THE ROYAL SOCIETY

December 2020

Royal Society Submission to the National Data Strategy Consultation

Lessons from COVID-19: Data readiness, pathfinders, and lighthouse projects

Addendum to joint National Academies submission

Key Points:

- Timely access to good data is essential in crisis situations, and the COVID-19 pandemic has highlighted issues relating to timeliness, cost and quality. The implementation of the National Data Strategy should consider how data, and in particular private sector data, could be more effectively shared safely with government and researchers in crisis situations.
- The National Data Strategy should seek to implement lessons from the pandemic responses, carrying out projects that will lay the foundations for future data readiness, by establishing the governance mechanisms, infrastructure and appropriate take up of technologies to support data-enabled response to future emergencies.

1. Introduction

1.1. The Royal Society is the National Academy of science for the UK. Its Fellows include many of the world's most distinguished scientists working across a broad range of disciplines in academia, industry, charities and the public sector. The Society draws on the expertise of the Fellowship to provide independent and authoritative advice to UK, European and international decision makers.

The Society's fundamental purpose, reflected in its founding Charters of the 1660s, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity. Our strategic priorities therefore are to promote excellence in science; to support international collaboration; and to demonstrate the importance of science to everyone.

1.2. This document is an addendum to the joint National Academies' submission made to the government's National Data Strategy consultation, giving more detail on data for COVID response. It summarises a recent report by the Data Evaluation and Learning for Viral Epidemics (DELVE) group. DELVE was formed in response to the pandemic with a remit to explore specific policy questions in a data-driven manner, and was convened by the Royal Society. The DELVE Initiative was established with the ambition that data science could play a role in helping develop policy responses to the COVID-19 pandemic, by identifying lessons from the responses of other countries or by combining datasets to generate novel insights.

1.3. The Royal Society also convened the Rapid Assistance in Modelling the Pandemic group (RAMP), and the experience of that group in securing access to data to support their work mirrors some of the challenges to accessing data, set out by DELVE.

2. Data readiness and COVID-19

2.1. In its November 2020 report *Data Readiness: Lessons from an Emergency* DELVE found that across the COVID-19 pandemic, data availability and quality have proven a bottleneck to understanding the current state of the UK's health and economy and to providing policy advice for rapid decision making.

2.2. This bottleneck has affected a wide range of data types, from early challenges with access to testing data that would enable estimates of COVID-19 incidence, Rt and mortality rates across the country, to the difficulty of obtaining data about monetary transactions that would help estimate the effects of the pandemic on GDP, to issues with the use of mobility data that would enable assessment of how movement of people across the UK population contributes to disease spread.

2.3. However, today technologies that can acquire data are pervasive. Data is continually produced by devices like mobile phones, payment points and road traffic sensors. This creates opportunities for nowcasting of important metrics such as GDP, population movements and disease prevalence, which can be used to address the challenges outlined above and design policy interventions that are targeted to the needs of specific sectors or localities.

2.4. The data collected as a by-product of daily activities is different to epidemiological or other population research data that might be used to drive the decisions of state. These new forms of data are happenstance, in that they are not originally collected with a particular research or policy question in mind but are created through the normal course of events in our digital lives, and our interactions with digital systems and services.

2.5. This happenstance data pertains to individual citizens and their daily activities. To be useful it needs to be anonymized, aggregated and statistically calibrated to provide meaningful metrics for robust decision making while managing concerns about individual privacy or business value (see section 3 in the full report). This process necessitates particular technical and domain expertise that is often found in academia, but it must be conducted in partnership with the industries, and public sector organisations, that collect or generate the data and government authorities that take action based on those insights. Such collaborations require governance mechanisms that can respond rapidly to emerging areas of need, a common language between partners about how data is used and how it is being protected, and careful stewardship to ensure appropriate balancing of data subjects' rights and the benefit of using this data. This is the landscape of data readiness; the availability and quality of the UK nation's data dictates our ability to respond in an agile manner to evolving events.

2.6. There are a range of ways in which policymakers could seek to increase access to data. For example, government could encourage the adoption of a duty to share certain types of data at times of national crisis. (This point is set out in more detail in the National Academies joint response.)

2.7. Considering these points, DELVE made three recommendations:

- Government should update the statutory objective of the Office for National Statistics (ONS) to accommodate trustworthy access to happenstance data to generate national and local statistics. Such statistics are required on very short time frames to facilitate fast decision-making for the nation in the rapidly evolving circumstances of a national emergency.

- The ONS should collaborate closely with the Information Commissioner's Office (ICO) to formulate a standardized qualification for data access, equivalent to a 'data driving license' that would demonstrate trustworthiness and ensure that qualified experts can get rapid access to different data types with the appropriate standardized ethical and legal training in place.

- Government should fund interdisciplinary pathfinder data projects. These projects should require collaborations between industries, run across government departments and integrate different academic expertise. Each project should target a specific policy question. Beyond the pathfinder role, the projects will leave a legacy in the form of expertise and guidance in understanding the stages of the data-sharing pipeline.

3. Pathfinders and lighthouses

3.1. The National Data Strategy lists 'lighthouse projects' between the Cabinet Office and the ONS as one of the actions that will drive implementation of the strategy. These lighthouse projects should be designed to achieve the outcomes set out by DELVE, to put in place the mechanisms, guidance and expertise that are needed to respond in the future.

3.2. This action also echoes calls in the Royal Society's 2019 *Protecting privacy in practice* report which is cited in the strategy. That report argues there is an opportunity for government itself to lead by example in demonstrating the utility of these approaches, and that government should consider how the use of PETs could unlock new opportunities for data analysis, including opening up the analysis of sensitive datasets to a wider pool of experts whilst fully addressing privacy and confidentiality concerns. These lighthouse projects can also be a means to establish appropriate and trustworthy use of these technologies, to improve privacy-preserving use of data in the public interest.

3.3. The response to COVID-19 underlines the importance of establishing the skills, guidance and technologies in government ahead of any future emergency. The DELVE report on *Data Readiness* shows that data sharing has worked best for organisations that have invested in the infrastructure and governance mechanisms to enable data sharing between multiple parties in a secure and privacy-aware way. It is important that the implementation of the National Data Strategy enables this within the UK.

For further information please contact public.affairs@royalsociety.org

ANNEX: DELVE report. Please note that the version on the DELVE website should be referred to as the most up to date version: *Data Readiness: Lessons from an Emergency*

Data Readiness: Lessons from an Emergency

Summary

Responding to the COVID-19 pandemic has required rapid decision-making in changing circumstances. Those decisions and their effects on the health and wealth of the nation can be

better informed with data. Today, technologies that can acquire data are pervasive. Data is continually produced by devices like mobile phones, payment points and road traffic sensors. This creates opportunities for nowcasting of important metrics such as GDP, population movements and disease prevalence, which can be used to design policy interventions that are targeted to the needs of specific sectors or localities. The data collected as a by-product of daily activities is different to epidemiological or other population research data that might be used to drive the decisions of state. These new forms of data are happenstance, in that they are not originally collected with a particular research or policy question in mind but are created through the normal course of events in our digital lives, and our interactions with digital systems and services. This happenstance data pertains to individual citizens and their daily activities. To be useful it needs to be anonymized, aggregated and statistically calibrated to provide meaningful metrics for robust decision making while managing concerns about individual privacy or business value. This process necessitates particular technical and domain expertise that is often found in *academia*, but it must be conducted in partnership with the *industries*, and *public sector organisations*, that collect or generate the data and *government* authorities that take action based on those insights. Such collaborations require governance mechanisms that can respond rapidly to emerging areas of need, a common language between partners about how data is used and how it is being protected, and careful stewardship to ensure appropriate balancing of data subjects' rights and the benefit of using this data. This is the landscape of data readiness; the availability and quality of the UK nation's data dictates our ability to respond in an agile manner to evolving events.

Key points

Across the COVID-19 pandemic, data availability and quality have proven a bottleneck to understanding the current state of the UK's health and economy and to providing policy advice for rapid decision making. This bottleneck has affected a wide range of data types, from early challenges with access to testing data that would enable estimates of COVID-19 incidence, Rt and mortality rates across the country, to the difficulty of obtaining data about monetary transactions that would help estimate the effects of the pandemic on GDP, to issues with the use of mobility data that would enable assessment of how movement of people across the UK population contributes to disease spread.[1]

This report builds on the experience of the Royal Society's DELVE group in providing rapid turnaround, data-driven answers to policy questions, and the challenges in access to data that arose during this work. It sets out the challenges we found that were common to these data sharing activities, focussing on the use of *new data modalities*, arising from, for example mobile phone usage or electronic payments data, rather than the classical challenges of collecting surveillance data. It then considers recommendations for addressing these challenges and creating an environment that supports the safe and rapid use of a range of different types of data in policymaking.

[1] Health data is another important data modality that we do not discuss in detail in this report. There are particular issues with health data, and we are aware of ongoing work from HDR UK to highlight these challenges.

Challenges in accessing and analysing non-traditional data sources to inform COVID-19 policy

- The pipeline for data from discovery and acquisition to decision is fundamentally different for *happenstance* data created by everyday interactions, than for data collected specifically for research, but the two types of data are rarely differentiated in data sharing discussions. Access to happenstance data relies on successful interactions across the public sector, private sector and academia to establish trustworthy frameworks for data sharing that enable data access while managing concerns around security and privacy.
- The lack of a common language for understanding data quality or the action needed before a dataset can be used in analysis holds back research collaborations. Guidelines for data quality that do exist focus on intentions - or set out the desired end-point - rather than providing mechanisms for improving data accessibility. This contributes to the prevalence of perceived barriers to data sharing.
- The pipeline for data from discovery and acquisition to sharing includes ethical, legal, technical, commercial and quality issues. But a lack of a coherent vocabulary for identifying and resourcing issues means projects are often inappropriately resourced and staffed.
- Where successful multiparty data sharing collaborations between academia, the private sector and government have been established at pace, they have often relied on relationships that existed before the pandemic, which could be rapidly adapted to pandemic needs. In the long-term, new initiatives are needed to cultivate such partnerships, so that data sharing efforts can be repurposed at times of crisis, in ways that encourage public dialogue about data use. Creating such partnerships will require further effort to build capability at all levels of decision-making within government, if it is to promote the trustworthy use of data in policymaking.
- Many of the barriers to data sharing that have been experienced during the pandemic are long-standing issues in data policy, including the lack of common technical standards, concerns about privacy or security breaches, and the lack of organisational capabilities. Together, these contribute to a cluster of coordination problems between organisations that are felt acutely in efforts to share happenstance data, due to the need for multiparty coordination.

