



## EUROPE

CHILDREN AND FAMILIES  
EDUCATION AND THE ARTS  
ENERGY AND ENVIRONMENT  
HEALTH AND HEALTH CARE  
INFRASTRUCTURE AND  
TRANSPORTATION  
INTERNATIONAL AFFAIRS  
LAW AND BUSINESS  
NATIONAL SECURITY  
POPULATION AND AGING  
PUBLIC SAFETY  
SCIENCE AND TECHNOLOGY  
TERRORISM AND  
HOMELAND SECURITY

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis.

This electronic document was made available from [www.rand.org](http://www.rand.org) as a public service of the RAND Corporation.

Skip all front matter: [Jump to Page 1](#) ▼

### Support RAND

[Browse Reports & Bookstore](#)

[Make a charitable contribution](#)

### For More Information

Visit RAND at [www.rand.org](http://www.rand.org)

Explore [RAND Europe](#)

View [document details](#)

### Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND electronic documents to a non-RAND Web site is prohibited. RAND electronic documents are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see [RAND Permissions](#).

This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

# TECHNICAL REPORT

---

## Data strategies for policymaking

Identifying international  
good practice

Final report

Philipp-Bastian Brutscher, Jan Tiessen,  
Ala'a Shehabi, Daniel Schweppenstedde,  
Claire Celia, Christian van Stolk, Tom Ling

Prepared for the European Commission Directorate General for  
Health and Consumers (DG SANCO)

The research described in this report was prepared for the European Commission. The opinions expressed in this study are those of the authors and do not necessarily reflect the views of the European Commission.

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

**RAND**® is a registered trademark.

© Copyright 2010 European Commission

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the European Commission.

Published 2010 by the RAND Corporation  
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138  
1200 South Hayes Street, Arlington, VA 22202-5050  
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665  
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom  
RAND URL: <http://www.rand.org>  
RAND Europe URL: <http://www.rand.org/randeurope>  
To order RAND documents or to obtain additional information, contact  
Distribution Services: Telephone: (310) 451-7002;  
Fax: (310) 451-6915; Email: [order@rand.org](mailto:order@rand.org)

# Preface

---

The European Commission Directorate General for Health and Consumers (DG SANCO) commissioned RAND Europe to provide support in developing a comprehensive data strategy for DG SANCO that meets the needs of increasingly evidence-based policymaking in the future.

This final report presents the findings of RAND Europe's analysis. It contains some of the analytical thinking around the concept of a data strategy as well as findings from 13 international case studies on international data-management practices.

This report is likely to be of interest to officials involved in data management in the public sector and to a wider audience interested in the foundations of evidence-based policymaking.

RAND Europe is an independent not-for-profit research organisation that aims to serve the public interest by improving policy making and informing public debate. Its clients are European governments, institutions and firms with a need for rigorous, impartial, multidisciplinary analysis. This report has been peer reviewed in accordance with RAND's quality-assurance standards.

For more information about this document or RAND Europe, please contact:

Jan Tiessen

RAND Europe  
Westbrook Centre  
Milton Road  
Cambridge CB4 1YG  
United Kingdom  
Tel. +44 (1223) 353 329  
tiessen@rand.org



# Contents

---

Preface.....	iii
List of Figures.....	ix
List of Tables.....	xi
List of Acronyms .....	1
<b>Executive Summary.....</b>	<b>1</b>
Background and rationale.....	1
Developing a conceptual framework for data strategies .....	1
Lessons for DG SANCO .....	3
Identifying and prioritising data needs.....	3
Collecting data .....	3
Data management .....	4
Data analysis .....	5
Data use.....	5
Revising the linear model .....	6
<b>CHAPTER 1 Introduction.....</b>	<b>1</b>
1.1 Evidence-based policymaking and the need for data .....	1
1.2 Challenges for DG SANCO.....	1
1.3 DG SANCO’s efforts to develop a comprehensive data strategy .....	2
1.4 Research question and research approach.....	3
1.5 The case studies.....	3
1.6 This report .....	4
<b>CHAPTER 2 A conceptual framework of data strategies.....</b>	<b>5</b>
2.1 Defining data .....	5
2.1.1 The data, information and knowledge wisdom hierarchy .....	6
2.1.2 Typologies and categories of data .....	8
2.2 Identifying data needs .....	9
2.2.1 Exploring issues in data management.....	10
2.3 A data-strategy framework.....	11
<b>CHAPTER 3 Identifying and prioritising data needs.....</b>	<b>13</b>
3.1 Defining data needs: centralised or decentralised?.....	13
3.2 Systematic and ad hoc definition of data needs.....	15

3.3	Prioritising data needs: not at all, bottom up or top-down?.....	16
3.4	Summary/Conclusion.....	18
3.5	Lessons for DG SANCO .....	19
<b>CHAPTER 4</b>	<b>Data collection.....</b>	<b>21</b>
4.1	External vs. internal data collection.....	21
4.2	Primary and secondary data .....	24
4.3	Assuring quality in data collection .....	24
4.4	Summary/Conclusion.....	26
4.5	Lessons for DG SANCO .....	26
<b>CHAPTER 5</b>	<b>Data management.....</b>	<b>29</b>
5.1	Who is responsible for data management.....	29
5.2	Data verification .....	32
5.3	Privacy and identifiable data .....	33
5.4	Summary/Conclusion.....	34
5.5	Lessons for DG SANCO .....	34
<b>CHAPTER 6</b>	<b>Data analysis.....</b>	<b>37</b>
6.1	Data – information – knowledge .....	37
6.2	Who analyses data? .....	37
6.3	Performing analysis.....	40
6.4	Assuring quality in data analysis.....	40
6.5	Reporting the analysis.....	42
6.6	Summary/Conclusion.....	43
6.7	Lessons for DG SANCO .....	43
<b>CHAPTER 7</b>	<b>Data use.....</b>	<b>45</b>
7.1	Data users.....	45
7.2	What are data used for?.....	46
7.3	Addressing users .....	47
7.4	Summary/Conclusion.....	49
7.5	Lessons for DG SANCO .....	49
<b>CHAPTER 8</b>	<b>Revising the linear model.....</b>	<b>51</b>
8.1	Identifying and prioritising data needs.....	51
8.2	Data collection .....	53
8.3	Data management .....	54
8.4	Data analysis.....	54
8.5	Summary/Conclusion.....	56
8.6	Lessons for DG SANCO .....	56
<b>CHAPTER 9</b>	<b>Conclusions and recommendations.....</b>	<b>57</b>
9.1	The key lessons for DG SANCO.....	57



Identifying and prioritising data needs..... 57

Collecting data ..... 57

Data management ..... 58

Data analysis ..... 58

Data use58

Revising the linear model ..... 59

9.2 Conducting research into public-sector data strategies ..... 59

9.3 Implementing data strategies: the maturity of information-management systems ..... 60

**REFERENCES ..... 63**

Reference list ..... 65

**APPENDICES ..... 71**

**Appendix A: Methodology ..... 73**

Literature review ..... 73

Interview template ..... 74



# List of Figures

---

Figure 0.1: Data-strategy framework.....	3
Figure 2.1: The DIKW hierarchy.....	6
Figure 2.2: Data-strategy framework.....	12
Figure 3.1: Identification and prioritisation as part of a comprehensive data strategy .....	13
Figure 4.1: Data collection as part of a comprehensive data strategy.....	21
Figure 4.2: Trade-off internal/external data collection.....	23
Figure 5.1: Data management as part of a comprehensive data strategy .....	29
Figure 6.1: Analysis as part of a comprehensive data strategy.....	37
Figure 7.1: Data use as part of a comprehensive data strategy.....	45
Figure 8.1: Essential feedback loops in a comprehensive data strategy .....	51
Figure 9.1: The Gartner information-management maturity model .....	61



# List of Tables

---

Table 0.1: Overview of case studies.....	2
Table 1.1 Research approach.....	3
Table 1.2: Overview of case studies.....	4



# List of Acronyms

---

AHW	Animal Health and Welfare Delivery
DEFRA	Department for Environment, Food and Rural Affairs (UK)
DGCCRF	Direction générale de la concurrence, de la consommation et de la répression des fraudes (France)
DG SANCO	European Commission Directorate General for Health and Consumers
DoD	Department of Defence (US)
EHO	European Health Observatory
EIM	enterprise information management
EFSA	European Food Standards Agency
FSA	Food Standards Agency (UK)
HiT	Health Systems in Transition
IMF	International Monetary Fund
INC	Instituto Nacional del Consumo (Spain)
NAO	National Audit Office (UK)
OECD	Organisation for Economic Cooperation and Development
PHAC	Public Health Agency Canada
RRG	Dutch Regulatory Reform Group (joint organisation of the ministries of Finance and Economic Affairs)
SAMHSA	Substance Abuse and Mental Health Services Administration (US Department of Health and Human Services)
SCA	Swedish Agency for Consumer Affairs (Konsumentverket)
SIS	Statistical Information System
WHO	World Health Organisation





# Executive Summary

---

## Background and rationale

In recent years the European Commission has gradually moved towards a more evidence-based policymaking approach. The underlying assumption is that policy decisions based on systematic evidence produce better outcomes.

With the move towards evidence-based policymaking, the need for high-quality data – which can include statistical and qualitative data, but also existing research – has increased. The data needs for the increasingly formalised and institutionalised impact-assessment process are an example of how data needs of modern policymaking have increased.

DG SANCO has already done much to respond to this increased need – including developing its Consumer Markets Scorecard, its health economics database and its health statistics database. Despite these efforts, DG SANCO feels it needs to explore the possibility of having a more coordinated data strategy in order to be able to put the best evidence from research at the heart of policy development, implementation and enforcement.

## Developing a conceptual framework for data strategies

Drawing on the literature and 13 case studies of data strategies in different organisations dealing with data (for an overview of the organisations, see Table 0.1 below), we developed a simple framework that allows DG SANCO to think about its own data strategy in a systematic way. We discussed the corresponding key elements and drew out possible lessons for DG SANCO.

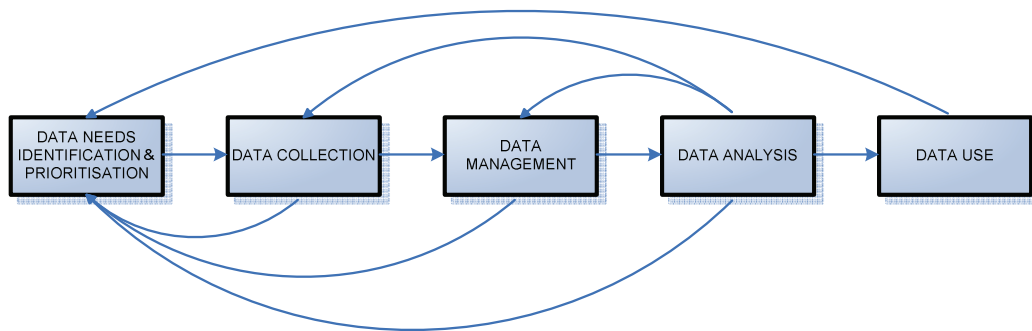
**Table 0.1: Overview of case studies**

	<b>Comparator</b>	<b>Country</b>	<b>Policy area</b>
1	Department for Environment, Food and Rural Affairs (DEFRA)	UK	Animal health
2	DG Enterprise	EU	General
3	DG Environment	EU	General
4	DG Trade	EU	General
5	Direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF)	France	Consumer
6	European Health Observatory (EHO)	International organisation	Public health
7	Food Standards Agency (FSA)	UK	Consumer
8	Instituto Nacional del Consumo (INC)	Spain	Consumer
9	Konsumentverket, Swedish Agency for Consumer Affairs (SCA)	Sweden	Consumer
10	Organisation for Economic Cooperation and Development (OECD)	International organisation	General
11	Public Health Agency Canada (PHAC)	Canada	Public health
12	Regulatory Reform Group (joint organisation of the ministries of Finance and Economic Affairs) (RRG)	Netherlands	General
13	World Health Organisation (WHO)	International organisation	Public health

The framework we developed represents an idealised, sequential process of developing a data strategy ‘from scratch’. It comprises five key elements (for an illustration see Figure 0.1: below):

1. identification and prioritisation of data needs
2. data collection
3. data management
4. data analysis and communication
5. data use.

The simple idea underlying the data-strategy framework is that the identification and prioritisation of data needs feeds into and determines the collection of data; which in turn feeds into data management (including data storage, verification and compilation); which again feeds into data analysis; which finally informs policymakers and so feeds into use of the data. At the same time we do expect there to be feedback mechanisms that do not follow the sequential model, but link different elements directly.



**Figure 0.1: Data-strategy framework**

## Lessons for DG SANCO

Using the data-strategy framework developed in the initial stage of the research, the key lessons learned and the recommendations for DG SANCO may be summarised as follows.

### Identifying and prioritising data needs

There are several ways of identifying data needs. We distinguished between decentralised and centralised approaches to data identification and discussed the advantages and disadvantages of a systematic approach to identifying data needs, as opposed to an ad hoc approach.

When it comes to the prioritisation of data needs, we found that organisations differ in terms of the degree of formality with which this is done. Some organisations do not prioritise their identified data needs at all; others do so in a very informal way; still others use a formal process. Formal approaches, in turn, may be categorised into top-down and bottom-up approaches.

#### *Lessons for DG SANCO*

The main lessons for DG SANCO regarding the identification of data needs that follow from our discussion include the following:

1. Data needs should – where possible – be identified in a systematic way, using appropriate methods to overcome the cognitive limitations that can often act as a barrier to the identification of data needs.
2. To the extent that subject-related and methodological expertise sits with individual policy units, a decentralised system of identifying data needs is generally preferable to a centralised one.
3. A decentralised system of identifying data needs can be complemented with a top-down way of prioritising data needs, which ensures high-level buy-in to empirical work. At the same time, it allows the collection of the data needed to answer the most pressing policy questions.

### Collecting data

When it comes to collecting data, we distinguished between organisations that collect data internally, externally and by means of a downstream agency; and organisations that collaborate with other organisations in data collection.

We discussed some of the pros and cons (in terms of costs and timeliness) of organising data collection in one of these ways. With regard to the question of what data to collect, or where to collect them from, we distinguished between primary data collection, a hybrid type of data collection (which relies on Member States and/or agencies to collect data) and secondary data collection.

We found that the hybrid form of data collection is the most popular of these – in particular if, as is often the case, primary data collection is not available. Regardless of which approach is chosen, basically all organisations stress the importance of having quality controls in place to organise the data collection. The main approaches used include peer review, pilot studies and transparency with regard to the collection approach, coverage, response rates and other essential characteristics of the collection process.

#### *Lessons for DG SANCO*

The lessons for DG SANCO regarding the collection of data that follow from our discussion include the following:

4. The organisation of data collection should take costs into account and rely primarily on collaborative collection efforts. Only where this is not possible on external (internal) data collection efforts if the same type of data is collected with low (high) frequency.
5. Similarly, it is advisable to rely on secondary data collection when possible and to collect primary data only where secondary data are not available/suitable. At the same time, the case study on the UK Department for Environment, Food, and Rural Affairs (DEFRA) illustrates that generally fewer relationships / secondary data sources are preferable to many.
6. Probably the most promising approach to assure the quality of data to be collected is to use a peer-review process at the planning and possibly the pilot stage of the data collection. There is good reason to try to involve people responsible for the data analysis in this process.

#### **Data management**

In the context of data management, we reviewed the advantages and disadvantages of a centralised/decentralised data-storage system and compared how data verification is handled across the organisations we reviewed.

In addition, we discussed the role of data compilation in the context of the work of different organisations and outlined the main strategies of dealing with identifiable data: not to collect identifiable data in the first place, or, if identifiable data is collected, to make sure that the data are anonymised before being disseminated and that the requestor of the data is legally bound to use the data sensibly.

#### *Lessons for DG SANCO*

From our discussion, again several interesting points for DG SANCO emerge. These include the following:

7. DG SANCO could use three main criteria to evaluate whether a centralised or decentralised data-storage system is more appropriate: i) Is there an overlap in the data needs of different policy units; ii) Are data needs rather limited in scope? iii) How timely do the data have to be?

8. Owing to the danger of encountering severe bias if data verification is designed in the wrong way, DG SANCO could benefit from a formal system for this, which could be done either in-house or by out-sourcing.
9. To the extent that DG SANCO intends to use its data for research purposes, there is good reason to take data compilation seriously. WHO, PHAC and OECD offer good examples of how data compilation can be organised (see Box 8).
10. Given the high risk associated with dealing with identifiable data, DG SANCO may want to adopt a policy of collecting data at the lowest level of identification possible, distributing identifiable data only in exceptional cases.

### Data analysis

For the analysis stage of data strategies, we reviewed and compared the different organisational approaches to data analysis in terms of who carries out the analysis, how they do so, and how the analysis is quality assured and communicated.

The main distinctions across the organisations in analysing data were the size and capacity of the organisation to carry out data analysis, the cost of having in-house analytical research competencies versus outsourcing, and benefits that are accrued to the organisation if analysis is done in-house (eg the accumulation of in-house expertise and streamlining the data strategy from beginning to end).

#### *Lessons for DG SANCO*

For the data-analysis stage of a data strategy, DG SANCO could take heed of the following points:

11. DG SANCO could evaluate its internal capability to carry out data analysis, identify where there are gaps in its knowledge, compare the costs of expanding its analytical capabilities relative to the cost of outsourcing / using external contractors, and measure the value added and the benefits of having in-house expertise in data analysis.
12. The merits of a centralised or decentralised approach to data analysis need to be assessed against DG SANCO's existing organisational structure. Good practice examples for either way of organising data analysis are EHO and DG Enterprise.
13. Good practice in quality assurance for data analysis involves transparency and communication of methodologies (principles, assumptions and methods) and systematic reviews of methodologies. OECD and PHAC provide standards that could be adopted by DG SANCO.

### Data use

In terms of data use, as the final element of our data-strategy framework, we found that there are four main users of data and analysis: the general public, government and public bodies/agencies, internal clients and academics.

These organisations use data for four main purposes: to inform policy; to monitor, review and evaluate; for academic use; and for internal use.

*Lessons for DG SANCO*

14. DG SANCO may want to consider data use and communication as part of an integrated data strategy and;
15. Identify the main users of its data-analysis products to ensure the wide and effective use and uptake of DG SANCO work
16. Understand how and for what purpose the end users use the analysis – eg by using SANCO internal workshops
17. Develop mechanisms to reach and involve users – tested examples include Scenario Logic and Scenario Gaming workshops.

