github's users are software developers, an increasing proportion is made up of data analysts, data scientists, and researchers, many of whom share their code freely under permissible licenses. Another platform Stack Overflow allows users of open source software to ask questions that can be answered by other users. In 2018 the platform had over 100 million users, with 2 million out of 2.5 million questions answered successfully. Again, whilst mainly used by software developers, Stack Overflow and it's sister site Cross Validated (a home for questions related to statistics and machine learning) have a thriving community of data analysts and scientists. When faced with a new problem for which a solution does not exist, a public sector data analyst working with an open source language (for instance Python or R) can look on Github or Stack Overflow (or elsewhere online) to reuse or adapt a solution that others have developed for the same or a similar problem. Given the amount of material that now exists on these sites and others, it is a challenging problem indeed that cannot be at least partly solved within ten minutes and access to a search engine.

The second way in which open source software can assist with reuse, is if analysts across the public sector are able to publish their work openly for others to reuse. This is precisely the situation with RAP and the SIAL and SIDF layers for the IDI. One reason why RAP has

1 Evaluation of the Isaacus project's data lake solutions in research use
2 https://gds.blog.gov.uk/2017/01/31/what-you-can-learn-from-making-data-user-centred/
3 Errors in Operational spreadsheets
4 What We Don't Know About Spreadsheet Errors Today: The Facts, Why We Don't Believe Them, and What We Need to Do
5 spreadsheet risk management and solutions conference
6 Improving how we manage spreadsheet data - Data in government
7 https://octoverse.github.com/

been so successful is that the prototype was published openly on Github under a permissible license that allowed anyone with an internet connection to scrutinise, adapt, and reuse the tool for their own use case. Clearly not all public sector code can be shared openly, but often it is not the logic enshrined in the code that is sensitive, it is the data on which the logic operates, and these two can easily be decoupled.

Open source tools help prevent vendor lock-in

Another way in which using open source tools can aid with reuse is by preventing vendor lock-in. Findata provides a good example of this. The data lake infrastructure uses an open source technology called Apache Hadoop. This technology is supported by all of the major cloud computing suppliers (Amazon Web Services, Google Cloud Platform, Microsoft Azure), and can also be deployed on physical hardware within a national or public sector run data centre. If the decision is made to change the hosting option, it would be a relatively straightforward undertaking to reuse everything that has been built by deploying it to a new host. Not only does this give public sector organisations great flexibility in where their data is stored and processed, but it can help to keep the cost of the infrastructure competitive by ensuring that it is possible to switch suppliers.

Build and iterate

We have noted that working with open source software facilitates the reuse of code to solve analytical problems, but there are other ways in which the examples in the case studies have built on prior work. Both the IDI and Findata were built on a number of projects that had been completed over the preceding years. The IDI prototype, for instance, was created from data integration efforts completed for various projects prior to Cabinet approval for a cross-government data integration service in 2013. The infrastructure underlying Findata was trialled in precursor projects orchestrated by health administrations across Finland, and evaluated openly by a third party. These were valuable projects in their own right, and the lessons learnt were able to inform the implementation of Findata.

Architecture and Hosting

Choose the right data storage option

Since the inception of the internet, and the general availability of larger quantities of data than ever before, there has been somewhat of an explosion in the types of data storage solutions and infrastructure available to public sector organisations. There is however no 'one size fits all' solution for data infrastructure, and organisations need to make well informed choices about which infrastructure to use and where to deploy it. Poorly informed decisions can be costly.

While 'big data' solutions can seem appealing, many public sector organisations do not have big data, and will likely never have big data by today's standards. This is because administrative data often conforms to a fairly homogenous format that can be stored easily and managed using tried and tested technologies. Furthermore, cloud suppliers are able to scale these traditional technologies in ways that were not previously possible, making it even easier for organisations to store ever larger quantities of data, with ever decreasing effort.

Of the three case studies which involve a data storage solution, the IDI and KOKE projects use traditional proprietary database solutions, whilst Findata is built upon an open source 'big data' solution. Health data stored by Findata in particular can fall into the realm of big data because it can include images, and video from medical imaging devices. Such data are difficult to store and analyse with traditional solutions. Furthermore, Findata followed three precursor projects which tested the technology, and was subject to independent and

public scrutiny. Such systems are however significantly more complex than simpler more traditional technologies; recognition of this complexity and the related skills gap was an outcome of the precursor projects.