Recommendations

- Government should update the statutory objective of the Office for National Statistics (ONS) to accommodate trustworthy access to happenstance data to generate national and local statistics. Such statistics are required on very short time frames to facilitate fast decision-making for the nation in the rapidly evolving circumstances of a national emergency.
- The ONS should collaborate closely with the Information Commissioner's Office (ICO) to formulate a standardized qualification for data access, equivalent to a 'data driving license' that would demonstrate trustworthiness and ensure that qualified experts can get rapid access to different data types with the appropriate standardized ethical and legal training in place.
- Government should fund interdisciplinary pathfinder data projects. These projects should require collaborations between industries, run across government departments

and integrate different academic expertise. Each project should target a specific policy question. Beyond the pathfinder role, the projects will leave a legacy in the form of expertise and guidance in understanding the stages of the data-sharing pipeline. Priority areas for pathfinder projects include:

- Nowcasting of economic metrics: At least one of these pathfinder projects should create a close collaboration between Cabinet Office and Treasury around nowcasting of classical economic metrics (such as GDP) from happenstance data (e.g. payments data). Efficient resourcing and strategic implementation of data sharing projects will only be possible if Treasury and Cabinet Office are aligned on plausible benefits and costs of data sharing pro jects.
- Mobility data: Another project should drive a step-change in the use of mobility data for public policy. To achieve this, the ONS should act as the trusted body to convert happenstance data into high-frequency population mobility statistics. One pathfinder project should produce daily views of population mobility between geographic regions, aggregated from origin to destination counts from mobile phone operators.

This paper has drawn on evidence available up to 12 November 2020. Further evidence on this topic is constantly published and DELVE will continue to develop this report as it becomes available. This independent overview of the science has been provided in good faith by subject experts. DELVE and the Royal Society accept no legal liability for decisions made based on this evidence.

1. Background

DELVE's remit and data science for COVID-19 policy

The Royal Society's DELVE group was formed in response to the COVID-19 crisis with a remit to explore specific policy questions in a data-driven manner.[2] The DELVE Initiative was established with the ambition that data science could play a role in helping develop policy responses to the COVID-19 pandemic, by identifying lessons from the responses of other countries or by combining datasets to generate novel insights. Such analysis requires access to data, which could come from both official statistics, or from so-called happenstance data, generated as a by-product of daily activities. Drawing from a multidisciplinary team of domain experts in policy, public health, economics, education, immunology, epidemiology, and social science, alongside statisticians, mathematicians, computer scientists and machine learning scientists, DELVE set out to provide advice and analysis that could feed into live policy decisions.

By the middle of August 2020 DELVE had completed five reports: on the use of face masks by the general public; the test and trace system; the control of hospital and health-care acquisition of the disease; the risk associated with a return to schools; and the economic aspects of the COVID-19 crisis. These reports have grappled with a wide range of questions, including:

- Should face masks be worn to help contain COVID-19, and who should wear them where and when?

[2] Royal Society convenes data analytics group to tackle COVID-19 https://royalsociety.org/news/2020/04/royal-society-convenes-data-analytics-group-to-tackle-covid-19/

- Which kinds of households and firms would be most impacted by non-pharmaceutical interventions, and by how much?
- Are there gaps in emergency fiscal measures introduced to cushion the impact of lockdown?
- What is the impact of future case burdens on ICU and ventilator capacity?
- Can early signals of the effect of policy interventions such as lockdown measures and the easing of such measures be obtained?

The breadth of these questions highlights the intersections of expertise that were required as well as the diversity of data sources that were potentially in scope of these efforts.

In March 2020, DELVE formed an initial list of datasets we felt would be informative in addressing policy questions. This data included mobility data (e.g. the aggregated movements of mobile telephones between different cell towers) and payments transaction data (e.g. aggregated credit card payments being made in different geographic regions). Such data originates from individual citizens but is collated by commercial companies in the course of their normal business. The data is invaluable in determining, for example, an estimate of the effect of lock down measures on GDP in real time[3] or for aiding localized estimation of the disease reproduction number (R) through understanding the movement of people between localities.

Despite the potential benefits of using such data in the COVID-19 response, access to many forms of data has been problematic across the pandemic. Recognising the shortcomings of the UK's COVID-19 data readiness, in an evidence session with the Commons Science and Technology Committee in July, Government Chief Scientist Patrick Vallance explained that "an important lesson across not just pandemics but every emergency, is data; you must have the data flows and you must understand data ownership and how data is going to get to people for the information you require".[4]

DELVE's work across the pandemic illustrates this lesson, and the action that is needed to ensure the UK is better able to use its data resources in response to emergencies in future. It also illustrates the scale of the successes that have been achieved elsewhere, which may offer lessons for the UK's data strategy. For example, the French interbank network, Groupement des Cartes Bancaires (CB), made nearly five billion payment card transactions from approximately 70 million cards issued by all banks in France available for analysis[5]. Their partnership with analysts complied with the EU GDPR (Article 89), and produced an intricately detailed view of the impact of COVID-19 on consumer spending and business sales on a daily and weekly scale. We examine such examples of coordinated collaboration to enable data use in Boxes A and B in this report. While there are examples of the successful deployment of similar data types to support COVID-19 policy development in the UK, the scale of

[3] Economic Aspects of the COVID-19 Crisis in the UK
https://rs-delve.github.io/reports/2020/08/14/economic-aspects-of-the-covid19-crisis-in-the-uk.html

[4] UK Science, Research and Technology Capability and Influence in Global Disease Outbreaks
https://committees.parliament.uk/work/91/uk-science-research-and-technology-capability-and-influence-in-global-disease-outbreaks/publications/, accessed 23 October 2020

[5] Consumers' Mobility, Expenditure and Online-Offline Substitution Response to Covid19: Evidence from French Transaction Data. By David Bounie et al, 2020
https://www.telecom-paris.fr/consumers-mobility-online-substitution-covid-19
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588373

collaborations established elsewhere demonstrates the wider opportunities at stake.[6]

We revisit three representative questions the DELVE group considered below. Each illustrates different types of challenges that data scientists and policymakers might encounter when using data in policymaking. These scenarios cover:

- **COVID-19 prevalence in different localities across the UK**: Important statistics, such as local COVID-19 reproduction rate, are currently derived from case numbers collected through official national studies. Happenstance data, such as high-quality origin to destination mobility data, can be used to refine these estimates, enabling more targeted policy responses at a local level. Designing such tailored responses requires ready access to multiple data sources, which in turn relies on strong pre-existing relationships and connections across the public sector, industry and academia.
- **International comparisons on the impact of COVID-19 on different communities**: International comparisons can help policymakers understand what types of policy interventions might be helpful in managing the COVID-19 pandemic, or benchmark performance of approaches taken in the UK. Performing such analyses requires access to data of sufficient quality to enable cross-national comparisons. Data published by governments across the world may have different levels of accessibility, different formats, or cover different population samples. Before analysis, it must first be wrangled into comparable forms. A common understanding of data readiness could help speed up this process.
- **The pandemic's economic impact**: Official statistics play an important role in assessing the health, wealth and wellbeing of the nation. At times of crisis, these can change at pace. While vital, these official measures may not be responsive enough to inform rapid policy development. Alternative ways of measuring economic activity exist - for example, consumer behaviour data can indicate patterns of economic growth or constriction - but they rely on access to data held by private companies.

Case Study 1: What is the COVID-19 reproduction rate in a local area?

Important statistics, such as local COVID-19 reproduction rate, are currently derived from case numbers collected through official national studies. Happenstance data, such as high-quality origin to destination mobility data, can be used to refine these estimates, enabling more targeted policy responses at a local level. Designing such tailored responses requires ready access to multiple data sources, which in turn relies on strong pre-existing relationships and connections across the public sector, industry and academia.

DELVE's report on the economics of lockdown demonstrates the importance of deploying localised responses to the pandemic. The use of such measures is now a central component of the government's approach to managing the spread of COVID-19. However, it remains challenging to accurately measure the disease's reproduction rate - the R number - at a local level: the ability to estimate R diminishes with reduced transmission, and transmission reduces as disease numbers drop as they tend to when we focus on smaller local regions.

[6] Examples of these successes include: the use of Mastercard Spending Pulse data (further information at: https://www.mastercardservices.com/en/solutions/spendingpulse) and Retail Location Information data (further information at: https://cityinnovatorsforum.com/assessing-and-forecasting-the-economic-impact-of-covid-19/)

The primary information source for estimating reproduction number is the number of COVID-19 cases in a local area. Through these numbers it is possible to estimate daily changes in the number of cases, and hence the rate of reproduction. This data is provided in a clean and curated form by Public Health England and NHSX on the UK Government's COVID-19 dashboard[7].

When disease prevalence is low in a particular region, it becomes more difficult to estimate R for that location. One way of improving the accuracy of these estimates is to consider neighbouring regions: by understanding the aggregate movement of people between regions, we can hope to understand cross infection between regions and use our understanding of cross infection to improve our estimate of R. One way of tracking this movement of people is through origin to destination mobility data.

Origin to destination mobility data is available within any mobile phone company, and corresponds to the movements of their customers between their cell phone towers. In the UK, this data is not publicly available for analysis, and has been difficult to access for research purposes.

In contrast, the Spanish Government has supported a collaboration between Spain's three main mobile phone operators (Orange, Telefónica, Vodafone) and the Instituto Nacional de Estadística (INE, the Spanish Office for National Statistics) that makes available data describing the daily flow of people from origin to destination between more than 3,000 local districts. This data is derived from location information from more than 80% of the mobile phones in Spain and can be accessed without prior registration.[8] Access to this data has allowed Spanish authorities to include origin to destination mobility in their estimates of local R through a metapopulation model[9], from which correlations between regional mobility and COVID-19 reproduction also become clear. Furthermore, interconnected communities could be detected from granular mobility flows; these then supported decisions regarding selective confinement of certain geographical areas, depending on their epidemiological situation. The more self-contained the mobility of a geographic area is, the smaller the epidemiological impact of confining such a region would be, as most mobility is internal and not to other geographical areas. Lastly, the data is informative in simulating the effect that decisions on population mobility have in the future evolution of the pandemic. We consider factors that contributed to the Spanish mobility data's existence in Box A (Coordinated collaboration to enable data use).

Case Study 2: How much more are the elderly at risk in different countries?

International comparisons can help policymakers understand what types of policy

[7] GOV.UK: Coronavirus (COVID-19) in the UK
https://coronavirus.data.gov.uk/

[8] Información estadística para el análisis del impacto de la crisis COVID-19 / Datos de movilidad (Statistical information for the analysis of the impact of the COVID-19 crisis / Mobility data)
https://www.ine.es/covid/covid_movilidad.htm

[9] Análisis del grupo de trabajo de ciencias de datos para el COVID-19 Comunitat Valenciana (Analysis of the data sciences working group for COVID-19 Comunitat Valenciana)
http://infocoronavirus.gva.es/documents/170024890/170025022/Informe+Movilidad+gva+Mayo+2020.pdf/5b043319-eed9-4a66-8477-d214ffe11c39, accessed 23 October 2020

interventions might be helpful in managing the COVID-19 pandemic, or benchmark performance of approaches taken in the UK. Performing such analyses requires access to data of sufficient quality to enable cross-national comparisons. Data published by governments across the world may have different levels of accessibility, different formats, or cover different population samples. Before analysis, it must first be wrangled into comparable forms. A common understanding of data readiness could help speed up this process.