**Revising the linear model**

Whilst in the early chapters we implicitly assume a linear model of data strategies – moving from identifying and prioritising data needs to collecting data, managing data, analysing data and using it – in the final chapter of this report we explicitly discuss some of the nonlinearities involved in data strategies.

More specifically, we argue that all (later) stages of a data strategy have bearing on the identification and prioritisation of data needs; that data management and analysis influence data collection; that data analysis influences data management; and that data use has some bearing on the analysis stage of data.

*Lessons for DG SANCO*

Several lessons for DG SANCO follow from our discussion in this chapter:

18. DG SANCO may want to structure its identification and prioritisation of data needs into three parts: i) identification of data needs; ii) feasibility; and iii) prioritisation. Policy analysts would be involved in part i), people responsible for collection and management in part ii) and senior management in part iii).
19. A formal system of workshops that allow people from the different stages to interact with one another – in particular those from data collection, management and analysis and data analysis and management – could be beneficial to DG SANCO.

## 1.1 **Evidence-based policymaking and the need for data**

The European Commission has gradually introduced a more evidence-based approach under the presidency of Jose Manuel Barroso. With it, the need for high-quality data and associated data management has increased. Indeed, as evidence-based policymaking gains in importance, data quality and management have come under greater scrutiny.<sup>1</sup> The need for high-quality data and data management has become apparent in impact assessments, which have increased in importance and number, and in the need to understand how European policy affects consumers and markets, which requires increased market monitoring.

## 1.2 **Challenges for DG SANCO**

This emphasis on evidence-based policymaking also holds a number of challenges for DG SANCO. There is a risk that poor data quality could undermine the findings of important evaluations and impact assessments, and indeed may lead to questions about the robustness of decisionmaking. This could lead to DG SANCO being less effective in initiating and promoting important advances in public health, animal health and welfare, food safety and consumer protection.

DG SANCO also has a particular concern about the use and management of high-quality data, given the aim of better monitoring of the internal market from a consumer perspective.

That said, much has already been done within DG SANCO to collect appropriate data (eg Consumer Markets Scoreboard, European Health Indicators). However, the level of data use varies between DG SANCO directorates. This is a result of different data needs and historical circumstances as well as each directorate's own capacity. Nonetheless, at least some of the types of data needed to meet the requirements of the impact-assessment process are fairly similar, regardless of the policy context. Therefore, DG SANCO has found it useful to look at having a more coordinated data strategy, in order to exchange knowledge and best practice better within the DG.

---

<sup>1</sup> For a discussion on the rise of evidence-based policymaking, see eg Sanderson, Ian (2002); as well as Rieper *et al.* (2009).

The overall challenge, then, is to improve data quality and management where necessary, practical and appropriate; and the vision is to have relevant, accurate, timely, accessible, compatible and coherent data available to decisionmakers to enable them to place evidence alongside other legitimate factors to support decisionmaking and inter-institutional and stakeholder engagement.

### 1.3 **DG SANCO's efforts to develop a comprehensive data strategy**

Against this background DG SANCO's Strategy and Analysis Unit, Unit 02, is working on designing and preparing a comprehensive 'Good Practice Model for Data Strategy', comprising all DG SANCO's policy areas. To develop this strategy DG SANCO initiated a project in two phases:

1. In a first **mapping phase**, DG SANCO intended to analyse the current state of data handling and strategising, within DG SANCO as well as outside it. For this purpose DG SANCO identified five work packages:

**WP 1 – External mapping of good practices.** A review of existing data practices in other relevant administrations and, if possible, the identification of 'good practices', adapted to the specific role and needs of DG SANCO.

**WP 2 – SANCO data needs.** When? Who? What for?

**WP 3 – SANCO data sources.** What data do DG SANCO already have? Mapping and reviewing available sources and data available in-house.

**WP 4 – Key partnerships.** The main relationships on which SANCO relies to get the data it needs have to be listed and studied in more detail to identify the nature of the links and the ways to improve cooperation with key partners such as the European Food Safety Agency (EFSA) or Eurostat.

**WP 5 – Data quality.** Develop the main criteria/indicators of quality of DG SANCO's data and explore common practices, for example on enforcement.

2. A **strategy phase** in which the findings of these work packages are synthesised and developed into a coherent and comprehensive strategy that can then be transferred into an action plan to be implemented at DG SANCO.

As part of this process, DG SANCO commissioned RAND Europe to conduct research into work package 1, the external mapping of data strategies.



## 1.4 Research question and research approach

The key objective of this research project is thus to identify and evaluate good practices of data strategies for policymaking and develop recommendations for which elements could be used to build a data strategy within DG SANCO.

In the context of this research and this report, data are defined as quantitative, statistical information. Data in the context of policymaking thus mean for example macroeconomic statistics such as unemployment rates, health information such as morbidity and mortality rates, information about enforcement activities such as number of inspections, and also survey results about consumer perceptions or consumption patterns.

To address this question, RAND Europe applied a comparative case-study approach. This builds on a common structure and template for individual case studies which allows the capture of a range of practices in a systematic way. The results are presented along the lines of common themes and issues, rather than as individual cases. This allows for a clearer presentation of the lessons that can be learned across all cases we looked at. Broken down into manageable tasks, the research approach of RAND Europe consisted of the eight project tasks listed in Table 1.1 below. Further methodological detail is provided in Appendix A.

**Table 1.1 Research approach**

<b>Task</b>	<b>Objective</b>
Task 1: Kick-off meeting	Setting up and piloting the research template and selecting case studies
Task 2: Selecting the case studies	Identifying comparators that are likely to offer practice relevant to DG SANCO
Task 3: Developing a comparative framework	Identifying a series of criteria that will allow us to identify areas of interest to DG SANCO
Task 4: Desk-based research	Population of template through document review.
Task 5: Mid-term meeting	Sharing findings with DG SANCO, identifying any gaps or difficult areas, drawing out emerging themes for analysis and identifying areas for further investigations
Task 6: Refining case studies	Refining the findings and filling any gaps
Task 7: Assessing good practice	Internal workshop to gain some consensus on what good practice is and formulate recommendations for DG SANCO
Task 8: Reporting	Draft interim and final reports
Task 9: Project management	To keep the project on time and budget and communicate effectively with DG SANCO

## 1.5 The case studies

This research builds to a very large extent on case studies. The selection of the case studies was conducted in close cooperation with the SANCO study team and the steering group for this project. The following criteria were considered:

- coverage of all of DG SANCO’s policy fields
- organisations that are similar to DG SANCO

- geographical spread
- the availability of data and information/research on initiatives of public-sector organisations
- evidence of interesting or innovative practice.

We selected a total of 13 organisations to conduct case-study research (listed in Table 1.2 below). This was a change to the original research design, which had foreseen a smaller number of cases but more in-depth scrutiny of a selection of cases. As the availability of supporting documentation as well as the willingness of officials to participate in research interviews was relatively poor, we increased the sample of case studies in order to gather an appropriate body of evidence.

**Table 1.2: Overview of case studies**

	<b>Comparator</b>	<b>Country</b>	<b>Policy area</b>
1	Department for Environment, Food and Rural Affairs (DEFRA)	UK	Animal health
2	DG Enterprise	EU	General
3	DG Environment	EU	General
4	DG Trade	EU	General
5	Direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF)	France	Consumer
6	European Health Observatory (EHO)	International organisation	Public health
7	Food Standards Agency (FSA)	UK	Consumer
8	Instituto Nacional del Consumo (INC)	Spain	Consumer
9	Konsumentverket, Swedish Agency for Consumer Affairs (SCA)	Sweden	Consumer
10	Organisation for Economic Cooperation and Development (OECD)	International organisation	General
11	Public Health Agency Canada (PHAC)	Canada	Public health
12	Regulatory Reform Group (joint organisation of the ministries of Finance and Economic Affairs) (RRG)	Netherlands	General
13	World Health Organisation (WHO)	International organisation	Public health

While the key findings from the case studies are integrated into the main body of the text, you will find textboxes throughout this report that refer to particularly interesting practices from the case studies.

## 1.6 This report

This final report presents the findings of the research. It is structured into nine main parts. The Introduction (Ch. 1) is followed by a brief review of the literature on data strategies and a description of our conceptual framework (Ch. 2). Chapters 3–8 then discuss the key elements of our framework in more depth, drawing mainly on our case studies. This is followed by a chapter in which we discuss some of the nonlinearities (and feedback loops) between the key elements of our framework (Ch. 9). The methodology for our literature review and case-study selection and the study/interview template are presented in Appendix A of this report.

## CHAPTER 2 **A conceptual framework of data strategies**

---

In this chapter we develop the conceptual framework for a data strategy. In order to inform our thinking, we conducted a review of the literature (see Appendix A for details).

Our review resulted in only very limited information and evidence about data strategies. Most of that information focused either on the data needs within the private sector – in particular on ensuring process learning, learning from customer feedback, and market intelligence – or on the technical aspects related to building a database (as emphasised in the literature on data warehousing, for example). Whilst this information was useful in terms of gaining a better understanding of what data are, how data needs can be identified and what issues there are regarding data management, it was not sufficient to provide a comprehensive basis for our discussion.

At the same time, there is a growing body of literature about the wider issues of knowledge and information management, which also touch upon questions of data management in the private sector (see eg Wiig 1997 for an introduction). The narrower remit of this research allowed no systematic inroads into this body of literature, and thus we used only selected, relevant texts to inform our thinking, in particular about the definition of data as well as the organisational consequences of knowledge and data management.

As a result of the focused research on data strategies in public-sector organisations, we have selected three topics that have been discussed in the literature (in the same context) and have explored possible implications for an overall data strategy arising from these issues. The three topics are:

- defining data
- identifying data needs
- exploring issues in data management.

In the following section, we briefly discuss the three topics in turn.

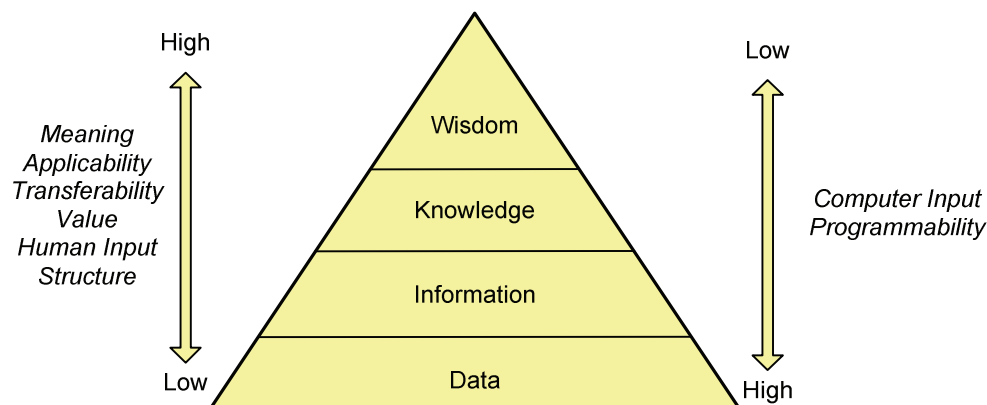
### 2.1 **Defining data**

The literature uses two broad approaches to define data. Firstly, what we call the analytical approach defines the term by contrasting it with information and knowledge. The second approach, which is more pragmatic, tries to circumvent some of the problems of the analytical approach by identifying typologies of data and thereby narrowing down the

meaning of the term. Both approaches provide interesting points for our thinking about data strategies.

### 2.1.1 The data, information and knowledge wisdom hierarchy

This study is interested in ‘data’ as such. Probably one of the most prominent ‘analytical’ approaches to defining data is to contrast them with three other concepts: information, knowledge and wisdom. Represented in the knowledge pyramid or the DIKW hierarchy, this definition was originally developed by Ackoff (1989) and refined, adapted and extended subsequently.<sup>2</sup> Figure 2.1 shows the commonly used pyramid, as well as the variables usually thought to change as you move between the levels of the pyramid.



**Figure 2.1: The DIKW hierarchy**

SOURCE: Rowley (2007)

Analysing the definitions used in recent textbooks about knowledge management and information systems, Rowley (2007) identified the following definitions for these four key elements:

1. **Data** are primarily discrete, objective facts about or descriptions of things, activities and transactions. Data themselves lack meaning, value and structure and also context and interpretation. Data are syntactic entities, they are patterns with no meaning and they just form the input into an interpretation process (Aamodt and Nygård 1995).
2. **Information** in turn is often defined in relation to data. As such, information is structured, interpreted data, which makes it relevant in a specific context or for a specific purpose. Information is data that have been given meaning by putting them into context. One way to think about data and information is in terms of an encrypted message for which you need a key in order to make sense of it. The encrypted message represents data, and the key the ability to interpret these data and so to transform them into information.
3. **Knowledge** is learned information. Knowledge is data and or/information that have been organised and processed to convey understanding, experience, accumulated learning and expertise. While data are a property of things, knowledge resides with people – it is information incorporated in an agent’s

<sup>2</sup> See Rowley (2007) for a comprehensive overview.

reasoning resources. As an example, imagine receiving an encrypted message for which you possess the key and from which you extract the following information: ‘The cat is tired.’ Unless you possess enough contextual background knowledge to realise that the message refers to something more than an exhausted cat – possibly a Mafia boss – you may not be in a position to react in an adaptive way (Boisot and Canals 2003).

4. While the definition of knowledge is less stringent and clear cut than those for data and information, definitions of **wisdom** are even less so. A common theme in those definitions is that wisdom includes the ability to transfer knowledge and experience between domains, to exercise judgement and to include values.

In addition to this classification into categories, the literature also discusses certain attributes or characteristics that change between the different levels of hierarchy:

1. The **structure and order** increases as you move from data to information and beyond.
2. The **human input increases**, at the expense of computer input, the closer you get to knowledge and wisdom.
3. Data are thought of as having little or no **meaning** and value in themselves. Meaning and value are only generated by contextualising and structuring them as you move up the hierarchy.
4. **Applicability, actionability and transferability** are, finally, other categories of usefulness in organisational terms that change along the hierarchy.

There are three main implications flowing from this with regard to data strategies.

According to the definition, data as such are meaningless. Only in the context of existing knowledge can they be transformed into information and later on into knowledge. This suggests that a data strategy should be selective and should be based on what we know already. Consequently, a data strategy should only include what we are able to make sense of.

1. A data strategy is helpful only if it is selective in terms of the data collected, stored, analysed and used.

The flip side of the argument above is that, as the stock of knowledge increases (and also the ability to elaborate on existing information and knowledge), so does the ability to make sense of data. Thus the relationship between data, information and knowledge is not static, but dynamic. As our stock of knowledge increases, so does what we can make sense of in terms of data and, as a consequence, so does what we want to collect. This means that a data strategy will need to be flexible.

2. A data strategy is a dynamic process, which should be adjusted to new insights and thinking.

As an example, only with the discovery that nuclear radiation has severe long-term health implications has it become meaningful and important to measure radiation in different areas. Similarly, with the insight that ‘output over number of workers’ is not a good

measure for productivity, we now collect output over total hours worked in order to consider part-time workers.

Finally, the hierarchy also indicates that a focus on data and the development of a data strategy must be seen within the wider context of an information or knowledge strategy – so a concern about data should be linked to wider questions of what knowledge an organisation requires.

3. A meaningful data strategy will need to engage with the wider questions of what information, and ultimately what knowledge, an organisation requires to be successful.

### 2.1.2 Typologies and categories of data

One of the main problems of the analytical definition of data is that it is very abstract. As a consequence, it appears that ‘data’ is either used too ambiguously – as a collection place for multiple significations – or too narrowly.

One way to deal with definitional problems is to take a pragmatic approach and survey the landscape in order to identify different typologies of data. The resulting definition may not be fully satisfactory. That is, the boundaries between these typologies may be indistinct and such an approach may not satisfy those determined to establish a single correct meaning of ‘information’. But if the principal types of data can be identified then, as Buckland (1991) argues, some progress is made.

There are a number of different typologies of data. The most prominent ones include typologies by purpose, subject area and importance. As an example, Gustafson *et al.* (1992) identify three purposes of data strategies, to support:

- operational decisions (eg through monitoring)
- managerial decisions (eg through evaluation)
- strategic decisions (eg in policy formulation).

In internal documents DG SANCO comes up with a (slightly) different typology of data by purpose, differentiating data to:

- help drive and promote change by actors outside the Commission
- help assess and develop own policies and proposals
- monitor implementation of action by Member States or at Community level
- signal potential future problems or threats.

In terms of subject area, DG SANCO discusses (*inter alia*) the following approaches to identify subject areas:

- controlled vocabularies (ie closed list of vocabularies)
- taxonomies (which arrange terms into hierarchies)
- faceted classification (which works by identifying facets into which terms are divided).

Finally, Browne and Ramesh (2002) provide an example of a data typology that is based on importance. They divide data into:

- essential data
- rapid collection items
- periodic collection items
- low priority items.<sup>3</sup>

4. Probably the most important implication flowing from the pragmatic definition of data is that it opens up the possibility of multiple data strategies.

Because there is more than one type of data, it is conceivable that there are also different strategies to deal with these data in an efficient and effective way. As an example, ‘essential data’ may have to be selected with a higher frequency than low-priority items – while low-priority items may have to be reviewed more often than essential data. A pragmatic, meaningful distinction of data types will thus be an important starting point for a data strategy.

## 2.2 Identifying data needs

The literature about data needs is much dispersed. There are two broad themes, however, that can be condensed from the literature: identifying drivers of data needs and obstacles to identifying data needs. The main drivers of data needs that come out of the literature, which include Goodhue *et al.* (1992) and Nemati *et al.* (2002), are as follows:

- an organisation’s present strategic agenda and the corresponding drivers, processes and outcomes
- a future strategic agenda (with drivers, processes and outcomes)
- an organisation’s operational objectives.

Further drivers are (Nemati *et al.* 2002):

- context – such as the economic, social and environmental situation
- public policy action – which comprises interventions by public authorities.

Earlier we argued that data needs are likely to change as the stock of knowledge changes, and so certain pieces of information start making sense, whereas others stop being informative. One idea following from the identification of drivers of data needs is that data needs also change with changes in strategy (present and future), changes in context and public policy action. As a consequence, a data strategy is likely to have to go through cycles of revision.

---

<sup>3</sup> The classification is based on a survey of stakeholders in which they assign a score between 1 and 3 to each data item. The data items are then ordered, based on the average rating and the difference between the lowest and highest ratings.

