

eurostat 



# Handbook on measuring data in the System of National Accounts

2025 edition

 MANUALS  
AND GUIDELINES





# **Handbook on measuring data in the System of National Accounts**

**2025 edition**

Prepared by staff of the IMF and EU. The views expressed in this book belong to the authors. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Union or of the International Monetary Fund.

Nothing contained in this book should be reported as representing the official view of the European Union or of the IMF, member governments, or any other entity mentioned herein.

Luxembourg: Publications Office of the European Union, 2026



© European Union and International Monetary Fund, 2026

The Commission's reuse policy is implemented under Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39, ELI: <http://data.europa.eu/eli/dec/2011/833/oj>).

Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed, provided appropriate credit is given and any changes are indicated.

For any use or reproduction of elements that are not owned by the European Union, permission may need to be sought directly from the respective rightholders. The European Union does not own the copyright in relation to the following elements:

cover © AdobeStock/Image Craft

Collection: Manuals and guidelines

Theme: Economy and finance statistics

Print	ISBN 978-92-68-29885-5	ISSN series 2363-197X	doi:10.2785/9142603	KS-01-25-039-EN-C
PDF	ISBN 978-92-68-29884-8	ISSN series 2315-0815	doi:10.2785/2811750	KS-01-25-039-EN-N

# Contents

<b>Foreword</b> .....	<b>6</b>
<b>Acknowledgements</b> .....	<b>7</b>
<b>Executive summary</b> .....	<b>8</b>
Conceptual clarifications regarding the measurement of data.....	9
Recommendations associated with the compilation of data in the system of national accounts.....	9
<b>List of Abbreviations</b> .....	<b>12</b>
<b>Introduction</b> .....	<b>13</b>
The purpose and composition of this handbook.....	13
<b>1. The conceptual boundary of data in the 2025 System of National Accounts</b> .....	<b>14</b>
How does the 2025 SNA define data?.....	14
How the treatment of data in the 2025 SNA differs from that in 2008 SNA.....	15
The specific characteristics that data must exhibit to be considered a produced asset within the System of National Accounts.....	16
Separating the production of data from production of other intellectual property products.....	18
Computer Software, Artistic Original and Mineral exploration.....	18
Research and Development, and Artificial Intelligence.....	19
Is all data output capital formation? The treatment of short-lived data.....	21
The incorporation of data in product and industry classifications used in economic statistics.....	23
Central Product Classification (CPC).....	23
Industrial Classification of Economic Activity (ISIC).....	24
The production of data by all sectors across the economy.....	24
<b>2. Compiling a nominal estimate of Data output</b> .....	<b>27</b>
Introduction.....	27

<b>Outline of Sum-of-cost valuation approach</b> .....	<b>29</b>
How the sum of costs methodology is applied to data production.....	30
Detailing aspects of the sum of costs formula for data in more detail.....	31
<b>The choice of occupations in the compilation of own account data output</b> .....	<b>32</b>
Selection of occupations through expert knowledge.....	33
Systematic approach to identifying occupations.....	34
Calculating the Labor costs associated with these occupations ( <i>W</i> ).....	36
<b>The need for involvement rate adjustments and recommendation on such adjustments</b> .....	<b>37</b>
<b>Deriving the mark-up for non-labor cost</b> .....	<b>39</b>
<b>The use of information on market transactions of data</b> .....	<b>42</b>
Calculating an estimate of net imports of data for compiling estimates of GFCF.....	43
<b>Additional points of clarification when compiling a nominal estimate of data output</b> .....	<b>44</b>
The potential of data being consumed within one year .....	44
Reflecting the incorporation of new data points to existing data assets.....	45
<b>The production of quarterly estimates of data output</b> .....	<b>45</b>
<b>Annex 2.1. Summary of estimation methodology</b> .....	<b>48</b>
<b>Annex 2.2. Using natural language processing to better understand data intensity for certain occupations</b> .....	<b>50</b>
<b>Annex 2.3. Production of Special Internet Survey in Japan</b> .....	<b>52</b>
<b>Annex 2.4. Key words used for web scrapping job vacancies</b> .....	<b>55</b>
<b>Annex 2.5. Using the CROM for Selecting Data-Relevant Occupations and Computing Involvement Rates</b> .....	<b>57</b>
<b>Annex 2.6. Standard list of data producing occupations and involvement rates</b> .....	<b>62</b>
<b>Annex 2.7. Numerical example of standard non-labor cost mark up</b> .....	<b>66</b>
<b>Annex 2.8. Summary of recommendations for compiling nominal estimates of data</b> .....	<b>67</b>
<b>Annex 2.9. The treatment of transactions in data copies</b> .....	<b>71</b>
<b>3. Creating volume estimates of data</b> .....	<b>73</b>
<b>Introduction</b> .....	<b>73</b>
<b>Deflation options presented in the SNA</b> .....	<b>73</b>
<b>Additional consideration in identifying a price index for data measurement</b> .....	<b>74</b>
<b>Use of an established output price index covering similar product(s) or producing industries</b> .....	<b>76</b>
<b>Use of an input price index specific to data production</b> .....	<b>77</b>

Incorporating a quality adjustment to price index.....	78
Final considerations on the compilation of volume estimates .....	78
Annex 3.1. Volume estimates of data asset: The case of Pakistan.....	80
Annex 3.2. U.S. Bureau of Economic Analysis development of adjusted input price indexes for data and database assets.....	82
Annex 3.3. Summary of recommendations.....	83
<b>4. Creating Capital Stock estimates.....</b>	<b>85</b>
Introduction.....	85
Background and guidance on the PIM and compiling capital stock.....	85
Extending countries current approach to PIM parameters to Data assets .....	87
Estimating average asset lives of data.....	89
Annex 4.1. Capital stock estimates of data asset: Case of Pakistan.....	92
Annex 4.2. Obtaining information on asset life via surveys: the experience in Japan.....	94
Annex 4.3. Summary of recommendations.....	95
<b>5. Incorporating estimates of data production into existing national account outputs, including backcasting.....</b>	<b>96</b>
Introduction.....	96
Reporting of data within the national accounts.....	96
Backcasting.....	99
Annex 5.1. Summary of recommendations for incorporating estimates of data production into existing national account outputs .....	103
<b>References.....</b>	<b>105</b>

# Foreword

“Data” are often referred to as “the new oil”, and some even consider data as “the world’s most valuable resource”. Nonetheless, “data” were not explicitly included in national accounts until the recent endorsement of the 2025 System of National Accounts (2025 SNA), the global standard for the compilation of national accounts worldwide. This is a sign of how sudden and intense the digital revolution has been!

The inclusion of “Data” as a new class of produced assets is one of the most significant innovations in the 2025 SNA, expected to have a significant impact on the main macroeconomic indicators, including the level of gross domestic product (GDP) and gross national income (GNI).

This handbook, developed by an international task team jointly led by Eurostat and the International Monetary Fund, provides recommendations on data sources and methods for the compilation of data as an asset in national accounts. The handbook is at the same time groundbreaking and forward looking. With its two sets of “standard” and “advanced” recommendations, it allows all countries to produce reliable and comparable estimates of data assets with existing data sources, while providing indications for future development of more sophisticated approaches. The handbook is also an information document for users who seek to understand what is represented in the estimates of data output and capital formation of data.

This Eurostat-IMF Handbook on measuring data in the System of National Accounts is the result of a collegial effort from many experts from different world regions. We trust that it will facilitate a smooth and swift introduction of data as an asset in national accounts, thus enhancing the measurement of economic activities worldwide.

**Mariana Kotzeva**

Director-General  
Eurostat

**Bert Kroese**

Chief Statistician, Data Officer and Director of the Statistics Department  
International Monetary Fund

# Acknowledgements

This handbook has been developed by the Eurostat-IMF task team on Measuring Data as an Asset in National Accounts. The task team has been chaired by Erich Strassner (IMF). Andreas Dollt (Eurostat) has ensured the secretariat. John Mitchell, consultant to the IMF and Eurostat, has acted as the lead editor.

The following experts have contributed to the development of the handbook through their participation in the task team:

Kylia Wall (Australian Bureau of Statistics), Elisa Huber and Tobias Großbauer (Statistik Austria), Amanda Sinclair and Brenda Bugge (Statistics Canada), Sebastian Rebora (Banco Central de Chile), Dong Jiaxin and Yang Lin (National Bureau of Statistics of China), Cinthya Maria Marin Rojas, Gabriela Saborio Munoz and Mariana Carrillo Reyes (Banco Central de Costa Rica), Ralph Bøge Jensen (Statistics Denmark), Alejandro Donado, Daniel Seeger and Yannik Weigelt (Federal Statistical Office of Germany), Brian King and Kenneth Kennedy (Central Statistics Office - Ireland), Keisuke Yamagishi, Takeshi Sakuramoto and Yosuke Kono (Cabinet office of Japan), Nicky Kuijpers and Rick de Kruijf (Statistics Netherlands - CBS), Sindre Midttun and Steinar Todsén (Statistics Norway - SSB), Attiq-ur-Rehman and Liaqat Ali (Pakistan Bureau of Statistics), Chris Davies, Clíodhna Taylor, David Matthewson, Johnny Mudge, Laura Williams, Skevi Pericleous and Vigna Nanthakumar (UK Office for National Statistics - ONS), Jose Bayoán Santiago Calderon (USA Bureau of Economic Analysis - BEA), Nicola Massarelli (Eurostat), Eurydice Fotopoulou and Jennifer Ribarsky (International Monetary Fund), Pete Harper and Peter van de Ven (ISWGNA), Bram Edens, Jorrit Zwijnenburg and Ng Li Hiang (Organisation for Economic Co-operation and Development), Benson Sim (United Nations Statistics Division).

Eurostat and the IMF would like to thank all people that contributed by providing comments and suggestions for the improvement of this handbook.

# Executive summary

The production of data is a regular process for most companies and organizations across the economy. The digitization of so many facets of the economy has made the collection of facts and information cheaper and easier than ever before while at the same time providing ample evidence of the benefits that data can provide organizations. The production of massive amounts of data has created brand-new business models reliant on data, while increasing the risk that many traditional enterprises will be usurped by more productive and efficient, data focused producers.

While the impact of data on the economy is indisputable, it has only recently been explicitly recognized as an output of production and asset within the System of National Accounts (SNA), the internationally agreed standard on how to compile measures of economic activity.

The inclusion of data into the SNA production and asset boundary was formally endorsed at the 2024 UNSC (United Nations Statistical Commission, 2024) as part of the planned publication of the 2025 system of National Accounts (2025 SNA). However, the commission also “emphasized the importance of addressing conceptual uncertainties and stressed the importance of continuing to develop implementation guidance on the new recommendations to facilitate the implementation of the 2025 SNA in an internationally comparable way” (United Nations Statistical Commission, 2024).

In response, the Intersecretariat Working Group on National Accounts (ISWGNA) formed the Eurostat-IMF task team on Measuring Data as an Asset in National Accounts. The task team, made up of representatives from national statistical authorities and international organisations was tasked with providing “recommendations on how to estimate data as an asset in the national accounts by exploring data sources and methods. [In addition] the TT should develop concrete guidance for countries on how to compile the requested data in line with the 2025 SNA” (Task team on measuring data in the National Accounts, 2023). This handbook forms the primary output of the task team, and its publication coincides with the official endorsement of the 2025 SNA.

In [chapter 1](#), the handbook provides clarity around the definition of data for the purpose of the SNA. This includes clarifications on which types of data should be considered as an output of production, who is likely to be producing it, and where the line exists between the production of data and the production of other goods and services, including other Intellectual property products (IPPs). The clarifications are offered to provide insight to users to better comprehend the concept behind the data estimates, while also helping guide compilers to better understand what kind of information and source data are likely to be useful in the compilation of data and how it might be reported in the accounts.

[Chapter 2](#) details the methodology involved in compiling nominal estimates of data output and GFCF using the sum of cost methodology. It includes the standard source material and recommendations that have been gradually refined following testing by the task team. Unlike the clarifications included in [chapter 1](#), the recommendations in [chapter 2](#) (as well as the remainder of the handbook) offer both standard and advanced recommendations. The task team agreed that a standard set of compilation recommendations would greatly assist in promoting international comparability as countries began the task of compiling estimates of data output and data GFCF for the first time. At the same time, the inclusion of advanced recommendations reflects the task team’s desire to see the methodology continually developed and improved, so that users are able to have greater confidence in the accuracy of the estimates. Advanced recommendations also provide countries the ability to use already available source material and compilation practices, including those beyond what is required in the standard recommendations.

Chapters 3-5 cover compilation requirements such as the deflation of nominal estimates, the creation of depreciation and capital stock estimates through the perpetual inventory method (PIM), recommendations on backcasting and other compilation challenges. Similar to chapter 2, chapters 3-5 include both standard and advanced recommendations.

Prior to the formation of the Eurostat-IMF task team on Measuring Data as an Asset in National Accounts, several countries had already compiled experimental estimates of data investment. While this work formed the basis of the discussion, additional testing, source data gathering and research by task team members refined the recommendations to arrive at a final set of conceptual clarifications and compilation recommendations. A summary of these is presented below.

In summary, this handbook has been written to serve two main purposes. First, it is a compilation guide to assist economic statisticians with compiling outputs consistent with the 2025 SNA. Second, it is an information document for users who seek to understand what is represented in the estimates of data output and GFCF of data being produced by statistical offices.

To that end, the handbook contains case studies from countries that participated in the task team which often provided the basis for the advanced recommendations. While these case studies offer real world reference material which other countries can learn from, they should be viewed as just the initial step in the continual improvement of data estimates.

## Conceptual clarifications regarding the measurement of data

**Definition of data:** Information content that is produced by accessing and observing phenomena, recording and storing information elements from these phenomena in a digital format and that provides an economic benefit when used in productive activities.

**Treatment of non-digital data:** For the purpose of the 2025 SNA, only digital data is considered within the 2025 SNA production and asset boundary.

**Treatment of Ancillary data:** If the data is not providing a direct economic benefit to the entity, it is considered outside of the 2025 SNA production and asset boundary.

**Reporting of combined data and databases asset:** While data and databases are conceptually different assets, both assets are usually produced or purchased together. Therefore, it is practically difficult to compile separate estimates for each. As such, for reporting purposes, data and databases are combined into a single detailed intellectual property (IP) product called data and databases.

**Separately identifying data even when used in the production of other IPP:** It is recommended that the value of produced data is separately identified when capitalized. The value should be excluded from the cost of own account production of other goods and services regardless of how data dependent the final output is. Two exceptions exist, including data integral to the production of Research and development and data assembled in a database created solely as a step in the production of an AI computer program and cannot be re-used. In these situations, the value of the produced data should be included in the value of the final Research and Development or AI (computer software) product.

**The production of data across all sectors of the economy:** Data can be produced and used in production by all sectors of the economy including the non-market sector.

**The limit to expenditure associated with data production:** The incorporation of additional information content or improving the data's quality at either a granular or aggregate level is considered expenditure on data production. Analysis of the data to obtain insights or using the information contained in the data in productive activities is considered the production of a good or service other than data.

## Recommendations associated with the compilation of data in the system of national accounts

**Valuation approach:** Data produced on an own account basis is valued using the already established sum-of-cost methodology, starting from the labor costs related to occupations involved in the creation of data.

**Choice of occupations:** In the absence of other sources, national statistical authorities (NSAs) should use the list of occupations provided in this handbook for the compilation of data output. If the standard list of occupations are used, these must be complimented by the same or very similar involvement rates to those listed in this handbook. If possible, NSAs are encouraged to derive a list of occupations (and involvement rates) through an objective and systematic approach to better determine which occupations are most likely to be involved in data production. Importantly, occupations should be considered for the list if the occupation involves tasks which explicitly contribute to adding value to the production of data and the worker undertakes these tasks as an integral part of their role.

**Involvement rates:** NSAs are recommended to apply the same or very similar involvement rates to those listed in this handbook in order to determine the degree to which each occupation is involved in the creation of data. The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data. An advanced recommendation is for NSAs to develop and use involvement rates (and list of occupations) derived through a more systematic and objective means specific to their economy.

**Non-labor costs:** It's recommended that non-labor costs are incorporated into the final estimate via a single mark-up applied to labor costs. Such a mark-up represents the costs of inputs, any adjustment associated with taxes less subsidies on production, the depreciation of fixed capital used in production, as well as a return to capital (operating surplus). A single mark-up, based on the ratio of total gross output applied against remuneration of employees from the "Computer programming, consultancy and related activities" (ISIC 62) and "Information service activities" (ISIC 63) – or similar available aggregate - is applied to total labor costs. An advanced recommendation is for NSAs to apply multiple mark-ups to the estimate separately so that differences in depreciation and operating surplus across occupations and industries can be applied more accurately and transparently.

**Adjustment for short lived data:** When feasible, countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment. If such an adjustment is made, countries are encouraged to publish (perhaps as a memorandum item) information on the level of data treated as short lived in order to improve the comparability across countries. If information on short lived data is unavailable, the standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalised expense and should be classified as GFCF, with no adjustment made to represent short lived data. For the purpose of estimating capital stocks based on a perpetual inventory method (PIM), NSAs should select a service life that reflects whether an adjustment to exclude short lived data has been made to the nominal estimate.

**Market transactions:** Data assets that are purchased with exclusive rights are treated as an outright sale of the data asset by the seller to the buyer. However, assuming that the transaction is not a cross-border or inter-sector one, similar to sales of other assets, this transaction would net off and not impact the overall level of GFCF for a specific sector. Data assets that are purchased without exclusive rights are treated as a purchase of a copy and contributes to the GFCF of the purchaser (and aggregate GFCF) if it satisfies the necessary conditions of GFCF, (i.e. use in production for more than one year).

**Price index used:** Any price index used to deflate nominal estimates of data must reflect the price change observed in both the labor and non-labor costs involved in data production. In addition, its recommended that the price index appropriately accounts for the technological and quality improvements that have been observed in the production of digital products over the past several years. As a standard position, compilation can be undertaken using an output price index based on an alternative but similar product. The advanced recommendation is to compile chain volume estimates of data output using a pseudo output price index. This can be created by aggregating appropriate input price indexes and weighted to reflect the actual input costs included in the sum of cost calculation. An adjustment to reflect quality and productivity improvements made to the final output would be added to transform the input price index into a pseudo output price index.

**Quality adjustment applied to price index:** It is recommended that an aggregate adjustment to reflect quality and productivity improvements is made to the price index used. Such an adjustment may be calculated as

- The difference in growth between the input price index for data and the output price index for a similarly produced product where market prices are available.
- The growth between the calculated difference in Input price and output price index for similarly produced products where market prices are available.
- The total factor productivity estimates for industries that contain a large amount of the occupation identified as data producers.