One aim of the DELVE group was to make international comparisons to inform the UK response to the disease. Detailed daily COVID-19 case and mortality data are provided for different nations by Johns Hopkins University[10], the European Centre for Disease Prevention and Control (ECDC)[11] and the World Health Organisation (WHO)[12]. These resources are useful in many analyses seeking to track the spread of the virus. However, these data do not provide the breakdown of such cases by age and sex, so they are not useful for understanding the variation in fatality rates across these different strata. COVID-19 affects different age groups in different ways, and understanding this differential impact is important in developing policy responses. DELVE has created data resources that can be deployed for this type of analysis, but in doing so has encountered various challenges in bringing such data together.

This case study does not draw directly on happenstance data, but is included here to highlight the breadth of practical issues around data readiness.

Discoverability

For some countries (e.g. Belgium and Germany), COVID-19 data by age and sex is available on their national statistics websites. Where data exists for many other countries, it is not easily discoverable, for example being published on websites that are not linked to from any official sources. An example is Austria: at the time of writing we were unable to locate their COVID-19 by age dashboard by following links from their official government statistics websites. India doesn't seem to provide any official source for age-disaggregated data, but there is an unofficial crowd-sourced project called COVID-19 India that does so[13].

Accessibility

Even when the data can be found, it does not mean it can be accessed. Roughly speaking, there are three levels of accessibility that we encountered:

1. **Machine-accessible data.** This data is the easiest for scientists to access. It can be readily loaded into analysis packages or spreadsheets such as Excel. For example, in Germany the Robert Koch Institute[14] and in Belgian Epistat[15] provide clearly

[10] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
https://github.com/CSSEGISandData/COVID-19

[11] European Centre for Disease Prevention and Control: COVID-19 data
https://www.ecdc.europa.eu/en/covid-19/data

[12] WHO Coronavirus Disease (COVID-19) Dashboard: Situation by Country, Territory & Area
https://covid19.who.int/table

[13] COVID-19 India: A crowdsourced initiative
https://www.covid19india.org/

[14] Robert Koch Institute: Infectious Disease Epidemiology

annotated “comma separated values” files. These contain time series of COVID-19 cases and deaths by age and sex. Even within this type of high-quality data, there may be work required to wrangle data into forms suitable for analysis. For example, the French Institute for Demographic Studies (INED) have collected COVID-19 death data from sixteen different countries[16]. These are provided as Microsoft Excel spreadsheets. However, inconsistent formatting and layouts across different releases of these spreadsheets (for example, cells which were previously merged becoming unmerged) break the processing pipeline for other aggregating sites or data scientists analyzing the data.

2. **Machine readable, but not machine accessible, data.** This is data that exists in electronic form, but in formats that are designed for human eyes. One common example of this data is tables in portable document format (PDF) - an open standard for documents that is designed for ease of human reading. When data is presented in these formats it is reformatted for the convenience of our eyes, but these changes bring challenges to the computer programs used for data analysis. As a result, information can be lost and the data is challenging to access electronically.[17] One way of managing these difficulties is to write a secondary program - called a ‘parser’ - to extract the data from the PDF file. Depending on the layout of the tables in the document, different parsers must be written. If data is interrupted by column or page breaks this can also lead to problems for the parsers. This means that different PDF reports present different challenges. For example, the Italian EpiCentro[18] formats their tables in a consistent way, so a single parser can be written to collate age and sex data from them. In contrast, the Spanish Ministry of Health, Consumption and Social Welfare[19] have frequently changing layouts and formats. This data requires considerable software engineering experience alongside manual effort to extract. This type of expertise is often not available to scientists or statisticians. To illustrate the difference in effort required for loading this data we include in Appendix A examples of python code that show the difference between loading data from a CSV and data from a PDF. The process of writing these additional parsers adds to the time taken for analysis, and the complexity of the resulting code.
3. **Machine inaccessible.** The final format is one where direct human intervention is required, as it is not possible to feed the data directly into any form of computer program. For example, in Portugal’s reports[20], data is presented only in the form of charts, meaning that manual parsing is the only option to extract values.

---

https://www.rki.de/EN/Home/homepage_node.html

[15] EPISTAT Belgian Infectious Diseases Dashboard
https://epistat.wiv-isp.be/

[16] Institut national d'études démographiques (INED): Demography of COVID-19 deaths
https://dc-covid.site.ined.fr/en/data/

[17] Best Practices for PDF and Data: Use Cases, Methods, Next Steps by Thomas Forth and Paul Connell. The ODI. Available from https://www.w3.org/community/pdf-open-data/odi-report-best-practices-for-pdf-and-data/

[18] Istituto Superiore di Sanità: L'epidemiologia per la sanità pubblica (Epidemiology for public health)
https://www.epicentro.iss.it/

[19] Ministry of Health, Consumer Affairs and Social Welfare (Spain)
https://www.mscbs.gob.es/en/home.htm

[20] For example: COVID-19: Relatório de situação, 2 September 2020

When attempting international comparisons, a key challenge is integrating data produced in different formats by different countries.

Inconsistent data

Sometimes data is discoverable and accessible, but is not consistent. For example, the sum of deaths by age and sex for Germany appears to be shifted when compared to the aggregated data provided by the European Centre for Disease Prevention and Control as shown in the figure below. This raises questions about the reliability of different data sources, which needs to be accounted for in any analysis.

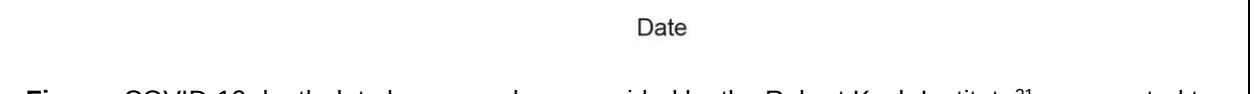

Figure: COVID-19 death data by age and sex provided by the Robert Koch Institute[21], aggregated to daily counts and plotted against the statistics provided by the European Centre for Disease Prevention and Control.

Ephemeral reporting

Some countries provide a historical time series of COVID cases and deaths by age and sex. Others, like Australia[22], report only the latest age-disaggregated data through a dashboard that is refreshed daily. To collect this data a 'web scraper'[23] needs to be written and run daily. If this web-scraper is not written at the point of first publication - at the beginning of the pandemic - the data that has previously been put online is lost to public analysis. While there may still be ways of accessing those datasets, a potentially useful public resource is no longer available.

Languages

All these issues of data collection are compounded by the language challenges when accessing data from across different countries. Search engines don't necessarily match English search queries to websites in other languages. Furthermore, automatic translation of websites is handicapped when text in websites is embedded in images, as is often the case with figures. Where data is published as part of hard-formatted PDF reports, considerable manual intervention is required to translate possibly informative tables or graphs.

Curating high-quality data sources

There are examples of public sector resources that have been made available by national resources in a form that favours rapid analysis, with these resources often coming from countries that had previously experienced major disease outbreaks.

---

https://covid19.min-saude.pt/wp-content/uploads/2020/09/184_DGS_boletim_20200902.pdf (accessed 23 October 2020)

[21] CSV files and table with the Covid-19 infections and mortality per day (time series) https://hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6, accessed 23 October 2020

[22] COVID-19 cases by age, group and sex https://www.health.gov.au/resources/covid-19-cases-by-age-group-and-sex, accessed 23 October 2020

[23] A web scraper is an automated program for regularly accessing a webpage and extracting data from that webpage.

For example, the Mexican General Directorate of Epidemiology[24] provides extensive open data. Their dataset includes tests, hospitalisation, geographic information, travel history and comorbidities for individual (anonymised) patients with IDs that can be tracked over time. This system was developed after 2009's H1N1 outbreak, after which a national digital strategy was developed with the aim of ensuring that there was clear responsibility and political leadership for the use of data to improve public services, and systems to ensure fast and effective cross-government coordination at times of crisis.[25] This has helped create an environment in which there is a single source for government data, backed by a clear methodology for creating and using that data. Countries with similar experience of previous disease outbreaks have also implemented coordinated data management systems that were better placed to respond to the challenges of COVID-19. In contrast, in the UK a range of different agencies have different responsibilities for collecting different forms of data across England, Wales, Scotland and Northern Ireland.

Case Study 3: How have COVID-19 policies affected consumer behaviour, and what are the implications for economic growth?

Official statistics play an important role in assessing the health, wealth and wellbeing of the nation. At times of crisis, these can change at pace. While vital, these official measures may not be responsive enough to inform rapid policy development. Alternative ways of measuring economic activity exist - for example, consumer behaviour data can indicate patterns of economic growth or constriction - but they rely on access to data held by private companies.

Official statistics on household consumption and economic growth have long been used by policymakers to track the economic health of the nation. These typically arrive at quarterly frequency and are aggregated at regional or national scale. While important in measuring the medium- and long-term impact of the pandemic, these are not sufficient to identify changes in patterns of economic activity over weeks or months, at a scale needed to identify where local policy interventions may be needed. Alternative data sources are needed to monitor economics responses, providing signals that indicate the immediate impact of Government actions and highlight where the economic costs of disease and lockdown are being born most highly.

One of the most powerful types of data for conducting this kind of analysis is large-scale payments data collected by banks and other financial service providers. Several countries provide examples of how such data can be harnessed to effectively understand the economic consequences of the COVID-19 crisis: these include Denmark; France; Portugal; Spain; Sweden; and the USA. The data sources in these countries include those provided by private banks to researchers; data analytics companies that collect and organize financial data; and national payments systems operated by consortia of large banks. In the first six months of the pandemic, the examples of researchers successfully using large-scale payments data were all born from existing collaborative relationships between organisations and research groups. It is only in the third quarter of 2020 that we see examples of research on large-scale payments

[24] Gobierno de México. Datos Abiertos - Dirección General de Epidemiología (Government of Mexico. Open Data - General Directorate of Epidemiology) https://www.gob.mx/salud/documentos/datos-abiertos-152127, accessed 23 October 2020

[25] Legal Framework for the National Digital Strategy of Mexico http://www.oecd.org/gov/mexico-legal-framework.pdf

data being facilitated by national statistics offices; in the case of Statistics Portugal (Instituto Nacional de Estatística), it was fully anonymized data from more than 6 million people[26]. The examples from other countries are insightful, and we examine them further in Box B.

In contrast, it seems that no financial institution in the UK that has made large-scale payments data widely available for policy analysis and research during the COVID-19 pandemic. There have been studies that use data from limited samples of households from financial apps like Money Dashboard, which showed the change in weekly consumer spending in different sectors in the UK between 2019 and 2020[27]. Similarly, categorized samples of daily firm-to-firm transactions from the CHAPS payments system gave indications of daily changes to staple, delayable (like DIY and household goods) and work-related spending prior to and during lockdown[28]. CHAPS is a sterling same-day payment system that is used for instance to settle high-value wholesale payments. The samples are small: around 15,000 financial app users in the UK and 90 UK companies, and not the kind of large-scale data available in other countries. While work in this area is continuing, no normalised process for accessing payments data has yet been established.