The idea is that when the range (between the lowest and highest ratings) is small, that suggests that there is a consensus about the importance of the data item across different issues. When the range is high, that suggests that for some issues the data item is important and for others it is not as important. A data item is ‘essential’ if it is important for all issues, ‘rapid collection’ if it is important to most issues, and so on.

5. An interesting implication flowing from these drivers is that again a data strategy is a dynamic concept and that a sustainable strategy will need to be able to accommodate changes in data needs and the context of data.

The second relevant theme from the literature on identifying data needs was the obstacles to identifying data needs. Those include the following:

- Cognitive limitations to the identification of what is really needed. This includes the problem that people are sometimes not aware of how they make decisions and what information they use in those decisions. Another example is that decisionmakers sometimes do not recall in detail occasions on which they needed more information.
- Problems of feasibility in collecting, managing, analysing and using data in terms of costs. This relates to resources, burdens on third parties when it comes to data collection and political costs.

6. An implication flowing from the barriers to addressing data needs is that it might be useful to use special methods to overcome cognitive limitations and identify data needs.

Examples of methods that have been suggested in the literature to overcome the cognitive limitations listed above include what-if analyses, scenario response tasks, devil's advocacy, decision maps and note boards. (see Browne and Ramesh 2002 for an overview).

### 2.2.1 Exploring issues in data management

The final broad topic discussed in the literature related to data strategies concerns issues in data management. In particular, the objectives of data management are frequently discussed. Some examples are listed below:

The US Department of Defense identifies the following objectives (DoD 2003):

- Make data visible – so that users can discover the existence of data assets through catalogues, registries and other search services.
- Make data accessible – so that descriptive information about the data (metadata) is provided and the data are stored in a way that allows users to access them.
- Institutionalise data management – to incorporate data approaches into an organisation's processes and practices. This can be achieved through sustained leadership, advocacy, training or education.
- Make data understandable – so that users can comprehend the data, both structurally and semantically, and readily determine how the data may be used for their specific needs.
- Enable data to be trusted – so that users can determine and assess the authority of the source.
- Support data interoperability – for example by providing metadata that allow mediation or translation of data between interfaces as needed.
- Be responsive to user needs – to allow that perspectives of data consumers and producers to be incorporated into data approaches via continual feedback.



Another example of objectives of data management is provided by the OECD (Samuelson and Thygesen, 2004). The objectives of its Statistical Information System (SIS) are to:

- ensure efficiency of data and metadata collection, validation, processing, storage and dissemination
- ensure data quality, eliminating errors and incoherencies and shortening statistical publication cycles
- enhance the accessibility and visibility of the OECD's statistical outputs.

A final example is the US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration (SAMHSA, 2007). It identifies the following data-management objectives:

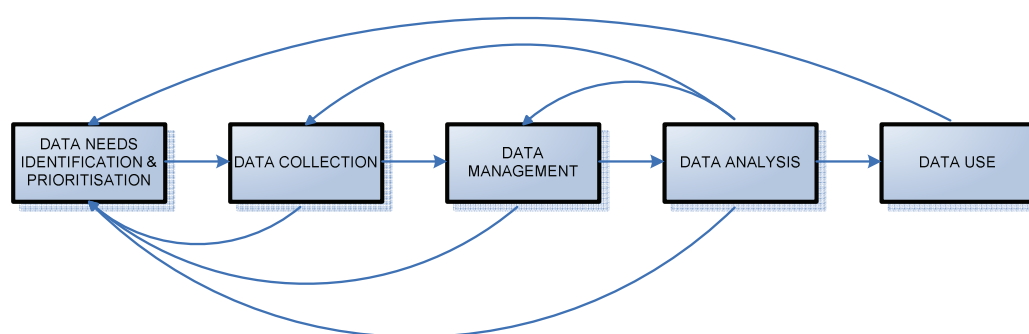
- periodically assess data needs and data quality
- improve ease of access and user interface, while ensuring confidentiality
- ensure analysis of data is responsive to clinical, programmatic and policy needs
- increase dissemination of data and analysis.

Again there are a number of implications flowing from these findings. One of the points we have made previously in the context of defining data (arguing that only data that can be made sense of should be selected). The main point here is that a data strategy should be selective because the collection, storing, analysis and use of data is costly and should therefore be weighted against its benefits.

7. The first implication is that a data strategy must be selective in order to manage the costs of collecting, storing, analysing and using data.
8. The second implication is that a good data strategy is unlikely to be linear (moving from identifying data needs to data use), but allows for feedback loops (eg from consumers and producers).
9. A third major implication is that quality assurance and communications / user engagement are important parts of a data strategy.
10. Finally, this strand of the literature points to the importance of embedding a data strategy in the wider organisations by making it a visible part of the organisation, building capacity and triggering cultural change in an organisation.

### 2.3 **A data-strategy framework**

On the basis of this discussion of the main themes emerging from the literature, we developed a simple data-strategy framework. In this section we outline its key stages. The main purpose of this framework for the research progress was to provide a basis for the data collection for our case studies and to structure the cross-case analysis in the next chapter. It allows us for example to capture information for different stages of a data strategy – such as the identification of data needs – consistently across the case studies and to identify best practice in different contexts.



**Figure 2.2: Data-strategy framework**

We have divided the framework into five key elements, which represent an idealised, sequential process of developing a data strategy from scratch. The framework is illustrated graphically in Figure 2.2 above. The key stages of this framework are as follows:

- 1) **Data-needs identification and prioritisation.** The first stage of the process is to define what data are needed, and to prioritise the data that should be collected at a later stage. The definition of data needs will include considerations about the policy objectives, future uses of data and data availability.
- 2) **Data collection.** Once data needs and priorities are established, an organisation needs to consider how the desired data can be collected. This stage includes questions about data sources, data gathering and frequency of data collection, amongst others.
- 3) **Data management.** Data collected will need to be managed within the organisation. This may just include storing it in databases, but could also involve stages of data processing such as harmonising and standardisation.
- 4) **Data analysis.** In the subsequent stages, data are structured and analysed to become useful information that can be used in a policy context. In the case studies, we were particularly interested in what type of analysis was performed in the organisations.
- 5) **Data use.** Closely related to the analysis of data is the use of the data. This stage covers the question of what is done with the data: is it produced for internal processes, for the general public or for selected stakeholders?

In the following chapters we discuss each of the key elements of our conceptual framework in more depth. The discussion will be based mainly on the case studies. At the end of each chapter we draw out some key lessons for DG SANCO in their effort to develop a data strategy.



**Figure 3.1: Identification and prioritisation as part of a comprehensive data strategy**

In this chapter we discuss the first element of our data-strategy framework: the definition and prioritisation of data needs. From the definition of data we inferred that not all information is relevant. Similarly, our discussion of data needs suggested that data are costly (in terms of collection, management, analysis and use) and should be limited for that reason. But how do we decide which data to collect and which data not to collect? In this section we discuss how different organisations identify their data needs and how they prioritise these needs. We complement/contrast our findings from the case studies with those from the literature whenever possible.

### 3.1 **Defining data needs: centralised or decentralised?**

The case studies suggest a number of different approaches to identifying data needs. One broad distinction we can make is between organisations that identify data needs in a centralised way – say by a strategy and/or statistics unit – and organisations in which this is done in a decentralised way, for example by several policy units. The majority of the cases we reviewed rely on a decentralised mechanism. Examples include the OECD and FSA. One example where a centralised mechanism is used to identify data needs is RRG.

One advantage of organising the identification of data needs in a decentralised way is that, because the expertise and knowledge of a particular subject sits with different policy units, as discussed earlier, these units are often best suited to identify what data are needed in the sense of what data can be made most sense of. Having said that, it is important to note that methodological knowledge can be as important for identifying data needs as subject-related expertise. This suggests that if this expertise sits outside the individual policy units, it should be consulted to decentralise efforts to identify data needs, in order to make the most of these efforts.

On the other hand, the case may be made for organising the identification of data needs centrally, with senior management involvement to ensure high-level buy-in to empirical research from the outset and to ensure that the data collected are relevant and can help

address the key policy problems. Combining a decentralised way of identifying data needs in DG SANCO with a top-down prioritisation mechanism could be one way to combine the advantages of both approaches – making use of decentralised expertise and generating high-level buy-in at the same time. We shall discuss this option in more detail below.

#### BOX 1

**Example 1) Centralised identification of data needs.** Most of the organisations analysed within the study used either a decentralised or hybrid approach to identify data needs. **Dutch Regulatory Reform Group (RRG)** represents an example where the use of a centralised process of data-need identification is more or less driven by the size and the remit of the organisation. RRG's mission is centred on the provision of information about the business perception of regulation and administrative burden.

The mission of the group is therefore focused on outside actors and does not require input from other parts of the organisation. Furthermore, RRG is a working group of the Dutch ministries of Economic Affairs and Finance and therefore a decentralised actor interlinking two bigger organisations. The extent to which a further decentralisation of the identification process provides additional benefits (eg economies of scale, specialised knowledge from different ministerial units) is therefore limited.

**Example 2) Decentralised identification of data needs.** Several examples for a complete decentralised process could be identified within the study. **The European Health Observatory (EHO)** applies a bottom-up approach by consulting Member States to meet their data requirements. They assess their own data needs and EHO collects the information for that. EHO provides an overview of the needs, and on the basis of this information Member States make their decision about what kind of research EHO should conduct.

Furthermore, consultation takes place with policymakers who have first-hand experience of what kind of data are needed. Additional information is gathered within consultation processes involving main actors in the EHO partnership, for example the World Bank and WHO.

**Canada's Public Health Agency (PHAC)**'s decentralised approach involves periodical static research tasks, combined with the flexibility to adapt to inevitable changes in data needs. Many data needs were identified and prioritised years ago; others have been identified recently by the programme areas and their stakeholders. PHAC has also started to develop assessment criteria to evaluate existing surveillance systems and to identify gaps and needs. For example PHAC uses a template with a list of criteria for identifying data needs. These criteria include considerations such as the size of the target population affected, evidence of economic or other impacts (ie impacts on family, community, health care system and other community support systems), whether the data would link with PHAC's priority areas, identifying how often the data would need to be collected and why, and stating the sample size needed and the target population. We envisage that some of these criteria could apply to DG SANCO's work, in particular to its public health area. For the development of large public health surveys, an agency-wide consultation process takes place that represents the first element of a decentralised approach. Information becomes streamlined and recommendations for main decisions of the senior management are prepared.

As a second element, PHAC also adapts its data need to constant changes of preference in order to address the current and future information needs of its main users and stakeholders. Adjustment to needs that evolve over time should be established through several feedback mechanisms: liaison and consultation arrangements with federal and provincial officials, ad hoc consultation with advocacy groups, studies for assessing the use of current databases and the potential for new uses/users (see PHAC Data Quality Framework, definitions of the dimension 'relevance' and the characteristic 'adaptability').

**WHO Regional Office for Europe** and the **UK Food Standards Agency (FSA)** provide examples of the more or less internal decentralised identification process of data needs. Within FSA policy teams mainly drive the identification process, while the central strategic unit also contributes to the decisionmaking process. The WHO Regional Office for Europe involves managers, technical and senior staff when preparing the biannual programmes. Under certain conditions this process may be led by the respective programme areas due to the need for special knowledge input.

**Example 3) Consultation.** A special example of a relatively wide consultation process is that of the **UK Department for Environment, Food and Rural Affairs (DEFRA)**. When developing a long-term data-need strategy, which finally led to the development of the RADAR system, DEFRA established a wide consultation process in which practically everybody could take part (academia, delivery bodies, veterinary surgeons, etc).

On the basis of these inputs, a six-year strategy plan was developed indicating on which data objectives DEFRA should focus. In 2009 a revision of the strategy was started in order to assess the need for a change of goal. Alongside the involvement of stakeholders in the long-term planning, changes of data needs are assessed on an ad hoc basis. DEFRA researchers get short notice of developments through their network relations with research institutions, delivery bodies and so on.

### 3.2 Systematic and ad hoc definition of data needs

Another way to think about the identification of data needs coming through from our case studies is whether data needs are identified systematically (every six months, say) or on an ad hoc basis (whenever a need arises). From interviews with key staff in the organisations we reviewed, we learned that whether data needs are identified systematically or on an ad hoc basis depends to some extent on the nature of the work done by the corresponding organisations.

For some organisations for example, it is possible that (in certain situations) data needs occur with no warning that would permit the identification of these needs beforehand in a systematic way. One example is that of impact assessments at DG Enterprise. Even though impact assessments do not generally occur overnight (but are preceded by a rather lengthy process of preparation), the preparation period is typically too short – and data needs too specific – to identify data needs in advance.

Having said that, in most of the cases we have studied there is an attempt to identify data needs systematically whenever possible. One reason for this is, as has been pointed out by one of our interviewees, that a systematic identification of data needs (in regular meetings, say) allows the collection of time series of data (not just single data points) while allowing

the fact that data needs may change over time – as discussed earlier – to be taken into account. Another reason which came through in our literature review is that a systematic identification of data needs will allow for a (more) comprehensive identification of data needs.

The underlying reasoning of the latter point is that often it is difficult to identify data needs comprehensively on an ad hoc basis. For example people, even experts, are often not aware of how they make decisions and what data they use in those decisions. In other words, people think but often do not know how they think. Gustafson *et al.* (1992) argue: ‘The process of thinking for the most part is an unconscious process. Thus, when you ask someone how they made their decision, they may list a number of pieces of information, they claim they paid attention to but in reality they may have made their mind up based on a much smaller set of information.’

This suggests that, even though it is not always possible, organisations have a reason to try to use a systematic approach to identify their data needs. As discussed, identifying data needs at regular intervals allows for the construction of time-series data while taking into account that data needs may change over time. In addition, a systematic approach has the advantage of allowing the identification of data needs (more) comprehensively. The reason for the latter is that it allows the use of proven methods to overcome some of the cognitive limitations that often make it difficult to identify data needs. A list of such methodologies is provided in Box 2.

#### **BOX 2**

**Example 1) Informal ways of data-need identification.** **DG Trade** identifies new data needs via unit meetings and follows up increasing requests of users on particular subjects. Statistical teams assess the possibility of fulfilling user needs and compiling new files, tables and databases where appropriate.

The **European Health Observatory (EHO)** also applies a more a less informal process when collecting feedback from Member States and key stakeholders (policymakers and partnership members) through different organisational and network channels.

#### **Example 2) Formal ways of data-need identification.**

The **Spanish Consumer Institute (INC)** established consultation processes on the basis of regular meetings between representatives of INC and regional governments in order to discuss developments in data needs.

The **WHO Regional Office for Europe** uses a two-way system consisting of established working groups, ad hoc exchanges with country experts and consultations with country representatives during the working process in order to channel and discuss information about data needs and the feasibility of research efforts in terms of budget planning and administrative burden for member countries.

### **3.3 Prioritising data needs: not at all, bottom up or top-down?**

When it comes to the prioritisation of data needs, we find several different approaches coming out from our case studies. The first distinction we can make is between

organisations where no prioritisation is done, organisations where prioritisation is done on a rather informal basis, and organisations where data needs are prioritised in a formal way. (Examples may be found in Box 3.)

The differences in the degree of formality of prioritisation may be explained to some extent by the difference in the nature of the work done at these organisations. Some have a very narrow remit and their data priorities follow straight from their mission. Other organisations have a broad remit, and so for them, because they are bound by resource restrictions, a prioritisation of data needs is important.

Which data are essential and need to be collected? Which data would it be nice to have but not essential? What data are of low priority – and should be collected only if resources allow it after all data with higher priority have been collected? The organisations we reviewed that do prioritise data needs may be divided into those that use a top-down approach and those that use a bottom-up mechanism to do that. By ‘top-down approach’ we mean one in which a centralised actor identifies needs centrally derived from a small set of policy objectives ‘Bottom-up’, in contrast, defines an approach in which decentralised lower-level actors define their needs independently, and the organisation-wide needs are then identified in a prioritisation exercise.

### BOX 3

**No data prioritisation.** The **Spanish Consumer Institute (INC)** does not prioritise data as such because data are collected on a day-to-day basis by the regional governments. Once a year data are transmitted from the regional level to INC, where further analysis is conducted. Nevertheless, this process is formalised through bilateral agreements between INC and regional level.

The **French General Directorate for Competition Policy, Consumer Affairs and Fraud Control (DGCCRF)** has a very broad policy and regards all data collected as of prime importance. RRG distinguishes between two forms of data without prioritising one of them: survey data and qualitative data collected via telephone interviews based on a smaller sample of businesses. Hence, data priority follows straight from RRG’s mission.

**Formal data prioritisation – top-down approach.** In the case of the **Swedish Agency for Consumer Affairs (SCA)**, the municipalities as main data collectors are legally obliged to provide SCA with the required data. The internal strategy and analysis unit decides on the type of data needed in order to respond to commissions from the central government.

**Formal data prioritisation – bottom-up approach.** The **UK Department for Environment, Food and Rural Affairs’ (DEFRA)** Prioritisation Project aims to build an evidence base of information about animal diseases and health welfare issues in the UK, and was established to meet the prioritisation goal of the Veterinary Surveillance Strategy developed in 2003 on the basis of an large consultations process between DEFRA and stakeholders.

The scope of the project was extended in 2004, and now covers all aspects of DEFRA’s policies related to animal diseases, health and welfare issues. Since information is also collected ad hoc by researchers through exchange with external actors, the approach may be defined as being mainly bottom up.

One very clear example of a top-down approach was suggested by Gustafson *et al.* (1992). In the organisation they researched, leaders are asked to rate the importance of each information item previously identified. They are asked: 'For the issue X, assign a score between 1 and 3 to each information item, reflecting the need for the information. A score of 1 means the information has low priority, a score of 2 means that the information is nice to have but not essential, and a score of 3 means that the item is essential for addressing the issue.' When finished, the data items rated highest across all issues are considered most relevant to the organisation's needs.

We find that most of the organisations which use a formal approach to prioritising data needs rely on a top-down approach. The reasons for this include, as was indicated in several interviews, the possibility of assuring the quality of data needs (which have been identified mostly in a decentralised way). As mentioned earlier, organising the identification of data needs in a decentralised way brings, along with several advantages, a disadvantage. This is that it does not automatically ensure the buy-in of senior management nor does it ensure that the data will allow the key questions faced by decisionmakers to be addressed.

Allowing for a top-down prioritisation may be one way of dealing with this problem (by giving senior management staff the possibility of double-checking and assuring the quality of the identification of data needs). In addition, a top-down prioritisation may help to consolidate data requests from different sources – and so reduce the danger of unnecessary duplication.