**Additional deflation consideration:** *The intention of any recommendation in this handbook is not to overrule any existing regulations, rather the handbooks' goal is to assist countries compile the most accurate estimates of data output possible. It is the view of the task team that the introduction of a quality adjustment on top of an input price index is conceptually appropriate and would improve the accuracy of the final estimate. It is accepted that countries will continue to adhere to other frameworks and standards that oversee the compilation of their national accounts.*

**Parameters used in PIM (Excluding Average service life):** *Countries should apply the same parameters in the compilation of depreciation and net capital stock of data as applied currently to other IPPs. However, countries should aspire to continually collect additional information on different assumptions and parameters to refine and improve the estimates of depreciation and capital stock being compiled.*

**Average service life:** *In the absence of other information, countries should apply as standard, an average service life not longer than 5 years for data assets, this chosen service life reflects the likely inclusion of short-lived data. Any service life chosen should reflect whether an adjustment to exclude short lived data has been made to the nominal estimate. Ideally, countries should aspire to break up the estimate of data investment by industry in order to allow for different service lives to be applied based on the industry producing the data.*

**Backcasting:** *In the absence of additional information to the contrary, this handbook recommends incorporating a time series for data investment to at least the period covering 1995 – 2005. A standard recommendation is to use an already available series that displays a correlation with the growth observed in the compiled annual estimates of data output and GFCF to populate the back series. The newly constructed back series should not mirror the indicator series exactly as the end points are likely different between the two. Rather a decision should be made as to when the production and GFCF of data should commence and the indicator series used to 'guide' the back series to this point. If the required data sources are available, an advanced recommendation is to compile a back series via the bottom-up approach. However, adjustments should be made to the chosen occupations and involvement rates over time to reflect the evolution of data production over the previous periods.*

# List of Abbreviations

<b>ABS</b>	Australian Bureau of Statistics
<b>AI</b>	Artificial Intelligence
<b>BEA</b>	United States Bureau of Economic Analysis
<b>CFC</b>	Consumption of fixed capital
<b>CPC</b>	Central Product Classification
<b>CPI</b>	Consumer Price Index
<b>CROM</b>	Competence-Relevant Occupation Methodology
<b>DBMS</b>	Database management system
<b>ESA</b>	European system of accounts
<b>ESRI</b>	Economic and Social Research Institute
<b>ESCO</b>	European classification of Skills, Competences, Qualifications and Occupations
<b>FSO</b>	The Federal Statistical Office of Germany
<b>GDP</b>	Gross domestic product
<b>GFCF</b>	Gross Fixed Capital Formation
<b>IFRS</b>	International Financial Reporting Standards
<b>IMF</b>	International Monetary Fund
<b>IP</b>	Internet Protocol
<b>IPD</b>	Implicit Price Deflator
<b>IPP</b>	Intellectual Property Products
<b>ISCO</b>	International Standard Classification of Occupations
<b>ISIC</b>	International Standard Industrial Classification
<b>NAICS</b>	North American Industry Classification System
<b>NLP</b>	Natural language processing
<b>NSA</b>	National Statistical Authorities
<b>OECD</b>	Organisation of Economic Co-operation and Development
<b>OP</b>	Observable Phenomena
<b>PIM</b>	Perpetual Inventory Method
<b>PBS</b>	Pakistan Bureau of Statistics
<b>R &amp; D</b>	Research and development
<b>ROE</b>	Remuneration of employees
<b>SNA</b>	System of National Accounts
<b>UNSD</b>	United Nations Statistical Division

# Introduction

## The purpose and composition of this handbook

- I. This handbook has been written to serve two main purposes. First, it is a compilation guide to assist economic statisticians compile outputs related to the production of data that are consistent with the SNA. With data included as an explicit asset within the 2025 SNA production and asset boundary, countries will be required to compile estimates of data production and use in their own economies. This handbook assists countries achieve this.
- II. The second purpose of this handbook is to provide information to users who seek to understand what is represented in the estimates of data output and gross fixed capital formation (GFCF) of data being produced by statistical offices. As well as providing conceptual clarity, the handbook can help users better understand why the measurement of data is undertaken in the specific manner that it is.
- III. To achieve both these aims the handbook will provide clear recommendations which countries can follow in order to produce estimates of the production and stock value of data. This allows for a consistent baseline methodology to be introduced by all countries provided that a minimum set of source data is available. Discussions and presentations of the practical compilation recommendations begin from [chapter 2](#). However, in [Annex 1.1](#) of this chapter, there is a summary of the recommended sum of cost methodology that countries should aim to follow in the construction of estimates. The handbook will also include case studies by countries of their compilation methodology, some of which go beyond the baseline or standard recommendations. As well as offering real world reference material which other countries can learn from, these case studies provide a foundation for the overall continual improvement of data estimates, leading to increased confidence in the estimates by users.
- IV. [Chapter 1](#) of this handbook will explore the conceptual boundaries of data for the purpose of measurement within the confines of the SNA. [Chapter 2](#) focuses on the compilation of a nominal estimate of data, introducing the sum of cost methodology for estimating own account production of data. As will be further covered, for a range of reasons, this is seen as the most important perspective in ensuring that a comprehensive but internationally comparable estimate of data output is compiled. It presents the various inputs required to successfully implement the methodology while also covering additional compilation aspects such as the production of quarterly estimates.
- V. [Chapter 3](#) covers the deflation of the nominal estimates, to produce volume estimate of data production consistent for inclusion in headline aggregates of GDP and GFCF. [Chapter 4](#) discusses the use of the Perpetual Inventory Method (PIM) for deriving estimates of capital stock and depreciation. [Chapter 5](#) presents remaining compilation challenges, including backcasting data to incorporate the estimates into existing aggregates.

# 1

## The conceptual boundary of data in the 2025 System of National Accounts

### How does the 2025 SNA define data?

- 1.1. The production of data is a fundamental process for most companies and organisations across the economy. The digitisation of so many facets of the economy has made the collection of facts and information cheaper and easier than ever before while at the same time providing ample evidence of the benefits that data can provide. Therefore, while the fundamental practice of observing and recording information is not a new business process, digitalisation has created a substantial increase in the information that can be collected and used, as well as expanding the number of companies and organizations actively involved. This transition has resulted in an unprecedented amount of data being produced and used in the modern economy. The production of massive amounts of data has created brand-new business models reliant on data, while increasing the risk that many traditional enterprises will be usurped by more productive and efficient, data focussed producers.
- 1.2. Although much has been written about data, including describing it as being like oil, sunshine, or water, among countless other descriptions, and while its impact on the economy is indisputable, it is largely absent in economic statistics, including previous iterations of the System of National Accounts (SNA). To ensure that the SNA is reflective of the modern economy, as well as properly understanding the benefits that data can provide, data needs to be appropriately measured and incorporated to existing standards. A first step in this process is defining exactly what is meant by “data.”
- 1.3. Due to the many possible understandings of what data is, or is not, and potential for misinterpretation, for the purpose of economic measurement a detailed and extensive definition is required. Within this handbook and the 2025 System of National Accounts (2025 SNA) data is considered as **“information content that is produced by accessing and observing phenomena, recording, and storing information elements from these phenomena in a digital format and that provides an economic benefit when used in productive activities”** (2025 SNA §22.22).
- 1.4. Such a technical definition differs from the perspective of the proverbial “person in the street.” For many, data is a simpler concept even though it can refer to many different things. In fact, when used by most people, the term data is broadly indistinguishable from ‘information’ and can cover anything ranging from a single fact or point of knowledge up to large datasets from which numerous insights can be drawn from. These perspectives are not wrong and are not necessarily inconsistent with the SNA definition provided, which includes describing data as information content. Data, from the SNA perspective can be produced on a single item (the personal information of an individual) or on whole economies (the GDP for an entire country) lending further alignment with the common understanding of data. Where the two perspectives differ is the additional caveats regarding being produced in a digital format and providing an economic benefit. As will be discussed further below, these two caveats are added to ensure both consistency with the existing treatment of assets in the SNA as well as making a concept of data which is feasible to measure for statistical

compilers. However, it is important to acknowledge that different terminology or definition may need to be used when obtaining information on data from enterprises. This is discussed in [Box 2.4](#).

## How the treatment of data in the 2025 SNA differs from that in 2008 SNA

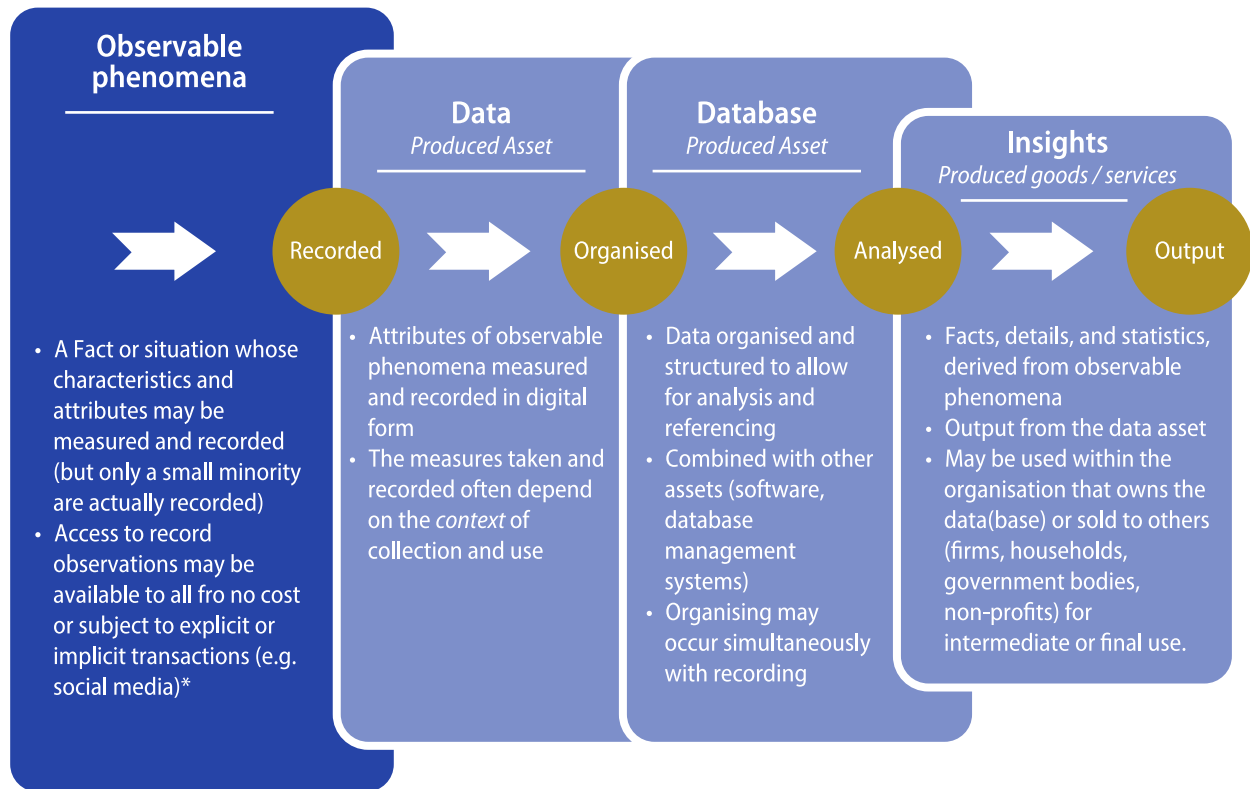
- 1.5. Data was not explicitly defined or categorized as an asset in the 2008 System of National Accounts (2008 SNA). It has been argued that within the 2008 SNA, data, if included at all, was implicitly included in estimates of purchased goodwill (Calderón & Rassier, 2022) <sup>(1)</sup> making it a non-produced asset. However, this treatment reflected the slightly different viewpoint of how the term data was interpreted to those tasked with drafting the 2008 SNA.
- 1.6. The discussions leading up to the 2008 SNA viewed data simply as information (Ahmad, 2005), meaning that ‘data’ existed even before it had been accessed, recorded and stored. This interpretation of data included the embodied information content of what is now typically referred to in the new lexicon of data value chains as the information content of ‘observations’ or ‘observable phenomena’ (OPs) <sup>(2)</sup>. This consideration of data as simply embedded information is one reason why the authors of the 2008 SNA, in an attempt to limit the possibility of implicitly “capitalising knowledge” (Ahmad & van de Ven, 2018), chose to limit the value of databases to only include the cost of preparing data in a format that conforms to the “database management system (DBMS)” while excluding the cost of acquiring or producing the data (2008 SNA §10.113).
- 1.7. The 2025 SNA considers data the information content that comes from accessing and observing phenomena, recording, and storing information elements from these phenomena rather than the embedded information contained in the phenomena themselves. In simple terms, the concept of data, for the purpose of economic measurement and analysis has moved from the first to the second box in the data value chain represented in [Figure 1.1](#) <sup>(3)</sup>. This movement introduces a clear element of production (as defined in the SNA) <sup>(4)</sup> to the process as well as a clear line where non-produced information embodied in non-produced facts and situations (regarded as data in 2008 SNA) can be transformed into information content, produced by accessing, recording, and storing information elements from observable phenomena (regarded as data in 2025 SNA).

<sup>(1)</sup> It is considered “implicitly” as data was not explicitly considered a non-produced asset. Rather when an explicit transaction in goodwill was made, it was considered that some of the value of the goodwill (a non-produced asset) was derived from the value of the data contained within business.

<sup>(2)</sup> In Mitchell, Ker & Leshner (2022) Observable Phenomena (OPs) are defined as “a fact or situation, whose characteristics and attributes may be recorded”.

<sup>(3)</sup> This is visibly represented in [Figure 1.1](#) by the movement from the dark blue box representing items that are non-produced into the lighter coloured blue box representing output from production.

<sup>(4)</sup> The technical definition of production is ‘an activity, carried out under the responsibility, control and management of an institutional unit, which uses inputs of labor, capital, and goods and services to produce outputs of goods and services’. In this case the output is the data product, while the input is the resource and labor used to access, observe, and record the information elements.

**FIGURE 1.1****Data-information chain from a System of National Accounts perspective**

(Mitchell, Ker, & Leshner, 2022).

- 1.8. [Figure 1.1](#) separates data from databases which accurately reflects the 2025 SNA, as the standard has explicitly identified and defined data as a separate asset from the complementary database asset that already exist within the 2008 SNA production boundary. However, the two assets, Data and Database are still unequivocally intertwined. So much so that there is limited expectation that data and databases will be reported separately. Paragraph 22.25 of SNA 2025 make this point quite clear by stating “*despite their conceptual difference, data and databases are difficult to measure separately because they are produced with similar inputs and because transactions prices generally reflect the combined value of the database and the data. For reporting purposes, data and databases are therefore combined into a single detailed intellectual property product (IPP) called data and databases*” (2025 SNA §22.25).
- 1.9. This handbook recommends that **within the SNA, outputs and investments (GFCF) in data should be reported together with databases as a single IP product, however, this should be reported separately from computer software**. This separation of software presents a difference from current reporting based on the 2008 SNA. This challenge as well as the overall reporting of data will be further discussed in [chapter 5](#).

## The specific characteristics that data must exhibit to be considered a produced asset within the System of National Accounts

- 1.10. The addition of the caveats regarding the digital nature of data and the provision of economic benefit allows for data to be conceptualised and measured in a way that is not only feasible for statistical offices but also consistent with other fixed assets in the SNA. The SNA considers produced fixed assets to be “*assets that have come into existence as outputs*”

*from production processes" (2025 SNA §11.11) and that "are used repeatedly or continuously in production processes for more than one year. The distinguishing feature of a fixed asset is not that it is durable in some physical sense, but that it may be used repeatedly or continuously in production over a long period of time, which is taken to be more than one year." (2025 SNA 11.13).*

- 1.11. In today's economy, there are countless examples of data being created as an output of a production process and subsequently being used in business processes '*repeatedly or continuously*' over a period of one year. Simple examples include sales data to assist with forecasting demand, customer information as part of loyalty programs, or cookies collected from websites and used to personalize your searching experience. As such, the suggestion that data is used in the economy, and it should, for the purpose of the SNA, be considered a produced asset, has been widely supported.
- 1.12. However, it has also been acknowledged that there exists data, which due to either the way it is created or used, may not meet the concept of a produced output as defined by the SNA. To compensate for this, the 2025 SNA definition of data brings in several aspects that must be fulfilled for data to be considered an output of production and as an asset. This includes being on a digital format and providing an economic benefit to the owner.
- 1.13. It is well established that non-digital data exists and has been used previously in production. However, within the modern economy, due to the greater resources required to process and transfer non-digital data, it is considered that this would make up a small and inconsequential amount of the overall data used in production. Furthermore, the inclusion of non-digital data is considered a significant measurement burden for countries and not commensurate with the influence of this data on the economy. **Therefore, while acknowledging that non-digital data exists, for the purpose of the 2025 SNA, only digital data is considered within the 2025 SNA production and asset boundary.**
- 1.14. In addition, most businesses generate data that is not directly relevant to the primary production of the business. Although this ancillary data may be captured digitally, **if the data is not providing a direct economic benefit to the entity, it is considered outside of the 2025 SNA production and asset boundary.** It is not practical to explicitly list what type of data this might entail as it will be different from business to business and may change over time. However, if the data is not used to derive insights or information which may increase the level or efficiency of production (a.k.a., providing an economic benefit), then any costs associated with its generation should be considered as a current input cost of production of the underlying output of the entity rather than a capital cost (GFCF of data).
- 1.15. A delineation between the generation of these different types of data (produced data providing an economic benefit and data that is considered ancillary) can be implied through the occupations chosen as contributing to the sum of cost of data production. While elaborated further in [chapter 2](#), it is recommended that **occupations should be considered as data producing if the occupation involves tasks which explicitly contribute to adding value to the production of data and the worker undertakes these tasks as an integral part of their role.** Such a recommendation combined with the general theory of expected return on investment suggests that any occupations specifically tasked with producing data are doing so with the intent that the data produced will be utilised in productive activities and thus should be included.
- 1.16. This assumption extends to the non-market sector. **Like all other assets in the economy, data can be produced and used in production by all sectors of the economy including the non-market sector.** For example, the production of data is often done to improve the mechanisms of a government's taxation or social security systems. The workers involved in occupations producing this data are not undertaking their work to generate a direct financial return, but rather this is an example of the government undertaking data activities to conduct their production (government final consumption expenditure) in an efficient manner.
- 1.17. Overall, this delineation is not always clear cut. It is possible that data may have been kept for record keeping originally, only for businesses to later realize its value. For this reason, the list of data producing occupations and each occupation's involvement rate (see [chapter 2](#)) should be regularly reviewed to ensure that it reflects the current approach to data production.
- 1.18. While the fundamental concept of excluding ancillary data is not considered controversial and, in fact, merely brings data in line with other assets included within the 2025 SNA asset boundary, the practical implementation is

challenging. It is not feasible to classify data as productive or ancillary based on the actual type of data produced due to it being estimated at an aggregated level. However, a delineation based on the type of occupation creating the data is considered more achievable.