One of the natural facilitators for expanded use of payments data in the UK is the ONS, which began to establish relationships with payments providers even before the crisis began. While financial institutions may be understandably concerned about their responsibility to carefully-manage the personal data of their customers, that other countries operating within the GDPR have succeeded in setting in place data sharing arrangements to make use of such data suggests that there are approaches that can enable use of this data.

Understanding different data types

The COVID-19 response has demanded rapid creation of new data systems. The early stages of the pandemic were characterized by uncertainty about the incidence and prevalence of COVID-19 at both local and national levels, raising question marks over the apparent and actual hospitalization and mortality rates. These questions have been addressed through increasing the scale and scope of the UK's testing program[29]. Similarly, probability surveys that we speculated may be necessary, for instance to assess the prevalence of COVID-19, were either enacted as proposed or modified appropriately as prioritizations changed[30]. Such

[26] How do People Respond to Small Probability Events with Large, Negative Consequences? Martin Eichenbaum et al, 2020 https://www.kellogg.northwestern.edu/faculty/rebelo/htm/portugalcovid.pdf

[27] Hacioglu, S., D. Kaenzig, and P. Surico. 2020. "Consumption in the Time of COVID-19: Evidence from UK Transaction Data." *CEPR Discussion Paper* No. 14733 https://cepr.org/active/publications/discussion_papers/dp.php?dpno=14733, accessed 23 October 2020

[28] Second quarter 2020 Bank of England Speech (Andy Haldene) https://www.bankofengland.co.uk/-/media/boe/files/speech/2020/the-second-quarter-speech-by-andy-haldane.pdf, accessed 23 October 2020

[29] Coronavirus (COVID-19): Scaling up our testing programmes. Department of Health and Social Care, published 04 April 2020 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/878121/coronavirus-covid-19-testing-strategy.pdf, accessed 23 October 2020

[30] As an example, the REACT-1 (Real-time Assessment of Community Transmission) Study tests a representative sample of 120,000 randomly selected individuals every month to obtain an unbiased estimate of how many people in the population are infected with COVID-19.

core studies are foundational to implementing any public health response. These areas are not the focus of this report. Our aim is to characterize the new data modalities, which we come by through happenstance data, that could be used to improve policy responses.

So-called happenstance data is not originally collected with a particular research or policy question in mind but is created through the normal course of events in our digital lives, and our interactions with digital systems and services. This happenstance data pertains to individual citizens and their daily activities. To be useful it needs to be anonymized, aggregated and statistically calibrated to provide meaningful metrics for robust decision making. Compounding these technical issues associated with the use of such data are differences in understanding of what data might usefully be deployed for which purposes: unlike traditional forms of research data, collected with a clear purpose in mind, happenstance data is repurposed from another function. It may not be obvious to researchers or policymakers what forms of data might be useful, or to those holding data that they have an asset that could be helpfully deployed in response to a crisis.

Deploying these happenstance data resources at times of crisis requires careful cooperation across the private sector, which collects much of the relevant data, the public sector, which can provide an institutional base for data aggregation and governance, and academia, which can supply resources to analyse it. These multiparty data access agreements need to be developed in accordance with data protection legislation and wider concerns about data security, privacy, and value to different parties. The scenarios discussed above illustrate the potential usefulness of two types of happenstance data - origin to destination mobility data and card transaction data - in developing different types of policy response.

Despite the usefulness of this data and the good intent of some commercial partners, these forms of data remain difficult to access in the UK. In contrast, the relative success of some other countries in creating structures to use this data for public good suggests there are policy integrations, practices or relationships that could increase the accessibility of this data. The sections that follow draw from DELVE's experiences in aggregating and analysing COVID-19 policy-relevant data. Through collaborators in those countries we were also able to compare the challenges experienced in their data sharing journey with ours. Those experiences inform the key points outlined at the top of this report and the recommendations that follow.

---

https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/the-react-1-programme/, accessed 23 October 2020

2. Challenges in Data Sharing

Happenstance data can be very difficult to share.

Firstly, happenstance data is often collected at large scale and potentially across many distributed devices. There can be technical challenges to assimilation and sharing of the data for analysis. Even if the data can be shared, the quality and comprehensiveness of the underlying data may not be sufficient to address a given policy question.

Secondly, the data originates from the activities of individual citizens whose privacy and personal data rights must be respected. Careful stewardship of this data is required, based on collaborations across organisations to establish research questions, agree ways of working and set in place formal contracts that together enable its safe and rapid use. Such collaborations take time to establish and rely on pre-existing relationships.

Thirdly, such data is often commercially sensitive, or is perceived to be commercially valuable. This creates business concerns that discourage data sharing across organisations.

Finally, once the quality of the data is understood and inadequacies in the data collection have been accounted for, there remain statistical challenges in the interpretation of the data. In particular population-level studies collecting bespoke research data will be analysed to take account of study design, including potential confounders, bias and study size to minimise the threat of invalid inference and present findings that acknowledge the limitations.

Addressing these challenges requires effort to:

- Streamline the data preparation pipeline
- Build capability for long-term data sharing
- Create incentives or duties to promote responsible data sharing

Streamline the data preparation pipeline

DELVE's experience during the Covid-19 pandemic is that the process by which data scientists, statisticians and other researchers obtain data is frequently the most time-consuming part of attempting to arrive at data-driven answers to any question. This is either because of the difficulty in accessing data, or the process of collecting, cleaning and combining various diverse sources of data.

Establishing a common language

Happenstance data is subject to various levels of "data accounting". Financial data corresponds directly to monetary units and is subject to regular audit, both from corporate entities and their accountants. Other forms of data, with much less direct value, are rarely subject to careful audit. There, data may have missing values or outlier values that stem from incorrect calibration or inappropriate thresholding. In the most extreme case, the data may have been improperly recorded, or vital values may have been deleted from log files to reduce storage requirements.

Data quality can improve when data sharing guidelines are used. Some attempts to formulate such guidelines for the use of data in research can already be found emerging from the Bioinformatics community. A range of organisations have also been active in developing and promoting the 'FAIR' principles for data use, seeking to ensure that open data is made

"Findable", "Accessible", "Interoperable" and "Reusable".[31] Each of these terms indicates a desired state for data by which an individual data set can be judged, in particular around the meta data and the licensing of the data set. These principles describe a useful destination where we would like our datasets to be, but they are less helpful when it comes to characterizing the nature of the journey to get there, what capabilities and resources will be needed, what the milestones should be monitored. In particular, they are of little help when it comes to resourcing a project, either from the perspective of the generators of the data set, or from the point of view of the putative users of the data.

The pipeline for data from discovery and acquisition to decision is fundamentally different for happenstance data created by everyday interactions, than for data collected specifically for research, but the two types of data are rarely differentiated in data sharing discussions. Instead, researchers, civil servants and organisations often believe they are discussing similar types of data or data features, when in fact each is referring to data being held in a different state, which requires different types of processing or management to make it useful. A common language is needed - one that enables common goals to be built around the processing of data - in order to make faster progress in negotiations around data access.

To fill this gap, a system is needed to create a common reference point for both the generators of the dataset and the putative users to discuss its maturity and suitability for analysis. DELVE's approach to filling this gap was the development of a data readiness framework – a data maturity marker that helped coordinate and prioritise the efforts of our research software engineers in bringing datasets to a state of availability.

Case Studies 1 to 3 tell different tales of data readiness. To provide the language with which we could describe the state of data, we broadly classify data into a spectrum of Bands of readiness.[32] The starting point for anyone wishing to provide an answer could be in any of these Bands:

Band D: a blank slate, where the data does not exist at all;

Band C: from 'we've heard via the grapevine that data exists', to verifying its existence and obtaining access to it;

Band B: data is accessible, but it is not clear whether it is appropriate for the question or faithfully described;

Band A: data is readily accessible, documented and useful for a specific task.

The data Band depends on the context in which a question is asked, and the role of those analysing it. Data could be in Band A for the group who owns it, but the same data might be in Band C for someone who doesn't have clearance to access it, or appear in Band B when published online in a report.

The categorization is useful because it frames the expectations on whether any data-driven

[31] See, for example, the 2016 G20 Leaders' Communique from the Hangzhou Summit, paragraph 12, available at: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967 and the work of the International Science Council's Committee on Data (CODATA) https://codata.org/initiatives/working-groups/fair-data-expert-group/

[32] Data Readiness Levels, by Neil D. Lawrence (2017) https://arxiv.org/abs/1705.02245

conclusions could be drawn, and the time and effort involved in the process. The effort is usually spent at the transitions between the Bands: between C and B, finding means to access the data digitally, and between B and A, defining questions that could be answered from the data. As an example of a transition, one might hear that some entity has potentially useful data (Band C), then follow a long path of negotiating digital access, only to discover that the data would not ultimately be useful for the original question.

Implementing a framework for data readiness in DELVE projects

DELVE set out to address a series of research questions in areas where science advice was needed to inform COVID-19 policy development. For each of these questions, DELVE researchers began investigating the data that might be available, labelling sources as being in Band A, B, C or D. This process highlighted areas where rapid data analysis might be possible.

In Case Studies 1 to 3 there are examples where data has to be collected from scratch specifically for the question, and others where data has to be brought together from a diversity of existing sources. Over time, some datasets have remained at a single data readiness level, while others have moved between levels. For example:

- Weather data provided by the Met Office Informatics Lab, which was ingested into the DELVE Global COVID-19 Dataset as country-level population-weighted daily averages of temperature (min, mean, max), precipitation, specific humidity, shortwave radiation, wind speed and shortwave radiation (as a proxy for sunshine), has remained at the boundary between Bands A and B in our classification. This was on the basis that the data was available and well-curated, but its use depended on the research question such data could be deployed to address.
- Both temporal origin to destination mobility data and card transaction data remained in Band C since the start of the pandemic.
- Some data moved from non-existent into Bands B or A as testing capacity increased, or as other institutions shared data more widely. An example of Band C data that became available during the course of the pandemic is the Human Mortality Database's Short-term Mortality Fluctuations (STMF) time series[33], which makes historical weekly mortality statistics available for 34 countries.

From the data scientist's point of view, we encountered two patterns at play when sourcing data during the course of the COVID-19 pandemic. The patterns differ because the alignment of ownership differs. We will look at them individually.