### 3.4 **Summary/Conclusion**

In this section we reviewed the main approaches to identifying data needs and prioritisation. We distinguished between decentralised and centralised approaches to data identification and discussed the pros and cons of a systematic approach to identifying data needs versus an ad hoc approach.

When it comes to the prioritisation of data needs, we found that organisations differ in terms of the degree of formality with which this is done. Some do not prioritise their identified data needs at all; others do this in a very informal way; still others use a formal process. Formal approaches, in turn, may be categorised into top-down and bottom-up approaches.

Overall we found that there is some argument for using a systematic approach to identify data needs (and so to ensure up-to-date time-series data). In part a reflection of the remit of the organisations we reviewed, the majority of cases also preferred a decentralised data-needs identification mechanism to a centralised one. Because data needs are identified in a decentralised way, our interviews provided arguments to complement this with a formal, top-down prioritisation mechanism.



### 3.5 **Lessons for DG SANCO**

The main lessons for DG SANCO regarding the identification of data needs which follow from our discussion include the following:

- Data needs should – where possible – be identified in a systematic way, using appropriate methods to overcome the cognitive limitations that can often act as a barrier to the identification of data needs.
- To the extent that subject-related and methodological expertise sits with individual policy units, a decentralised system of identifying data needs is generally preferable to a centralised one.
- A decentralised system of identifying data needs can be complemented with a top-down way of prioritising data needs, which ensures high-level buy-in to empirical work. At the same time, it allows the collection of the data needed to answer the most pressing policy questions.





**Figure 4.1: Data collection as part of a comprehensive data strategy**

Once data needs are identified and prioritised, the next challenge is to collect the corresponding data. This, however, raises a number of questions. Who should be responsible for the data collection? How often should the data be collected? What sources should be used for the data collection? We describe in this chapter how the organisations we reviewed approach these (and other) questions. As before, we complement/contrast our findings from the case studies with those from the literature whenever possible.

#### 4.1 **External vs. internal data collection**

When it comes to data collection, the first question is who should collect the data. One broad differentiation that comes through in the case studies is between organisations that handle data collection internally and those that outsource data collection to external contractors. Some organisations are best described as hybrid. They rely on downstream organisations and/or collaboration with other organisations (such as statistical offices) for their data collection. Box 4 provides examples for each type of organisation.

##### **BOX 4**

**Internal data collection.** The **French General Directorate for Competition Policy, Consumer Affairs and Fraud Control's (DGCCRF)** local officers who are in charge of processing the data and entering them into the central database system are also responsible for collecting the data. There is no protocol that guides data collection efforts.

The **Spanish Consumer Agency (INC)** collects data for household surveys on its own, although data are mainly collected by the autonomous regional governments.

**External data collection.** The **Dutch Regulatory Reform Group (RRG)** uses contractors to collect data, and different, decentralised teams within RRG analyse the data. The process of administering surveys and interviews is subcontracted to external companies.

**Hybrid data collection.** The **UK Department for Environment, Food and Rural Affairs (DEFRA)** collaborates with a wide range of organisations which participate in common efforts to collect data (eg the Animal Health Agency, Veterinary Laboratory Agency, border inspection authorities and local authorities).

In the case of Animal Health and Welfare Delivery (AHW), the landscape has become too complex (see Eves Review report (Eves 2006) and the DEFRA Action Plan (DEFRA 2008)) and the anticipated need for reform may have an impact on the way data are collected by DEFRA in the future.

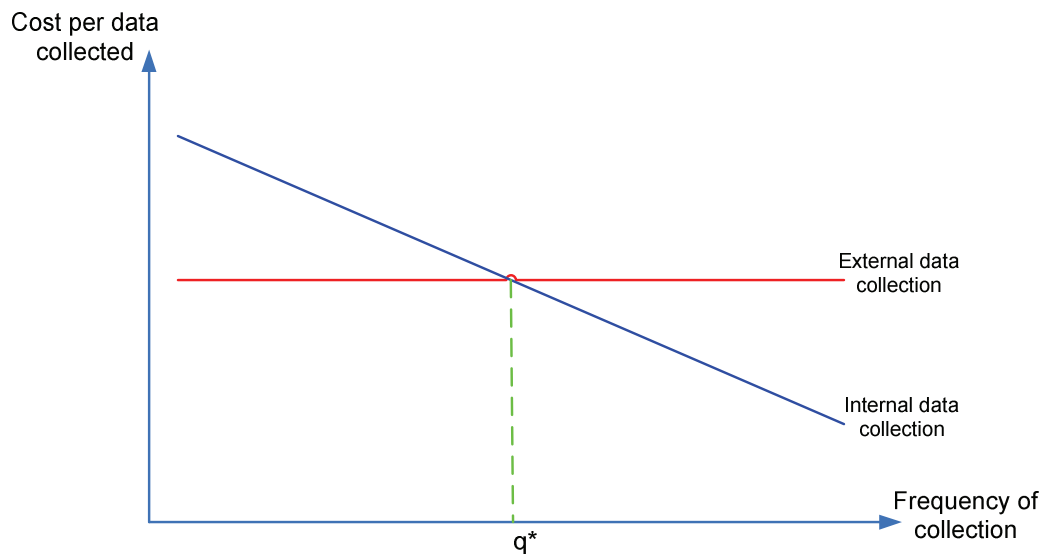
**DG Enterprise and DG Environment** use a similar set-up for collecting data. In both DGs, policy units are responsible for their sector-specific data-collection efforts and analysis, but they collect very few data on their own.

In the case of DG Enterprise, data are collected by Member States and used internally. Additional data come from Eurostat and OECD.

DG Environment relies mainly on data from a downstream organisation, the European Environment Agency (apart from data from Member States and Eurostat). Unlike DG Enterprise, policy units do also collect additional data, for example for their own impact assessments. We find that most organisations (of the ones we reviewed) use a hybrid form of data collection (that is, they rely on downstream organisations and/or collaboration with other organisations). Because collaboration may not always be possible, some organisations rely on a combination of a hybrid mechanism and internal handling of data for which no collaboration is possible. Others handle data collection exclusively internally. Only a few organisations, such as FSA and RRG, use contractors for their data collection.

One way to make sense of this finding is in terms of costs. As mentioned earlier (in the literature section) and as pointed out by several interviewees, data collection comes with a cost – for the organisation collecting the data as well as for the respondents who have to spend time in providing the data. A collaborative approach allows the costs of data collection to be spread over several organisations and the burden on respondents (who have to answer the same or similar questions only once) to be reduced. This explains, at least in part, why a collaborative approach is often the preferred option.

For many organisations, however, data needs are too specific to share one data collection across several organisations. In these cases the other modes of collecting data – which are handling data collection internally, through a downstream agency, or externally (through contractors) – are the main modes. One interviewee suggested that which mode to go for (still) depends mainly on costs. Because of the high costs of establishing expertise to collect data internally – and/or in an agency – these modes of data collection tend to be economical only if data have to be collected relatively frequently. A simple graphical representation of the idea is provided in Figure 4.2.



**Figure 4.2: Trade-off internal/external data collection**

To the extent that different types of data – such as qualitative data and quantitative data, microlevel data and macrolevel data – require different skill sets and accumulating these skills is costly, we should add two caveats to this view: i) collecting data internally or by means of an agency is likely to be economical only if *similar types* of data have to be collected relatively frequently; ii) to the extent that the demand for certain types of data within an organisation can be higher than for others, the economical solution may be to collect data in high demand internally, or by means of an agency, and data in lower demand externally.

Taking into account the practical difficulties and transaction costs of commissioning external data collection might however undermine this argument.<sup>4</sup> The more specific and the more complex ad hoc, non-recurring data requests are, the more time consuming it will be to communicate the demands to a third party effectively, while it tends to be relatively easy to contract out the gathering of standard information.

Another factor influencing the decision about whether to handle data collection internally (or delegate it to an agency) or externally is the question of timeliness. One interviewee reported that data collection in their organisation is often handled internally because the time span between the availability of data and the event or phenomenon described has to be very short. To the extent that this requirement often goes hand in hand with short planning schedules, this case favours having more flexible data-collection-related expertise in-house, and also having in-house the ability to respond to unexpected data needs in a timely manner.

---

<sup>4</sup> The authors are grateful to David Brackfield for pointing them towards this argument during the discussion of this report in a workshop organised by DG SANCO.

## 4.2 Primary and secondary data

Another question that arises when it comes to the collection of data is where to collect the data from. The main categories coming from our case studies are data collected directly from individuals, businesses and so on; data collected from these individuals, businesses and so on through an agent; and data collected from organisations that have already collected the data for their own administrative and/or statistical or research purposes from the primary data providers. The first category is typically referred to as primary data collection, the second as a hybrid form of data collection and the third as secondary data collection.

Most of the organisations we have studied use an agent – such as Member States or statistical offices – to collect data for them. Many use primary data collection, sometimes in combination with agents. Only a few organisations rely on data that have already been collected by other organisations (secondary data collection). That few organisations rely on secondary data collection may mainly reflect the fact that data needs are often very specific, so that there is little room to build upon the data collection efforts of others.

In fact, organisations such as PHAC, OECD and FSA prescribe that once data needs have been identified, a check has to be made on whether relevant data exist elsewhere. The rationale is that secondary data collection is typically cheaper and less risky (eg in terms of not getting the right response rate) and imposes a lower burden on respondents. A limitation of secondary data analysis – and possibly a further reason why it is not used more often – is that an organisation has little influence on whether legal (eg privacy), ethical and policy requirements are met.

## 4.3 Assuring quality in data collection

A related issue is that of quality assurance. Most organisations have an explicit quality framework in place. Traditionally quality at the data-collection stage has been defined in terms of data accuracy (looking at coverage, sampling, non-response, response, etc).<sup>5</sup> Today most of the organisations we reviewed define quality in a much broader way. Additional dimensions include: relevance (Do the data address their purpose?), timeliness (Are the data up to date?), coherence (Are the data logically connected and mutually consistent?) and trustworthiness (Are the data to be trusted?). Some examples of quality frameworks are listed in Box 5.

### BOX 5

**Example 1) Quality framework:** The UK Department for Environment, Food and Rural Affairs (DEFRA) uses its so-called RADAR Data Quality Framework to frame how data quality will be determined, described and reported within its veterinary surveillance system. The basic approach is that users get sufficient information to enable them to judge for themselves whether or not the data are of sufficient quality for their intended use.

---

<sup>5</sup> See OECD, 2003.

The framework is based on the quality categories defined for the European Statistical System and the Office of National Statistics. In the first phase RADAR aims at 'indicating' rather than 'measuring' quality. Data quality is assessed against six 'categories': relevance, accuracy & precision, timeliness & punctuality, accessibility & clarity, comparability, and coherence. (RADAR is to be developed and released in phases between 2005 and 2013.)

**Example 2) Quality framework.** The **Organisation for Economic Cooperation and Development's (OECD)** Quality Framework and Guidelines for OECD Statistical Activities draws on several concepts of statistical data quality (eg the International Monetary Fund – IMF, Eurostat, Statistics Canada and other national statistical offices) and identifies seven dimensions: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence.

OECD tried to define quality in terms of user needs rather than simply equating it with accuracy. Furthermore, the framework encompasses not only guidance on how to assess the quality of statistical information, but also extensive definitions of statistical managerial and technical processes.

It is based on four elements: a definition of quality and its dimensions, an outline of a procedure for assuring the quality of proposed new statistical activities, a procedure for evaluating the quality of existing statistical activities on a regular basis, and a set of broad principles for statistical activities covering all phases of the production process.

In terms of how quality-assurance processes are implemented at the data-collection stage, we find that most organisations use a peer-review process at the planning stage. Some draw on in-house expertise to do so; others use external specialists. Often the people responsible for the data analysis are asked to review the data collection. This helps to ensure that the data collected are useful to the people who work with them – while imposing the least cost on the organisation and the lightest burden on respondents. People analysing the data are also well positioned to judge the necessary level of identification. For legal/privacy issues the general rule is to keep this level as low as possible.

Other approaches to operationalise quality assurance that came through in our case studies (which can be complemented with the peer-review process) are to use pilots and transparency. The idea of pilots is to collect a fraction of the data up front. That can then be checked for response and non-response errors before the whole sample is selected, allowing adjustment of the data-collection approach if problems occur. The pilots can be reviewed by the people responsible for the data collection, external reviewers, or the people responsible for the data analysis. Being transparent about the approach with which data have been collected is the approach to ensure quality that is mentioned most often.

Being transparent provides an incentive to make data collection as strong as possible – in order to avoid criticism. In addition, transparency addresses the issue of data quality by making it more trustworthy. Examples of how quality assurance is operationalised at some of the organisations we reviewed is provided in Box 6.

#### **BOX 6**

**Quality Assurance 1.** The seven dimensions set out in **DEFRA's RADAR Data Quality Framework** are operationalised by 'quality indicators'. The categories and indicators are used to set up a quality statement against which pertinent data quality

indicators are reported. The templates (one template for each category and the corresponding indicators) are meant to be used as a guide – that is, not every indicator must be reported for each data set.

**Quality Assurance 2.** OECD quality checklists for assuring the quality of new statistical activities and for evaluating the quality of existing statistical activities (as well as the broad principles covering the whole process) represent the main pillars of the quality-assurance system.

The checklists for existing activities cover seven steps and the broad principles (captured by a separate checklist for general issues): the definition of data requirements in general terms; evaluation of other data currently available; design and planning of statistical activity; extraction of data and metadata from databases within and external to OECD; implementation of specific data and metadata collection mechanisms; data and metadata verification, analysis and evaluation; data and metadata dissemination.

The quality checklists also incorporate elements of feedback mechanisms and ensure mainly that all relevant guidelines are taken into account when the statistical managers make their review. To support that process, the checklists additionally provide links to guidelines in key areas and examples for good practice.

#### 4.4 **Summary/Conclusion**

In this section we reviewed some key questions regarding the collection of data. We distinguished between organisations that collect data internally, externally and by means of a downstream agency; and those that do so in collaboration with other organisations.

We discussed some of the pros and cons (in terms of costs and timeliness) of organising data collection in one of these ways. When it comes to the question of what data to collect, or where to collect them from, we distinguished between primary data collection, a hybrid type of data collection (which relies on Member States and/or agencies to collect data) and secondary data collection.

We found that the hybrid form of data collection is the most popular one – in particular if, as is often the case, secondary data collection is not available. Regardless of which approach is chosen, all organisations stress the importance of having quality controls in place to organise the data collection. The main approaches used include peer review, pilot studies and transparency with regard to the collection approach, coverage, response rates and so on.

#### 4.5 **Lessons for DG SANCO**

The lessons for DG SANCO regarding the collection of data needs which follow from our discussion include the following:

- The organisation of data collection should take costs into account and rely primarily on collaborative collection efforts. Only where this is not possible on external (internal) data collection efforts if the same type of data is collected with low (high) frequency.



- Similarly, it is advisable to rely on secondary data collection when possible and to collect primary data only where secondary data are not available/suitable. At the same time, the case study on the UK Department for Environment, Food, and Rural Affairs (DEFRA) illustrates that generally fewer relationships / secondary data sources are preferable to many.
- Probably the most promising approach to assure the quality of data to be collected is to use a peer-review process at the planning and possibly the pilot stage of the data collection. There is good reason to try to involve people responsible for the data analysis in this process.





**Figure 5.1: Data management as part of a comprehensive data strategy**

In this chapter, drawing on the literature and the case studies reviewed, we discuss the third element of our conceptual framework: data management. The main constituents of data management are data storage, data verification, data compilation and access. Below we briefly discuss each element in turn.

### 5.1 **Who is responsible for data management**

In terms of data storage, one of the main questions emanating from our interviews is whether to store and manage data centrally (by a designated statistics unit) or decentrally (by individual policy units, say) – possibly trying to link them through a distributed network with a common interface. We found an almost even split among the organisations we reviewed between those which store their data centrally and those which store data in a decentralised way. Box 7 provides a few examples.

This finding reflects to some extent the open debate in the literature about the advantages and disadvantages of the two approaches. A centralised system, on the one hand, allows the sharing of the development and running costs across different policy units. It allows the consolidation of overall data efforts. Typically, because of the broader responsibility and better career progression prospects for them, it manages to attract better people to work on the system than decentralised systems do.

On the other hand, centralised data-storage systems are often characterised as more rigid and less responsive to individual needs and concerns (eg of different policy units) than decentralised systems, and the system software for a centralised system tends to be more complex and resource consuming. In addition, if there is a problem with the centralised system, often the entire system is in danger (as opposed to just part of the system, as happens with a decentralised data-storage arrangement).

Analogously, decentralised systems have the advantage that they can often be more responsive to the local problems of individual policy units. They allow for a closer interaction between data management and users and are typically run on less-resource-

consuming programmes. The flip side is that decentralised systems tend also to be prone to problems of interoperability and – even if hooked up to a distributed network – tend to be expensive as a result of duplicated efforts in system development and operation.

#### BOX 7

**Centralised systems of data storage.** The **WHO Regional Office for Europe** keeps certain data centralised – for example information on communicable diseases, which is stored and managed within one central system. The Health Info Subgroup as the central statistical unit is responsible for the whole range of tasks, such as data collection, storing, processing and production of outputs for internal usage.

The **UK Department for Environment, Food and Rural Affairs (DEFRA)** stores data centrally in its RADAR ‘data warehouse’. Relevant DEFRA teams have password access and the chief information officer performs a cross-departmental function, for example by providing standardised definitions, ensuring that information can be linked.

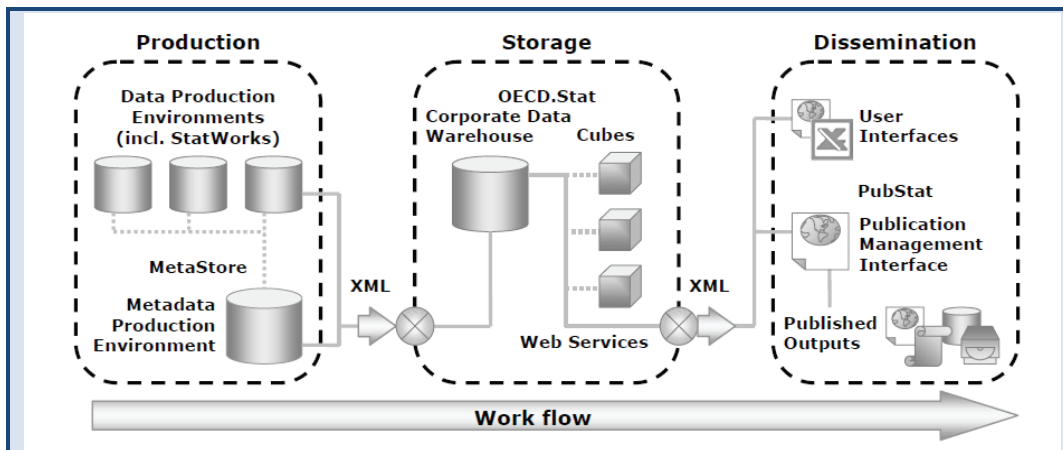
**Decentralised data storage.** Within the **UK Food Standards Agency (FSA)** data are usually kept by relevant policy teams. Data collectors assess and decide whether information is of broader value and has to be shared with colleagues. There is no systematic mechanism in place to determine which data are shared across FSA. Nevertheless, efforts are being made to develop corporate IT support eventually. This would include a share-point system to facilitate data sharing, which would represent a shift towards a more distributed system of data storage.