## Separating the production of data from production of other intellectual property products

- 1.19. The introduction of a new asset class, predominately measured through the sum of cost methodology brings inherent risks of double counting. These risks are not new, in fact, in 2010, following the introduction of R & D into the asset boundary the OECD cautioned compilers that it was *“important to ensure, in using the sum of costs approach to valuating of IPP assets produced on own account, that the same costs are not included in the valuation of more than one asset”* (OECD, 2010). Such advice remains fundamental today, albeit with additional IPP assets to consider. The following paragraphs will cover the conceptual differences between the IPP categories, while [chapter 2](#) will further discuss the avoidance of double counting.
- 1.20. When describing the conceptual boundary of data, it is worth clearly stating that just because something is stored on a computer does not mean it should automatically be considered data as defined in the 2025 SNA. As pointed out by the OECD, the concept of data often takes on a quantity perspective when it refers to *“Internet Protocol (IP) traffic or the volume of digitised information stored on servers and other hardware”* (OECD, 2022). As such, additional clarifications are provided to separate data, as defined in the 2025 SNA from other intangible assets.
- 1.21. Most digital content such as videos, photos, emails will likely not only fail to meet the characteristics of data they are likely to not even be considered a fixed asset as defined in the SNA (*used repeatedly or continuously in production processes for more than one year*)<sup>(5)</sup>. Those that do meet the asset test may fall into other, already existing SNA intangible asset classes. These include artistic originals, computer software, mineral exploration and evaluation, and research and development. Since these intangible assets usually generate their value from the information content contained in them and are almost always stored in a digital manner (and for some it’s a necessity) they do share many conceptual similarities with data. What distinguishes them from data is how they are produced and the type of information content they possess.
- 1.22. It should be noted that this handbook will not discuss separating data from the existing IPP of databases. As covered earlier in the chapter despite their conceptual difference it is expected that the compilation of estimates of data production and database production will be done in combination and subsequently reported as a single IPP category of *data and databases*.

## Computer Software, Artistic Original and Mineral exploration

- 1.23. In order to separate these intangible assets, one must look at the different characteristics of the information that each contains including the source of this content. Computer software *“consists of computer programs, program descriptions and supporting materials for both systems and applications software”* (2025 SNA 11.112). Artistic originals are *“Entertainment, literary and artistic originals consist of the original films, sound recordings, manuscripts, tapes, models, etc., on which drama performances, radio and television programming, musical performances, sporting events, literary and artistic output, etc., are recorded or embodied”* (2025 SNA 11.119). The information content of both assets differs greatly from that of data produced by *‘accessing and observing phenomena; recording and storing information elements from these phenomena in a digital format.’* The information contained in computer software has a relatively narrow remit, focused on ensuring the correct running of computer programs on appropriate hardware. While the information contained within artistic originals can vary greatly, the fact that this information may be stored digitally, does not make it data per

<sup>(5)</sup> While there is no explicit rule stating that data cannot take the form of audio and visual files, it is thought that the vast majority of data will be alphanumeric. However, the changing use of data may impact this in the future, as AI models increasingly use audio and visual data for training. This is particularly evident in applications like computer vision, where image and video data are essential, and in speech recognition systems, which rely heavily on audio data.

se (as per the 2025 SNA definition). It has been well established that a key requirement for artistic originals is that the work should have ‘primary artistic intent’ (OECD, 2010). Information content in data can extend well beyond a set of instructions in a programming language or artistic endeavor. Furthermore, the usually systematic manner in which data is produced contrasts to the thoughtful or creative production of computer software and artistic originals. Therefore, while data shares some similar conceptual ground with software and artistic originals, there is still clear conceptual difference between the asset categories.

- 1.24. Software can be created for the specific purpose of producing data, or the ability to produce data can be a significant consideration in the development of certain software. As discussed in [chapter 5](#), the purpose of the software does not remove its fundamental nature and classification and therefore, expenditure on, and production of, software should remain as such regardless of its level of contribution to data production.
- 1.25. Mineral exploration and evaluation contain information content on the “*exploration for petroleum and natural gas and for non-petroleum deposits and subsequent evaluation of the discoveries made*” (2025SNA §11.108). Producing data on mineral deposits could be considered a subset of data production. However, the information elements collected as part of mineral exploration cover a very specific topic with a very specific use, meaning that the phenomena searched for and observed to produce the information are also specific. Thus, to maintain consistency with the existing assets already covered in the SNA, expenditure on producing data on petroleum and natural gas and non-petroleum deposits should remain as mineral exploration and evaluation <sup>(6)</sup>. Additionally, there remains a large amount of expenditure unrelated to data production that still contributes to the estimate of mineral exploration and evaluation.

## Research and Development, and Artificial Intelligence

- 1.26. 1.16. The line between data, and research and development (R & D) is not quite as clear cut as other IPP. Fundamentally, R & D includes *expenditure on creative work undertaken on a systematic basis in order to increase the stock of knowledge* (2025SNA §11.105). A key component of this research aimed at increasing knowledge is generating information content by recording information elements from observable phenomena (often undertaken in controlled circumstances). Such activities could include data production (as defined in the 2025 SNA).
- 1.27. However as outlined in the Frascati manual (OECD, 2015) not all data generated by organizations is automatically considered as expenditure on R & D. Rather to qualify as R & D expenditure, the work must satisfy five core criteria <sup>(7)</sup>, several of which can serve as a means to delineate between the two asset classes. The first is the notion of creative and/or novel work. While some methods of data production (particularly the collection aspect) should certainly be considered as innovative (or even creative) most data production is undertaken for a well-established specific purpose (creating economic benefits) rather than a novel one. Did sales go up or down? How many people accessed the website, how long was the package in the depot before being delivered? Expenditure related to obtaining information on this kind of routine business-related questions shows that the data is being collected for a specific purpose, unrelated to R & D and are thus, excluded from R & D.
- 1.28. Additionally, but just as important, expenditure on ‘general data collection’ is also excluded from the measurement of R & D unless the data is being collected *‘solely or primarily for the purpose of R & D’* (OECD, 2015). Most data produced by both market and non-market organizations is unlikely to fulfil this criterion as the information is collected as a by-product of “routine” production rather than for the specific purpose of expanding the ‘stock of knowledge.’ This is not to say that accurate and organized data on production and consumers are not beneficial to the company. They certainly are, which is why businesses invest in ensuring they can gain insights from them and why the SNA considers them an asset. However, the purpose of this data production is considered different to the purpose of R & D.
- 1.29. The exclusion of general data collection from R & D fits with the newly proposed framework of measuring data assets. The conceptual boundaries of data as defined for the 2025 SNA specify that once the information content (obtained through accessing, observing, and recording information elements from OPs) is fit for use in productive activity, the

<sup>(6)</sup> This situation already exists in the SNA, in that mineral and petroleum exploration is arguably a subset of R & D, however, since it covers a very specific area of research, it is separately identified and classified.

<sup>(7)</sup> The five criteria are that the activity must be: novel, creative, uncertain, systematic, and transferable and/or reproducible (OECD, 2015).

production of the data asset is finished. This data asset could theoretically be used for R & D, in which case the expenditure on doing this (i.e., analyzing or testing the data to gain insights) would be capitalized <sup>(8)</sup>, or it could be used as an input to production of another good or service. Either way, while the **incorporation of additional information content or improving the data's quality at either a granular or aggregate level is considered expenditure on data production, analysis of the data to obtain insights or using the information contained in the data in productive activities is considered the production of output of a good or service other than data** <sup>(9)</sup>. This conceptual treatment is an important consideration when choosing occupations as outlined in [chapter 2](#), for instance occupations that heavily *use* data may not contribute significantly to *producing* data.

- 1.30. A reason for having this conceptual endpoint in data production is due to a desire to separately identify the production of data from the output that the data asset is contributing to. A significant driver for the explicit inclusion of data in the 2025 SNA was that expenditure contributing to an input that is used repeatedly in the production of other goods and services should be identified and treated appropriately (i.e., capitalized). Since most data is produced with a particular final use in mind, the 2008 SNA did not explicitly identify much of this expenditure and simply recorded this as intermediate costs in the production of another product.
- 1.31. This treatment of separately identifying and capitalizing data should be applied regardless of if the data dependent product is subsequently consumed (such as advertising, logistics, finance, retail) or capitalized itself (production of other IPP).
- 1.32. Two areas where conceptual cross-over may exist is the production of data 'integral to the production of Research and development' or 'data assembled in a database created solely as a step in the production of an AI computer program and that cannot be re-used.'
- 1.33. Since 'general data collection' is excluded from the R & D, and for the various reasons outlined earlier, much of the data produced in the economy does not meet the characteristics of R & D, it is not expected that expenditure on data 'integral to R & D' is significant. However, since this caveat already exists within the well accepted international manual on R & D statistics, the Frascati manual (OECD, 2015), it is prudent to include it also in the guidance for measuring data output.
- 1.34. Like data, Artificial Intelligence (AI) is being explicitly identified for the first time in the 2025 SNA. In the framework, it is envisioned to separately identify the production of AI software as a subset of computer software <sup>(10)</sup>. It is well established that the production of AI is fundamentally dependent on the generation of data that the AI algorithms can use to train and learn. This data can be sourced from various locations, ranging from publicly available free data to datasets design to specifically train the software to undertake certain functions.
- 1.35. The 2025 SNA recognizes the important role that data plays in production of AI software, noting that "data is often acquired and organized in a database for the specific purpose of training an AI software program. In addition, AI programs often use data to generate their output" (2025 SNA §22.36). The SNA acknowledges the point previously made in the handbook that it is preferable to separately identify the production of data from the output that the data asset is contributing to. As such, it suggests that "the value of the data used to train an AI software product or to help AI software to generate its output should be recorded separately from the value of AI software, as the data could have multiple uses" (2025 SNA §22.36). However, it also notes that "data assembled in a database created solely as a step in the production of an AI computer program and that cannot be re-used may be included in the costs of producing AI programs" (2025 SNA §22.36). This tacitly acknowledges that the single use nature of the data limits its value as a 'data asset' while maintaining its value from the perspective of AI computer software production.
- 1.36. In light of the recommendation, a particular question that will impact the respective estimates of each is how much AI is produced based on already existing data sources compared to specifically assembled data, "used solely as a step in

<sup>(8)</sup> Conceptually, in such a circumstance, the value of the own account R & D, measured via the sum of costs, would need to include a depreciation cost associated with the data asset, being used as an input.

<sup>(9)</sup> While this statement is true for many examples, it is conceivable that the original data asset could be used to produce additional data output, however this would still not involve analysing or using the data in productive activities.

<sup>(10)</sup> The SNA is defining AI as 'capabilities of a computer program, or system controlled by a computer program, of recognition, reasoning, communication, and prediction emulating human recognition, reasoning, and communication' (2025 SNA §22.35).

the production of AI" (2025 SNA §22.36). Due to its relatively recent emergence, the production of AI is an area that will require more research.

- 1.37. Therefore, **it is recommended that the value of produced data is separately identified when capitalized. The value should be excluded from the cost of own account production of other goods and services regardless of how data dependent the final output is. Two exceptions exist, including data integral to the production of Research and development and data assembled in a database created solely as a step in the production of an AI computer program and cannot be re-used. In these situations, the value of the produced data should be included in the value of the final Research and Development or AI (computer software) product<sup>(1)</sup>.** Ideally, countries should investigate to see if a noticeable amount of expenditure related to "single use" data is included in the sum-of-cost method when estimating own account production of R & D and the computer software subset of Artificial intelligence. Existing practices related to compilation of own account R & D and software will need to be reviewed to ensure that the same expenditure is not recorded twice and classified to the own account production of both assets.
- 1.38. It is envisioned that in the next update of the Frascati manual, this inclusion of data collection considered integral to R & D may be removed entirely from the R & D asset boundary reflecting the fact that a standalone data and database asset category is now established in the SNA.

## Is all data output capital formation? The treatment of short-lived data

- 1.39. The concept that data assets exist and are being used in the modern economy has been strongly and broadly agreed to. Where there is less certainty is what proportion of data output is actually fulfilling the characteristics of a fixed asset as outlined in the SNA, often referred to as the capitalization rate of data. While a large amount of this handbook will focus on the idea of measuring data assets used in production, fundamentally, the inclusion of data in the SNA begins with measuring the output of data. Since output is simply goods and services produced by an establishment, data is first an output of production, which can then (and often will) be considered a fixed asset. The SNA is clear that expenditure on products (or in the case of own account production - own account output) should only be capitalized if the product produced is involved '*consistently and repeatedly*' in production for more than one year.
- 1.40. Within current compilation practices, most expenditure is earmarked as either capital formation or not based on the type of product being purchased or produced. A business buying a car is considered to undertake 100% capital formation, while expenses related to a business buying a pencil is considered a current expense with 0% capitalized. That is not to say that a car will automatically have a service life of greater than a year, or that a pencil will be fully consumed within a year. Rather, realistic, and informed assumptions are made based on the normal characteristics of the inputs.
- 1.41. These informed assumptions seem to be applied consistently by the SNA and international corporate accounting standards (International Financial Reporting Standards (IFRS)). Assets classified as such in the SNA (Structures, Software, Machinery, and equipment) are also considered assets within the accounting standards. This makes collection of the expenditure through business surveys a relatively easy task as the expenditures have already been separated in the company accounts. This is not (yet) the case for data expenditure which, while not explicitly excluded, is not (yet) considered an asset within the international accounting standard. This in turn suggests that such data expenditure is included in current expenditure in business accounting.
- 1.42. Currently this is somewhat inconsequential as compilers have focused on producing estimates of data investment at the aggregate level (a.k.a. the supply side), using existing data sources other than business surveys. While this will hopefully change in the future, it currently leaves the decision on the exact capitalization rate of data in the hands of statistical compilers. As such, several countries in their initial estimates of data GFCF have made an adjustment to the final nominal estimate of own account output of data to represent the data that is consumed within one year and thus should not be capitalized. However, as noted by the NSAs who have made these adjustments, they are, at the moment,

<sup>(1)</sup> The costs associated with using the data assets will feed into the capital service costs associated with the own account production costs of other goods and services.

considered quite arbitrary as information on the percentage of data used within one year is so far not readily available from businesses or any other source.

- 1.43. There is general agreement that the long-term trend of data collection and use has shifted from one entirely based around stable stocks of digital information – databases of names and other well-defined personal and business data – to one which is more about real-time flows of often unstructured data (Eurostat-OECD, 2019). This has meant that a large amount of the value placed on data is driven by the time sensitive nature of data. Such a phenomena would mean that an increasing amount of data is used quickly and – importantly – only once in production, meaning expenditure related to its production should be treated as an intermediate cost.
- 1.44. Conversely, over time, the cost of producing and, more importantly, storing data, has declined precipitously. This combined with introduction of new digitalisation concepts such as generative AI, has meant that even if used once, data is often stored for re-use as businesses have found that the overall quantity of data at their disposal can be just as important as the quality.
- 1.45. Broadly, countries are encouraged to continue to seek information from data producers on what an appropriate capitalisation rate may be. **When feasible, countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment.**
- 1.46. However, since this information is considered unavailable in most countries, it is the view of the task team on measuring data in the national accounts that **a standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalised expense and should be classified as GFCF, with no adjustment made to represent short lived data.** More context on this recommendation is provided in [Chapter 2](#).
- 1.47. Such a decision should not be viewed in isolation but also in conjunction with the recommended average service life assigned to data assets. This is discussed in [chapter 4](#) and reflects the known uncertainties that still exist regarding how data is used. Additionally, it is possible to incorporate the retirement of a large cohort of data assets within the first year of existence through the Perpetual Inventory Method (PIM) – See [chapter 4](#). This would alleviate the need for an adjustment based on unavailable data. This treatment would also align with the proposal that was supported by the global consultation and expert group when considering the incorporation of data as an asset in the 2025 SNA (ISWGNA, 2023).
- 1.48. This measurement challenge is acknowledged in 2025 SNA which suggest an approach consistent with the recommendation given here. Specifically the SNA suggest that “When feasible, the cost of production of data whose service life is clearly short (e.g., data that is stored for only a short time) should be treated as intermediate consumption rather than fixed capital formation” (2025 SNA §22.30) before adding that “the information needed to separately identify the costs of producing the short-lived data and the costs of producing the long-lived data is often unavailable” so therefore “When the separate cost of producing the short-lived data is unknown, a relatively short average service life that reflects the inclusion of the data with a service life shorter than a year may be used” (2025 SNA §22.31). This handbook supports this approach as will be discussed in [chapter 4](#).
- 1.49. A final consideration regarding the capitalization rate is where any data considered as short-lived data would actually be reflected in the national accounts. While allowing for exceptions, the SNA points out that “It is unusual to record goods and services used as intermediate consumption within the same establishment” (2025 SNA §7.131). When applied to the production of data this would mean that any produced data which is not considered an asset will not be shown explicitly as output, rather the value of the output will feed into the overall value of the subsequent product produced by the organization. Such a practice is relatively normal within the accounts. For example, many large organizations have specific people producing accounting services, but these are not shown separately since they are entirely consumed in the production of the main output of the organization. Theoretically, it could be argued that this is essentially how data is being treated currently, i.e., as an input, produced on an own account basis, but entirely consumed in the production of other goods and services.
- 1.50. If this convention is followed, any adjustment applied to data produced on an own account basis will not be explicitly visible in the accounts, as the proportion of produced data considered to be used within a year is consumed and

embedded in other products. Therefore, within the Supply and Use Tables, any difference between the level of data output and the level of data GFCF will be limited to only purchases of data between two separate establishments in which the data was subsequently and entirely consumed within a year.

- 1.51. When the recommendation on capitalization is combined with the established convention of not separately identifying output entirely consumed internally, the recommendation concerning the capitalization rate for own account data takes on the following order.
- I. Since it is likely that a proportion of data has a short life, when feasible, **countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment.**
  - II. If an adjustment is undertaken, **in order to follow standard SNA conventions and be consistent with the treatment of other own account output consumed internally, the proportion of data consumed within a year should not be explicitly identified as data output. Rather the expenditures should be recorded as input costs into the production of the relevant product.**
  - III. **If information on short lived data is unavailable, the standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalised expense and should be classified as GFCF, with no adjustment made to represent short lived data.**

## The incorporation of data in product and industry classifications used in economic statistics

- 1.52. Data (as defined in the 2025 SNA) has not been explicitly identified in many of the existing statistical classifications that complement the SNA, such as Central Product Classification (CPC) and International Standard Industrial Classification (ISIC). However, in unison with the update to the SNA revisions are also occurring across various statistical classifications and standards. This includes the CPC and ISIC, the most recent versions of which, incorporate new categories on a range of issues, including the production of data.