- **Retroactive**: A science or policy question is asked, and then there is a "downward pass", owned by the scientist, to trace data and bring it into Band B and determine its usefulness. The process takes time, as the data owner is not the "owner" of the question who desires it to be in Band B and potentially A.
- **Proactive**: There are data signals which are universally useful in the COVID-19 pandemic. The insight is "owned" by the data owner, who proactively curates it and shares it publicly in Band B. The Google COVID-19 Community Mobility Report[34] is one

[33] Short-Term Mortality Fluctuations Dataseries (STMF) in the Human Mortality Database https://www.mortality.org/Public/STMF_DOC/STMFNote.pdf, accessed 23 October 2020

[34] Google COVID-19 Community Mobility Reports https://www.google.com/covid19/mobility/

such example. Given a question, a researcher could iterate over such datasets and gauge their usefulness for the question in a short space of time.

Assuming that the process is retroactive, the challenges associated with data access and use change as we move across the data readiness levels. For example:

Bands D to C: A common starting point for the public research community is hearsay data – data which is stated to exist but not seen firsthand – which in our classification is somewhere in Band C. Data readiness becomes a function of the *discoverability* and *accessibility* of data. When data is in private hands, a researcher might engage in long discussions with a legal team, sign an access agreement, obtain a version of the anticipated data, only to find out that the data is not appropriate for the particular question. The pattern might look slightly different if data is sensitive and held by a public institution, but the resonating message is similar. One doesn't know in advance how useful data is for a question until one sees the data and tries to answer a question from it.

Bands C to B: The process of finding, accessing and seeing data can take months. Not all clean data (in Band B) is useful, and the process of discovering data and determining its usefulness is often repeated. There are more efficient and streamlined means to determine whether some data is useful for a particular question, without having to ever obtain full access to it. If data is not publicly accessible, the practice of:

- making metadata descriptors of the data available;
- making "dummy" datasets, with the same fields as the original but with a handful of representative but entirely made-up rows;

would already facilitate marking some data as false candidates.

Bands B to A: Whether data is useful for the original question depends on its kind. It could be data that's collected purposefully for the question, or it could be data that were generated as an uncontrolled by-product of other processes. A dataset can only be considered in Band A once a question or task is defined, and that is determined by its existence in Band B. Any of many *happenstance* datasets could be considered Band A for a specific question if it is accessible and its properties understood. The litmus tests of the quality of data-driven answers that could be obtained during an emergency or pandemic like COVID19 are therefore:

- The breadth of national data that is already in Band B, and
- The efficiency of overarching processes by which Bands D and C data could be brought into Band B.

Within DELVE projects, the data readiness bands provided a common language in interdisciplinary collaborations; from the mutual understanding that hearsay data is only hearsay data, to the properties that make it fit for purpose. As most effort is usually spent at the transitions between the Bands, it frames the challenges to readying data. Foremost, Case Study 1 highlighted the downstream impact of the practices of data producers and publishers: substantial "data wrangling" is required to bring data from online dashboards and PDFs into research-friendly formats. A coherent vocabulary about the readiness of data helps the resource planning, time estimation and staffing of projects.

Accreditation mechanisms

There are access barriers between qualified experts and publicly held data, but the steps to

access data held by the ONS have been simplified during the early months of the COVID-19 pandemic. Prior to the pandemic, a researcher had to travel to the ONS headquarters in Wales to do an induction followed by a test to access data already held in the Secure Research Service (SRS). After the induction, data could be accessed in approved "safe rooms", which, in the worst case, meant traveling to ONS headquarters again. Now, induction is done remotely, and access to SRS data is possible from a researcher's home office. The one caveat is that if data is shared with the ONS but is not held in the SRS, it can't be accessed through this protocol.

However, the processes to access national data are not streamlined. The access and induction process has to be repeated for data sources that could possibly be in Band A for a question, but are held by different public bodies. Iterative onboarding, without the guarantee that data, once accessed, might be useful, is a bottleneck when research *agility* is of the essence.

These challenges are further compounded by the multiparty nature of the data access arrangements required to use many forms of happenstance data. Such agreements require that data be aggregated from private sector organisations, standardised and stewarded by public sector bodies, and analysed by authorised third parties, while taking into account the requirements of data protection legislation and concerns about protecting individual rights, data security, and intellectual property.

In this respect, there may be lessons from recent developments in the governance of health data and the role of accreditation mechanisms in enabling rapid data access while continuing careful data stewardship[35]. One way to reach this goal is for the ONS to collaborate with the Information Commissioner's Office (ICO) to formulate their accreditation process into a standardized qualification for data access. The qualification could be equivalent to a "data driving license" that would ensure that qualified experts can get rapid access to *different* data types with the appropriate standardized ethical and legal training in place. The authorisation and authentication of researchers is emerging as an international best practice, with the ONS, HDR UK and other organisations already subscribing to the "five safes" framework[36].In practice the ONS would form part of the convening group that generalizes their framework to a *national qualification* that would include other data sets, with scope for other institutions to join later.

Build capability for long-term data sharing

Access to happenstance data will typically require the coordinated cooperation of many parties. In the case of mobility data and transaction data, for example, effective collaborations during COVID-19 have relied on the skills and expertise of individuals from across the public sector, private sector and academia. Boxes A and B outline the coordinated collaboration

[35] The Global Alliance for Genomics & Health (GA4GH) has approved the technical specifications for "access passports" and associated infrastructure for authentication and authorization. https://www.ga4gh.org/news/ga4gh-passports-and-the-authorization-and-authentication-infrastructure/

[36] UK Data Service. Regulating access to data: five safes: Safe people, safe projects, safe settings, safe outputs, safe data (https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes). This framework also includes the notion of accredited and auditable multi-tenant trusted research environments (TREs), which reduce data travel and simplifies multiple concurrent layers of security (https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf)

required to enable data use, and explore the requirements to set up such collaborations across different countries.

With notable exceptions, barriers to data sharing are perceived rather than real. Several factors likely contribute to this dynamic:

- Data scientists and data protection officers in organisations may perceive different risks associated with data use. This may be because the roles of data scientists and data protection officers are separated. Following our running example of card transaction data, a typical legal concern is that research would require explicit consent from each individual card holder. However, as analysis can often be conducted with fully anonymized subsamples, or aggregates in which no individual could be traced, the risk of disclosure can be directly managed. The challenge in making progress has two sides. On one side, GDPR can too easily be called upon where there is a lack of will to find a solution. On the other side, there is the uncertainty of being in breach of how data protection legislation is interpreted by regulators and Government.[37]
- While many organisations aspire to use data effectively, there may be a mismatch between these strategic aspirations and the business-as-usual practices that contribute to organisational data maturity. The actions described earlier in this report in relation to data readiness can contribute to an organisation's data maturity. In addition to these, there are wider considerations in relation to organisational competence for data use that can contribute to the ease with which data resources are deployed at times of crisis. Some of these are considered further in Addendum 1.

Overcoming these barriers to collaboration takes time, energy, and human resources, all of which are in high demand at times of crisis.

Box A. Coordinated collaboration to enable data use: using human mobility data to understand the impact of COVID-19 policies on the movement of people

Mobility data is informative to public health actions across early-, middle-, and late-stage phases of the COVID-19 pandemic.[38] [39] It is crucial to modelling (estimating) local COVID-19 spread. In our example in Case Study 1, we noted that a scientist in the UK can obtain counts of the daily flow of people from origin to destination between more than 3,000 districts in Spain at the single click of a button, with data running back to the start of the COVID-19 pandemic in Europe in March. The success story is the result of a coordinated effort between Spain's three main mobile phone operators (Orange, Telefónica, Vodafone) and the Instituto Nacional de Estadística. We distil key ingredients that contributed to the timely existence of Spanish origin to destination mobility data here:

[37] Data's value: how and why should we measure it? By Ben Snaith *et al*, 2018. Open Data Institute blog post
https://theodi.org/article/datas-value-how-and-why-should-we-measure-it/

[38] Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle by Nuria Oliver *et al* in Science Advances 2020 6 (23)
https://advances.sciencemag.org/content/6/23/eabc0764

[39] The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology by Kyra H. Grantz *et al* in Nature Communications 11:4961 (2020)
https://www.nature.com/articles/s41467-020-18190-5

Prior experience: As a result of their experiences in modelling the H1N1 flu outbreak in Mexico in 2009, researchers in Telefónica were aware of the value of mobility data[40] for measuring and modelling[41] the spread of an epidemic before COVID-19. The Mexican Government's 2009 response included three levels of lockdown measures (medical alert; closure of all schools and universities; a weeklong shutdown of all non-essential activities).

An early call to action: On 12 March 2020 an article in El País, titled *El valor de los móviles y la Covid-19* (The value of mobiles and COVID-19)[42], made the case for more effective use of mobile phone data. In it, Nuria Oliver explained a feeling of *déjà-vu* as cases surged in Spain. A decade ago, human mobility data from mobile phones were used to model the impact of the Mexican Government's response to H1N1. Mobility data could be of great value to help better plan resources, understand the effectiveness of various public measures to contain mobility, and more accurately predict the spread of COVID-19. The article set out an explanation of why aggregate human mobility could be useful and how this could be managed in a privacy-preserving way, and a call to action for public institutions, the private sector and civil society to collaboratively seek methods of making such data available. The call to action in El País further served to strengthen trust and understanding by engaging the broader public.

A pathfinder project: In 2019, the Instituto Nacional de Estadística (INE) started a pathfinder project with Orange, Telefónica and Vodafone, developing systems to analyse commuter mobility. At the end of 2019, data sharing agreements between these companies and the INE were already in place, and the processes, infrastructure and spatial aggregation functions were already defined. The INE initially paid the telecommunications companies for their data and contributions to the pilot project. The foundational data sharing agreements and infrastructure took more than a year to lay. A pandemic doesn't afford us this time. Quoting Nuria Oliver[43],

"If you have an epidemic with exponential growth, you have no time."

In March 2020, within a week of the publishing of the call to action in El País, presidents or former presidents of Orange, Telefónica and Vodafone, as well as representatives from central and regional governments and other experts were contacted to secure support for an extension of this arrangement. The existence of a pathfinder project meant that it could be swiftly repurposed for COVID-19.

A clear demand from a policy customer: Repurposing the pathfinder project was expedited by a clear first "customer" with a delineated demand. Before any of the mobility data was made public, the INE's first pilot region in March 2020 was Valencia. The Valencia Government, through its assembled COVID-19 data science team, had a clear demand for detailed mobility data.