**Combination of centralised and decentralised data-storage systems.** The **OECD** uses two parallel systems in which data can be classified for central usage or stored in decentralised databases. Data are handled by directorates and the respective managers, up to the point where they can be seen as OECD products. By fulfilling certain criteria, data can be changed in status and may be approved for storage in central databases. Final data are stored in OECD Stat (the corporate statistical system, which is managed by the Statistics Information Management and Support Division) and made available to users.

The OECD is developing a new SIS for improving the efficiency of the whole process of data collection, compilation and dissemination.

The architecture of SIS consists of three layers:

- production layer – for collection, validation, processing and management of statistical data and metadata
- storage layer – where validated statistics and related metadata are stored
- dissemination layer – for producing statistical publications and online/offline interactive statistical products.



Source: OECD (2009)

SIS consists of five single components, each reflecting the decentralised and centralised nature of the OECD information system.

- OECD Stat.** OECD Stat is the central repository where validated statistics and related metadata are stored. It enables OECD analysts to get data from a single online source. Directorates' data and metadata are made available within a corporate system, while preserving OECD's decentralised statistical activities.
- StatWorks.** OECD's data-production environment has been developed to provide a common environment to host production databases, making use of several software platforms. Nevertheless, directorates are not obliged to migrate all of their production environments to StatWorks, which further reflects the decentralised approach of SIS.
- MetaStore.** This corporate metadata facility provides data-set managers with common tools and standards for managing fragmented metadata maintained by different directorates across the OECD. A database manager can use it as an interface for managing metadata and as a storage area. It is also used to export metadata to OECD Stat.
- PubStat.** The main tool for producing traditional statistical productions and extracting metadata from OECD Stat is PubStat. By combining OECD Stat and PubStat, the OECD tries to streamline the production of statistical publications, for which 'horizontal' process data are drawn from separate data sets.
- Workflow system.** This automates the process and is used to monitor the transmission of data from one layer to another.

In the light of these arguments, the literature suggests (Kroenke and Hatch 1994; Robson 1997; and Hodgkinson 1996) that a centralised system for storing data is generally more appropriate if i) there is a significant overlap in the data needs of different policy units (both in terms of type of data and content); ii) the individual data needs of policy units are rather small (and do not justify separate staff and resources); and/or iii) the data need not be very timely (as centralised systems tend to be slower in turning around data). On the

other hand, a decentralised system is preferred if data needs vary significantly between policy units, if individual data needs are rather high; and/or if data have to be very timely.<sup>6</sup>

## 5.2 Data verification

Data verification involves a number of things across the different case studies. These include data-item edit checks to make sure that values are within valid ranges and respect the accounting identities; comparisons of data over time, across data sets and across countries, or to ensure the coherence of the data in terms of definition, classification, standards and so on. Sometimes, data verification also involves checking that metadata are up to date and presented in a coherent and consistent way, so that users understand the strengths and limitations of the data and whether the data adequately approximate to what they wish to measure.

Data verification is handled internally by only a few of the organisations we reviewed – most notably the OECD. Other organisations outsource data verification together with data collection and/or leave it (explicitly) to those analysing the data. One explanation for this is that the practice of editing data is very costly in terms of resources needed, timeliness and the increased response burden. At the same time, as one of our interviewees pointed out, there is a good argument for handling data verification internally: careless handling of data verification can lead to severe bias as a result of incorrect design or application of editing procedures.

The OECD provides the most comprehensive definition of data compilation:

data compilation involves the production of additional figures or indicators derived from data originally collected [...]. These comprise ratios, supra-national aggregates, seasonally adjusted or trend-cycle values, composite indicators or composite leading indicators. [...]. In some cases, the compilation phase also requires the estimation (imputation) of missing values, adjustments to harmonise data across countries and to overcome series breaks over time caused by changes in concepts or methodology.

The organisations we reviewed in our case studies vary significantly in terms of their data compilation efforts. Some – like DGCCRF and RRG – work only with raw data. Others – including DG Trade, OECD and WHO – invest significant resources in data compilation. This reflects, to some extent at least, the difference in the nature of the work of these organisations. As is explained in more detail in Box 8, DGCCRF collects data mainly for purposes of monitoring consumer satisfaction and so on, and this requires less adjustment than collecting data for, among other things, policy research purposes, as is the case with the OECD and WHO.

### BOX 8

**Data compilation efforts, example 1. WHO Regional Office Europe** applies a wide range of review and compilation techniques. Data collected are compared with data from WHO headquarters in order to develop a benchmark and get an understanding of whether data are in line with data from other organisations. Internally, data are

<sup>6</sup> Some organisations make use of a combination of centralised and decentralised systems. See Box 7 above.

compared with trends and checked for discrepancies. In the event of differences in the data sets, feedback mechanisms within the countries ensure that data are checked a second time for comparability and compatibility with already existing data. The single units produce outputs which are disseminated across the organisation. Other units might carry out additional statistical analysis using these outputs.

**Data compilation effort, example 2. Public Health Agency Canada (PHAC)** uses a quality framework to ensure high standards of data quality. The framework consists of five dimensions: accuracy, timeliness, serviceability, usability, and relevance. These are further operationalised into characteristics and criteria. The accuracy dimension includes the characteristic ‘compilation and/or summary values’, which is further operationalised by the criteria ‘compilation techniques’ and ‘assessment of summary statistics’. The framework entails a detailed definition of the criteria, providing a guideline for how to proceed when it comes to data compilation and assessment of summary values.

**Metadata compilation effort.** The **OECD** Quality Framework and Guidelines for OECD Statistical Activities provide analysts with clear parameters on how to compile metadata. According to the guidelines, ‘metadata compilation involves the preparation of summary or detailed metadata about data received or autonomously compiled by the Secretariat, in a format useful for the uses for which they have been collected’. The guidelines provide a definition of different types of metadata which are commonly used to describe data. Internationally accepted standards (in particular definitions contained in the SDMX Metadata Vocabulary and the OECD Glossary of Statistical Terms) are applied when compiling metadata, and the data are stored in the forthcoming common repository of metadata (MetaStore) according to the OECD’s metadata-storage guidelines.

Another form of data compilation is the compilation of metadata. It involves the preparation of summary or detailed metadata about data received or autonomously compiled, in a format appropriate for the uses for which they have been collected.

### 5.3 Privacy and identifiable data

When it comes to data dissemination (the last part of data management), one of the key issues that emerged from the literature and the case studies was the problem of how to deal with data which allow direct or indirect identification. Direct identification means that data elements allow an individual, business or organisation to be directly identified – for example through name, address and health service number. Indirect identification means that an individual, business or organisation can be identified when the data are used in combination with other data such as date of birth, gender, incidence of a rare medical condition and so on.

One way of dealing with this problem, which has been chosen by RRG and SCA, is not to collect identifiable data. This, however, is not possible in all contexts – for example if for research purposes individual health outcomes have to be tracked over time. The organisations we reviewed that do (have to) collect identifiable data all adopt a broadly similar approach. That entails an evaluation of whether identifiable data need to be disclosed for analysis purposes. Then, depending on the outcome of this evaluation, the procedure is:

- make sure that the data are appropriately deidentified before their dissemination, accounting for both direct and indirect identifiers
- determine i) the legal and contractual authority to do so and ii) whether a written agreement or contract with the data requestor/recipient is required.

The default option for all organisations is not to disseminate identifiable data, and to store identifiable data in a manner that avoids the risk of disclosure. The manager in charge of the data set containing identifiable data adopts all the precautions and measures necessary to avoid any unauthorised access to the data. An example of a case where this approach did not work well is described in Box 9.

#### **BOX 9**

In December 2008, Germany faced one of the largest data theft scandals ever to occur. Confidential information about over 10,000 credit card customers of the Landesbank Berlin (LBB), the country's largest creditor, was anonymously sent to the *Frankfurter Rundschau*, one of the most prestigious German newspapers. Customers' names, addresses, credit card numbers, bank account numbers, transaction information and even various PIN numbers were included in the microfiches that *Frankfurter Rundschau* received. Not only LBB clients but many customers of other banks were also affected by the data breach, which took place as a result of an enormous security failure within the bank's data management. LBB had hired AtosWorldline to do its accounting, and it is assumed that the data went astray during a transfer between the two companies via a courier service. LBB's practice of passing sensitive data to third parties was criticised as an 'Achilles' heel and enormous loss of control' by Thilo Weichert, director of the Independent Centre for Privacy Protection in Schleswig-Holstein.

## 5.4 Summary/Conclusion

In this section we discussed the third element of our data-strategy framework: data management. We reviewed the advantages and disadvantages of centralised and decentralised data-storage systems and compared how data verification is handled across the organisations we reviewed.

In addition, we discussed the role of data compilation in the context of the work of different organisations and outlined the main strategies of dealing with identifiable data. One way is not to collect identifiable data in the first place. If identifiable data have to be collected, the data must be deidentified before they are disseminated, and the requestor of the data should be legally required to use the data sensibly.

## 5.5 Lessons for DG SANCO

From our discussion, again several interesting points for DG SANCO emerge. These include the following:

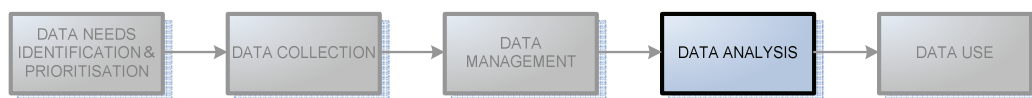
- DG SANCO could use three main criteria to evaluate whether a centralised or decentralised data-storage system is more appropriate: i) Is there an overlap in the



data needs of different policy units; ii) Are data needs rather limited in scope? iii) How timely do the data have to be?

- Owing to the danger of encountering severe bias if data verification is designed in the wrong way, DG SANCO could benefit from a formal system for this, which could be done either in-house or by out-sourcing.
- To the extent that DG SANCO intends to use its data for research purposes, there is good reason to take data compilation seriously. WHO, PHAC and OECD offer good examples of how data compilation can be organised (see Box 8).
- Given the high risk associated with dealing with identifiable data, DG SANCO may want to adopt a policy of collecting data at the lowest level of identification possible, distributing identifiable data only in exceptional cases.





**Figure 6.1: Analysis as part of a comprehensive data strategy**

The analysis of data is the last step within our framework before the data are actually used. This element is discussed in this chapter.

### 6.1 **Data – information – knowledge**

As discussed in the conceptual definition of data, data analysis is the step that adds meaning and structure to data and transforms them into information that can be ‘understood’ by actors (Aamodt and Nygård 1995). In turn, ‘learned’ information is what is considered ‘knowledge’. Giving data meaning therefore depends on how much sense we make of them and how much we know already. The organisation will need to think about how the analysis of data is done so that the data collected are transferred into useful information and ultimately into actionable knowledge.

This may lead the organisation to consider the broader knowledge framework – the principles, concepts and methods of knowing (or heuristics) which are appropriate to the organisation’s particular need (Callahan 2009) and how a data strategy would fit within that. Staying in line with the remit of this study, the key questions we asked were however more specifically focused around data analysis: who carries out data analysis, how the data are analysed, how quality in the analysis of data is assured, and how the analysis is communicated to serve the purpose(s) of the data strategy.

### 6.2 **Who analyses data?**

The case studies suggest there are different organisational models for organisations that wish to carry out data analysis. These models assign responsibility for data analysis to different parties/divisions who operate internally and/or externally to the organisation. Some organisations choose to carry out no analysis of the data they collect and compile, and their responsibility stops there.

The decision to carry out data analysis is often driven by consideration of the balance between capability (technical resources, experience and expertise), the costs of having in-

house analytical research competencies versus outsourcing, and the benefits that accrue to the organisation if analysis is done in-house (eg of accumulating in-house expertise and streamlining the data strategy from beginning to end).

Those organisations that do have internal analytical capabilities manage the data-analysis process in a centralised or a decentralised structure. A centralised structure would be one where data analysis is carried out in one central division, unit or department. DGCCFR for example has a central unit for analysing data. The advantage of this is that if the objective of the organisation is to inform policymakers, the organisation is able to have a streamlined analytical process and to have control over the type and nature of outputs produced to serve their organisational goals (see Box 10). This is most easily done within small organisations with a narrow remit and focus. Scaling this up in an organisation with a wider remit would be problematic.

#### **BOX 10**

Larger organisations, for example **WHO** or **OECD**, tend to specialise, and contain policy units that drive the analysis. Their organisational structure is modelled around a centralised data collection unit (eg the Health Surveillance Unit of WHO) that feeds the data requirements of decentralised policy units. These organisations would define their research agenda at the data-analysis stage, which would involve consideration of data needs as a loop back to the initial phase of a data strategy.

A decentralised organisational structure, particularly one that functions across different research domains, may need separate mechanisms for information sharing and transparency to allow knowledge to flow through and cross-fertilise. This is particularly important because data are not the unique property of any one unit and may be used in different ways by different units at the same time. In the case of the **Spanish Institute of Consumers (INC)**, the data obtained from autonomous regions are already preanalysed.

**1) Internal analysis of data (centralised).** The **French General Directorate for Competition Policy, Consumer Affairs and Fraud Control (DGCCRF)** has a central unit, which is part of the central administration, for analysing data. The data can be collected across the organisation, but they are analysed within the central administration.

Another form of centralised analysis is a central body with a network of partners. The **European Health Observatory's (EHO)** office in Brussels operates in this way, acting as the main data analyst (carrying out comparative studies and data analysis) within a partnership. EHO and its partners monitor health systems across Europe and collect and analyse data as a central unit. EHO produces Health Systems in Transition (HiT) profiles on a periodic basis.<sup>7</sup> In order to facilitate comparisons between countries, the

<sup>7</sup> HiT profiles seek to provide relevant information to support policymakers and analysts in the development of health systems in Europe. They are building blocks that can be used:

- to learn in detail about different approaches to the organisation, financing and delivery of health services and the role of the main actors in health systems
- to describe the institutional framework and the process, content and implementation of health care reform programmes
- to highlight challenges and areas that require more in-depth analysis
- to provide a tool for the dissemination of information on health systems and the exchange of experiences of reform strategies between policymakers and analysts in different countries.

profiles are based on a template for analysis,<sup>8</sup> which is revised periodically and provides detailed guidelines and specific questions, definitions and examples needed to compile HiTs, which is done centrally within EHO.

**2) Internal analysis of data (decentralised).** The **UK Food Standard Agency (FSA)** involves mainly involves policy teams in the analysis of data. The same situation prevails in **DG Enterprise** and the **Dutch Regulatory Reform Group (RRG)** where the policy units themselves (and specifically the desk officers) are in charge of analysis.

**3) Collaboration and external contractors.** **FSA** frequently commissions surveys with the help and support of external contractors who undertake part of the data collection or data analysis, or both.

The decision to employ external contractors is reached at the resource planning stage, when an FSA project officer estimates the level of internal staff resources required to undertake the work. If there are insufficient internal resources, an external contractor may be hired to carry out the work.

All potential contractors are required to give details of proposed analytical quality assurance (AQA) procedures to be used within the survey by completing a questionnaire and submitting a proposal in response to an invitation to tender. In appraising contractors' proposals, FSA requires that an FSA statistician and, preferably, an appropriate expert in analytical methodology (eg a member of a supporting working party and/or methodological subgroup) are included in the appraisal panel for analytical surveys. Their advice is also sought regarding whether potential contractors have within their teams, statistical expertise that is appropriate for the design, interpretation and reporting requirements of each survey.

A preferred proposal is identified through the appraisal process. When doing this, the panel members must consider whether the proposals represent value for money. The expected costs of the survey work (and resources) will already have been estimated. The FSA project officer and the contractor will then move on to the contract negotiation phase, in which details of the tasks and services are finalised.

Contract negotiation for survey projects can be problematic in that it may involve developing contracts with more than one external party. For example if a single contractor undertakes the sampling and analysis, this is likely to involve only one contract negotiation. If the contractor undertaking the sampling is different from the one undertaking the end analysis, this may result in multiple contract negotiations, which ideally have to be linked in order to create a cohesive protocol.<sup>9</sup>

On the other hand, organisations may choose to employ external contractors or to collaborate with other parties (eg in academia) to carry out data analysis, depending on ad hoc or systematic needs. The decision to contract out analytical services is mainly driven by resource or knowledge constraints. It may however also be driven by strategic growth ambitions, through for example forming organisational consortia (eg EHO) or collaborative partnerships, which are mostly with similar organisations, agencies or universities.

<sup>8</sup> <http://www.euro.who.int/Document/E88699.pdf>

<sup>9</sup> <http://www.food.gov.uk/multimedia/pdfs/fsasurveysguidance.pdf>

Managing external contractors involves an added resource burden for public organisations regarding procurement (eg contractual, legal and administrative time), but it may have benefits of learning through knowledge sharing, knowledge transfer and the receipt of expert independent advice. An illustration of FSA procedures regarding external contractors is given in Example 3 in Box 10 above.

### 6.3 Performing analysis

Analysis methods used may vary according to the inputs (data collected) that are used and the desired outputs (evidence, information and knowledge required to answer a policy question). Our case studies identified two general forms of data analysis: qualitative descriptive analysis and quantitative analysis. The form seems to depend on the remit or agenda of the organisation, a significant factor being whether it is a data provider whose role is to disseminate the raw data and their basic descriptive statistics, or whether its remit is to interpret the data and produce research outputs such as studies pertinent to the organisation's research interests and agenda. The latter would require a substantial increase in organisational capacity and technical expertise.