### Central Product Classification (CPC)

- 1.53. Previously, specific examples of data such as 'On-line directories and mailing lists' and 'Web search portal content' were included in the CPC. These products reflect the relatively concentrated focus on data which previously existed in the economy, whereby data was likely produced by dedicated data producers and used for narrow purposes rather than being a product produced and used by most firms throughout the economy. As such the categories had a relatively narrow definition which excluded much of the data that is produced in the modern economy. These categories still exist as separate products, representing the specific services of compiling and organizing information, although both are obviously very reliant on the data.
- 1.54. CPC version 3.0 explicitly defines a new group "*Data and data compilation*" consisting of two classes – 8371 "*data*" and 8372 "*compilation services of data*" to facilitate the classification of data output, regardless of the specific information content within the data. The definition of the data used in the CPC is "Original compilations of information content organized for retrieval and consultation, produced by accessing and observing phenomena."
- 1.55. The small but understandable difference that exists between this definition and that used for data within the 2025 SNA revolve around the specific caveats added to the SNA definition (i.e., data must be in digital format and provide an economic benefit) to ensure its consistency with the broader classification of fixed assets. The CPC does not require this as the central product classification must retain relationships with statistical classifications beyond the SNA, as well as provide a classification that can be used by policy analysts and businesses that use economic data for studying industrial activity (UNSC; Task Team on ISIC, 2024).

## Industrial Classification of Economic Activity (ISIC)

- 1.56. Since a single industry can produce a range of products, it is expected that data (as defined in the revised CPC) will be produced by almost all industries across the economy, likely reflected in the SUTs as an additional product on top of their primary product. Such a result would also reflect the expectation that most data is produced for own account use.
- 1.57. However, some producers are focused on producing data for sale or providing their data production service to other organizations. With data now being used for so many different facets of production, these data producing firms are expected to be more prevalent than previously. As such, this has been reflected in a new ISIC division 'Computing infrastructure, data processing, hosting, and other information service activities' being included in the revised ISIC (UNSC, Task Team on ISIC, 2024) with intent of mapping the CPC data product to this ISIC category. This division and accompanying group and class 'Computing infrastructure, data processing, hosting and related activities' include such activities as digitalisation of files (for further processing of data), provision of data entry services and data processing services.
- 1.58. It should be noted that many countries, do not directly use the CPC or ISIC. Rather statistical offices apply regional or national variations based on these international frameworks <sup>(12)</sup>. However, having these changes introduced into the CPC and ISIC will allow for the flow down of revisions into these local variations. Importantly, having data better represented in the product and industry classification is a vital complementary step in the quest for consistent estimates of data in the SNA. While this handbook focusses on providing a consistent methodology that can be applied across countries, the consistent classification of this output and investment is just as important to ensuring comparability across countries. The revised product and industry classification offers countries the opportunity to do so.

## The production of data by all sectors across the economy

- 1.59. Like other assets in the economy, **data can be produced and used in production by all sectors of the economy including the non-market sector**. Occupations listed as part of the sum of costs calculation (see [chapter 2](#)) would include those working for the general government and NPISH sectors. The data produced by these occupations may include both publicly available and non-available data since both are considered as contributing to the output of the government sector. For example, data compiled by security forces, that assists in the provision of public safety as well as taxation and social security databases, created to assist in the efficient delivery of government services, are clearly data investments (GFCF in SNA lexicon) undertaken by government that provide an economic benefit to its owner (the government) over future periods.
- 1.60. Like other assets that are publicly owned and made available to use with no direct charge to the users, the services produced by these data assets are consumed collectively and theoretically the value that the public places on these assets may extend well beyond the sum of costs it took to produce them. Despite this, data produced on an own account basis by the non-market sector should be valued using the sum of costs methodology, similar to own account data production undertaken by the market sector. This would include an adjustment to factor in a return to capital. Previously this adjustment was for the market sector only. However, the 2025 SNA has expanded such an adjustment to the sum of cost methodology for all sectors. The fact that some data produced by the government sector is publicly available, does not negate the ownership of the data by the data producer (the government). While they may not be able to obtain financial benefit from the data, the production and dissemination of the data represents government output.
- 1.61. Theoretically, data can also be produced by the household sector. However, since data for the purpose of SNA involves information content obtained through accessing and observing phenomena rather than simply anything saved digitally, this would exclude a substantial majority of videos, photos, blogs, and other self-published material from

<sup>(12)</sup> Examples include NAICS and NACE for industry and CPA and NAIPS for the product dimension.

being considered as data in the SNA <sup>(13)</sup>. Therefore, while conceptually possible, it is considered that the contribution of the household sector to the overall amount of data production would be minor compared to other sectors <sup>(14)</sup>.

<sup>(13)</sup> Influencers and other producers of social media content who have monetized their output would theoretically count as production and so their posts could be considered an asset, however they are likely more akin to artistic originals than data.

<sup>(14)</sup> One transaction that would not constitute production of the household sector involves monetary payments made in relation to participation in a survey or other data gathering tools. While these transactions are considered to be minor, the 2025 SNA explicitly includes these transactions as rent on other non-produced non-financial assets.

**ANNEX 1.1****List of recommendation associated with the conceptual boundaries of data**

Topic	Recommendations
<b>Definition of Data</b>	Information content that is produced by accessing and observing phenomena, recording, and storing information elements from these phenomena in a digital format and that provides an economic benefit when used in productive activities
<b>Non-digital data</b>	While acknowledging that non-digital data exists, For the purpose of the 2025 SNA, only digital data is considered within the 2025 SNA production and asset boundary
<b>Ancillary data</b>	If the data is not providing a direct economic benefit to the entity, it is considered outside of the 2025 SNA production and asset boundary
<b>Reporting of data and databases together</b>	While data and databases are conceptually different assets, it is practically difficult to compile separate estimates for each. As such, for reporting purposes, data and databases are combined into a single detailed intellectual property (IP) product called data and databases
<b>Separately identifying data when used in the production of other IP products, eExcluding sion of data integral to R &amp; D or created solely for AI which cannot be re-used</b>	The value of produced data is separately identified when capitalized. The value should be excluded from the cost of own account production of other goods and services regardless of how data dependent the final output is. Two exceptions exist, including data integral to the production of Research and development and data assembled in a database created solely as a step in the production of an AI computer program and cannot be re-used. In these situations, the value of the produced data should be included in the value of the final Research and Development or AI (computer software) product
<b>The production of data across all sectors of the economy</b>	Data can be produced and used in production by all sectors of the economy including the non-market sector.
<b>The limit to expenditure associated with data production</b>	The incorporation of additional information content or improving the data's quality at either a granular or aggregate level is considered expenditure on data production and is considered GFCF. Analysis of the data to obtain insights or using the information contained in the data in productive activities is considered the production of a good or service other than data
<b>Potential adjustments to account for short lived data</b>	<ol style="list-style-type: none"> <li>1. When feasible, countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment. If such an adjustment is made, countries are encouraged to publish (perhaps as a memorandum item) information on the level of data treated as short lived in order to improve the comparability across countries</li> <li>2. If an adjustment is undertaken, in order to follow standard SNA conventions and be consistent with the treatment of other own account output consumed internally, the proportion of data consumed within a year should not be explicitly identified as data output. Rather the expenditures should be recorded as input costs into the production of the relevant product.</li> <li>3. If information on short lived data is unavailable, the standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalized expense and should be classified as GFCF, with no adjustment made to represent short lived data</li> </ol>

# 2

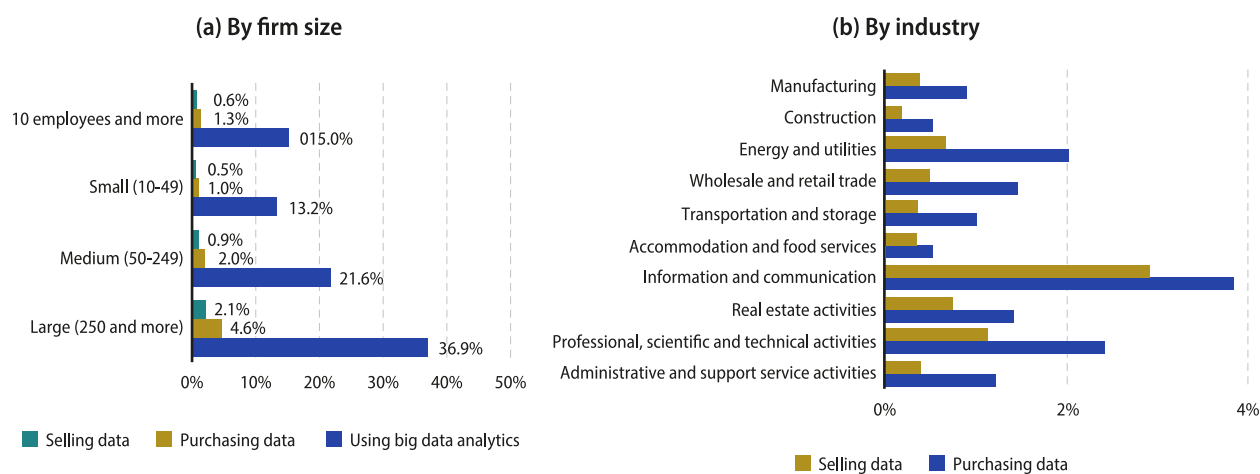
## Compiling a nominal estimate of Data output

### Introduction

2.1. Theoretically, the total output of a product in an economy includes a mix of that which is produced on an own account basis as well as the amount purchased in market transactions. However, unlike many other products in the economy, recent studies have shown that most data used in the economy is obtained on an own account basis. A 2019 ICT use survey from Eurostat (See [Figure 2.1](#)) showed that although 15.3% of large business were using data analytics, only 1.3% of these businesses were buying this data, meaning that the remaining firms were using data created by themselves. Similarly in 2022, the Japanese special internet survey revealed that of the workers involved in the production of data, nearly 75% of them produced the data for use “in-house” (Japanese Cabinet Office, 2022). These results provide important empirical evidence to the well-established opinion that most data used in the economy is created on an own account basis.

**FIGURE 2.1**

### Share of enterprises using, purchasing, and selling data, Europe, 2019

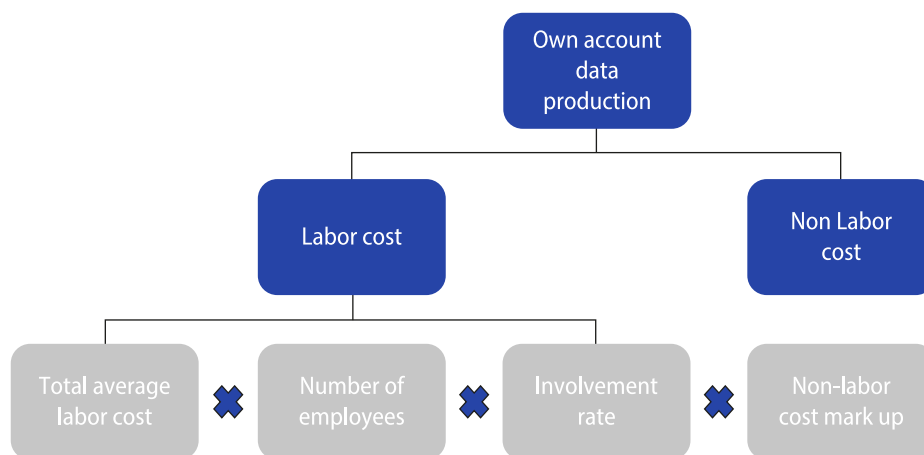


Source: (OECD, based on Eurostat, 2021).

- 2.2. Such a revelation makes logical sense since one of the central pillars of data's value is exclusivity: an organization possessing data that its competitors do not provides a clear point of comparative advantage and thus, value. While publicly available data or data that is not exclusive can still be used in production, its potential for adding value is greatly diminished. For this reason, it has been consistently observed that business models are becoming more dependent on proprietary data (Nguyen & Paczosi, 2020), a trend that is likely to continue with increases in legislation aimed at protecting consumer privacy that prohibits or limits the sale of third-party data (Corrado, Haskel, Iommi, Jona-Lasinio, & Bontadini, Data, Intangible Capital, and Productivity, 2023). This is further discussed in [Box 2.1](#).
- 2.3. As such, the focus of this chapter will be how statistical organizations should approach valuing the production of data created on an own account basis as this will form the vast majority of the final estimate of total data output.
- 2.4. It is recommended to value own account output of data via a sum of costs methodology. When producing a nominal estimate of data output using this method, different source components are required. These are shown in grey in [Figure 2.2](#) (along with the portion of the chapter that discusses the item). Some of these are already collected by statistical organizations, while others will either require new data collection, modelling, or both. Each of these source components will be discussed separately in this chapter and a summary of the recommendations is provided in [Annex 2.8](#).

**FIGURE 2.2**

### Simplified representation for deriving nominal estimates of own account data output



- 2.5. Primarily, this chapter will focus on the compilation of annual estimates of data output. Preliminary estimates of data output have focused on obtaining annual estimates. This reflects the fact that much of the inputs used to derive the estimate are available only on an annual basis. Similar to other estimates within the National Accounts, annual estimates are considered of superior quality to quarterly data, especially when subsequently benchmarked to other measures, including within the supply and use tables. However, quarterly estimates will be required as part of the standard compilation of quarterly national accounts. At the conclusion of the chapter, the subject of compiling quarterly estimates based on the previously derived annual estimates will be covered.
- 2.6. Unlike the recommendations provided in chapter 1 concerning the conceptual boundary of data, the recommendations in this chapter are presented in two forms. The first is a standard option for National Statistical Authorities (NSAs) which do not have more specific sources or information on data production. This is complemented with a more advanced set of recommendations that countries should attempt to work towards.

- 2.7. Due to the compilation of data output being in its infancy, there is a shortage of information on assumptions used to compile the nominal estimates. As such, NSAs are encouraged to continue to research various aspects to improve the quality and robustness of the output. In this vein the recommendations labelled as advanced will likely involve the incorporation of information that is not yet available for many countries.
- 2.8. At the same time, clearly defined standard recommendations are provided to promote international comparability in the estimates, an important requirement of any estimate included within the SNA. These recommendations are also included in these chapters and should be considered as a standard methodology that NSAs can use to compile their estimates.
- 2.9. Finally, while this chapter focuses on, and provides standard recommendations for, compiling nominal estimates of data, in practice, NSAs are likely to be compiling a combined estimate of data and databases. As pointed out in the 2025 SNA *“Despite their conceptual difference, data and databases are difficult to measure separately because they are produced with similar inputs and because transactions prices generally reflect the combined value of the database and the data. For reporting purposes, data and databases are therefore combined into a single detailed intellectual property product (IPP) called data and databases”* (2025 SNA§22.25) This is discussed further in [chapter 5](#), however its well understood that while this chapter will only refer to the compilation of data, the methodology put forward can be applied for the measurement of both assets if it is intended that they will be reported together as a single item <sup>(15)</sup>.

## Outline of Sum-of-cost valuation approach

- 2.10. **It is recommended that data produced on an own account basis is valued using the already established sum-of-cost method, starting from the labor costs related to occupations involved in the creation of data.** This method is used regularly by National Statistical Authorities (NSAs) for other purposes. Attempts to value data and its impact on the economy have been extensive and there are examples of other valuation methods beyond those outlined in the SNA <sup>(16)</sup>. However, many of these (often put forward by academia or private organizations) are considered either not consistent with the overall SNA framework or not practical for consistent implementation across countries. For example, the extremely low level of market transactions in data combined with the heterogeneous nature of data, makes the valuation of data using market transactions unfeasible. Similarly, due to the individual nature of data and its value being so dependent on context, calculating present value based on future returns is considered similarly unpracticable. A more extensive discussion on how output is valued in the SNA and what this means for data measurement is provided in [Box 2.1](#).

### BOX 2.1 HOW THE SNA VALUES OUTPUT AND WHAT THIS MEANS FOR MEASURING DATA

*Since most data is produced on an own account basis and cannot be valued using the conventional practice of recording of market transactions, it is worth briefly covering how the System of National Accounts (SNA) records such output, and how it relates to data.*

The 2025 SNA is not prescriptive in its choice of valuation methods for measuring output or asset, it is only directive on the valuation principle. That is, to make the SNA the powerful analytical tool that it is, it uses a single accounting unit, money terms. Therefore, while the 2025 SNA does suggest primarily using values “at the actual price agreed upon by the transactors” (2025 SNA §3.60) thereby making exchanged prices the basic reference for valuation in the SNA, several other valuation approaches are mentioned as appropriate or acceptable in certain circumstances.

One of those circumstances is when there is an absence of market prices, a situation that is certainly applicable for data assets. For example as shown in results published by Eurostat, the vast majority of

<sup>(15)</sup> While fundamentally the methodology is the same for both assets, additional consideration may be required on the occupations chosen and involvement rates if the production of databases is shared with the production of data.

<sup>(16)</sup> An excellent summary of the many different ways that data can be valued is presented in (Coyle & Manley, 2022).

data that is used in production is produced by corporations and organizations themselves. This results in only a small amount of market transactions in data relative to the amount being used in production (see [Figure 2.1](#)).

Importantly, even if a greater number of market transactions existed, the extreme heterogeneity of data prevents their use as proxy for own-account produced data. The highly contextual and independent nature of data means that the collected prices are not nearly as representative of other (non-observed) transactions as is the normal case for the measurement of many other goods and services.

This absence of market prices and the difficulty in using them when they are available, means that compilers must search for alternative valuation methods. The 2025 SNA provides several options for valuing output or assets where market prices are absent. These include.

- Estimating a value according to costs incurred (2025 SNA §3.60)
- By referring to market prices for analogous goods or services (2025 SNA §3.60)
- Estimating a discounted present value of future returns expected from a given asset (2025 SNA §3.61)

The very low level of market transactions in data combined with the heterogeneous nature of data, makes the second option unfeasible. The third option is often used within the national accounts, for example when valuing natural resources. However, while data is often referred to as the “new oil,” from a valuation perspective there are clear differences between the characteristics of data and natural resources that impact the ability to accurately forecast future returns. These include the near limitless stock of future data, the lack of homogeneity in data products as well as the highly contextual nature that data is used in production. As such, the creation of a future earnings forecast that would allow for an accurate estimate of the current value of all data assets within the economy was considered unrealistic.

Overall, due to these data characteristics the 2<sup>nd</sup> and 3<sup>rd</sup> valuation options are considered untenable and as such **it is recommended that data output produced on an own account basis is valued using the already established sum-of-cost method.**

## How the sum of costs methodology is applied to data production

2.11. The 2025 SNA provides a simple description of the sum of costs approach (see [Figure 2.3](#)). It lists the value of output as equal to the sum of the following items: intermediate consumption, remuneration of employees, depreciation and other taxes (less subsidies) on production <sup>(17)</sup>. It is necessary to also include an estimated mark-up to account for the producers’ net operating surplus, also referred to as the return to capital. This last addition was previously reserved solely for market producers, but the 2025 SNA considers that such a return to capital is an expense for all producers, both market and non-market <sup>(18)</sup>.