Interoperability is key to breaking down silos

The IDI is a good example of a concerted effort to bring together datasets from various Government departments, and to store them on one common Integrated Data Infrastructure. Despite the IDI gathering around 550 public sector datasets together in one place, it does not automatically solve the issue of interoperability. This problem arises because organisations tend to have different processes for managing, collecting, and using data. The SIAL was built to address this problem: ironing out the idiosyncrasies of data from 14 different agencies, all of which likely have subtly different ways of representing reality in their data. This is in part why the SIAL is successful: anyone who uses the IDI immediately faces this interoperability problem, and it usually only needs to be solved once.

One way to help solve these issues is to encourage organisations to conform to the same standards in their own business processes, so that when data from two organisations are brought together, they already have similar characteristics. The UK Government registers initiative[8] is a good example of this. Key pieces of data infrastructure from lists of countries to lists of Government organisations are curated by a custodian and made publicly available via an easy to consume service with an API.

RAP also deals with the interoperability problem. The UK government does not yet have an integrated data infrastructure like the IDI, but agencies do share data between each other. The prototype RAP for instance was built on data collected by the Department for Digital, Culture, Media, and Sport (DCMS) and the Office for National Statistics (ONS). These data arrive to the analytical team replete with the idiosyncrasies of each agency, and in multiple formats. RAP deals with the interoperability problem by developing a software layer - like the SIAL - in which various data sources are manipulated into a common format before they are used in analysis.

Public trust is paramount

Breaking down silos in public sector data storage and use starts with legislation. This can be seen in New Zealand's IDI and Finland's Findata. Both required an act to be brought into law to provide a specific legal basis for the activities undertaken by the services. In each of these examples, suitable weight was given to the issues of privacy and data security to ensure that the services were fit for purpose, and importantly have the public's trust. Examples of poor practice abound, with some of the most concerning breaches of security and privacy happening in the health sector[9]. As the value of large quantities of public data increases with the sophistication of the tools and techniques that can be applied to it, it is easier than ever for organisations to put public data at risk. This must be met with a proportionate response that does not unduly restrict the potential public benefit that can be derived from these data and techniques.

Building capability

As organisations become more dependent on streams of data to understand the world and make decisions, it is critical that the public sector keep pace with developments by building capability, upskilling the workforce where possible, and bringing in new talent where it is not. Capability is an underlying theme in all the case studies we examine.

8 https://www.gov.uk/government/publications/registers/registers
9 https://medconfidential.org/for-patients/major-health-data-breaches-and-scandals/

Developing data capability can reduce the need to outsource technical work

The SIAL and SIDF tools were developed by highly skilled data scientists who were able to build the tools to meet their own, and others' needs internally. These resources were then shared openly allowing others to benefit from the work. Building this kind of capability can allow organisations to solve more of their analytical and infrastructural problems internally without the need to outsource. Conversely the future implementation of the Estonian KOKE system is being reviewed due to the expense and time taken to make changes to the system (which must be outsourced), although this may have more to do with the system being based on proprietary tools rather than a lack of in house capability.

Analysis of the precursor projects to Findata noted that the capability to deal with the highly technical data infrastructure was an early constraint. In the event, the management of at least part of this infrastructure was outsourced to a private sector consultancy, but for this highly complex system to be utilised to the fullest extent, it will likely require upskilling of operators in the day to day use of the technology.

Successful data projects rely on a mix of subject matter knowledge and data expertise

Often, public sector problems are highly complex and require a significant amount of experience to understand these complexities. In the case study of the Estonian KOKE project, the interviewees noted that finding staff who could combine data expertise with the requisite understanding of tax affairs had been challenging. Rather than trying to teach this critical business knowledge to new data analysts, they preferred to upskill existing subject matter experts with the skills required to analyse the data themselves. The RAP project is similar, analysts within existing Government departments are generally upskilled in place. Indeed, since it relies on open source technology, the greatest restrictions to wider deployment are usually limitations on the use of open source tools, and the necessary skills and capabilities among the analysts in those organisations. RAP's success owes a great deal to the efforts of its proponents to make learning resources easily available to other analysts. Few such public sector initiatives can boast an ebook[10] and a massively open online course[11], but just as important is the cross-Government community of analysts which support its adoption.

Recruitment and retention of highly skilled analysts can be hard

Highly skilled data analysts, scientists, and engineers are in demand across all sectors, and the public sector may find it difficult to compete with the salaries and benefits that are available to the most skilled. Providing good opportunities for development can help fill these skill gaps by upskilling existing public servants, and by attracting more junior data professionals who aspire to develop these skills. One reason for the popularity of RAP is that it has allowed analysts to develop skills that are highly sought after, and use tools that are in demand across all sectors.

10 https://ukgovdatascience.github.io/rap_companion/
11 https://www.udemy.com/course/reproducible-analytical-pipelines/