Nearly all organisations interviewed conduct at the basic level descriptive analysis of the data in qualitative reports (eg percentages, frequency tables and trend charts). A few carry out quantitative data analysis that is more complex, sophisticated and targeted (eg constructing indicators, regression analysis, etc). In PHAC for example, the specialised areas that require more technical expertise would use economic analysis or mathematical modelling. DEFRA uses data sets for a variety of stakeholders, and statistical analysis is performed on specific subjects – such as blue tongue disease, for example. On the other hand, DG Trade produce their data in the form of tables, files and databases in order to have a descriptive view, and little further analysis is performed. DG Trade rely on different units that are interested in the data to analyse them.

### 6.4 Assuring quality in data analysis

Many of the organisations we reviewed were conscious of the need to ensure the quality of the analysis being carried out, particularly as it is held up to external scrutiny if disseminated through public channels. Those that produce more complicated quantitative analysis recognise that the interpretation of the data that they produce needs to be accurate, robust and reliable. Those who carry out only a basic descriptive analysis of data have a minimised risk of introducing error since they do not transform or manipulate the data to a significant extent in the first place.

Our case studies suggest there are four main elements for good practice in quality assurance.

Firstly, there is peer-review approval for publication, either by an internal or external expert(s) (the latter is preferred) who reviews the work rigorously and checks its accuracy from a methodological perspective. This is considered a robust provision for quality assurance and validation of analysis. DG Environment for example stipulates that impact

assessments are scrutinised by the Impact Assessment Board, which is seen as a quality assurance body.

Secondly, transparency of analysis through communication and wide dissemination (eg through publishing methodology, description of data quality and limitations of analysis, and presentation to stakeholders for feedback) opens the door for critique and improvement.

Thirdly, the use of guidelines and/or protocols systemises the quality-assurance process in an organisation and identifies the benchmark of quality (this may either be internally formulated or externally recognised or defined standards).

Finally, ensuring the data strategy involves a systematic review of the organisation's data-analysis activities (assumptions, principles and methods used) to identify and assess the strengths and gaps or weaknesses in the organisation's ability to conduct data analysis. Ideally, the quality-assurance process should contain all four elements to ensure the validation and credibility of the analysis carried out. Box 11 gives two examples of guidelines for the quality assurance of data analysis, in OECD and PHAC.

Those organisations that conduct only basic descriptive analysis, such as RRG, did not carry out quality assurance on their analysis.

#### **BOX 11**

**Formal framework for quality assurance 1.** The issue of quality management is dealt with by all directorates in OECD, who follow the organisation's Quality Framework and Guidelines for statistical activities. They also have a systematic review of activities, in which directorates take turns to review their statistical activities with the statistics directorate, IT and publishing. The Quality Framework has the following guidance on the quality assurance of data analysis:

For products that include primarily, or only, analytical results, documentation should be provided on both the source data and the method of analysis. The requirements for documentation for the source data are similar to those for other statistical products. Documentation of the methods of analysis may be incorporated into the product, either as part of the presentation of the analytical results in the body of the report, or in separate text boxes. Such text boxes should also include summary information on:

- the data sources
- key features of the methodology and accuracy of the source data pertinent to the analysis
- analytical objectives, concepts and variables
- analytical methods used, their assumptions and caveats
- statistical significance of the results and any relevant conflicting or corroborating results
- appropriate use of the results

a brief description of the possible impact of accuracy issues, assumptions and caveats (OECD 2003).

**Formal framework for quality assurance 2. Public Health Agency Canada (PHAC)** has in its policy for collection, use and dissemination of public health data the following guidelines for data use and dissemination:

**On data use – analysis and interpretation:**

- Determine and consider the data quality in analysing and interpreting the data.
- When generating forecasts or projections, use mathematical models with clearly articulated and publicly available specifications and with reproducible results.

**On data dissemination – publication and distribution:**

- Include appropriate descriptions of the underlying sources of the data and information used, the concepts and the methodology in statistical and research products, as well as indicators of data quality, in order for users to determine the product's fitness for their particular uses.
- Have the data product, publication, report or presentation reviewed before dissemination, ensuring that all requirements of the agency's policy for the publication of scientific and research findings are met, including the consideration of policy implications, in consultation with Strategic Policy Directorate, as appropriate.
- Ensure that an appropriate communications strategy is in place for a timely and effective dissemination, in consultation with Communications Directorate.<sup>10</sup>

## 6.5 Reporting the analysis

There are several channels of communication that are used to disseminate the findings of data analysis. In the case studies we reviewed, the following channels were the ones most commonly employed:

- Reports – main channel of communication for public and policymakers. Most produce printed publications, as well as electronic versions of their reports on their web site.
- Web site for dissemination of reports and any other information related to analysis (eg references, links, related studies, etc).
- Intranet for internal communication – for example DEFRA has interactive reports and maps for common problems on the DEFRA intranet.
- Presentations to stakeholders.

Lastly, one interviewee felt strongly that databases full of tables, graphs and other information which no one knows how to interpret is of little use, and that, worse still, policy mistakes can be made because policymakers are basing decisions on data without

---

<sup>10</sup> Source: Policy for the Collection, Use and Dissemination of Public Health Data, Public Health Agency of Canada.



understanding what those data mean. Some were of the view that a lot of data is produced and a lot less is analysed; for instance, ‘we produce a lot of data and less frequently we produce important analysis and even less frequently do we generate strategic analysis that is essential for both the organisation and countries’. Policymakers and researchers need to be ‘context sensitive’ about the sorts of research methodologies and the types of evidence best suited to different circumstances (Marston and Watts 2003). At the analysis stage the data strategy should support and enable policymakers and researchers by being context sensitive in this way.

## 6.6 Summary/Conclusion

In this section we discussed the fourth element of our data-strategy framework: data analysis. We reviewed and compared the different organisational approaches to data analysis in terms of who carries out the analysis, how they do so, and how the quality of the analysis is assured and communicated.

## 6.7 Lessons for DG SANCO

For the data-analysis stage of a data strategy, DG SANCO could take heed of the following points:

- DG SANCO could evaluate its internal capability to carry out data analysis, identify where there are gaps in its knowledge, compare the costs of expanding its analytical capabilities relative to the cost of outsourcing / using external contractors, and measure the value added and the benefits of having in-house expertise in data analysis.
- The merits of a centralised or decentralised approach to data analysis need to be assessed against DG SANCO’s existing organisational structure. Good practice examples for either way of organising data analysis are EHO and DG Enterprise.
- Good practice in quality assurance for data analysis involves transparency and communication of methodologies (principles, assumptions and methods) and systematic reviews of methodologies. OECD and PHAC provide standards that could be adopted by DG SANCO.





**Figure 7.1: Data use as part of a comprehensive data strategy**

The final element of our data-strategy model is data use. Improving the use of evidence in policy will involve ‘effective *dissemination* of evidence to where it is most needed and the development of effective means of providing wide access to knowledge’ (Marston and Watts 2003). Ultimately, the use of data and whether they are accessible and exploited by the end user will qualify the value and cost of the data strategy and form part of an integrated data strategy that will take into account the big picture right from the start. In analysing the case studies, we were concerned at this point with who uses the data and what the information is used for. At the same time we are aware that consideration of future data use is an integral part of the identification and prioritisation phase.

## 7.1 **Data users**

The general public, the government and parliamentarians, internal clients and academics are the main users of the data produced and analysed by the organisations we reviewed. There are other users, of course, such as independent researchers, stakeholders in the public health community and the media. Most of the organisations strategically target one or more of these groups – for example WHO is most interested its work reaching policymakers and academics, so its data products are tailored towards the needs of these two groups.

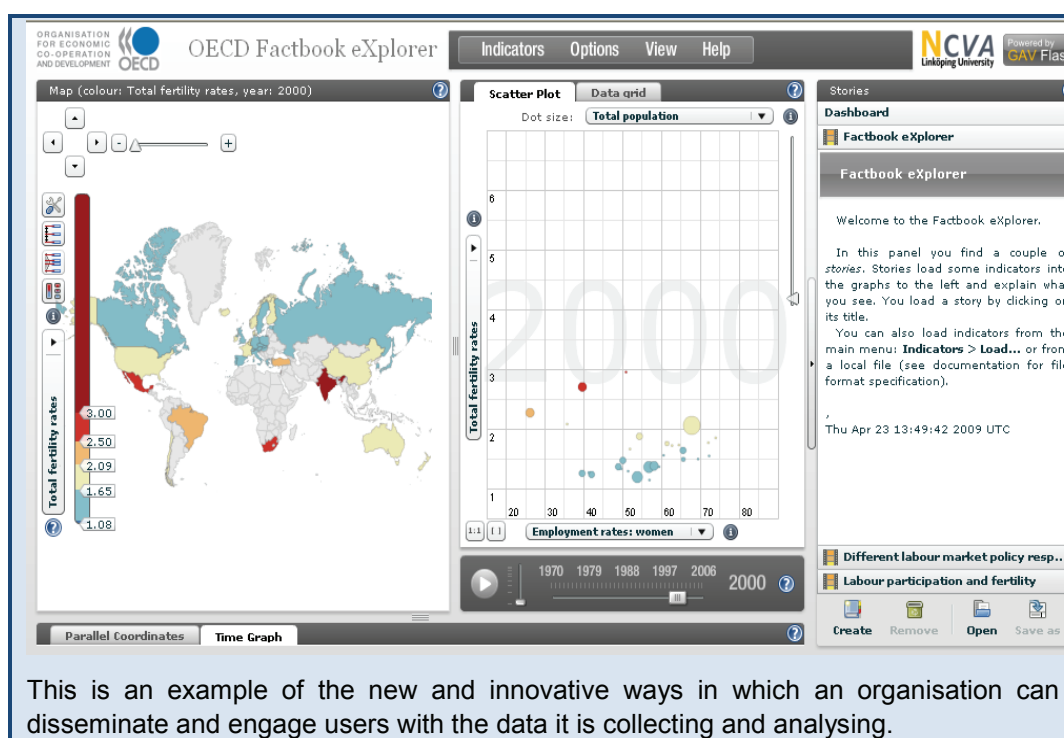
### **BOX 12**

**Presenting Data: OECD Factbook Explorer.**<sup>11</sup> A new OECD web site presents statistics in an animated and interactive fashion. The BBC<sup>12</sup> calls it ‘An exciting stats website’ and it is an example of how data that are user friendly and engaging to users can be visualised. It also illustrates how organisations can invest in new technology and new types of multimedia skills to convey data in new and interesting ways.

---

<sup>11</sup> <http://stats.oecd.org/oecdfactbook/>

<sup>12</sup> <http://news.bbc.co.uk/1/hi/magazine/8130554.stm>



## 7.2 What are data used for?

The organisations we reviewed were of four types: national governmental bodies, supranational organisations, international organisations and regulators. These bodies may use data in similar or different ways at the strategic level. National governments and their agencies mostly use data to inform policy, as evidence to evaluate, justify and recommend existing or potential policy choices. Supranational organisations, such as the European Commission, use data for the same reasons as a national government – for evidence-based policymaking, but across countries – and therefore need comparative cross-country data to garner consensus among members with the focus on the standardisation of definitions, frequency and other factors. Regulators, on the other hand, need data to monitor, recommend and enforce standards set by the government. Overall, we can say that data are used in the following ways:

- **To inform policymaking.** Governments and public agencies such as FSA use data to explain courses of action taken to the public and to stakeholders. Similarly, RRG stressed that one of its uses of data is to create political pressure and momentum for reducing the administrative burden on businesses. EHO, which focuses on using secondary data in its health system monitoring, states that its priority is that broad policy in health-system-specific issues should make use of its work for comparative analysis and evidence-based policy.
- **To monitor, review and evaluate policies.** Monitoring can also be a reason for collecting information. INC reports for example that it uses data to monitor and assess the success of implementing new legislation. RRG uses data to monitor the

regulatory burden that new legislation and regulation place on businesses (both before and after the enforcement of legislation and regulation). EHO uses the data it collects and analyses to monitor at European level, whilst WHO and OECD take a global outlook on issues and collect comparative data across countries. The main users of their data are other countries and national organisations who use it for benchmarking. One of the ways in which WHO uses its data is to show whether planned interventions are successful in countries of intervention. The European Commission directorates use data for impact assessments, reviews and ad hoc studies.

- **To communicate and inform the wider public.** Most of the organisations make their data and analysis freely available online through their web sites, and consider this to be their main channel of dissemination. WHO keeps tabs on who accesses their web site to download data and how many times their reports and data are cited in other publications.
- **For academic (research and teaching) purposes.** WHO has strong interactions with some academic groups who use their information and material for both training/teaching and research; these services are requested systematically by academics.
- **For their own internal use and to support the organisation's work and management.** The SCA also uses its data for internal purposes such as planning and priorities for legal departments. In particular, the consumer affairs agencies use the data they collect, or that are collected on their behalf, to inform both the management of operations as well as to spot emerging policy problems

### 7.3 Addressing users

Given that an organisation knows who the users are and how they intend to use data analysis, the next issue is what an organisation can do to reach out to the users. Our case studies reveal several examples of outreach mechanisms that organisations use to target the main intended users of their work (potentially including stakeholders like industry and consumers). These are workshops, focus groups, interviews with policymakers, conferences, and client surveys to understand users' perceptions of their work in order to evaluate whether the organisation's work is having an impact and in what way. EHO for example targets policymakers direct and will carry out specific research as inputs to policy briefs, ministerial meetings such as the Czech presidency, committee meetings and conferences.

#### **BOX 13**

**Dissemination tools.** The European Health Observatory (EHO) uses several channels for dissemination to reach its target audience: policymakers. These channels are publications, web and list servers and face-to-face tools listed below.

### 3. Dissemination

#### Reaching policy makers

---

- **Publications**
  - HiTs
  - Joint OBS/Open University Press book series
  - Observatory series (in-house)
  - Policy briefs
  - Newsletter: *Euro Observer* (approx 5000 copies)
  - Journal: *Eurohealth* (approx 5000 copies)
  - Translations (Russian, Spanish,...)
  - Journal articles (approx 80 per year)

[www.observatory.dk](http://www.observatory.dk)



### 3. Dissemination

#### Reaching policy makers

---

- **Web and listserve**
  - 150,000 separate visitors
  - Close to 1 million visits
  - Over 150,000 downloads
  - List serve approx 3,200 subscribers from 98 countries

[www.observatory.dk](http://www.observatory.dk)



### 3. Dissemination

#### Reaching policy makers

---

Finally, one notable comment made by an interviewee was that large organisations need to understand the value of a comprehensive data strategy that delivers a systematic collection and analysis of evidence to monitor and evaluate policies, in order to avoid their tendency to ‘work in silos’. However, another interviewee felt that there would be limited benefits in

a systematic approach for collecting information as an organisation may require very specific information and need it only once. Overall, it seems that the best option is to systematise the data-analysis process to some degree, and also to have the flexibility to deliver to meet ad hoc data-analysis needs as new questions of interest in relation to policy arise.

#### 7.4 **Summary/Conclusion**

In this chapter, as the final element of our data-strategy framework, we have discussed the use of data. We identified four main users of data and analysis: the public, government and public bodies/agencies, internal clients, and academics. They use it for four main purposes: to inform policy; to monitor, review and evaluate; for academic use; and for internal use.

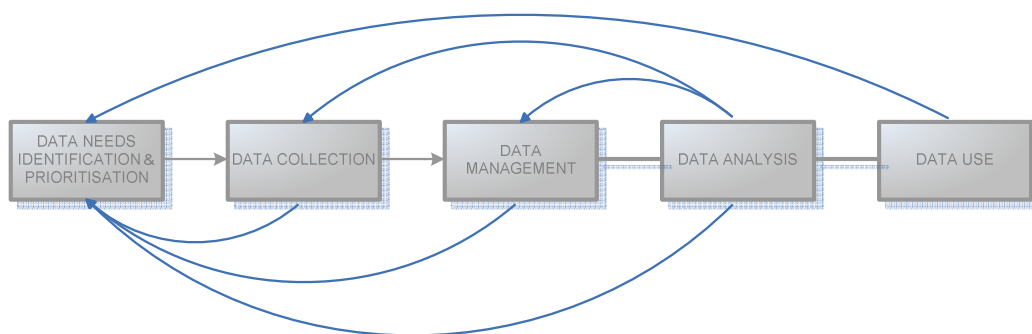
#### 7.5 **Lessons for DG SANCO**

If DG SANCO is to consider data use as part of an integrated data strategy, it may choose to implement the following procedures:

- Identify the main users of its data-analysis products to ensure the wide and effective use and uptake of DG SANCO work
- Understand how and for what purpose the end users use the analysis – eg by using SANCO internal workshops
- Develop mechanisms to reach and involve users – tested examples include Scenario Logic and Scenario Gaming workshops.







**Figure 8.1: Essential feedback loops in a comprehensive data strategy**

In the previous chapters we discussed the five key elements of our framework: identifying and prioritising data needs, data collection, data-management, analysis, and use. In our discussion we assumed a linear relationship between the key elements – with the identification of data needs feeding into data collection, data collection feeding into data management, data management feeding into analysis, and finally management feeding into use. In reality the picture is much more complex. Data collection, management, analysis and use, for example, are likely all to have a bearing on the identification of data needs. In this chapter we discuss some of these nonlinearities.

### 8.1 **Identifying and prioritising data needs**

Our case studies and earlier discussion suggests that all stages feed into the identification of data needs. We argued (in Chapter 4) that there is good reason to involve data users in the identification of data needs: to ensure buy-in from users at an early stage of the process and also that the data collected will be relevant.

Data analysis also informs the identification of data needs. People involved in the data analysis typically are the best to judge what data are needed to answer the questions of interest (to users) in the most robust way. In addition, in practice – as pointed out by one of our interviewees – analysis often shapes the expectations of users in terms of what they want to know (eg by identifying relevant policy problems), and so indirectly has a bearing on the identification of data needs.

Data management is relevant to the identification of data needs insofar as one of the key functions of data management is to ensure compatibility of existing and new data. We referred to this earlier as ‘data verification’. Involving people responsible for data

management in the identification of data needs can help to balance the trade-off between having the most relevant data (at a specific moment in time) and having long time-series of data (which are relevant beyond the current situation).

Not all data that are relevant can be easily collected. As an example, to evaluate the effectiveness of border-control policies it would be interesting to compare the level of human trafficking at different points in time and in different locations. Collecting data on human trafficking however is tremendously difficult. People don't register to smuggle humans; smugglers are hard to identify – and even if identified they are typically not interested in sharing information. This suggests that the feasibility of data collection has an important bearing on the identification of data needs.