<sup>(17)</sup> The 2025 SNA has revised certain terminology compared to the 2008 SNA, these include Compensation of Employees, which is now referred to as Remuneration of Employees and Consumption of Fixed Capital which is now referred to as Depreciation. Conceptually there has been no change to either concept.

<sup>(18)</sup> Previously, net operating surplus for non-market producers was considered zero by convention.

**FIGURE 2.3****Formula for Sum of Costs approach (2025 SNA §7.141)**

Gross Output	=	Intermediate Consumption
	+	Remuneration of employees
	+	Depreciation
	+	Other taxes on Production
	-	other subsidies on production
	+	Mark up representing return to capital

2.12. All countries that have produced estimates of data output have used this sum of costs approach when constructing estimates of data output and gross fixed capital formation (GFCF) of data. Ideally, information is available on all cost components involved in the creation of data, but country practices have shown that this is often difficult to obtain at that level of detail. For that reason, most NSAs have focused on capturing the most important cost element, i.e., remuneration of employees, and to then derive the remaining elements of the sum of costs approach via a mark-up.

2.13. [Figure 2.4](#) provides an example of this practical approach from work by the United States Bureau of Economic Analysis. In this formula, for each occupation ( $\omega$ ), industry ( $i$ ), and year ( $t$ ), the labor cost is calculated by multiplying the annual number of employees ( $H_{\omega,i,t}$ ) by the average total remuneration ( $W_{\omega,i,t}$ )<sup>(9)</sup>. Additionally, an occupation-specific time-use factor ( $\tau\omega$ ) that reflects the actual time-effort that the occupation allocates to data-related activities is applied. A final parameter ( $\alpha$ ) represents a mark-up that reflects other costs (not included in the wage bill) including capital costs and intermediate consumption, and operating surplus (Calderón & Rassier, 2022).

**FIGURE 2.4****Sum of costs approach presented mathematically**

$$C_i = \alpha \sum \tau\omega W_{\omega,i,t} H_{\omega,i,t}$$

Source: (Calderón & Rassier, 2022)

**Detailing aspects of the sum of costs formula for data in more detail**

2.14. The use of this equation in compiling nominal estimates of data can be broken down into four key input requirements. Although terminology differs across the various work streams undertaken by countries so far, each of the components are fundamentally the same. In this handbook the four components, as outlined in [Figure 2.2](#) and [Figure 2.4](#), will be referred to as:

- **The relevant occupations ( $\omega$ )**, that is a list of the occupations considered to be involved in producing data output. Importantly, the specific data input concerns the number of employees working in these occupations.
- **The labor costs of these occupations ( $W$ )**. As will be discussed further down, in the established work this labor cost can be calculated in different ways depending on the data sources available, but conceptually it should cover all labor costs associated with occupations chosen (including those beyond remuneration of employees).

<sup>(9)</sup> Although the equation represents total remuneration, as discussed in section 2.2, it is likely that the starting point will be average annual wage. As such an adjustment to incorporate 'other' labor costs will need to be included at some stage.

- **The involvement rates specific to each occupation ( $\tau$ ).** All NSAs apply some form of time use adjustment to acknowledge that each worker in the stated occupations is unlikely to spend 100% of its time producing data. This adjustment seeks to appropriately capture the proportion of their labor that is *actually* contributing to producing data output.
- **The mark-up representing non-labor cost ( $\alpha$ ).** While conceptually these additional costs of production could be calculated and summed to the labor cost estimate, and if the information is available, countries should apply these known costs to the labor cost estimate. However, since this information will likely be unknown for most countries it is recommended that this non-labor expenditure is estimated by applying a proportional mark-up to the labor cost value.

2.15. For completeness, the final section of this chapter will cover additional clarifications regarding specific compilation questions that go beyond the previously presented formula, but which NSAs should be aware of in order to produce useable estimates consistent with the standard recommendations.

## The choice of occupations in the compilation of own account data output

**Overall Recommendation:** NSAs should use a list of occupations as the foundation for deriving an estimate of labor costs involved in producing data. Such a list should only include occupations that undertake data producing tasks as an integral part of their role. While occupations will be added and removed over time, reflective of changes in the way that data is produced, in order to ensure that methodological decisions do not create breaks in the time series, this list should remain broadly consistent across periods. Furthermore, it is important to publish the metadata of what occupations have been included (and excluded), so that users have a good understanding of how the results have been derived.

**Standard recommendation:** In the absence of other sources, national statistical authorities (NSAs) should use the list of occupations provided in this handbook for the compilation of data output. If the standard list of occupations are used, these must be complimented by the same or very similar involvement rates to those listed in this handbook. The occupation list is presented at the International Standard Classification of Occupations (ISCO – 08) four-digit level for the compilation of data output.

**Advanced recommendation:** NSAs are encouraged to derive a list of occupations (and involvement rates) through an objective and systematic approach to better determine which occupations are most likely to be involved in data production. Examples include the use of machine learning or survey information as well as the use of key words when reviewing occupation classifications or job advertisements.

2.16. The choice of occupations is an important foundation stone in the construction of nominal estimates of data output. It has been observed from initial research that the list of occupations considered as possible data producers is broader than those involved in the production of other IPP assets <sup>(20)</sup>. While there has been some consistency in selected occupations within the studies completed to date, there is also some variance, with a large number of occupations included by only one or two countries, often with a small involvement rate. Such a result displays the potential for divergent estimates of data being produced across countries despite using the same fundamental methodology.

2.17. Although several other possibilities exist (e.g., surveying businesses, applying key word searches of occupation classifications, or using occupation lists created by other countries), the current approaches applied by NSAs who have already produced estimates of data are limited to the following two:

- A selection of occupations within a statistical classification through expert knowledge of the analyst or by using some key words.

<sup>(20)</sup> As will be discussed in chapter 2.3, while there appears to be more occupations involved in the production of data, these occupations appear less specialised with lower involvement rates than those occupations contributing to Software, or research and Development.

- Creating a list of occupations through nominating specific tasks or key words that are associated with the production of data and then review job advertisements or occupation classifications to identify occupations that include these specific tasks <sup>(21)</sup>.
- 2.18. Regardless of the manner in which the list of occupations is created, it is desirable to select the occupations at the most detailed classification level possible. The more detailed the class level, the more accurate the estimate of data can potentially be as there is less scope for including workers who are undertaking tasks unrelated to data production. Ideally, at least the 4<sup>th</sup> digit of the ISCO classification (or equivalent) should be used. That said, it is important that any list of occupations can be complemented by additional data on employment, either labor costs associated with the occupation or, as a minimum, the number of workers within the class. Such a requirement may dictate the level of detail of the occupation list. In this way, the actual data source required is not a list of occupations but actually the number of employees and self-employed people working in these occupations.

## Selection of occupations through expert knowledge

- 2.19. A selection of occupations through expert knowledge can be produced relatively easily. Analysts review tasks considered as contributing to the production and attempt to match these with tasks undertaken by specific occupations. It's important to note that **occupations should be considered for the list if the occupation involves tasks which explicitly contribute to adding value to the production of data and the worker undertakes these tasks as an integral part of their role.** Some occupations may technically be involved in the production of data; however, their role is an extraneous one, at best, occurring due to circumstance rather than fundamentally adding value to the output of data. In these circumstances, the cost of their labor should *not* contribute to the overall production costs. More information on this is provided in [Box 2.2](#).
- 2.20. The SNA research guidance note endorsed by the AEG following global consultation (ISWGNA, 2023) outlined the following broad tasks as contributing to the production of data:
- planning, preparing, and developing a data production strategy,
  - accessing, recording, and storing information embedded in observable phenomena,
  - processing, and cleaning the data to allow for use in productive activities.
- 2.21. As discussed, reasonable differences of opinion can exist in the selection of data-producing occupations. Due to this, a survey was undertaken by the task team on measuring data whereby members submitted their respective opinion on which occupations (and their involvement rate) are involved in the production of data. The results of this survey along with occupation lists derived through more systematic approaches have been aggregated, reviewed, and tested by task team members. The final outcome is the list is displayed in [Annex 2.6](#).
- 2.22. This list forms the basis of the standard recommendation for NSAs compiling estimates of data for the first time. The list is not designed to be a comprehensive list of data-producing occupations. Rather, the task team on data took the view that any list used for compilation of estimates, and created to encourage international comparability, should be limited to occupations which are *likely* involved in the production of data rather than *possibly* involved. NSAs are encouraged to seek more information on data producing occupations and build on the list as required.
- 2.23. It is acknowledged that when producing estimates, there may be slight variations on this list due to differences in the nature of occupation classification used, and the level at which data is collected by NSAs. In addition, NSAs may arrive at a different list of occupations through the collection of additional information (see next section). That said, in the absence of other information, **national statistical authorities (NSAs) should use the list of occupations provided in this handbook for the compilation of data output. If the standard list of occupations are used, these must be complimented by the same or very similar involvement rates to those listed in this handbook. The occupation list is presented at the International Standard Classification of Occupations (ISCO – 08) four-digit level for the compilation of data output.**

<sup>(21)</sup> This process not only identifies data producing occupations but is able to also provide a systematic measure of the occupations' level of data intensity or time factor spent producing data (referred to as involvement rate in this handbook) such information is also required (see chapter 2.3).

## Systematic approach to identifying occupations

- 2.24. The identification of occupations using Machine Learning techniques begins by identifying tasks or key words as being related to the production of data. Blurbs from job advertisements are then reviewed to match the identified tasks or key words with those mentioned in the job advertisements. This process allows for specific occupations to be identified as involved in the production of data based on the prevalence of the tasks and key words. To demonstrate, [Annex 2.3](#) contains the list of key words that were used by Statistics Canada when they applied this approach to determining which occupations should be considered as data producing.
- 2.25. The most significant advantage of this approach is the removal of the subjective element from the selection by of occupations and ensuring that selections are based on real world evidence rather than on any assumptions, potentially improving the accuracy and consistency of the estimates across countries and over time. Evidence of this is the much larger number of occupations that are selected through this method than through manual analysis and research.
- 2.26. Machine Learning also has the advantage of selecting occupations in a more robust manner. The systematic approach has the potential to detect changes occurring in the tasks undertaken by occupations in a more robust manner than via personal knowledge and opinion. Occupations that may not produce data currently but might in 5 years' time, will be picked up quicker using this method than the more subjective expert knowledge approach. This feature is important as while it is desirable to have a certain level of occupational consistency across periods to ensure that breaks in the series are not introduced, this desire must be weighed against the need for the list to appropriately reflect changes in the occupations actually producing the data.
- 2.27. However, the use of machine learning also creates additional challenges, namely the resources required to make it operational, which may make it unavailable to many countries. Additionally, by removing the "common sense check," the results produced can sometimes be unrealistic and extremely sensitive to method changes. For example, slight changes in the tasks or key words selected can result in a significantly different selection of occupations, which would lead to considerably different estimates of data output. As such it is likely that some human refinement would be needed on top of any occupation list created using systematic approaches.
- 2.28. An alternative method to derive a list of occupations considered as data producing occupations involves selecting tasks or competencies considered as involved in producing data and matching them to those listed in the occupation classification or similar occupational data sources. The Federal Statistical Office of Germany (FSO) developed such an approach as part of the task teams work. This approach referred to as the Competence-Relevant Occupation Methodology (CROM) is outlined in [Annex 2.4](#).
- 2.29. Despite this need for additional resources, early testing of the different approaches provides confidence that a more objective selection of occupations will improve the accuracy and robustness of the estimate. As such, the advance recommendation is for **NSAs to derive a list of occupations (and involvement rates) through an objective and systematic approach to better determine which occupations are most likely to be involved in data production.**

### BOX 2.2 WHEN IS AN OCCUPATION ACTIVELY PRODUCING DATA VS BEING PASSIVELY INVOLVED IN DATA PRODUCTION?

The labor cost component used as part of the formula displayed earlier (see [Figure 2.4](#)) should include labor costs related to occupations associated with producing data. However, it is not straightforward to determine the relevant occupations. If followed scrupulously, many occupations may be considered as being *associated* with the process of accessing, recording, and/or storing information elements.

For example, today many electronic goods that contain internet connectivity include data tracking software that feeds information elements from observable phenomena associated with the product, back to the original producer who use these information elements to create data. In this situation, it could be argued that the salesperson who facilitated the sale of the product has assisted in the

producer being able to access and record observable phenomenon. As such, a (exceedingly small) portion of their wage could be considered expenditure related to the production of data. Based on such an interpretation many retail workers, such as cashiers in retail shops might be considered as producers of data since the information elements involved with scanning purchases are contributing to the production of data for the retail shop.

The inclusion of such employees considered as being passively involved in the production of data (as auxiliary activity) is not the intention of the measurement framework and compilers should not follow such an interpretation. Rather, **occupations should be included if their job involves tasks which explicitly contribute to adding value to the production of data and the worker undertakes these tasks as an integral part of their job. This is opposed to occupations that complete tasks involved in the production of data in a passive manner where the added value is incidental to their primary task** such as the cashier and retail worker. These occupations are essentially involved in the data creation chain only due to the circumstances surrounding how certain information elements are recorded (i.e., digital scanning of purchases).

The data produced by the retail store (quantity and price of products purchased, etc.) are indifferent to how the data is collected. The cashier does not quality ensure the collection of information elements that make up the data, rather their primary role is to ensure payment for the purchases. Importantly from a national accounts' perspective, the value that the producer places on the data does not differ depending on how the information is collected and recorded. As such, from a data valuation perspective it would appear conceptually inconsistent to apply a greater level of labor costs in the creation of one type of data just because a cashier recorded the sale in person compared to the value assigned to data of purchases made online.

If, however, the collection of information elements associated with certain OPs involves specifically tasked human intervention, then the labor costs associated with this labor input should be included. As the data producer has made a conscious decision to observe, collect and record OPs using human intervention rather than through digital means, this labor input is explicitly adding value to the data output by collecting the information elements from the OPs.

At the other end of labor intervention, often the accessing, recording, and storing of information embedded in observable phenomena occurs automatically or at least without explicit human intervention. However, this does not mean that no labor costs should be recorded. Rather the labor cost will likely reflect the initial and later stages of data production, focusing on the collection strategy as well as the cleaning of data to ensure it is fit for purpose. Additionally automatic systems still require regular maintenance and review which brings in further labor costs. Because of this it is likely that most of the reduction in the real price of data production has come due to the decline in labor as an input as more systems become automated, however it would be disingenuous to suggest that data can be produced without labor contribution at some point of the data value chain.

While the line of active vs passive contribution to the data value chain is not always clear, the occupation list included in this handbook aims to include occupations which not only contain tasks associated with the production of data but for which these tasks are an integral part of their job. While this may include specific actions of accessing and recording information elements, it will more likely involve tasks involved with the planning and implementation of data strategies and collection process that allow for recording and storing to happen automatically.

## Calculating the Labor costs associated with these occupations (W)

**Overall Recommendation: The annual nominal estimate of own account data output must reflect non-direct labor costs as well as wages and self-employed income. Since information on average annual wage at an occupational level is unlikely to be available every year, annual movements in wages should be used to move forward total labor costs.**

**Standard recommendation: NSAs are recommended to use average annual wage for each of the occupations selected as a starting point in estimating the total labor costs associated with own account data output for a specific year. For annual estimates of labor costs outside of periods when average annual wage is available, the standard recommendation is to use an appropriate indicator to move forward the nominal labor costs at an aggregated level. This assumes that the labor costs of data producers are changing at the same rate as changes in wages for the broader economy.**

**Advanced recommendation: NSAs are recommended to use average annual wage for each of the occupations selected as a starting point in estimating the total labor costs associated with own account data output for a specific year. Ideally, for annual estimates of labor costs outside of periods when average annual wage is available, total labor costs should be moved forward with an appropriate indicator at the individual occupational level.**

- 2.30. Conceptually, the final labor estimate used in the calculation of data output must include expenses beyond the direct wages and salaries paid to employees or mixed income earned by a self-employed business owner. How this is done will differ across countries, often dependent on the exact data source available to each country.
- 2.31. These additional costs, represent other non-direct labor costs to which workers are entitled, such as pension contributions or any taxes payable associated with the employee. Since households are unlikely to know this information, this amount may need to be added as an adjustment on top of the basic labor cost (i.e., wages received).
- 2.32. Such an adjustment is already in place in many of the estimates produced by NSAs. For example, the German Federal Statistical Office (FSO), in their work, notes that “In addition to wages and salaries we consider in our calculation non-direct labor costs, such as the costs of the associated human resource management and financial control, social security contributions [ . . . ] As no direct information on these costs is available, they are taken into account as a mark-up factor on wages and salaries” (Destatis, 2024). While Australia, Canada and the USA all observe that the mark-up applied in their work represents (among other things) ‘non-direct labor costs’ (Statistics Canada, 2019), ‘non-direct salary’ (Smedes, Nguyen, & Tenburren, 2022) and ‘employee benefits (not included in the wage bill)’ (Calderón & Rassier, 2022). All four of these NSAs have correctly accounted for non-direct labor costs but have done so as part of the overall mark-up applied to labor costs which mainly represents non-labor costs.
- 2.33. An adjustment is necessary since the data source for most NSAs will be wages and salary information usually collected from the household perspective. The most obvious advantage of this is that wage estimates can be stratified by occupation as households are able to provide both sets of information <sup>(22)</sup>. This is in contrast with wage data obtained from employers (collected in the form of wages paid) which is often only able to be collected as an aggregate and therefore delineated by industry rather than occupation. For small business, with a similar set of employees, an assumption can be made between industry and occupation, but this issue is more problematic for large entities which employ a range of occupations.
- 2.34. If the data source is from the household perspective (i.e., population census or household survey) it is important that the wage estimate used covers only employee wages and salary and/or own unincorporated business income, but excludes other income received (i.e., investment income, government benefits). This is usually possible as ‘other income’ is normally separately identified.
- 2.35. Some NSAs have access to tax data which may allow for actual wage estimates to be used rather than average wage. Such data would remove the need to multiply an average wage by number of workers, thereby arriving at the total

<sup>(22)</sup> It should be noted that NSA should remain vigilant regarding the volatility occasionally displayed in these detailed level occupation data. It may be that additional data cleaning may be required if data is being skewed by outliers.

wage bill in a single step. Depending on the scope of this data, it's likely that an adjustment would still be required to turn this wage estimate into total labor costs.