---

[40] Nuria Oliver: what big data and the Mexican pandemic taught us. Talk at WIRED 2013 https://www.youtube.com/watch?v=H5_FeuuS-zs

[41] Using Big Data to fight pandemics https://www.telefonica.com/en/web/responsible-business/article/-/blogs/using-big-data-to-fight-pandemics, accessed 23 October 2020

[42] El valor de los móviles y la Covid-19 (The value of mobiles and Covid-19) by Nuria Oliver published in *El País* https://elpais.com/elpais/2020/03/12/opinion/1584016142_423943.html, accessed 23 October 2020

[43] Personal interview

A trusted third-party to manage governance concerns: The INE played a key role in acting as a trusted intermediary to negotiate data governance issues. Mobility data potentially poses a number of governance challenges:

- To be useful, such data has to have a sufficient level of granularity to understand changes in commuter behaviours. However, this granularity needs to be balanced against the need for aggregation such that it does not risk exposing private data. To achieve this balance, municipalities with less than 5,000 inhabitants were grouped with neighbouring regions, while regions with more than 75,000 inhabitants were subdivided into smaller cells. Telecommunications companies contributed daily cell-to-cell mobility matrices, without any typical extrapolation based on their coverage. The INE combined the data from the separate parties, and handled privacy concerns by only published cell-to-cell mobility counts of it was larger than 100.
- Data in kind, given generously, could stand in competition with business products of the same or similar data[44]. Origin to destination mobility data from mobile signals, when appropriately aggregated, anonymized and analyzed, can provide valuable and actionable information to businesses. Data sharing, if not properly managed, therefore creates the risk of diluting the business value or business model of the company originally holding the data. The presence of a trusted intermediate to manage data sharing activities can help mitigate this risk.

Box B. Coordinated collaboration to enable data use: using transaction data to understand the impact of policy interventions on economic activity

To understand the nation's economic health, policymakers typically rely on publicly available aggregate measures, such as statistics from national accounts. However, new data sources could offer alternative measures. Transaction data in particular offers a promising way of measuring the economic impact of policy responses to COVID-19, as this allows real-time analysis at a high frequency, with the ability to examine heterogeneity in behavioural changes across the population through types of spending and customer characteristics. Access to this type of data can, however, be challenging: transactions data is both personal and potentially commercially sensitive.

Different countries have established a range of collaborations to enable access to such data:

France: An intricately detailed view of the impact of COVID-19 on consumer spending and business sales on a daily and weekly scale was possible as a result of coordinated cooperation between France's national interbank network, Groupement des Cartes Bancaires (CB). This made nearly five billion payment card transactions from approximately 70 million cards issued by all banks in France available for analysis[45]. Their partnership with analysts complied with the EU GDPR (Article 89), suggesting that the same should be possible in the UK.

[44] An origin to destination mobility example is "Flux Vision" from Orange Business Services https://www.orange-business.com/en/products/flux-vision, accessed 23 October 2020

[45] Consumers' Mobility, Expenditure and Online-Offline Substitution Response to Covid19: Evidence from French Transaction Data. By David Bounie et al, 2020 https://www.telecom-paris.fr/consumers-mobility-online-substitution-covid-19 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588373

Portugal: SIBS (Sociedade Interbancária de Serviços) is the main provider of point of sale terminals and on-line payments in Portugal, and manages the integrated banking network in Portugal. SIBS made aggregate transactional data publicly available early in the COVID-19 pandemic. The data aggregates all individual transactions into monthly observations between 2018 and 2020; there is a monthly observation for each of 39 sectors of activity for each of 308 municipalities.[46] In the latter half of 2020, Statistics Portugal (Instituto Nacional de Estatística) made an administrative data set available for research. This includes monthly data on individual itemized consumer expenditures of over 6 million people that has been anonymized and aggregated, covering the period from January 2018 to May 2020.[47]

Spain: One of the world's largest financial institutions, Banco Bilbao Vizcaya Argentaria (BBVA), made 2.1 billion transaction records available to external collaborators to help gain an understanding of COVID-19's effect on expenditure in Spain. This project was led by BBVA's Chief Economist, and the data was detailed enough to disaggregate point of sale transactions across 76 sectors on a daily frequency.[48] Transaction data revealed insights that one would expect from other sources, for instance how much mobility patterns diverged during lockdown according to income in which poorer households travel more during the workweek.

Denmark and Sweden: Danske Bank, the second largest bank in Scandinavia, anonymized and aggregated a representative subsample spending from 860,000 individual-level bank accounts in Denmark and Sweden from 1 January 2018 to 5 April 2020. As the samples from the two countries -- with very different COVID-19 policies -- are similar in key sociodemographics, geographic concentration in urban or rural areas and local exposure to affected industries, a comparison between the economic impact of policy interventions becomes possible[49]. Importantly, all data processing and aggregating was done inside Danske Bank by authorized personnel, following GDPR. Danske Bank controls data access, and public researchers could formally apply to collaborate.

United States of America: The JP Morgan Chase Institute aggregated data from account balances and transactions from Chase checking accounts, debit cards, and credit cards for over 5 million individuals from January to May 2020. The project was in collaboration with public institution researchers.[50] The size of the data sample allowed for disaggregation of spending and saving patterns by income groups and different segments of the labour market.

---

[46] What and how did people buy during the Great Lockdown? Evidence from electronic payments. By Bruno P. Carvalho *et al*, 2020
https://ideas.repec.org/p/eca/wpaper/2013-307531.html, accessed 23 October 2020

[47] How do People Respond to Small Probability Events with Large, Negative Consequences? By Martin Eichenbaum et al, 2020
https://www.kellogg.northwestern.edu/faculty/rebelo/htm/portugalcovid.pdf, accessed 23 October 2020

[48] Tracking the COVID-19 Crisis with High-Resolution Transaction Data. By Vasco M. Carvalho et al, Cambridge-INET Working Paper Series No: 2020/16
http://www.econ.cam.ac.uk/research-files/repec/cam/pdf/cwpe2030.pdf, accessed 23 October 2020

[49] Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia. By Adam Sheridan *et al* in Proceedings of the National Academy of Sciences Aug 2020, 117 (34) 20468-20473
https://www.pnas.org/content/117/34/20468

[50] Initial Impacts of the Pandemic on Consumer Behavior: Evidence from Linked Income, Spending, and Savings Data, by Natalie Cox *et al*, Becker Friedman Institute working paper number 2020-82
https://bfi.uchicago.edu/wp-content/uploads/BFI_WP_202082.pdf

Amongst other things, the data pointed to large increases in liquid asset balances for American households throughout the income distribution, suggesting that stimulus and insurance programs played an important role in limiting the effects of labor market disruptions on spending.

There are American examples of coordinated cooperation, notably that of the Opportunity Insights Economic Tracker[51] [52], which relies on payments data from at least eleven third-party providers.

The common denominator between the successful examples of data sharing cited in this report is the pre-existence of collaborative relationships or simple mechanisms by which they could be initiated. There have been successful examples of data sharing frameworks that have been established at pace to enable access to data held in the private sector. The experiences of other countries, set out in this report, illustrate the opportunity to pursue this at scale across multiple stakeholders[53]

At a time of national crisis, with events proceeding at pace, there is very little time to build the foundational infrastructure for collaboration. It takes time to agree on data sharing agreements, and it takes time to build data infrastructure. This process is made easier if there has previously been 'in principle' agreements around the forms of data that can be accessed and for what purpose.

The value of pathfinder projects

Pathfinder projects between key segments of industry, academia and government establish ways of working and common agreement around the types of data that could be made available between all parties. Once they exist, they could be adapted to changing policy needs in an agile way. To be successful, such projects require resourcing and strategic leadership, with the experience of other countries suggesting political leadership can play an important role in aligning incentives between stakeholders.[54] Beyond the pathfinder role, these projects can leave a legacy in the form of expertise and guidance in understanding the stages of the data-sharing pipeline.

Through implementation, pathfinder projects provide examples of addressing governance issues by practical necessity. Some are:

- **Multi-party data sharing.** In the examples in Boxes A and B that go beyond prior academic-corporate relationships, data sharing has worked best for organisations that had already invested in the infrastructure and governance mechanisms to enable data sharing between multiple parties in a secure and privacy-aware way.

---

[51] Opportunity Insights Economic Tracker https://tracktherecovery.org/

[52] The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. By Raj Chetty et al, 2020 https://opportunityinsights.org/wp-content/uploads/2020/05/tracker_paper.pdf, accessed 23 October 2020

[53] See footnote 6 for examples of such data sharing arrangements.

[54] Legal Framework for the National Digital Strategy of Mexico http://www.oecd.org/gov/mexico-legal-framework.pdf

- **Protocols for contributors.** We noted in Box A that data given in kind could stand in competition with business products that use the same data, for example aggregated human mobility data. A pathfinder project will contextualize the criteria by which data is contributed, including timelines and compensation.
- **Data privacy.** In Box A the INE, as a trusted third-party, ensured that anonymized and aggregated data wasn't re-identifiable, even when combined with other datasets.

We recommend that pathfinder interdisciplinary data projects are funded. Each project should target a specific policy question,be linked closely to senior decision-makers in relevant government departments, and seek to build on existing work to develop trustworthy data access and linkage frameworks.[55]

At least one of these projects should involve a close collaboration between Cabinet Office and Treasury on nowcasting of classical economic metrics (such as GDP) from happenstance data like payments data. The project may not require any law change, but would require a dedicated effort from Government to institutionalise a process that will enable the same kind of collaboration between academia, government and the key private sector players as the examples in Box B.

A second pathfinder project should produce daily views of population mobility between geographic regions, aggregated from origin to destination counts from mobile phone operators. Many measures to contain COVID-19 targeted the movement of people, and high frequency data about general population mobility is required to inform everything from statistical models to the assessment of the impact of policy decisions. In light of such data's broad utility, the ONS should act as the trusted body to convert happenstance data into high-frequency population mobility statistics.

Accounting for data bias

A dataset is only as valuable as the *design* by which it was collected. Happenstance data is collected for a different purpose than the questions for which it provides a basis for answers to. This necessitates calibration through surveys or complimentary data, much like the REACT-1[56] studies calibrate and correct for the systematic underreporting of COVID-19 cases in the UK. The examples of mobility and transaction data in Boxes A and B preclude the movement and economic activity of anyone who is digitally disenfranchised. Without a careful correction, data will be biased by the uneven use of smartphones or credit cards.

Create incentives or duties to promote responsible data sharing

Data that *might* be useful for a question that's asked in an emergency is usually held by organisations, where it is considered an asset, costing up to millions of pounds to curate and store, and is an integral part of the organisations' operations or business. This kind of data is often *time-person-place data* and is of a sensitive nature, regulated by user agreements.

---

[55] This includes, for example, work by the ONS on the 'Five Safes', further information about which is available at https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme

[56] REACT-1 (Real-time Assessment of Community Transmission) Study. https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/the-react-1-programme/, accessed 23 October 2020

The barriers to data sharing between organisations have been well-explored prior to the COVID-19 crisis. A 2017 report by HM Treasury described how data is "an under-exploited asset"[57], and a range of forces can act as disincentives to data use. For example:

- A lack of common technical standards or the efforts needed to make data interoperable may contribute to a 'market failure' in data sharing[58] or result in data being of insufficient quality to share.
- Organisations may fear privacy or security breaches, or believe that the risks of sharing data outweigh the benefits.[59]
- Organisations may lack the skills or management systems to create data infrastructures that enable its use.[60]

Together, these issues create a cluster of coordination problems between organisations, resulting in misaligned incentives and incomplete contracts.[61]

Corporations and public institutions have indicated willingness to contribute data to projects seeking to help the UK Government in responding to COVID-19, either by making such data public or through establishing collaborations.[62]

For example, the Google COVID-19 Community Mobility Reports[63] can be used to analyse the relative change in the number of visits to certain places – such as transport hubs in a region – over a specified time period. While not providing the granularity of origin to destination mobility data, Google Mobility Reports data can be used to monitor daily changes in mobility across a region, for example over a period of lockdown, to understand how local populations are responding to policy interventions. This data is similar to aggregated, anonymized insights that are used in products such as Google Maps. Without the proactive sharing of such data, it is unlikely that the research community would be sufficiently aware of the existence of this data to seek to build analyses around it.