Our discussion suggests that for good data-strategy feedback, mechanisms should be in place to allow communication between the different stages. Most of the frameworks we studied do have such mechanisms in place. The main difference between them is the degree of formality. Some organisations tend to rely on relatively informal mechanisms (allowing for people of all stages to feed back their views to the data identification stage); others use more formal approaches. Box 14 gives some examples of these feedback mechanisms.

As the size of an organisation increases and informal channels tend to become less efficient/effective, a more formal system of integrating stakeholders from different stages of our framework seems preferable to an informal approach. A challenge for a formal approach is of course to integrate people from the different stages of the framework in a structured way. One approach to doing so – very similar to that used at OECD – is to divide the identification of data needs into several components, such as i) identification, ii) feasibility and iii) prioritisation, with different stakeholder groups responsible for different components.

We have already discussed the fact that people from policy units – who are typically responsible for the data analysis – are well placed to identify data needs; while senior management, which is one of the main user groups, can be used at the prioritisation stage to ensure that the most relevant data are collected. In terms of feasibility, it seems that people from data collection and management are well positioned to judge what data can be collected and what it makes sense to collect in the light of what has been collected in the past.

#### **BOX 14**

As discussed in the main text, if there are suggestions for data improvements, then the data-analysis process can inform the identification of data needs and data collection, provided staff involved in the analysis are consulted so that those who are responsible for prioritising and collecting data are informed.

This is often done either in a formal or informal way. In the analysis process, staff discover what data are missing, which helps them to identify gaps. If this is fed back to the data collectors/providers, they can then request that additional data formally, either through a written request which is then assessed, specific meetings or consultation for feedback purposes. In **Public Health Agency Canada (PHAC)** for example, analysts and staff who analyse commonly acquired data inform their data needs through a

feedback loop between them and the data providers, either at stakeholder meetings or by regular communication.

**DG Trade** has a system whereby specific requests (after data analysis) could be addressed to the statistical team, and then taken into account if relevant. However analysts may be distant from some of the data collectors and have less ability to inform the data needs and data collection when there is no formal mechanism. **The European Health Observatory (EHO)** uses a circle of experts working with Member States, and talks to policymakers direct about their data-analysis needs so that partners in the EHO consortium resolve their data priorities and respond to the policy needs directly.

## 8.2 Data collection

The main stages that were highlighted in our case studies as feeding into data collection are data management and analysis. Data management has a direct bearing on the data collection stage, because it influences the format in which data should be captured. Problems of data verification (which is part of data management) are another reason for feedback from data management to the data-collection stage.

The idea that data analysis has a bearing on data management flows from the notion that modern analytical techniques can help to address problems of data collection – and so reduce the cost of data collection significantly – an example is the econometric ‘discontinuity approach’. This permits the evaluation of the effect of a policy intervention without dividing the population into treatment and control groups. Instead it estimates the effect of an intervention on the basis of the typically much cheaper information about the rationale for selection/non-selection. (The approach is described in more detail in Box 15.)

### BOX 15

**The Discontinuity Approach.** One of the central problems of evaluating policy interventions is the issue of selection bias. From a research point of view, ideally there is no difference between those individuals, businesses or organisations that are part of an intervention and those that are not. This would allow the difference in performance between the individuals, businesses and organisations after an intervention to be attributed to the intervention only.

In many situations, however, it is costly to allocate individuals, businesses and organisations randomly to a treatment and control group to get this effect. An alternative way to separate out the effect of an intervention is the so-called discontinuity approach. The idea is to utilise a discontinuity in the probability of selection that occurs at a particular threshold, with respect to some index of quality to identify the treatment effect separately from the impact of quality.

To make this concrete, imagine that a review panel or committee for an intervention ranks each applicant from best to worst, and records each applicant’s ranking. The selection official then draws a line through this ranking, indicating that if quality were the only issue, proposals above the line would be funded and those below would not. The location of this threshold is recorded. The official then makes the actual award decisions.

Without further assumptions, this approach identifies the treatment effect at the threshold quality level. That is, we can compare the performance of individuals,

businesses and organisations with similar rank, where one has been part of the treatment group but not the other. If we then plot a regression line through both groups (treatment and non-treatment) we can estimate the treatment effect for each of the funded individuals or businesses (not only those at the threshold), and its overall average.<sup>13</sup>

### 8.3 Data management

To the extent that several people analyse or use the data collected, it makes sense to have feedback channels in place from the analysis and user stage to the data-management stage. The idea is that data compilation can help to reduce the duplication of resources in these situations – by providing for example supranational aggregates, seasonally adjusted or trend-cycle values, and/or composite indicators which are used by most users/analysts (who do not have to calculate these values/indicators individually).

One way of organising this is by having small workshops following the systematic identification of data needs. In these, people from the analysis stage and user stage decide which additional figures or indicators should be produced at the data-management stage in order to avoid high duplication costs. Of the organisations in our case studies, it appears that WHO typically follows this approach the closest.

### 8.4 Data analysis

There are several reasons for the involvement of users at the data-analysis stage. In the following we summarise three of the most prominent ones:

- User involvement is important because it is often the users who decide what factor of a particular problem is the most important and determine the criterion on which to base a decision. Is the idea to minimise costs? To maximise impact? Or is it to broaden the base of those being served by an intervention?
- Users have a bearing on the data-analysis stage also because they are less likely to lose sight of the problem. Some disciplines specify analytical routines in detail for many circumstances. This often encourages people to begin work on a policy problem because it lends itself to their favourite method. Users may be needed in this situation to make sure their most pressing problems are actually addressed.
- In order for an analysis to be adopted and/or useful, it must be understood by the users. Because different users have different backgrounds and so different things are more easily accessible to them than to others, user involvement in the analysis can add real value by signalling to the analyst what parts of their work to explain in more detail and what to take as granted (and not to waste a policymaker's time with).

---

<sup>13</sup> Jaffe 2002.

There are a number of methods that have been suggested in the literature to allow for increased user involvement in the data-analysis stage. Most prominent are the so-called Issue Analysis / Dinner Party approach, Scenario Logic and Scenario Gaming. The methods are briefly described in Box 16.

#### **BOX 16**

**Issue Analysis.** The Issue Analysis / Dinner Party™ approach was developed by the UK National Audit Office (NAO). The NAO issues analysis that allows the main issue and its sub-issues from a range of issues to be established in a structured way.

It uses different levels of issues and tests them for being mutually exclusive and collectively exhaustive.<sup>14</sup> The highest level would have the main issue or the research question and the lower levels will have more detailed questions and issues leading to concrete evaluation tasks. It is a way of structuring quite complex topics in a workshop.

**Scenario Logic.** Scenario thinking aims to identify new developments, risks or impacts which might otherwise have been missed. It is a means to stimulate more informed and deeper conversations about future strategic-direction policy issues.

Scenarios are based on a thorough analysis of trend data, but should also creatively demonstrate that current uncertainties could lead to different, but equally plausible, futures. RAND has developed a six-step approach to the development of scenarios.<sup>15</sup> This includes the following six generic steps, in each of which users can be involved:

- Step 1:** Specify the system and define the outcomes of interest.
- Step 2:** Identify external factors driving changes in the system.
- Step 3:** Identify system changes, connections between these factors and system changes and how the changes affect the outcomes of interest.
- Step 4:** Categorise the uncertainty of the factors and system changes.
- Step 5:** Assess the relevance of the uncertain factors and system changes.
- Step 6:** Develop scenarios using the factors according to the combinations in Figure 8.1.

The dimensions to consider are constructed principally among the high-uncertainty items with large effects, and distinguished by their end-of-spectrum values across these dimensions. The number of scenarios chosen will be a reflection of the number of strategic-policy issues under consideration.

**Scenario Gaming.** Scenario Gaming is a method used to draw systematically on the expertise of a number of different people in order to understand a problem that contains uncertainty. It is best used in situations of mid-level uncertainty, where not enough is understood to be able to rely on formal analytical modelling tools, but enough is understood to be able to frame the problem in a logical and consistent manner.

Seminar games provide an opportunity to engage in semi-real decisionmaking without having to suffer the consequences of bad or unlucky decisions. Ideas can be explored

<sup>14</sup> NAO 2003.

<sup>15</sup> RAND Europe 1997.

and chances can be taken without having to wait years to discover the results or having to fail. Moreover, in a seminar game one can have the experience of thinking things through from a perspective other than one's own. The core philosophy behind a seminar game is that by (virtually) doing, one gains understanding well beyond that which is obtainable by reading or hearing.

## 8.5 Summary/Conclusion

While in the earlier chapters we implicitly assumed a linear model of data strategies – moving from identifying and prioritising data needs, to collecting data, to managing data, to analysing data and to using it – in this chapter we have discussed some of the nonlinearities involved in data strategies.

More specifically, we argued that all (later) stages of a data strategy have a bearing on the identification and prioritisation of data needs, that data management and analysis influence data collection, that data analysis influences data management, and that data use has a bearing on the analysis stage of data.

Drawing on these arguments, we suggested that a formal process to integrate these nonlinearities into the data strategy should be used. The main implications for SANCO are outlined in the next section.

## 8.6 Lessons for DG SANCO

Several lessons for DG SANCO follow from our discussion in this chapter:

- DG SANCO may want to structure its identification and prioritisation of data needs into three parts: i) identification of data needs; ii) feasibility; and iii) prioritisation. Policy analysts would be involved in part i), people responsible for collection and management in part ii) and senior management in part iii).
- A formal system of workshops that allow people from the different stages to interact with one another – in particular those from data collection, management and analysis and data analysis and management – could be beneficial to DG SANCO.

To conclude this report, this final chapter lists again the key recommendations that have been developed for DG SANCO throughout this report. It then moves on to discuss some wider issues relating to the research process before concluding with some thoughts on the implementation of organisational information-management tools.

### 9.1 **The key lessons for DG SANCO**

To start this concluding chapter, we should first like to summarise once more the key recommendations emerging from this research for DG SANCO. Using the data-strategy framework developed in the initial stages of the research, they may be summarised as follows.

#### **Identifying and prioritising data needs**

The main lessons for DG SANCO regarding the identification of data needs that follow from our discussion include the following.

1. Data needs should – where possible – be identified in a systematic way, using appropriate methods to overcome the cognitive limitations that can often act as a barrier to the identification of data needs.
2. To the extent that subject-related and methodological expertise sits with individual policy units, a decentralised system of identifying data needs is generally preferable to a centralised one.
3. A decentralised system of identifying data needs can be complemented with a top-down method of prioritising data needs, which ensures high-level buy-in into empirical work. At the same time, it allows the collection of the data needed to answer the most pressing policy questions.

#### **Collecting data**

The lessons for DG SANCO regarding the collection of data that follow from our discussion include the following.

4. The organisation of data collection should take costs into account, and should rely primarily on collaborative collection efforts. Only where this is not possible should it rely on external (internal) data collection efforts if the same type of data is collected with low (high) frequency.
5. Similarly, it is advisable to rely on secondary data collection when possible and to collect primary data only where secondary data are not available/suitable. At

the same time, the case study on the UK Department for Environment, Food, and Rural Affairs (DEFRA) illustrates that generally fewer relationships / secondary data sources are preferable to many.

6. Probably the most promising approach to assure the quality of data collected is to use a peer-review process at the planning and possibly at the pilot stage of the data collection. There is good reason to try to involve the people responsible for the data analysis in this process.

### Data management

From our discussion, again several interesting points for DG SANCO follow. These include the following.

7. DG SANCO could use three main criteria to evaluate whether a centralised or decentralised data-storage system is more appropriate: i) Is there an overlap in the data needs of different policy units; ii) Are data needs rather limited in scope? iii) How timely do the data have to be?
8. Owing to the danger of encountering severe bias if data verification is designed in the wrong way, DG SANCO could benefit from a formal system for this, which could be done either in-house or by out-sourcing.
9. To the extent that DG SANCO intends to use its data for research purposes, there is good reason to take data compilation seriously. WHO, PHAC and OECD offer good examples of how data compilation can be organised (see Box 8).
10. Given the high risk associated with dealing with identifiable data, DG SANCO may want to adopt a policy of collecting data at the lowest level of identification possible, distributing identifiable data only in exceptional cases.

### Data analysis

For the data-analysis stage of a data strategy, DG SANCO could take heed of the following points.

11. DG SANCO could evaluate its internal capability to carry out data analysis, identify where there are gaps in its knowledge, compare the costs of expanding its analytical capabilities relative to the cost of outsourcing or using external contractors, and measure the added value and the benefits of having in-house expertise in data analysis.
12. The merits of a centralised and a decentralised approach to data analysis need to be assessed against DG SANCO's existing organisational structure. Good practice examples for either way of organising data analysis are EHO and DG Enterprise.
13. Good practice in quality assurance for data analysis involves transparency and communication of methodologies (principles, assumptions and methods) and systematic reviews of methodologies. OECD and PHAC provide standards that could be adopted by DG SANCO.

### Data use

14. DG SANCO may want to consider data use and communication as part of an integrated data strategy.
15. It could identify the main users of its data-analysis products to ensure the wide and effective use and uptake of DG SANCO work.



16. It could aim to understand how and for what purpose the end users use the analysis – eg by using SANCO internal workshops.
17. Mechanisms to reach and involve users could be developed. Tested examples include Scenario Logic and Scenario Gaming workshops.

### Revising the linear model

18. DG SANCO may want to structure its identification and prioritisation of data needs into three parts: i) identification of data needs; ii) feasibility; and iii) prioritisation. Policy analysts would be involved in part i), people responsible for collection and management in part ii), and senior management in part iii).
19. A formal system of workshops that allow people from the different stages to interact with one another – in particular those staff involved in data collection, management and analysis and data analysis and management – could be beneficial to DG SANCO.

## 9.2 Conducting research into public-sector data strategies

Conducting comparative research into how public-sector organisations – mostly in the consumer, food safety, public health and animal health field – use data for policymaking meant charting a relatively new territory for analysis, resulting in a number of challenges for research that should be reflected on here.

Developing research into a previously under-researched field is likely to present difficulties in gathering relevant empirical evidence for a thorough and comprehensive analysis. For our research this meant collecting limited written documentation on how various organisations currently deal with their data-management issues. There were however marked exceptions – for example PHAC and OECD provided written documentation about the way in which they deal with data. The lack of written documentation resulted in a very strong reliance on key informant interviews in order to understand current practice in the various organisations.

During the key informant interviews, often with staff from statistical or analysis units within the organisations, we realised that the very notion or concept of a comprehensive data strategy – ranging from a definition of needs derived from the requirements of the policy process to the collection, management and analysis of data – is largely foreign to the organisations. There were however again exceptions, such as PHAC. Although PHAC had in place most of the elements we defined as part of a comprehensive data strategy, those elements were not necessarily considered as a strategic whole but were seen in isolation, and challenges were seen more on an operational level. In the light of this, DG SANCO's desire to have a comprehensive and systematic approach to deal with data is quite remarkable.

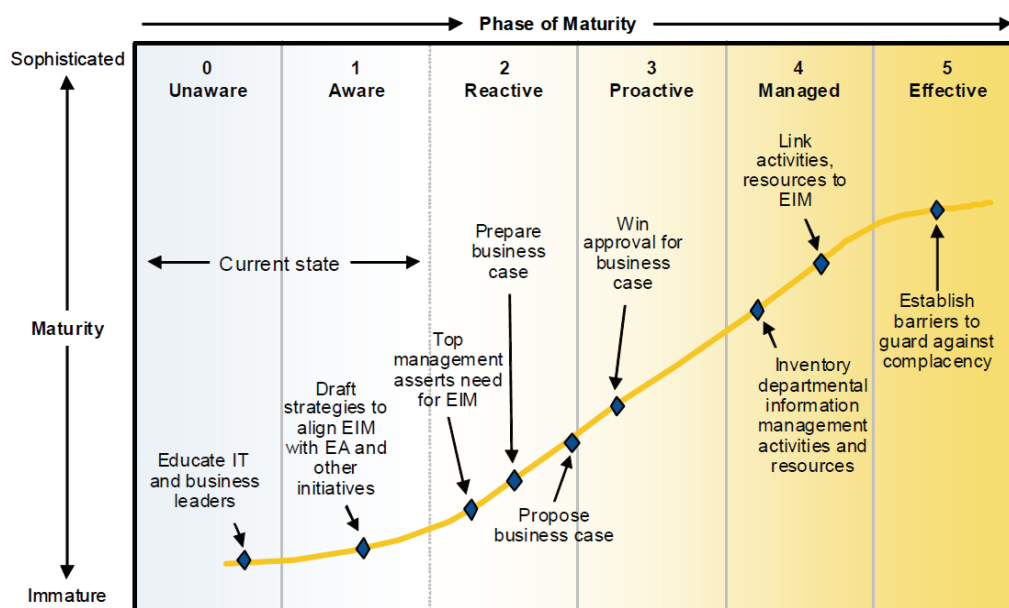
This report set out to define examples of good practice in data strategies which could be used in DG SANCO's efforts to develop its own comprehensive data strategy. Ultimately this calls for a judgement about what constitutes good practice and what is bad or 'normal' practice, based on an overall performance measure or standard. Given the fact that our assessment overwhelmingly relied on interviews and that some of the organisations

interviewed were somewhat early on in their development of a comprehensive data strategy, the statements made in this report and in particular the recommendations for DG SANCO should be seen as interesting practices that could be explored by DG SANCO, rather than as guidelines or best practice to be adopted. We considered the cases that had the more comprehensive approach to data and formal procedures, and that explicitly mentioned a data strategy, as those from which DG SANCO could learn most. The decision to describe interesting practice and the key choices and trade-offs that DG SANCO will need to consider in developing a data strategy is also a result of considering the transferability of practice from other organisations to DG SANCO. As a European Commission Directorate General, DG SANCO operates in a different political environment from Member States and international organisations. Its data needs are therefore likely to be distinct and we could not systematically control for context in our selection of case studies.

### 9.3 Implementing data strategies: the maturity of information-management systems

This research report will feed into DG SANCO's current activities in developing a comprehensive data strategy that will allow it to meet future challenges in policymaking. Designing and agreeing on a strategy is just the first leg of a longer journey of implementing and embedding a new strategy for using data, and ultimately information and knowledge, in policymaking.

To prepare for some of the challenges of this journey, a look at a widely used model in enterprise information management (EIM) may be insightful. Gartner, a consultancy and research company, developed a (normative) model to describe the stages and the maturity of EIM systems in the private sector, from which some conclusions may be transferable to DG SANCO (see Figure 9.1).