- 2.36. How NSAs then incorporate non-wage labor costs on top of the wage cost estimate will depend on the specific data sources used, as even from the household's perspective, there is a range of possibilities. Most NSAs use one of the following: population census, labor force surveys, household income and expenditure surveys, all with different levels of conceptual coverage and timeliness. As such, there is no definitive compilation recommendation on the data source that must be used for labor costs. However, for an estimate to be accurate **the annual nominal estimate of own account data output must reflect non-direct labor costs as well as wages and self-employed income**
- 2.37. Similarly, the data sources available in a country will also dictate how annual estimates of each occupation's wage is derived every year. Some NSAs will have information on wages at the occupation level available every year, either through a micro census, household earnings surveys or even tax data. Ideally, **NSAs should aim to project labor costs forward with an appropriate indicator at the individual occupational level. It is important that for compiling data output, such an indicator reflects both the price of the labor (wage) and the quantity of labor (number of employees)**, this will assist in more appropriately reflecting changes in the labor costs of data producers rather than economy wide changes in employment and wages.
- 2.38. In many countries this level of information will only be available at a more infrequent basis, perhaps as rarely as population census, undertaken every 5 or 10 years. In these circumstances the wage indicator available on an annual basis will likely only be at a whole economy or industry level. Therefore, **in the absence of occupation specific labor cost information, the standard recommendation is to use an appropriate indicator to move forward the nominal labor costs at the aggregate level, assuming that the labor costs of data producers are changing at the same rate as the broader economy wide changes in wages.**
- 2.39. Recommendations in existing material covering the compilation of IPP suggest that when units specialize in producing an IPP for sale, costs associated with acquisitions or the production of such products should be expensed (Eurostat-OECD, 2019). Often statistical compilers have taken this recommendation to mean that specific industries should be excluded when estimating own account capital formation on the expectation that these industries are predominately producing products for sale, which will be captured at a later date. The task team on data has taken the view that such a recommendation is *not* appropriate for the production of data and that **the sum of costs methodology includes the labor costs of all specified occupations regardless of the industry or unit they are working in which may include workers employed by units that specialize in the production and sale of data.** As noted later in the chapter, such an assumption may mean that adjustments relating to the production sale and purchase of data 'copies' may be required.
- 2.40. Such a recommendation reflects the task team on data's view that there is a broad distribution of businesses and industries producing data. Unlike other IPP such as computer software, the production of which is concentrated among relatively few industries, the production of data is widely observed across all industries. This characteristic combined with the scarcity of transactions (due to so much data being produced on an own account basis) resulted in the task team considering that there was a greater risk of undercounting data production by excluding certain industries than the potential double counting by not excluding these industries. As with all assumptions associated with the compilation of data, NSAs are encouraged to continue to monitor and gather information when possible.

## The need for involvement rate adjustments and recommendation on such adjustments

Overall Recommendation: **Involvement rates, representing the amount of time an employee or self-employed worker actually spends on producing data, are applied at the occupation level so that the actual labor cost associated with producing data is appropriately reflected.**

Standard recommendation: **NSAs are recommended to apply the same or very similar involvement rates to those listed in this handbook in order to determine the degree to which each occupation is involved in the creation of**

**data. The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data.**

Advanced recommendation: **NSAs are encouraged to develop and use involvement rates (and list of occupations) derived through a more systematic and objective means specific to their economy, such as business surveys or results from machine learning.**

- 2.41. It is well accepted when compiling estimates of own account output of a single good or service that the entire output from a single worker is unlikely to be contributing entirely to the single good or service. This may be due to the specific requirements of the workers role, their skill limitations or the characteristics of the industry or organizations that the occupation is placed in. Regardless of why it is occurring, some form of an adjustment is required so that this phenomenon is appropriately factored into the final estimate of output.
- 2.42. Such an adjustment is already recommended in the compilation of output of other IPPs. For example, the application of involvement rates in the compilation of own account computer software is recommended with the final Eurostat-OECD report on land and other non-financial assets <sup>(23)</sup>.
- 2.43. No two workers, despite having the same occupation are likely to contribute exactly the same amount of time to the production of a single product. This may be the case within a specific industry but may particularly hold for people occupied in different industries.
- 2.44. The previously mentioned Eurostat-OECD report mentions this phenomenon, also noting that firm size may play a role in the level of involvement that an employee might have. As such, it recommends making the adjustment at the most granular level possible, since “workers in specific industries may spend more time on own-account software and database production, and workers in larger enterprises may be able to spend more time on own-account software and database production than those in smaller firms” (Eurostat-OECD, 2019). However, conceptual accuracy needs to be managed with practical implementation, and it is not always feasible to apply a ratio or adjustment at such a detailed level.
- 2.45. An extremely simplistic approach would be to apply a single adjustment to the total labor cost estimate. However, this would appear to be an unnecessarily broad approach and given the fact that the results will be based on occupational data, it would be reasonable that such an adjustment would be done at this level of detail, at a minimum. Therefore, **it is recommended that, an involvement rate adjustment is applied at the occupation level rather than at the level of total labor cost for the economy as a whole.**
- 2.46. That said, while there is consensus on the need for such an adjustment and the level at which such an adjustment might be made, the adjustment rates themselves must still be determined. Data on these proportions is very difficult to capture, with many of the NSAs who have produced estimates already using involvement rates sourced via subjective expert opinion, best guesses or applying upper and lower bound involvement rates resulting in upper and lower bound estimates of data output and data GFCF. Such an approach is understandable for standalone research or experimental output; however, such a range estimation is not suitable for the inclusion of data in the core national accounts. Rather, for inclusion within the SNA production and asset boundary, NSAs will need to produce a single estimate using the best available information.
- 2.47. While the information on involvement rates has been difficult to source, there are several examples of work that have been undertaken to inform compilers. The Japanese Cabinet Office uses information from a special internet survey to help guide estimates of how much time workers are spending on data related work. This survey, including preliminary results, is presented in [Annex 2.3](#).
- 2.48. Alternatively, the OECD and others have used natural language processing (NLP) on job advertisements to estimate the involvement rate / data intensity of occupations as well as identifying the occupations themselves. This work has tended to produce slightly lower involvement rates than those estimated via expert opinion. However, it is important to note that these lower rates tend to be applied to a larger number of occupations, thereby producing similar overall estimates of labor costs. As such it is not recommended to mix and match selection methods (i.e. applying higher rates,

<sup>(23)</sup> The report uses the terminology ‘time factors’ for involvement rates but the two are interchangeable.

selected via expert opinion with the larger number of occupations derived through a systematic approach). More information on the OECD's work is documented in [Annex 2.2](#).

- 2.49. If possible, **NSAs are encouraged to obtain and use estimates of involvement rates, generated via systematic approaches, such as a business survey or via machine learning.** These are considered preferential to involvement rates sourced from expert opinion.
- 2.50. However, it is well accepted that obtaining the information required to produce these involvement rates can be quite financially and resource intensive. In this circumstance, in order to promote international comparability, it is the standard recommendation that NSAs apply the involvement rates accompanying the standard occupation list provided in this handbook. For this purpose, the Task Team on data have generated these rates based on involvements rates currently applied by NSAs estimating data estimates combined with empirical evidence coming from machine learning, key words associated with occupation competencies and survey data.
- 2.51. The Task Team agreed upon several key points when determining the standard set of involvement rates. These are reflected in the list provided in [Annex 2.6](#) and include the following:
- I. It was agreed that an involvement rate of 100% should not be assigned to any occupation. It was acknowledged that it is unrealistic to suggest that 100% of a worker's time is always undertaken on a single productive output. A maximum involvement rate recommended for any occupation is around 70%.
  - II. As discussed in the previous section, the task team on data decided that the standard list of occupations should contain only occupations that were *probably* contributing to data production (rather than *possibly* contributing) see [Box 2.2](#).
  - III. Involvement rates and occupations chosen through more systematic methodologies (surveys, web scrapping, etc.) suggest that while the number of occupations contributing to data is quite broad, and those occupations with high involvement rates are relatively high (i.e., 60 – 80%), the level of involvement in data production declines precipitously after these initial occupations. This evidence has also been taken into consideration with the formation of the standard list of occupations and involvement rates.
- 2.4. **The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data.** NSA should not apply one aspect of the standard methodology (occupation list or involvement rates) and combine them with a list or set of rates obtained separately as this significantly impact the international comparability of the estimates.
- 2.5. Since some occupations may also be involved in the production of other goods and services for which the output is determined via the sum of costs approach (particularly other IPP), **NSAs should ensure that costs are only taking into consideration once, i.e., the production stemming from a single occupation should not exceed 100% of the workers time. Overall, it is recommended that the compilation of estimates for the various IPP assets (including data) is undertaken in a holistic way in order to avoid double counting.**

## Deriving the mark-up for non-labor cost

Recommendation: **It's recommended that non-labor costs are incorporated into the final estimate via a single mark-up applied to labor costs. Such a mark-up represents the costs of inputs, any adjustment associated with taxes less subsidies on production, the depreciation of fixed capital used in production, as well as a return to capital (operating surplus).**

Standard recommendation: **A single mark-up is applied representing the non-labor costs of data production, including a return to capital. The single mark-up, based on the ratio of total gross output applied against remuneration of employees from the "Computer programming, consultancy and related activities" (ISIC 62) and "Information service activities" (ISIC 63) – or similar available aggregate - is applied to total labor costs.**

Advanced recommendation: NSAs are encouraged to investigate potential data sources that may provide more detailed information on the non-labor expenses involved in the production of data. **An advanced recommendation is for NSAs to apply multiple mark-ups to the total labor cost estimate separately so that differences in non-labor**

**production cost, depreciation and operating surplus across occupations and industries can be applied more accurately and transparently.**

Basic method: **A basic alternative, which should be seen as least desirable and only to be used if specific industry level information is not available, is to use the same mark-up as that applied to similar IPP assets compiled in existing compilation via the sum of costs.**

- 2.54. The final nominal estimate of data output must contain expenses beyond the cost of labor. These non-labor costs include expenditure on other types of intermediate consumption, depreciation of assets used in production of data and the net operating surplus from the production (a return to capital).
- 2.55. To date, no countries have been able to obtain expense information on data production, direct from data producers. Existing business surveys are usually designed to provide expense information at an industry level and not at the level of specific activities. In addition, operating surplus is a conceptual expense, so is usually calculated on a residual basis, which is not possible when there is no final price due to the absence of market transactions.
- 2.56. Therefore, most NSAs are currently estimating non-labor expenditure related to own account data by applying a mark-up to the labor cost. These mark-ups are usually derived on the basis of information on specific industries within the annual Supply-Use or Input-Output Tables. **Representing non-labor costs through the use of arbitrary mark-ups, created without any empirical evidence is not recommended.**
- 2.57. The main assumption associated with such a methodology is that the output of data exhibits a consistent production function, in so much that a consistent amount of non-labor input is required for each unit of labor input. This is not an outrageous assumption and is considered quite acceptable in lieu of actual data on expenditures on non-labor inputs. Such an assumption has already been recommended for the compilation of other own-account IPP output (Eurostat-OECD, 2019). Due to this, the recommendation regarding a mark-up representing non-labor costs revolves more around the estimation of such a mark-up.
- 2.58. The standard recommendation for NSAs is to select a defined set of industries that are known to contain a large amount of labor costs associated with data producing occupations. While the exact industry classification varies across countries, of the work published so far, countries have used derivatives of “Computer programming, consultancy and related activities” and “Information service activities.” These equate to industries 62 and 63 respectively, of the international standard of industrial classification (ISIC Rev. 4).
- 2.59. As such, **in the absence of more detailed information on non-labor expenses, a ratio of total gross output applied against remuneration of employees from the “Computer programming, consultancy and related activities” (ISIC 62) and “Information service activities” (ISIC 63) – or similar available aggregate - is applied to total labor costs** <sup>(24)</sup>. A numerical example showing this calculation is provided in [Annex 2.7](#).
- 2.60. Ideally, NSAs should aspire to have specific information relating to the production of data, including more detail regarding the companies (or industries) producing it. Such information could be added to the labor costs to better represent the costs of producing data. **NSAs are encouraged to investigate potential data sources that may provide more detailed information on the non-labor expenses involved in the production of data. If possible, NSAs are encouraged to apply multiple mark-ups to the total labor cost estimate separately so that differences in non-labor production cost, depreciation and operating surplus across occupations and industries can be applied more accurately and transparently.**
- 2.61. More recent work has focused on better understanding the costs involved in producing data and how these may differ across industries. Such work reflects the belief that the use of capital in data production and the expected return on any capital investment in data may be more closely aligned to the cost of production and asset use in the industry actually creating the data. The proportion of operating surplus is consistently higher for certain industries, reflecting the heavily dependence on capital. The investment decisions of firms in these industries may be different to those in industries where labor costs make up a larger share of value added.

<sup>(24)</sup> If information is not available at the two-digit ISIC division level, a similar ratio can be created for an equivalent industry at the single digit section level such as Section J Information and communication.

Such an idea is in contrast to the economic concept put forward by the use of ratios from the information and communication industry which suggest that when producing data, non-labor costs, including the cost of using assets in production are tied more to the asset being created rather than the producer creating it and so are likely very similar regardless of which industry is producing the data. Through its standard recommendation to use ratios derived from the information and communication industries, this handbook implicitly makes a recommendation for the latter. However, the former assumption is not without economic merit and countries may wish to investigate further, if possible. **A basic alternative, which should be seen as least desirable and only to be used if specific industry level information is not available, is to use the same mark-up as that applied to similar IPP assets compiled in existing compilation via the sum of costs.**

### BOX 2.3 IMPROVING TRANSPARENCY BY APPLYING SEPARATE AND SPECIFIC MARK-UPS COVERING RETURN TO CAPITAL AND DEPRECIATION

The recommendation provided earlier suggests using a single mark-up to cover all non-labor expenses. This includes expenses occurred via actual transactions (i.e., purchase of inputs used in production taxes on production paid and subsidies on production received) and conceptual expenses incurred by the producer but for which no actual transaction takes place. These include depreciation and the return to capital.

The depreciation expense reflects the cost associated with the decline in the value of capital as it is used in the production over multiple accounting periods. While this usually refers to capital items 'wearing out,' in the case of data production it may more accurately reflect the obsolescence of the data asset since technological innovation is such an important component of data production. Since it is expected that data producers are heavy users of capital goods in their production of data, it is a fundamental expense for data producers.

The return to capital reflects the opportunity costs of using capital in the production of data, i.e., diverting capital from other uses that yield an economic return. As non-market producers do not aim to maximize their profits, the SNA previously adopted the convention that their cost of capital should not include a return to capital. However, the convention is not being carried forward in the updated SNA2025, meaning that this is being regarded as a relevant component for both market and non-market producers.

As suggested by the standard recommendation, the use of a single mark-up reflecting all of these expenses is an acceptable approach when compiling a sum of costs estimate. However, it is conceivable for the expenses to be broken down into separate mark-ups (or additions) reflecting each component so that more nuance can be added at the occupation and industry level. This would not only provide more accurate results but also offer more transparency to users.

Information on the depreciation and the opportunity costs of using capital in the production of output is not usually covered in surveys. Additionally, while business accounts often include entries reflecting depreciation, these amounts are often aimed at maximising profit and so reflect the countries tax policies rather than accurately reflecting the value of the services provided by the capital asset.

Due to this, to derive a mark-up for depreciation and for a return to capital on an industry level, most NSAs rely on ratios for industries derived from SUTs to use as a proxy for inclusion into the estimate. In this way, it is similar to the single mark-up recommended previously. However, since the mark-up is covering more specific expenses, the ratio is calculated using more specific outputs from the SUTs. For example, a ratio representing depreciation may be calculated by applying the level of depreciation to the number of current expenditures listed in the SUTs (i.e., intermediate consumption plus remuneration of employees) for specific industries.

Similarly, a ratio representing the return to capital may be calculated by applying the level of operating surplus for a specific industry for a specific year against the capital stock for that specific industry and year. This calculated mark-up would be more reflective of the return to capital for that industry and likely more indicative of the expected return to capital that the producer in that industry might expect from an investment in data.

## The use of information on market transactions of data

- 2.62. Sections 2.1 to 2.4 discussed the compilation of own account data. Conceptually this makes up only one portion of expenditure on data within the economy. Purchases, both domestic and cross border, can also contribute to the total estimate of data output and data GFCF. NSAs should attempt to complement estimates of own account data production with values of monetary purchases of data.
- 2.63. However, a practical consideration is prevalent for the collection of market transactions. Currently, International accounting standards do not recognize the capitalization of data in the same way as more traditional corporate assets. The SNA even points out that ‘accounting conventions and valuation methods used at a micro level typically differ from those required by the SNA’ (2025 SNA §1.84). One need only to look at the large divergences that exist between the calculation of depreciation for corporate tax purposes compared to that used for the SNA to understand that, since respective accounting systems serve different purposes, deviations should be expected and somewhat warranted.
- 2.64. Although this different treatment is not a conceptual concern, it does create a practical challenge. By not specifically identifying data as a capital asset, producers of data would be unlikely to separately identify any expenditure on data. Rather, it would be aggregated with other current costs of production and so businesses, when asked, would be unlikely to accurately reflect expenditure on data like they can with other assets captured in both the SNA and international accounting standards. This, combined with the rarity of such transactions, means that traditional business surveys are unlikely to capture information on data easily.
- 2.65. Against this background, **unless the transaction relates to a purchase of a copy, information on market transactions in data, while useful, should be viewed as a complementary data source which can improve countries industry and sector splits of data investment, rather than a fundamental component of the compilation process.** Such a recommendation mirrors one provided by the OECD (2010) for measuring databases, which stated “the focus should be on measuring own-account database GFCF and that purchases of databases or database services only be recorded as GFCF on an exceptional basis, if and when such sales come to light” (OECD, 2010). It also reflects the agreed position that data sold in market transactions covers only a fraction of the total output occurring in the economy while acknowledging that this data is likely unobtainable for many NSAs.
- 2.66. This recommendation should not be interpreted as a suggestion to ignore information on the purchases of data and databases made via a market transaction. Rather, it provides reassurance that the absence of robust or comprehensive information on the level or growth of purchases of data and databases should not be seen as detrimental to the quality of the overall estimates being produced. The following paragraphs will detail how information on market transactions can be incorporated into the aggregate and industry estimates of data output and GFCF.
- 2.67. While information on market transactions may be limited for many NSAs, the transactions that are captured should be reflected in the accounts in a manner similar to transactions in other types of IPP (or assets). To do this, it is necessary to determine the length that the data is expected to be used for, as well as if the data is purchased on an exclusive basis or not.
- 2.68. **Data that is purchased and used for less than one year is regarded as being consumed and treated as intermediate consumption.** These costs of intermediate consumption are included as a cost of production as part of the organization’s overall output (i.e., logistics, advertising services, etc.).