---

[57] The economic value of data: discussion paper. By HM Treasury, August 2018 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/731349/20180730_HMT_Discussion_Paper_-_The_Economic_Value_of_Data.pdf, accessed 23 October 2020

[58] https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf

[59] Data's value: how and why should we measure it? By Ben Snaith *et al*, 2018. Open Data Institute blog post https://theodi.org/article/datas-value-how-and-why-should-we-measure-it/

[60] Data sharing: implications for policy and practice. Extract from "Towards trusted data sharing: guidance and case studies" at reports.raeng.org.uk/datasharing. Royal Academy of Engineering https://cdn.instantmagazine.com/upload/12506/data_sharing_-_implications_for_policy_and_practice.2ea38cb146d4.pdf, accessed 23 October 2020

[61] See pages 9 and 10 in "The Value of Data" by the Bennett Institute for Public Policy, Cambridge, 2020 https://www.bennettinstitute.cam.ac.uk/media/uploads/files/Value_of_data_summary_report_26_Feb.pdf

[62] For example the Emergent Alliance https://emergentalliance.org

[63] Google COVID-19 Community Mobility Reports https://www.google.com/covid19/mobility/

In many cases, access to a full dataset is not required to be able to draw insights from it. It is instead often sufficient to share useful aggregates under a common interface, alongside information about relevant pre-processing functions and detailed data descriptors.[64] In the Spanish mobility example, we've seen that separate daily mobility matrices were computed inside each telecommunications provider. The daily totals were shared with the INE who, as a trusted third party, combined and truncated them (so that very low origin to destination counts could not be used to infer individual movements by any means) to give national matrices.

The common thread between these successful collaborations is the pre-existence of relationships between researchers, policymakers and companies, considered in section c, below. In the absence of such 'bottom-up' collaborations, alternative approaches are needed to shift institutional behaviours towards data sharing.

There are a range of ways in which policymakers could seek to increase access to data, for example, by imposing a duty to share certain types of data at times of national crisis.

The Office for National Statistics (ONS) is responsible for collection and publication of statistics related to the UK's economy, population and society. Under 2017's Digital Economy Act, the ONS has "permissive and mandatory gateways to receive data from all public authorities and Crown bodies and new powers to mandate data from some UK businesses". It also has a cluster or responsibilities that require it to engage with businesses, for example to understand patterns of trade, and to obtain statistical information about the population of Great Britain.[65] In pursuit of these functions, the ONS has already established mechanisms that enable access to potentially sensitive data for research purposes.[66]

Recognising the importance of timely access to data, in April 2020 the Secretary of State for Health issued notifications to healthcare organisations, GPs, local authorities and arm's length bodies "that they should share information to support efforts against coronavirus (COVID-19)".[67] Where there is data that could be similarly vital in developing public health interventions, similar direction may be required in ensuring that organisations prioritise information-sharing activities.

Given the range of organisations that may hold 'happenstance' data, the ability to issue such a notice - to enable access to data that could support the COVID-19 response - would require expansion of the powers available to the ONS to access data held by organisations, beyond its current powers in relation to sample and survey data. If the ONS is to take this activist role in promoting data sharing at times of crisis, its statutory remit will need to be updated, to cover the compilation of happenstance data, or its aggregates, for use in policymaking on very short time frames. When circumstances are rapidly evolving in a national emergency, such data will contribute to fast decision-making for the nation.

---

[64] This information is required because data is likely to be biased, with the UK population not being represented uniformly, and further analysis should take into account these features of the original data.

[65] The relevant legislation around the Office for National Statistics's transparency and governance is further detailed at:
https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/relevantlegislation

[66] Office for National Statistics (2019), 'Accessing secure research data as an accredited researcher'
https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme

[67] For further information, see:
https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information

3. Conclusion

In the normal course of business, there can be misalignments between the generators and users of a given data set in terms of their objectives and incentives for data sharing. The COVID-19 pandemic has heavily impacted industries, individuals, their households and their family relationships. In some ways, this has removed a barrier to the process of data sharing; in our interactions with companies, we found them keen to help in addressing the important policy questions on which we were working. Agreements in *principle* on what data might be shared for which purposes were easy to make, but agreement in *practice* proved very difficult to implement.

While there are multiple stumbling blocks in addressing the challenges above, we were able to find examples of successful data sharing endeavours from other European countries that are currently subject to the same regulatory environment as far as data privacy and personal data rights are concerned. These countries also have the same commercial providers for their mobile phones and there is a considerable overlap between their banking systems and ours.

Core to the success of these efforts are teams of highly-skilled people from across academia, the public sector and companies, with skills covering data science, data governance and the development of digital tools for use in policymaking, who are able to rapidly create and scale research collaborations. Demand for people with such skills across sectors of the economy is high[68] and building capability in these areas is a significant challenge for governments.[69] In recent years, the UK Government has introduced a variety of initiatives aiming to build its digital capabilities, including professional development programmes and changes to recruitment.[70] However, it has also recognised that further work is required to equip civil servants with the skills they need to make effective use of data in policymaking.[71] Similar challenges exist across other sectors.[72] In this context, pathfinder projects can play a crucial role in building expertise at all levels in how best to manage and deploy data resources.

The COVID-19 pandemic has highlighted the best and worst practices in data access: our experience of obtaining similar data from different countries is of a wide range of practices, from screen-scraping individual PDFs to parsing well-documented CSV files directly with minimal effort.

Recent years have seen a variety of efforts to establish principles or guidelines for well-governed data sharing. However, many of these remain statements of good intention, rather than providing mechanisms for improving data quality and accessibility. DELVE used the data readiness bands, outlined in this report, as a common language for understanding progress in

---

[68] Dynamics of data science skills. Royal Society Report, 2019
https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/

[69] Institute for Government report: Making a success of digital government. By Emily Andrews et al
https://www.instituteforgovernment.org.uk/sites/default/files/publications/IFGJ4942_Digital_Government_Report_10_16%20WEB%20%28a%29.pdf, accessed 23 October 2020

[70] See paragraphs 130-131 in:
https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/1455/145509.htm

[71] "The privilege of public service". 2020 Ditchley Annual Lecture, by Michael Gove.
https://www.gov.uk/government/speeches/the-privilege-of-public-service-given-as-the-ditchley-annual-lecture

[72] Dynamics of data science skills. Royal Society Report, 2019
https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/

a data sharing project. Other efforts, such as the Oxford's Blavatnik School of Government's COVID-19 Government Response Tracker[73], show alternative means of compiling agile data resources in an emergency.

Ensuring these efforts are successful and sustainable will require trustworthy governance of data assets, both happenstance and traditional, creating frameworks that can give members of the public confidence that their data is being used safely and rapidly. Evidence from public dialogues on data and its use suggest that key concerns include the purposes for which data is used and how the benefits of its use are shared across society.[74] Such dialogues suggest, for example, that individuals are generally content for data to be used to improve healthcare services that result in public benefit.[75] Embedding such dialogue in the development of data governance frameworks can help develop and steward data assets in ways that maintain public confidence.[76]

In addition to the lessons from the COVID-19 response presented in this report, Government could take further action to bolster its planning for civil emergencies by reviewing the National Risk Register[77] through the lens of data readiness. Such a review would assess how data could be used in policy interventions for crisis response, the data resources that would need to be readily-available to enable such response, and the action needed to improve data readiness or organisational data maturity in those areas.

The availability and quality of the UK nation's data dictates our ability to respond in an agile manner to evolving events. If the UK is to effectively deploy its data resources and its expertise in data science to tackle future emergencies – whether from the continuing spread of COVID-19 or other rapidly-emerging policy challenges – action is needed to create a culture of careful stewardship of these resources. Such stewardship would seek to enable safe and rapid use of data to achieve public benefit, while managing concerns about individual rights and respecting the commercial value of these resources.

---

[73] Oxford COVID-19 Government Response Tracker https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

[74] See, for example, Royal Society and Ipsos MORI (2017) Public views of machine learning, available at https://royalsociety.org/topics-policy/projects/machine-learning/

[75] See, for example, work by Understanding Patient Data, available at https://understandingpatientdata.org.uk/sites/default/files/2018-08/Public%20attitudes%20key%20themes.pdf

[76] For further discussion of this, see work by the Royal Society and Centre for Data Ethics and Innovation on the role of public dialogue in policymaking, available at https://royalsociety.org/-/media/policy/projects/ai-and-society/RS-CDEI-Roundtable---Note-of-Discussions.pdf

[77] Cabinet Office (2017) National Risk Register of Civil Emergencies, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644968/UK_National_Risk_Register_2017.pdf

Appendices

Appendix A

Python code snippet for loading COVID-19 cases by age and sex for Belgium:

```
pd.read_csv('https://epistat.sciensano.be/Data/COVID19BE_CASES_AGESEX.csv
', parse_dates=['DATE'])
```

Python code snippet for loading COVID-19 cases by age and sex for Spain for a *single day*. Note that the coordinates listed in table_areas can change from report to report and need to be established manually.

```
url =
'https://www.mscbs.gob.es/en/profesionales/saludPublica/ccayes/alertasAct
ual/nCov-China/documentos/Actualizacion_54_COVID-19.pdf'

table_areas = [
 # (top left bottom right)
 [1.85*72, 0.8*72, 4.75*72, 7.65*72],
 [4.75*72, 0.8*72, 7.5*72, 7.65*72],
 [7.5*72, 0.8*72, 10.1*72, 7.65*72],
]

 url = base_url.format(report=report_number)
 with urlopen(url) as f:
   b = BytesIO(f.read())
 reader = PdfFileReader(b)
 num_pages = reader.getNumPages()
 found_page = -1
 date = None
 for i in range(num_pages):
   page = reader.getPage(i)
   text = page.extractText()

   if i == 0:
     match = re.search(r'.*(\d{2}\.\d{2}\.(?:\d{2}){1,2}).*', text)
     if not match:
       print("COULDN'T FIND DATE MATCH!")
       print(text)
       break
     date_text = match.groups()[0]
     if len(date_text) == 8:
       date_text = date_text + '20'
     date = pd.to_datetime(date_text, format='%d.%m.%Y')

   if re.search('.*edad y sexo.*', text):
     print(f'Found age data on page {i} in report {report_number}')
     found_page = i

 if found_page >= 0:
   print(f'Attempting to read table from page {found_page}')
   dfs = []
```

```
   for area in table_areas:
     b.seek(0)
     dfs.append(tabula.read_pdf(b, output_format='dataframe',
pages=found_page+1, area=area, pandas_options={'dtype': 'str', 'header':
None}, multiple_tables=False, stream=True)[0])
```

Addendum 1 Organizational Data Maturity

Prepared for the DELVE Initiative by Neil Lawrence, Jess Montgomery and Ulrich Paquet

One challenge that the DELVE group experienced was understanding how much resource was necessary to bring a particular data set up to the sufficient level of data readiness for use. This resource depends a great deal on what the data is, what its provenance is and who is managing it. Organizational Data Maturity is about the third of these three factors: who is managing the data.