### Figure 9.1: The Gartner information-management maturity model

SOURCE: Gartner, 2008 in Newman and Logan (2008)

In this model five stages of maturity are identified, and organisations are thought of as moving from stage 0 to stage 5 as they implement information-management systems. Slightly adapted to a public-sector organisation like DG SANCO, these stages could be described as follows (see Newman and Logan 2008):

- **Level 0: Unaware.** The organisation runs risks from undermanaged information in particular.
- **Level 1: Aware.** The organisation attains some awareness of information management.
- **Level 2: Reactive.** Business and IT leaders react favourably to the demand for consistent, accurate and faster information across business units. They take corrective measures to address immediate needs.
- **Level 3: Proactive.** The organisation perceives information as necessary for improved business performance, and therefore moves from project-level information management to comprehensive, organisation-wide EIM. Information supports process optimisation.
- **Level 4: Managed.** The organisation perceives information as critical for business. It has implemented significant portions of EIM, including a consistent information structure.
- **Level 5: Effective.** The organisation exploits information across the entire information-supply chain, with service levels agreements that are continuously reviewed.

To move an organisation from stage 0 to stage 5, the model suggests a number of actions, which are indicated in Figure 9.1. As DG SANCO moves further towards implementing its strategy, it may be worth considering some of the thinking related to the maturity model, while keeping in mind the differences between the private sector and the public sector. The rationale for implementing EIM in private-sector organisations is ultimately to become more profitable, through more efficient and leaner processes, or more innovative products, or a superior knowledge of the market and clients' demands. The benefits of implementing such a system in an organisation like DG SANCO are far more intangible and might not accrue to the organisation itself. These benefits might include generally better informed policies and less burdensome regulations for business, as well as more legitimacy and accountability.



## REFERENCES

---



## Reference list

---

- Aamodt, Agnar, and Mads Nygård, 'Different Roles and Mutual Dependencies of Data, Information, and Knowledge – an AI Perspective on their Integration', *Data and Knowledge Engineering*, Vol. 16, 1995, pp. 191–222.
- Ackoff, R.L., 'From data to wisdom', *Journal of Applied Systems Analysis* Vol. 16, 1989, pp. 3–9.
- Adelman, S., L. Moss and M. Abai, 'Data Strategy, an Excerpt from Data Strategy', *Information Builders Magazine*, Vol. 16, No. 1, Winter 2006, pp. 21–23.
- Alavi, Maryam, and Dorothy E. Leidner, 'Review: Knowledge Management Systems: Issues, Challenges, and Benefits', *Communications of the Association for Information Systems*, Vol. 1, Article 7, 1999. As of 27 August 2009: <http://aisel.aisnet.org/cais/vol1/iss1/7>
- Alavi, Maryam, and Dorothy E. Leidner, 'Review: Knowledge Management Systems: Conceptual Foundations and Research Issues', *Management Information Systems Quarterly*, Vol. 25, No. 1, March 2001, pp. 107–136.
- Boisot, Max, and Augusti Canals, 'Data, Information and Knowledge: Have We Got it Right?', *Internet Interdisciplinary Institute, Working Paper Series*, November 2003.
- Bresciani, Sabrina, and Martin J. Eppler, 'Gartner's Magic Quadrant and Hype Cycle', Institute of Marketing and Communication Management (IMCA), Università della Svizzera italiana, Faculty of Communication Sciences, Case No. 2, 2008.
- Brown, John S., and Paul Duguid, 'Knowledge and Organisation: A Social-practice Perspective', *Organization Science*, Vol. 12, No. 2, 2001, pp. 198–213.
- Browne, Glenn J., and V. Ramesh, 'Improving Information Requirements Determination: A Cognitive Perspective', *Information and Management*, Vol. 39, 2002, pp. 625–645.
- Buckland, M., Information as a thing. *Journal of the American Society of Information Science*, 42(5), 1991, pp. 351–360.
- Callahan, Shawn, 'Crafting a Knowledge Strategy', Report Anecdote Ltd., As of 27 August 2009: <http://www.anecdote.com.au/papers/CallahanCraftingKnowledgeStrategy.pdf>
- Chaug, Ta-Tao, and Surya B. Yadav, 'The Development of an Adaptive Decision Support System', *Decision Support Systems*, Vol. 24, 1998, pp. 73–87.

- Chauvel, Daniele, and Charles Despres, 'A Review of Survey Research in Knowledge Management: 1997–2001', *Journal of Knowledge Management*, Vol. 6, No. 3, 2002, pp. 207–223.
- DoD, 'Net-Centric Data Strategy', *Department of Defense, United States of America*, 9 May 2003.
- Davenport, Thomas H., David W. De Long and Michael C. Beers, 'Building Successful Knowledge Management Projects', *Center for Business Innovation Working Paper*, January 1997. As of 27 August 2009:  
[http://www.providersedge.com/docs/km\\_articles/Building\\_Successful\\_KM\\_Projects.pdf](http://www.providersedge.com/docs/km_articles/Building_Successful_KM_Projects.pdf)
- DEFRA, *Action plan in response to David Eves' Review of the Animal Health & Welfare Delivery Landscape - January 2008*, Department for Environment, Food and Rural Affairs, as of 27 August 2009:  
[http://www.defra.gov.uk/foodfarm/policy/animalhealth/deliver/action\\_plan180108.pdf](http://www.defra.gov.uk/foodfarm/policy/animalhealth/deliver/action_plan180108.pdf)
- Eves, David CB, *Review of the Animal Health and Welfare Delivery Landscape* A report by David Eves, London: DEFRA, June 2006.
- Fischer, Gerhard, and Jonathan Ostwald, 'Knowledge Management: Problems, Promises, Realities, and Challenges', *IEEE Intelligent Systems*, Vol. 16, No. 1, 2001, pp. 60–72.
- Freeman, Richard, 'The Health Care State in the Information Age', *Public Administration*, Vol. 80, No. 4, 2002, pp. 751–767.
- Galliers, Robert D., and Sue Newell, 'Back to the Future: From Knowledge Management to Data Management', In *Proceedings of the Ninth European Conference on Information Systems* (S. Smithson, J. Gricar, M. Podlogar and S. Avgerinou (eds.), 2001, pp. 609–615, Moderna Organizacija, Bled, Slovenia. As of 27 August 2009:  
<http://is2.lse.ac.uk/asp/aspecis/20010036.pdf>
- Goodhue, Dale L., Laurie J. Kirsch, Judith A. Quilliard and Michael D. Wybo, 'Strategic Data Planning: Lessons from the Field', *Management Information Systems Quarterly*, Vol. 16, No. 1, March 1992, pp. 11–34.
- Goodhue, Dale L., Judith A. Quilliard and John F. Rockart, 'Managing the Data Resource: A Contingency Perspective', *Management Information Systems Quarterly*, Vol. 12, No. 3, September 1988, pp. 373–392.
- Grover, Varun, and Thomas H. Davenport, 'General Perspectives of Knowledge Management: Fostering a Research Agenda', *Journal of Management Information Systems*, Vol. 18, No. 1, Summer 2001, pp. 5–21.
- Gustafson DH, Cats-Baril WL, Alemi F. *Systems to Support Health Policy Analysis: Theory, Models, and Uses*, Health Administration Press: Ann Arbor, Michigan, 1992.
- Hammer, Joachim, Hector Garcia-Molina, Jennifer Widom, Wilbert Labio and Yue Zhuge, 'The Stanford Data Warehousing Project', *Bulletin of the Technical Committee on Data Engineering*, Vol. 18, No. 2, June 1995, pp. 40–47.



- Hodgkinson, S. T. 'The Role of the Corporate IT Function in the Federal IT Organization,' in *Information Management: The Organizational Dimension*, M. J. Earl (ed.), Oxford: Oxford University Press, 1996.
- Hoffer, Jeffery A., Stephen J. Michaele and John J. Carroll, 'The Pitfalls of Strategic Data and Systems Planning: A Research Agenda', *Systems Sciences*, Vol. 4, 3–6 January 1989, pp. 348–356.
- Jaffe, A., 'Building Program Evaluation into the Design of Public Research Support Programs', *Oxford Review of Economic Policy*, 2002; Vol.18: pp. 22-34 2002.
- King, William R., 'Strategic Planning for Management Information Systems', *Management Information Systems Quarterly*, Vol. 2, No. 1, March 1978, pp. 27–37.
- King, William, R., and David I. Cleland, 'Information for More Effective Strategic Planning', *Long Range Planning*, Vol. 10, February 1977, pp. 59–64.
- Kroenke, D., & Hatch, R. (1994). *Management information systems* (3<sup>rd</sup> ed.). New York: McGraw-Hill.
- Lederer, Albert L., and Vijay Sethi, 'Key Prescriptions for Strategic Information Systems Planning', *Journal of Management Information Systems*, Vol. 13, No. 1, Summer 1996, pp. 35–62.
- Liu, Tony D., and X. William Xu, 'A Review of Web-based Product Data Management Systems', *Computers In Industry*, Vol. 44, 2001, pp. 251–262.
- Marston, Greg, and Rob Watts, 'Tampering with the Evidence: A Critical Appraisal of Evidence-Based Policy-Making', *The Drawing Board: An Australian Review of Public Affairs*, Vol. 3, No. 3, 2003, pp. 143–163.
- Martenson, Maria, 'A Critical Review of Knowledge Management as a Management Tool', *Journal of Knowledge Management*, Vol. 4, No. 3, 2000, pp. 204–216.
- National Audit Office, 'The Issues Analysis Dinner Party Approach, User Guidance', 2003.
- National Statistics Board, 'Best Practice Guidelines for the Development and Implementation of Formal Data / Statistics Strategies in Government Departments', *National Statistics Board*, March 2004.
- Nemati, Hamid R., David M. Steiger, Lakshmi S. Iyer and Richard T. Herschel, 'Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing', *Decision Support Systems*, Vol. 33, 2002, pp. 143–161.
- Newman, David, and Debra Logan, 'Gartner Introduces the EIM Maturity Model', Gartner, 5 December 2008.
- Organisation for Economic Cooperation and Development, 'Quality Framework and Guidelines for OECD Statistical Activities', OECD, 2003.
- Organisation for Economic Cooperation and Development, 'Management of Statistical Metadata at the OECD', 2.0, OECD, 6 September 2006. As of 27 August 2009: <http://www.oecd.org/dataoecd/26/33/33869551.pdf>

- Organisation for Economic Cooperation and Development, 'Data and Metadata Reporting and Presentation Handbook', OECD, 2007. As of 27 August 2009: <http://www.oecd.org/dataoecd/46/17/37671574.pdf>
- RAND Europe, Scenarios for Examining Civil Aviation Infrastructure Options in The Netherlands. Unrestricted draft, 1997.
- Rieper, O., Frans L. Leeuw and Tom Ling (eds.), *The Evidence Book*, 2009, New Brunswick (NJ): Transaction Publishers.
- Robson, W. (1997). *Strategic Management and Information Systems: an Integrated approach*, Pitman Publishing, London, England.
- Rowley, Jennifer, 'The Wisdom Hierarchy: Representations of the DIKW Hierarchy', *Journal of Information Science*, Vol. 33, No. 2, 2007, pp. 163–180.
- Samtani, Sunil, Mukesh Mohania, Vijay Kumar and Yahiko Kambayashi, 'Recent Advances and Research Problems in Data Warehousing', *Lecture Notes in Computer Science*, Vol. 1552, Proceedings of the Workshops on Data Warehousing and Data Mining: Advances in Database Technologies, 1998, pp. 81–92.
- Samuelson, Lee, and Lars Thygesen, 'Building OECD's New Statistical Information System,' OECD, 2 November 2004. As of 27 August 2009: [www.oecd.org/dataoecd/26/35/33869528.pdf](http://www.oecd.org/dataoecd/26/35/33869528.pdf)
- Sanderson, Ian, 'Evaluation, Policy Learning and Evidence-Based Policy Making', *Public Administration*, Vol. 80, No. 1, 2002, pp. 1–22.
- Sen, Arun, 'Design Support System: An Activity-oriented Design', *Journal of Information Science*, Vol. 7, No. 23, 1983, pp. 23–30.
- Storey, John, and Elizabeth Barnett, 'Knowledge Management Initiatives: Learning from Failure', *Journal of Knowledge Management*, Vol. 4, No. 2, 2000, pp. 145–156.
- Subramanian, Ashok, L. Douglas Smith, Anthony C. Nelson, James F. Campbell and David A. Bird, 'Strategic Planning for Data Warehousing', *Information and Management*, Vol. 33, 1997, pp. 99–113.
- SAMSHA, 'SAMHSA Data Strategy: FY 2007 – FY 2011', , Substance Abuse and Mental Health Services Administration, United States Department of Health and Human Services *SAMHSA*. 2007.
- Sundgren, Bo, 'Making Statistical Data More Available', *International Statistical Review*, Vol. 64, No. 1, April 1996, pp. 23–38.
- Ward, John, and Joe Peppard, *Strategic Planning for Information Systems*, 3rd ed., Chichester: John Wiley & Sons Ltd, 2002.
- Wende, Kristin, 'A Model for Data Governance – Organising Accountabilities for Data Quality Management', In M. Toleman, A. Cater-Steel and D. Roberts (eds.), Toleman, Cater-Steel & Roberts (eds), *ACIS 2007: Proceedings of the 18th Australasian Conference on Information Systems*, 2007, pp. 417–425. As of 27 August 2009: <http://www.alexandria.unisg.ch/publications/41455>

- Wiig, Karl M., 'Knowledge Management: An Introduction and Perspective', *The Journal of Knowledge Management*, Vol. 1, No. 1, September 1997, pp. 6–14.
- Zins, Chaim, 'Conceptual Approaches for Defining Data, Information, and Knowledge', *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 4, 2007, pp. 479–493.



## **APPENDICES**

---



# Appendix A: Methodology

---

## Literature review

A number of sources were used to identify relevant literature. Firstly, searches for the main relevant literature (journal-based or independent) were conducted through databases (including JSTOR, Elsevier, PubMed, Web of Science and Wilson Select Plus); peer review; and academic journals (such as Data Strategy, MIS Quarterly, Decision Support Systems, Data and Knowledge Engineering and others). Grey literature (ie reports and studies produced by professional associations, government, international organisations and other relevant bodies) was also consulted.

For the academic literature review, the following key terms were used: 'data strategy', 'data approach', 'data management', 'data needs', 'database strategy', 'database management', 'database needs', 'information management', 'information strategy', 'information needs', 'information systems management', 'information systems strategy', 'knowledge management', 'knowledge strategy', 'knowledge needs', 'evidence based policy making AND data', 'evidence based policy making AND information', 'data warehousing', 'information warehousing', 'knowledge warehousing' and others.

Following the search strategy described, we identified and read over fifty relevant abstracts in relation to data strategies. The articles/reports were then scanned for relevant information.

## Interview template

This section provides a copy of the interview template for our key informant interviews:

**To** :  
**From** : RAND Europe  
**Subject** : Interview questionnaire – DG SANCO Data-strategy project  
**Date** : 20 April 2009

---

This memo outlines the topics for our telephone interview. The questions in this questionnaire are designed both to capture the current practice of how your organisation uses data, and the underlying corporate strategies.

## Background

- Please provide us with a quick overview of your organisation and the role you and your unit have in this organisation?

## Data needs

In this first section of questions we want to learn about your organisation's data needs and what kind of data you are collecting.

- Could you give a brief overview of what kind of data is currently collected by your organisation?
- Could you distinguish between different categories of data?
- Who identifies data needs in your organisation? Senior management, end users, statistical units, strategy units?
- What are the reasons for collecting the different kind of data? Are there legal obligations; is it part of monitoring exercises or for strategic purposes?
- How do they identify data needs? Do you use consultation or feedback mechanisms? Are decisions made top down?
- How do you prioritise the data to be collected? Is it top down? Is it driven by availability of data or how often the data is needed?



## Data collection

In this section we want to learn about how data is collected in your organisation.

- Who is responsible for data collection in your organisation? Is data collected by a horizontal, specialised unit or is data collected on a decentralised basis?
- How do you collect data in your organisation? Do you:
  - o Collect own primary data? (Surveys, studies)
  - o Rely on secondary sources (such as a statistical office), if yes which ones?
  - o Collaborate with other organisations in collecting data, if yes, who do you collaborate with?
  - o Have a protocol that guides data collection efforts?
- How do you manage the demand for data collection by different groups in the organisation?
- How do you decide how often the same data is collected?
- Do you mostly collect data on a regular basis or more ad hoc for specific purposes or studies?
- How does the organisation ensure that data collection is not duplicated across the organisation?
- How do you ensure the quality of the data collection? Do you have quality protocols, a quality assurance process or designated staff?
- How does the organisation ensure data collection does not place an undue burden on third parties (e.g. businesses, Member States)?
- How does data collection inform the identification of data needs, is data collection staff consulted to inform data needs?

## Data management

In this section we want to learn about how the data is managed in your organisation (i.e. stored, maintained, updated) once the data has been collected.

- Who is responsible for storing, updating and maintaining data collected by your organisation?
- How do you ensure compatibility of data from different source (e.g. coherent terminology, methodology, etc)?
- Do you try to ensure compatibility with other organisations' data (e.g. other DGs, departments or international organisations)?

- Do you edit and transform the data you hold? If so how and on what grounds? How do you ensure the quality in this process?

## **Data analysis**

The next step in making data available for policymaking is the analysis of data.

- Who analyses data collected by your organisation:
  - o Centralised intelligence unit or decentralised policy units?
  - o Internal or also external?
  - o External contractors?
- How is the data analysed?
  - o Do you produce mostly descriptive accounts of the data?
  - o Do you perform further analysis, such as econometric analysis?
  - o Is the data analysed qualitatively?
- How is quality assured in the analysis of data?
- How is the analysis communicated to the end user?
- How does data analysis inform the identification of data needs and the data collection, are staff involved in data analysis consulted to inform data needs and/or collection?
- How does the later data use inform data analysis?

## **Data use and communication**

In this final section our questions centre on how analysed data is used and eventually communicated to internal and external parties.

- Who uses data and analysis in your organisation? Which third party uses data and analysis provided by your organisation?
- How are data and analysis used by these groups?
  - o To inform policymaking? (e.g. impact assessment)
  - o To monitor and evaluate policies?
  - o To communicate with the wider public?
  - o For academic purposes?

**Further information**

- Is there further information available to you that might be relevant to the issues discussed in this interview?
- Would you be able to share, in confidence, any additional information relevant to this research exercise?