- 2.69. **Data that is purchased and used for more than one year is regarded as an asset.** The complete treatment of this expense within the set of economic accounts depends on whether the data has been purchased on an exclusive basis or not.
- 2.70. **Data assets that are purchased with exclusive rights are treated as an outright sale of the data asset by the seller to the buyer. However, assuming that the transaction is not a cross-border or inter-sector one, similar to sales of other assets, this transaction would net off and not impact the overall level of GFCF for a specific sector.** In this scenario where GFCF is recorded following the purchase of data, if the transaction value is added to the nominal estimate of data output, an equivalent adjustment should be made to reflect the negative GFCF accompanying the sale of the asset. If the transaction is recorded (because it is cross sector or cross border), the value of any adjustment should equal the value of the transaction; this reflects that the transaction was undertaken at arm's length <sup>(25)</sup>.
- 2.71. **Data assets that are purchased without exclusive rights are treated as a purchase of a copy and therefore contributes to the GFCF of the purchaser (and aggregate GFCF) if it satisfies the necessary conditions of GFCF, (i.e., use in production for more than one year).** Since the seller retains the rights to the original, they have therefore not sold an asset. The value of the copies should be added to the producing industries' estimate of total data output, to acknowledge the data which is not produced on an own account basis.
- 2.72. Conceptually, the cost of producing the copies (which is equal to the value of the purchases) may also need to be removed from the producing industries' estimate of own account data production, to ensure that the same data is not capitalized twice, once as own account by the producing industry and then again by the industry purchasing the copies. A numerical example of these potential adjustments is provided in [annex 2.9](#).
- 2.73. While the SNA makes clear that the production of copies should not be considered GFCF (2025 SNA §11.99), NSAs will need to determine themselves if the current sum of cost methodology covering own account data production (considered as GFCF) may also be including the cost of producing copies and thus warrant an adjustment.

## Calculating an estimate of net imports of data for compiling estimates of GFCF

- 2.74. Unlike domestic market transactions which should be incorporated with caution to ensure that values are not counted twice, values of data imports and exports must be included to have a truly comprehensive estimate of data GFCF.
- 2.75. The challenge facing compilers gathering information on exports and imports of data is different to those faced in the recording of domestic transactions. Information on the export and import of IT services is usually already collected by most NSAs as part of the compilation of trade statistics. In this case, the challenge is estimating how much of this relatively broad category relates to data. EBOPS 2010 includes the classification of computer services, which is subsequently broken down into the two categories of 'computer software' and 'other computer services'.
- 2.76. Chapter 11 of BPM7 outlines the various classification contained within the services account. This includes the classification, "computer and information services." Since BPM7 is aligned with EBOPS 2010, it is not surprising that both contain a similar inclusions for this category such as 'Data recovery services, and provision of advice and assistance on matters related to the management of computer resources' and 'Data-processing and hosting services, such as data entry, tabulation and processing on a timesharing basis (UNSD, OECD, Eurostat, IMF, WTO, UNWTO, 2010). Both would appear to be likely places that purchases and sales of data would be placed. However, the broader 'computer and information services' category also includes a large number of other services that goes beyond data itself.
- 2.77. It is recommended that compilers undertake further analysis to try and ascertain an estimate of how much of the broad 'other computer services' category is likely made up of data. Such an analysis would allow compilers to make regular estimates of imports and exports which can be added to the level of GFCF derived through the sum of costs method. While such a method is unlikely to produce robust estimates of trade in data, it is expected that imported and exported

<sup>(25)</sup> Conceptually, an adjustment may need to be included in the revaluation account if the transaction value is significantly more (or less) than the valuation of the data asset as calculated via the sum of cost methodology. This would remove the possibility of selling more asset than has notionally been produced. It is expected that this would only be required in exceptional circumstances.

data makes up a relatively small amount of overall investment. Importantly, such an addition will provide comprehensiveness to the final nominal estimate of data GFCF.

2.78. It is worth noting that an adjustment for internationally traded data is required for the compilation of GFCF of data rather than output of data. Since the value of data output produced domestically includes data that is subsequently exported and should not include data produced by the rest of world, no adjustment for trade in data output is required. However, for estimates of data GFCF, an adjustment must be made to include GFCF by domestic firms undertaken by importing data and then exclude data produced domestically but ultimately exported.

### Summary of outputs of data

- Domestic data output = Own account data production for all domestic industries + purchase of domestically produced data copies.
- Net Imports of Data GFCF = Imports of data GFCF – Exports of data GFCF
- Domestic GFCF of data = Domestic own account data output + purchase of domestically produced data copies (that meet necessary conditions of GFCF) + Net Imports of Data GFCF

## Additional points of clarification when compiling a nominal estimate of data output

### The potential of data being consumed within one year

- 2.79. For most assets in the national accounts, respondents will be able to provide a good estimate as to if the purchase is intended to be capitalised on not, and these can be reflected in the survey forms. For assets produced on an own account basis and compiled at an aggregate level by a statistical office rather than based on information provided in survey forms, such a delineation is not always possible to make.
- 2.80. The task team on data accepts that it is not just possible, but likely that some expenditure on data production produces an economic benefit for a period less than one year. As such, **when feasible, countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment.**
- 2.81. However, the SNA acknowledges “the information needed to separately identify the costs of producing the short-lived data and the costs of producing the long-lived data is often unavailable.” (SNA 2025 §22.31) Due to this it is the view of the task team on measuring data that, **if information on short lived data is unavailable, the standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalised expense and should be classified as GFCF, with no adjustment made to represent short lived data.** <sup>(26)</sup>
- 2.82. The recommendation is made on the basis of two accepted actualities:
- I. There is limited empirical evidence which could provide guidance on the proportion of own account data produced that is consumed within one year.
  - II. Endorsing a recommendation implicitly encouraging countries to make subjective adjustments would negatively impact the international comparability of estimates.
- 2.83. Following the publication of the 2008 SNA, one particular piece of accompanying implementation guidance suggested that “not all database creation qualifies as GFCF, and as such, in the absence of any information on the proportion that does, it is recommended that it be assumed to be 50%” (OECD, 2010). While not providing any empirical evidence to support this recommendation, it essentially suggests that a 50% capitalisation rate is an appropriate starting point. Following discussions and consultation, the task team on data considers that, in the absence of new information, this

<sup>(26)</sup> Conceptually, since the recommendation is for own account data to be classified, this could include as an inventory, either as a finished good or work in progress. However, since this treatment appears unhelpful for both compilers and users of the data, by convention, it has been agreed that data inventories do not exist.

starting point for data, should be increased to 100%. This switch reflects the way that data is used in the modern economy. Not only is it cheaper and easier to store and use data in production than it was around the publication of the 2008 SNA, suggesting that data will be used not only more often but also for a longer period of time, the ability to use and re-use data in many different ways has also greatly increased the potential of data use in the economy.

- 2.84. The 2025 SNA points out that although the treatment to include all own account production as capital (in the absence of any relevant information) may cause producers GVA to be overstated it believes that “this disadvantage is outweighed by the advantage of capturing the potentially important value of the stocks of data whose useful economic life is a year or less as part of the measure of the stocks of the data assets” (2025 SNA §22.31).
- 2.85. As was discussed in [Chapter 1](#) and will be elaborated on further in [Chapter 4](#), the 100% capitalisation rate for own account production of data should be complemented with appropriate assumptions on the life length, the depreciation rate and profile when calculating capital stock estimates. A point also made in the 2025 SNA, which suggest that “if the measure of the production of data does not exclude all the data with a service life of a year or less, a relatively short assumption for the service life of data assets is likely to be appropriate” (2025 SNA §22.31).
- 2.86. While empirical evidence is limited, it is broadly agreed that a relatively larger proportion of data is rendered obsolete earlier in their service life than many other fixed assets in the national accounts. However, in lieu of statistically appropriate information providing guidance on this, the handbook recommends that this phenomenon is reflected in the service life applied rather than a potentially unsubstantiated adjustment to the nominal estimate.

## Reflecting the incorporation of new data points to existing data assets

- 2.87. **Expenditure undertaken to update a data asset with newly collected information should be considered as new investment (Gross Fixed Capital Formation - GFCF) rather than repair and maintenance.** The addition of new information likely “enhances [the data assets] efficiency or capacity or prolong their expected working lives and thus should be treated as new GFCF” (2025 SNA §7.251). This contrasts with just ensuring an asset maintains its current working life (the result of any repair and maintenance).
- 2.88. From a practical viewpoint this treatment also aligns with the previously discussed assumption that all own account expenditure on data should be considered as GFCF rather than intermediate consumption.
- 2.89. This recommendation of treating the expenditure as GFCF remains regardless of if the additional information is added to an existing data asset or used to create a new data asset <sup>(27)</sup>. The conceptual boundaries of a single data asset are inconsequential to the measurement of the level of data from the national account’s perspective. For example, the specific numerical number of cars, computers and new buildings are not recorded in the accounts. Even on survey forms, businesses are not asked how many computer software packages they purchased, instead, they are asked about their expenditure on assets. There is no need to treat data differently. Theoretically, it does not matter if a data producer considers their database a single data asset or millions of data assets joined together. From a National Accounts perspective, it is simply the value of production required to create this volume of data.

## The production of quarterly estimates of data output

**Recommendation: NSAs are recommended to develop quarterly indicators that can be used for both the interpolation and extrapolation of annual estimates.**

**Standard recommendation: NSAs are recommended to use already available quarterly information that displays a correlation with annual data investment estimates. No specific quarterly series is recommended as the choice will depend on data availability. Potential proxies include: 1) GFCF on assets relating to data production, such as computer hardware and software, 2) ROE for data producing occupations, or 3) ROE for industries heavily**

<sup>(27)</sup> This point also confirms the convention that inventories of data do not exist. Since a data asset could conceivably always be added too and therefore always be considered a work-in-progress, it has been agreed that this option is not a useful treatment for either compilation or analytical purposes.

**involved in data production. In the interim, NSAs are recommended to extrapolate annual estimates using a trended time series until a more suitable indicator can be identified.**

Advanced recommendation: **NSAs are recommended to obtain or develop specific quarterly information relating to business expenditure on data production, which can be used to extrapolate the annual estimate.**

- 2.90. All methods and compilation inputs discussed previously in this chapter have been annual in nature. While some of the inputs contributing to the sum-of-cost methodology may be obtainable on a quarterly basis, for many countries, the fine level of occupation and income data is only available annually, or even less frequently. However, there is a need to produce an estimate of data output and GFCF on both annual and quarterly bases regardless of when input information is available.
- 2.91. This is not a unique challenge in the compilation of the national accounts. This handbook recommends the standard practice advocated by other manuals in obtaining a quarterly indicator that is then benchmarked to the annual estimate of data <sup>(28)</sup>. Compilers should strive for a quarterly indicator which is highly correlated with the annual series of data production to minimise revisions. While not always possible, ideally both the quarterly and annual data would come from the same source.
- 2.92. Currently, there would appear to be three alternatives which could be used as a quarterly indicator to move forward annual estimates:
- I. Results from a specific quarterly survey seeking specific information on data production.
  - II. Quarterly information on labor costs of occupations considered as data producing.
  - III. An available quarterly indicator highly correlated to data production.
- 2.93. The first option would consist of a specifically designed quarterly survey that obtains information related to data production. For a range of reasons such as differences in sample size, reduced survey size, etc., it is expected that results from this survey would slightly differ from the annual estimates on a level basis. However, since the specific goal is to derive a correlated growth indicator rather than a level, this is of lesser concern. Indeed, the survey need not even ask expenditure questions, e.g., focusing on other business activity that may correlate with data production. Employee count, hours worked on tasks related to data etc., could be collected, and used if shown that they correlate with annual movements in expenditure on data production. It has been noted that initial discussions with survey responders have shown a lack of clarity surrounding the definition of data. Some approaches to overcome this challenge are discussed in [Box 2.4](#).
- 2.94. Since many organizations are attempting to reduce respondent burden, this first option may not be desirable. Instead, already collected quarterly data which matches that used in the compilation of annual estimates might be available. For example, if it is assumed that within the sum-of-cost methodology, the choice of occupations, the involvement rates, and non-labor cost mark-up do not change on a quarter-by-quarter basis, the most pertinent quarterly indicator would be the labor costs or labor participation of those occupation defined as data producing.
- 2.95. Ideally, labor participation or income data for specific occupations may be available on a quarterly basis. If these are not available, labor expenses from the producer side are a potential alternative that is widely available on a quarterly basis. However, due to the different behaviours of labor costs of the data producing occupations compared with economy wide movements in labor costs, it is not recommended to use an aggregate estimate of labor costs (i.e., economy wide Remuneration of employees (ROE)). An acceptable alternative would be ROE for the specific industries heavily involved in data production (e.g., ISIC 62, "Computer programming, consultancy and related activities" and/or ISIC 63, "Information service activities").
- 2.96. If quarterly labor costs at this lower industry level are not available, the third and final option is to use another quarterly indicator that shows some correlation with output of data. This may be output of software, computer hardware, or output of certain industries. The fundamental requirement is a correlation with the ultimate estimates of data produced

<sup>(28)</sup> The 2025 SNA specifically cites the proportional Denton method or the proportional Cholette–Dagum method as recommended benchmarking approaches.

on an annual basis <sup>(29)</sup>. Overall, it is important to develop quarterly indicators of data which allows for both the interpolation and extrapolation of annual estimates. So, while potentially at least, the best estimates will come from some form of data-centric quarterly survey, NSAs should ultimately choose, like the extrapolation of other annual estimates in the national accounts, a series which minimises revisions and provides the highest level of correlation.

#### BOX 2.4 COLLECTING INFORMATION ON DATA FROM QUARTERLY BUSINESS SURVEYS

Initial testing and engagement with businesses on the topic of data production has suggested that the statistical definition of data provided in [Chapter 1](#) may be unhelpful in obtaining information from businesses on data production. While the definition of data as *“Information content that is produced by accessing and observing phenomena; recording, and storing information elements from these phenomena in a digital format, which provide an economic benefit when used in productive activities”* is suitable for inclusion in a statistical framework such as the SNA, it is likely that other wording will need to be developed in order to obtain information from survey respondents.

At the time of writing, this is a step that has yet to be taken by many statistical organizations and, as such, developing best practices for obtaining information on data production is still a work in progress, although several possibilities exist. These include:

- *Testing various wording to achieve a greater response from survey respondents. By removing unnecessary language while still maintaining the critical points of the definition will likely result in a better understanding and response. Examples may include:*
  - What was the business’s level of expenditure on producing digitalized information content for use in productive activities? (Please include both labor and non-labor costs).
  - What was the business’s level of expenditure on digitally recording facts and information, including those created as a byproduct of production in order to use this information content at a later date? This includes both set up costs and costs to maintain any such system (please include both labor and non-labor costs).
- *Statistical offices could focus on asking about the inputs contributing to data production rather than expenses related to a data output. Survey questions could focus on expenditure by businesses on completing specific tasks related to data production. It’s likely that these tasks may be better understood by survey respondents, although this would need to be tested to ensure that respondents are still able to provide such information.*
- *Explicitly reach out to large data producers to better understand how NSA’s might obtain this information.*

An approach which may require more resources, but which would improve respondent understanding would be to tailor questions for each industry by including examples of data production occurring within that industry. Such examples would assist responders’ understanding of the questions.

Finally, since any information obtained on a quarterly basis is predominately used to obtain growth indicators, useful information may still be obtained even if it comes with some vagueness regarding the exact conceptual boundary of data from the SNA perspective. Due to this, it may be worthwhile for statistical organizations to simply ask for the level of expenditure on ‘data production’ and see what corporations are able to provide. A consistent approach, even if not conceptually perfect would still produce results that may be used for extrapolation purposes.

<sup>(29)</sup> Although the driving requirement is a correlation between the annual and quarterly series, the quarterly indicator must also be available in the required time frame and be consistent with the fundamental concepts of the SNA.

## Annex 2.1. Summary of estimation methodology

Step	Short description	Standard recommendation	Alternative/Advanced Methods & Approaches (where applicable)
<b>Compute own-account data</b>			
1	Identify occupations involved in data production	Occupation list compiled from task team	[1] Survey [2] Expert knowledge [3] Key words within statistical classification [4] Natural language processing (NLP) on job advertisements [5] Based on occupations selected by another country
2	Determine the number of employees in each occupational group	Survey / census data	
3	Determine the average wage in each occupational group	Survey / census data	
4	Determine the involvement rate for each occupational group	List of involvement rates compiled by task team	[1] Survey [2] Expert knowledge / best guesses [3] Natural language processing (NLP) on job advertisements [4] Based on involvement rates from another country. [5] Key word search using occupation classification
5	Calculate total labor costs	(number of employees) x (average wage) x Adjustment to include non-wage labor costs x (involvement rate)	
6	Calculate non-labor costs, includes <ul style="list-style-type: none"> <li>– Cost of inputs</li> <li>– Depreciation of assets used in production</li> <li>– Return to capital</li> </ul>	One single mark-up applied to labor costs <ul style="list-style-type: none"> <li>– Based on ratio from ISIC 62 and ISIC 63 (Total output / ROE)</li> </ul>	[1] Survey focussing on data production expenses [2] One single mark-up on labor costs [a] Based on industries with large amount of output from data-producing occupations [3] Separate mark-ups covering intermediate consumption expenses, taxes less subsidies on production, CFC, and return to capital  [a] Based on industries with large amount of output from data-producing occupations [b] Based on ISIC 62 and ISIC 63

Step	Short description	Standard recommendation	Alternative/Advanced Methods & Approaches (where applicable)
7	Compute domestic own account data output	(labor costs) step 5 + (non-labor costs) step 6	
8	Calculate purchase of data copies produced domestically		Survey
<b>Compute estimates of total domestic output of data</b>			
9	Calculate domestic output of data	= to total provided in step 7 + domestic purchase of data copies (step 8)	
10	Calculate total data output by industry and sector	Use indicator (i.e., labor costs, occupation count) to break up aggregate provided in step 9	Domestic data purchases can be used to adjust industry and sector totals.
<b>Compute purchased data</b>			
11	Calculate net imports of data (For Data GFCF only)	Apply ratios to trade data to delineate estimates of data imports and exports	[1] Survey [2] Use net imports from similar IPP assets
<b>Total data investment (GFCF)</b>			
12	Calculate total data investment	Own-account data (Step 7) + Purchase of data copies produced domestically (step 8) + Net imports of data (step 11)	
13	Calculate total data investment disaggregated by industries and sectors	Use indicator (i.e., labor costs, occupation count and domestic purchases) to break up total provided in step 12	Domestic data purchases can be used to adjust industry and sector totals.
<b>Compute volume estimates of data</b>			
14	Identify or estimate price indices for data	Use IPD for similar IPP assets, taken from SUTs	Compile data specific input price index based on input costs used in the production of data. Include adjustment for quality and productivity improvements
15	Compute price-adjusted total data investment	Total data investment in current prices (step 12) / price index (step 14)	
16	Compute price-adjusted total data output	Total data output in current prices (step 8) / data price index (Step 14)	

Step	Short description	Standard recommendation	Alternative/Advanced Methods & Approaches (where applicable)
<b>Compute capital stocks and CFC</b>			
17	Determine average service lives	Recommendation provided by task team (see <a href="#">chapter 4</a> )	[1] Tax lives [2] Company accounts [3] Statistical surveys [4] Administrative records [5] Expert advice [6] Other countries' estimates [7] Implicit service lives in depreciation rates [8] Based on service lives of similar assets
18	Determine the functional form of depreciation	Consistent with current country practice for deriving capital stock for other IPP	[1] Linear [2] Geometric
19	Specify the retirement pattern	Consistent with current country practice for deriving capital stock for other IPP	[1] Simultaneous exit [2] Linear [3] Delayed linear [4] Bell-shaped [a] Winfrey distribution [b] Weibull distribution [c] Gamma distribution [d] Lognormal distribution [e] Quadratic distribution
20	Calculate sufficiently long times series of total data investment and output	Use appropriate indicators and apply to total data investment (step 13) and total data output (step 9) to create time series	Use business indicators to move data back  Based on time series of similar (IPP) assets
21	Calculate capital stocks and CFC	Perpetual Inventory Method (PIM) Using information from step 17, 18, 19, 20	

## Annex 2.2. Using natural language processing to better understand data intensity for certain occupations

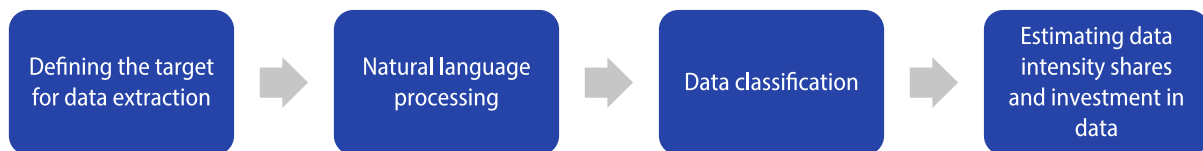
In 2023, the OECD generated occupation and industry level estimates of data intensity using natural language processing (NLP) (Schmidt, Pilgrim, & Mourougane, 2023). The work built off earlier work by Calderón & Rassier (2022) by expanding the methodology to a range of countries in order to test its suitability over time and location.