Advances in machine learning over the last 5-10 years have generated interest across sectors in the potential of advanced data analytics to enhance productivity and improve decision-making within organisations. Many companies aspire to be data driven in their decision making. But even within these organizations, the accessibility and availability of data may be limited. Similar challenges apply to a range of organizations, including government departments, the health service, local authorities and even academic fields.[78]

In support of these aspirations, a variety of approaches to assessing data maturity have emerged in recent years. These seek to help organisations understand how their current data management practices help - or hinder - the use of data in decision-making, and the interventions that can contribute to more effective deployment of organisational data resources.[79] Such interventions include technical measures (for example, adhering to data quality standards), organisational processes (for example, to share data across teams), or cultural change (for example, around how an organisation values or invests in managing its data).

The Interface with Science

This report discusses the actions needed to create data resources that can be readily deployed in data-enabled policy analysis. The cultural factors and operational processes that help create datasets that are ready for such deployment also contribute to an organisation's data maturity and its ability to generate inter-organisational business insights through use of data, and vice versa - data readiness and data maturity are interlinked. One consequence of organizational data maturity is therefore the potential to contribute evidence to scientific analysis that can contribute to policymaking. Such analysis requires resources at the "Band A" level of data readiness[80], and happens through data as an Application Programming Interface (API).[81]

Different types of API have been developed to facilitate the use of data to tackle COVID-19. These have varied from simply updating and republishing publically accessible CSV files, to

[78] We will refer to these entities as *organizations* in our text below. To reflect the hierarchical structures of these organizations, we will also refer to *departments* and *teams* as smaller sub-units of the wider organization.

[79] See, for example: https://datamaturity.esd.org.uk ; https://www.joelonsoftware.com/2000/08/09/the-joel-test-12-steps-to-better-code/ ; https://www.cio.com/article/3077871/the-four-stages-of-the-data-maturity-model.html ; https://www.oreilly.com/content/10-signs-of-data-science-maturity/

[80] Data Readiness Levels, by Neil D. Lawrence (2017) https://arxiv.org/abs/1705.02245

[81] An API is an interface that defines the mode of interaction between different software intermediaries. If an API to data is made available, it means that the data can be accessed programmatically, i.e. by the software directly, without the need for direct human intervention.

adding documentation and code, to that data processing functionality being shared between projects that use the data data, to controlling access in dedicated cloud compute environments. As one specific interface with the scientific community, the case study box below examines the Met Office Informatics Lab's API.

Case study: Met Office Informatics Lab COVID-19 Pangeo Environment

There is still uncertainty about the role of weather as a direct factor in COVID-19 transmission rates, or as an indirect factor via the ways it affects people's behaviour. As a result, there is a thin line between hypotheses becoming policy.[82] *To share weather data for understanding the interplay between COVID-19 and environmental factors, the Met Office instantiated a cloud API for research collaboration.*

The Met Office has world-leading weather forecasts, and as an organization ingests millions of data observations from around the world every day. To be useful for COVID-19 research, it was published as part of an API that also contained ancillary information that allowed weather data to be joined with other data sources. Such information included so-called shape files of different geographic regions, which were helpful to align atmospheric data to the granularity of COVID-19 reporting.

To implement the API, the Met Office Informatics Lab made hourly and daily global gridded weather data, including air temperature, precipitation, shortwave radiation (sunshine) and humidity available[83] through Microsoft Azure Blob Storage. Specific aggregations of this data for administrative regions in the UK, Italy and USA were included. The data interface with the broader scientific community was through a custom[84] Pangeo[85] environment, which included custom tools in Python and R, along with access to a Jupyter Lab Integrated Development Environment from which the data could be queried.

Data and compute access to the Pangeo environment was by request, followed by authentication through a data scientist's Github account. It meant that researchers never had to copy terabytes of data to their own machines, but run processing scripts "where the data lived". As one example, the DELVE Global COVID-19 Dataset[86] includes population-weighted weather data for every day for every country where COVID-19 statistics are reported by the ECDC[87].

---

[82] Misconceptions about weather and seasonality must not misguide COVID-19 response, by Colin J. Carlson et al, in Nature Communications 11:4312 (2020)
https://doi.org/10.1038/s41467-020-18150-z

[83] Met Office and partners offer data and compute platform for COVID-19 researchers
https://medium.com/informatics-lab/met-office-and-partners-offer-data-and-compute-platform-for-covid-19-researchers-83848ac55f5f, accessed 23 October 2020

[84] Met Office Informatics Lab: Our new Pangeo architecture
https://medium.com/informatics-lab/our-new-pangeo-architecture-bfc1b2b23658, accessed 11 November 2020

[85] A community platform for Big Data geoscience
https://pangeo.io/

[86] DELVE Global COVID-19 Dataset
https://rs-delve.github.io/data_software/global-dataset.html

[87] European Centre for Disease Prevention and Control: COVID-19 pandemic
https://www.ecdc.europa.eu/en/covid-19-pandemic

Data Maturity Assessments

Whether “Band A” data is public, or in a controlled environment like the case study above, or completely private within an organization, it is the result of healthy data management. Different organizations have different levels of data maturity, reflecting their capability in implementing such management systems. Even within organizations, maturity will differ between different teams and groups.

In order for data readiness to improve, it is important to assess what a particular organization’s level is, and the action needed to improve current practices. The table below suggests a loose framework through which organizations could gauge their data maturity, consider how their ways of working contribute to the ‘readiness’ of their data resources, and better connect their aspirations for data sharing to the action required to enable this.

To reflect the fact that a number of specific skill-sets, as well as cultural approaches to data are required, we suggest the term *data maturity* to reflect the ability of organizations, teams and individuals to efficiently process data. We have created a provisional data maturity model with five levels of increasing maturity *(i) reactive, (ii) repeatable, (iii) managed/integrated, (iv) optimized,* and *(v) transparent*. In the table below we summarize these different levels of Data Maturity as five levels. Although many similar models exist[88], the intent of these levels is to be a starting point that should be adapted according to the specific context that it is applied in.

Maturity Level	**Data Sharing**
1 Reactive	Data sharing is not possible or ad-hoc at best.
2 Repeatable	Some limited data service provision is possible and expected, in particular between neighboring teams. Some limited data provision to distinct teams may also be possible
3 Managed and Integrated	Data is available through published APIs; corrections to requested data are monitored and API service quality is discussed within the team. Data security protocols are partially automated ensuring electronic access for the data is possible.
4 Optimized	Teams provide reliable data services to other teams. The security and privacy implications of data sharing are automatically handled through privacy and security aware ecosystems.
5 Transparent	Internal organizational data is available to external organizations with appropriate privacy and security policies. Decision making across the organisation is data-enabled, with transparent metrics that could be audited through organisational data logs. If appropriate governance frameworks are agreed, data dependent services (including AI systems) could be rapidly and securely redeployed on company data in the service of national emergencies.

[88] The idea of Data Maturity is by no means unique, and many others exist: there is the Dell Data Maturity Model (https://www.cio.com/article/3077871/the-four-stages-of-the-data-maturity-model.html) as well as many Big Data Maturity Models. Classic models like the Capability Maturity Model examine software processes.

For data quality to improve, we must first empower organizations to assess the levels of data maturity across their teams and departments. Below we provide a set of indicators that can be used for assessing Data Maturity. It is inspired by the "Joel test"[89] for software development.[90]

Characterising Data Maturity

In this section we consider how organisations can assess their data maturity, by reviewing the ways in which best practice in data management and use is embedded in teams, departments, and business processes. These indicators are loosely themed according to the maturity level above. In practice, these characteristics would be reviewed in aggregate to give a holistic picture of data management across an organisation.[91]

1 Reactive

Data sharing is not possible or ad-hoc at best.

1. It is difficult to identify relevant data sets and their owners.
2. It is possible to access data, but this may take significant time, energy and personal connections.
3. Data is most commonly shared via ad hoc means, like attaching it to an email.
4. The quality of data available means that it is often incorrect or incomplete.

2 Repeatable

Some limited data service provision is possible and expected, in particular between neighboring teams. Some limited data provision to distinct teams may also be possible.

5. Data analysis and documentation is of sufficient quality to enable its replication one year later.
6. There are standards for documentation that ensure that data is usable across teams.
7. The time and effort involved in data preparation are commonly understood.
8. Data is used to inform decision-making, though not always routinely.

3 Managed and Integrated

Data is available through published APIs; corrections to requested data are monitored and API service quality is discussed within the team. Data security protocols are partially automated ensuring electronic access for the data is possible.

9. Within the organisation, teams publish and share data as a supported output.
10. Documentation is of sufficient quality to enable teams across the organisation that were not involved in its collection to use it directly.
11. Procedures for data access are documented for other teams, and there is a way to obtain secure access to data.

---

[89] Developed to assess the quality of work produced by software teams, the Joel Test is Joel Spolsky's series of 12 simple yes or no questions that can be used to identify areas for improvement in programming teams.
https://www.joelonsoftware.com/2000/08/09/the-joel-test-12-steps-to-better-code/

[90] This work also builds on previous engagement between the authors and others working in this area, including Daniel Marcu (2013). Machine Translation Maturity Models. Unpublished notes.

[91] To achieve this, each indicator could be assessed on a Likert scale.

4 Optimized

Teams provide reliable data services to other teams. The security and privacy implications of data sharing are automatically handled through privacy and security aware ecosystems.

12. Within teams, data quality is constantly monitored, for instance through a dashboard. Errors could be flagged for correction.
13. There are well-established processes to allow easy sharing of high-quality data across teams and track how the same datasets are used by multiple teams across the organisation.
14. Data API access is streamlined by an approval process for joining digital security groups.

5 Transparent

Internal organizational data is available to external organizations with appropriate privacy and security policies. Decision making across the organisation is data-enabled, with transparent metrics that could be audited through organisational data logs. If appropriate governance frameworks are agreed, data dependent services (including AI systems) could be rapidly and securely redeployed on company data in the service of national emergencies.

15. Data from APIs are combined in a transparent way to enable decision-making, which could be fully automated or through the organization's management.
16. Data generated by teams within the organisation can be used by people outside of the organization.