The study used NLP to review online job advertisement to derive the share of jobs involved in data production, referred to as “data intensity.” The basic methodology, outlined below in [Figure 2.5](#), identified specific text within job advertisements to determine not only if the occupation was data intensive but also the level of data intensity.

1. **Defining data-intensive jobs:** Job advertisements are a rich source of information to understand the labor demand. The data value chain put forward by Statistics Canada (Statistics Canada, 2019) and (Corrado, Haskel, Iommi, & Jona-Lasinio, The value of data in digital-based business models: Measurement and economic policy implications, 2022) is used as a conceptual framework to determine whether a job is data-intensive.
2. **Deploying Natural Language Processing:** The text data are cleaned of noise, and quality and consistency checks are deployed to check the properties of the data. Subsequently, the NLP algorithm performs the text feature extraction, which transforms text data into a mathematical object that can be classified.
3. **Classifying data:** The parts of the online job advertisement identified as data-related skills and tasks are classified into data entry, database, and data analytics related capabilities to allow for a breakdown by these types of data-related production activities.
4. **Deriving the data intensity by occupation and sector and estimates of investment in data:** In a final step, the data-intensive jobs are aggregated to derive a data intensity share by occupation and sector and matched onto national accounts data to calculate an estimate of investment in data at sector and economy level.

**FIGURE 2.5**

## Overview of the NLP approach



Source: (Schmidt, Pilgrim, & Mourougane, 2023)

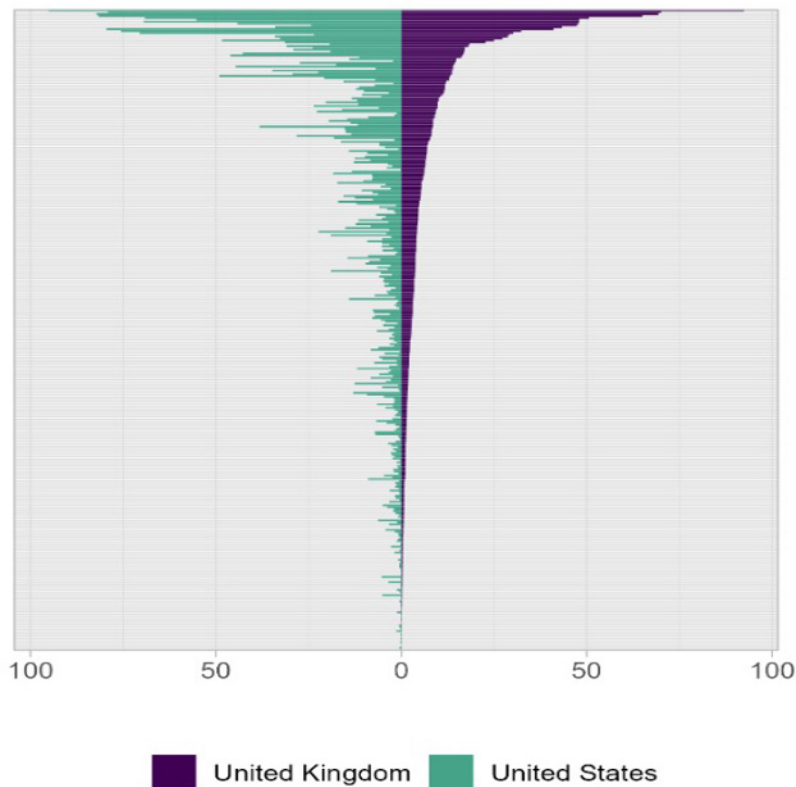
The final step of the project (combining data intensity shares with national account data to derive an estimate of investment in data) uses the same fundamental methodology as advocated in this handbook. Estimates of data investment were broadly in line with those already produced by countries. Many additional details on the process, including formula are available in Schmidt, Pilgrim, & Mourougane, 2023.

Overall, the work found that the high data-intensive occupations were very similar in all three countries for most professions, although these occupations in Canada and the United States exhibited slightly higher data intensity. Importantly the research supported the idea that data production was occurring across the economy with data intensive jobs found in all industries, albeit at different levels of intensity. That said, if this process is undertaken for the purpose of estimating data production, one would need to be careful to ensure that the specific text used to determine data intensity focusses on data production rather than simply utilising data as part of the role.

As shown in [Figure 2.6](#) in both the United States and the United Kingdom, the range of occupations that recorded some level of data intensity was significant, however both countries exhibit a long tail, with most occupations displaying a data intensity value of less than 15. These results were similar for Canada. Importantly, the results were relatively stable for the period studied (2012-2020) suggesting that the data source could be considered as suitable for use in compiling national account aggregates.

**FIGURE 2.6**

## Data intensity across occupation classes; United Kingdom and United States, 2020



Each row represents a single occupation. A rating closer to 100 suggest that the occupations is quite data intensive, while a rating of 0 reflects minimal data intensity.  
Source: (Schmidt, Pilgrim, & Mourougane, 2023)

Such evidence is useful for countries when selecting the specific data producing occupations and their involvement rates. This work suggests that while data production may be taking place across a large number of occupations, for many occupations, involvement rates could be quite low, with high involvement rates concentrated in a small number of specific occupations.

### Annex 2.3. Production of Special Internet Survey in Japan

In 2022, the Japanese Cabinet Office undertook the first iteration of the ‘Special Internet Survey in Japan’. The aim of the survey, which was sent to over 30,000 workers considered as engaging in data related occupations, was to gain greater insight into the characteristics of workers involved in the production of data and data analysis. The survey sought to determine the specific tasks that data producers undertake and how much of their working time is spent on data related tasks. Results from such a survey can prove beneficial to compilers trying to derive an estimate of both the level of capital formation and capital stock of data in the economy.

The survey included questions on the tasks listed below. For each of the eight specific questions regarding data producing tasks listed below, respondents were asked if they were.

- a) Engaged in the following tasks in your current job?
- b) How long they spent on each task? Please enter the rate for your total working hours (from 0 to 100 %).

1	Develop a plan to collect and utilize internal or external data
2	Prepare environment to produce data (e.g. guide and support the survey respondents)
3	Input or record data generated internally or externally (e.g. record information from business memos, surveys, customer inquiries)
4	Operate an application that collects data automatically (e.g., an app which collects search history or telephone history from smartphones)
5	Arrange and organize collected data for ease of use
6	Develop and operate a database
7	Analyse data (e.g., statistical analysis, create company management indicators)
8	Develop, manufacture, or maintain infrastructure or system to collect data automatically

The Cabinet office considers tasks 1-5 as contributing to the production of data, task 6 as contributing to the production of databases. Task 7-8, the analysis of data and tasks involved with data collection tools data are, by themselves, conceptually outside of the conceptual framework of data used in the SNA and may be considered as involved in the production of other services related to data but not considered as involved in producing data itself. That said, the information is still worth obtaining to assist in clarifying the line between data producing and data using occupations.

Many occupations may undertake several (or even all) task and so obtaining results from the survey do not answer all questions involved in the compilation of data output, however the survey goes a long way in providing valuable information that can be used to provide more confidence in the estimate being compiled.

## Initial Results from Japanese special internet survey

Responses to the following questions

- Have you engaged in the following tasks in your current job?
- How long have you spent on each task? Please enter the rate for your total working hours (from 0 to 100 %).

Develop a plan to collect and utilize internal or external data			
Planning		No.	%
<b>Total</b>		30,295	100.0
1	Yes	5,846	19.3
2	No	24,449	80.7
If yes, % of Time spent on task			
Responses		5,846	
Average %		24.63%	

Prepare and maintain environment for producing data (e.g. sending request and follow-up to the survey respondents, etc., process for rewarding points to respondents)			
Preparing and maintaining environment		No.	%
<b>Total</b>		30,295	100.0
1	Yes	5,004	16.5
2	No	25,291	83.5
If yes, % of Time spent on task			
Responses		5,004	
Average %		16.77	

**Entering or recording various types of data generated internally or externally (e.g. recording information from surveys or experiments, reading from store registers, entering business data, recording inquiries)**

Entering or recording		No.	%
Total		30,295	100.0
1	Yes	8,494	28.0
2	No	21,801	72.0

**If yes, % of Time spent on task**

Responses	8,494
Average %	25.07

**Database development or operation**

		No.	%
Total		30,295	100.0
1	Yes	3,236	10.7
2	No	27,059	89.3

**If yes, % of Time spent on task**

Responses	3,236
Average %	12.93

**Analysis of data (e.g. statistical analysis, creation of various management indicators, analysis of big data)**

		No.	%
Total		30,295	100.0
1	Yes	5,228	17.3
2	No	25,067	82.7

**If yes, % of Time spent on task**

Responses	5,228
Average %	17.35

**Operating applications where various types of data are automatically collected (e.g. search log aggregation tools for smartphone applications, applications for collecting order and communication records)**

Operating automated application		No.	%
Total		30,295	100.0
1	Yes	4,063	13.4
2	No	26,232	86.6

**If yes, % of Time spent on task**

Responses	4,063
Average %	12.61

**Arrange and organize collected business data (e.g. data related to sales performance, production runs, customers, comments on Social Media, web access logs) in a user-friendly manner**

		No.	%
Total		30,295	100.0
1	Yes	6,865	22.7
2	No	23,430	77.3

**If yes, % of Time spent on task**

Responses	6,865
Average %	16.93

**Either develop, manufacture, or support and maintain equipment and systems to automatically collect data (e.g., POS, search history, movement information)**

		No.	%
Total		30,295	100.0
1	Yes	2,925	9.7
2	No	27,370	9.3

**If yes, % of Time spent on task**

Responses	2,925
Average %	11.20

## Annex 2.4. Key words used for web scrapping job vacancies

Source: Statistics Canada

Data processing	Data communication	Forecasting	Clinical database development
Data hosting	Optical scanning	Data collection	Clinical research data accuracy and integrity
Data entry	Spreadsheet	Data architecture	Cloud security data protection and privacy
Statistical routines	Data operation	Data warehouse	Columnar databases
Monitor trends	Data evaluation	Data archiving	Conceptual data models
Data compilation	Data transformation	Big data engineering	Customer data integration
Data integration	Data manipulation	Data security	Customer service database
Data preparation	Time series	Data queries	Data acquisition
Charts	Data wrangling	Database performance	Data acquisition systems
Graphs	Data cleaning	Database migration	Data and safety monitoring board
Data summaries	Statistical analysis	Market analysis	Data buffers
Data reports	Financial trading	Data training	Data capture
Data support	Statistical consulting	Business intelligence developing	Data center hardware
Data survey	Data coding	Data advertising	Data compression
Data clerk	Big data analysis	3d seismic data	Data conversion
Data input	Data visualization	Accenture data governance framework	Data dictionary system
Data analysis	Statistical knowledge	Advanced data entry	Data documentation
Data analytics	Data dashboards	Assessment data	Data encryption
Machine learning	Data models	Billing data analysis	Data exploitation
Predictive modelling	Regression testing	Biological database search	DFHSM
Data automation	Data reporting	Business intelligence data modelling	Data governance
Database analysis	Econometrics	Cascading big data applications	Data integrity
Data administration	Aerial survey	Climate data analysis	Data lakes
Database design	Remote sensing	Clinical data abstracting	Data loss prevention
Database development	Data informatics	Clinical data analysis	Data management platform
Data management	Deep learning	Clinical data exchange	Data mapping
Data mining	Big data	CDISC	Data migration
Data monitoring	Programming languages	Clinical data management	Data mining industry knowledge
Data science	NoSQL database	Clinical data review	Data modelling
Survey interviewing	Analytic tools	Clinical data understanding	
Data recording	Quantitative analysis		
Research support	Business analytics		

Data modelling star	Data access object	Activity based costing	Clinical data interchange standards consortium (CDISC)
Data multiplex system	Open database connectivity (ODBC)	Ad hoc analysis	Cluster analysis
Data munging	Fiber distributed data interface (FDDI) master data management (MDM)	Ad hoc marketing	Clustering
Data operations	Amazon elastic map	Ad hoc market research	Competitive analysis
Data platform as a service	Data frame	Ad hoc reporting	Correlation analysis
Data pre-processing	Informatica	ADABAS	Data-driven testing
Data privacy	Teradata dba	Adaptable database system	Data communications
Data protection industry knowledge	Economics	Advanced encryption standard (AES)	Data engineering
Data protection planning	Panel data	Advanced statistics	Data flow diagrams (DFDS)
Data protection strategy	OLS	Analysis of variance (Anova)	Data management platform (DMP)
Data quality	Ordinary least squares	Artificial intelligence	Data services industry knowledge
Data quality assessment	Statistical visualization	Audit reports	Data taxonomy
Data security classification	Data officer	Balance sheet	Data trending
Study data tabulation model	Census	Balance sheet analysis	Data validation
Clinical data interchange standards consortium	Data scientist	Bayesian networks	Data verification
Snowflake schema	Wellhead plumbers	Benefits analysis	Data warehouse development
Data reservoirs	Amazon web service	Benchmarking	Data warehouse processing
Linear regression	AWS	Bill of lading	Database activity monitoring
Logistic regression	Hearing aid specialist	Biostatistics	Database administration
Supervised learning	Bioinformatics specialist	Bookkeeping	Database architecture
Teradata	Landfill meters	Budget forecasting	Database cloning
Metadata	Data structure	Budgeting	Database consolidation
Database tuning	LinkedList	Business analysis	Database conversion
Oracle database administration	Descriptive analytics	Business impact analysis	Database information accessing
Relational database management system	Diagnostic analytics	Business intelligence	Database information accessing
Electronic data interchange	Predictive analytics	Business intelligence reporting	Database maintenance
ActiveX data object	Prescriptive analytics	Business metrics	Database management
Database upgrade	Data exploration	Business modeling	Database marketing
Relational databases	Data exploration	Business process modelling	Database modeling
IBM infosphere DataStage	5500 reporting	Business systems analysis	Database optimization
Database partitioning	8d problem solving	Call-recording	Database performance management
Data acquisition	Accident analysis	Call volume/time analysis	Database programming
Database schemas	Accident reporting	Climate analysis	
	Audit data analytics		

Database responsibilities	Factor analysis	Qualitative analysis	Statistics
DataFlux	Financial data interpretation	Qualitative data analysis	Supplier database maintenance
Dataloggers	Financial forecasting	Quantitative analysis	Supply chain data analysis
Demand analysis	Flight data analysis	Quantitative data analysis	Syndicated market data
Demand forecasting	GoodData	Quantitative research	Test data management
Economic analysis	GPS data	Regression analysis	Time series analysis
Economic development	Health data analysis	Regression algorithms	Time series forecasting
Economic forecasting	Health data management	Sales database	Time series models
Economic models	Health databases	Sales metrics	Variance analysis
Enterprise data management	High-level data link control (HDLC)	Social data	Data center
Enterprise data services	Industry analysis	Statistical forecasting	Data center technician
Environmental data analysis	Market data	Statistical methods	
Environmental data management	Market research	Statistical modelling	
Environmental economics	Metadata design	Statistical process control (SPC)	
EpiData	Metadata standards	Statistical reporting	

## Annex 2.5. Using the CROM for Selecting Data-Relevant Occupations and Computing Involvement Rates

### Introduction

As part of the research into which occupations may be considered as data producing, The Federal Statistical Office of Germany (FSO) developed the Competence-Relevant Occupation Methodology (CROM) to select data-relevant occupations and compute involvement rates. This methodology was in response to potential weaknesses observed in occupation lists derived through expert knowledge or machine learning. This approach aligns with the recommendation in the handbook, that “countries should look to develop a list of occupations through an objective and systematic approach to better determine which occupations are most likely to be involved in data production.” The approach outlined below present a good example of how countries might use relevant skills, tasks or competencies listed within the occupation classification to derive a list of data producing occupations as well as determine an appropriate involvement rate based on the prevalence of the skills, task, or competencies assigned to that occupation.

Pleasingly, the results obtained from the CROM have produced a similar curve of occupation involvement rates as other work. This curve displays a long list of data producing occupations but with only a small percentage (around 11% in this methodology) having involvement rates of greater than 10% (See [Figure 2.8](#)). This is similar to the results obtained in Japan where only around 10% of occupation had an involvement rate greater than 10%.

It should be noted that thorough testing of this methodology has not been fully completed as the list of occupations and involvement rates have not yet been applied to relevant wage rates to compile an estimate of data output. The work so far has focused purely on developing an occupation list and involvement rate which can be compared with those compiled through other sources such as machine learning or surveys.

## The CROM methodology follows five steps

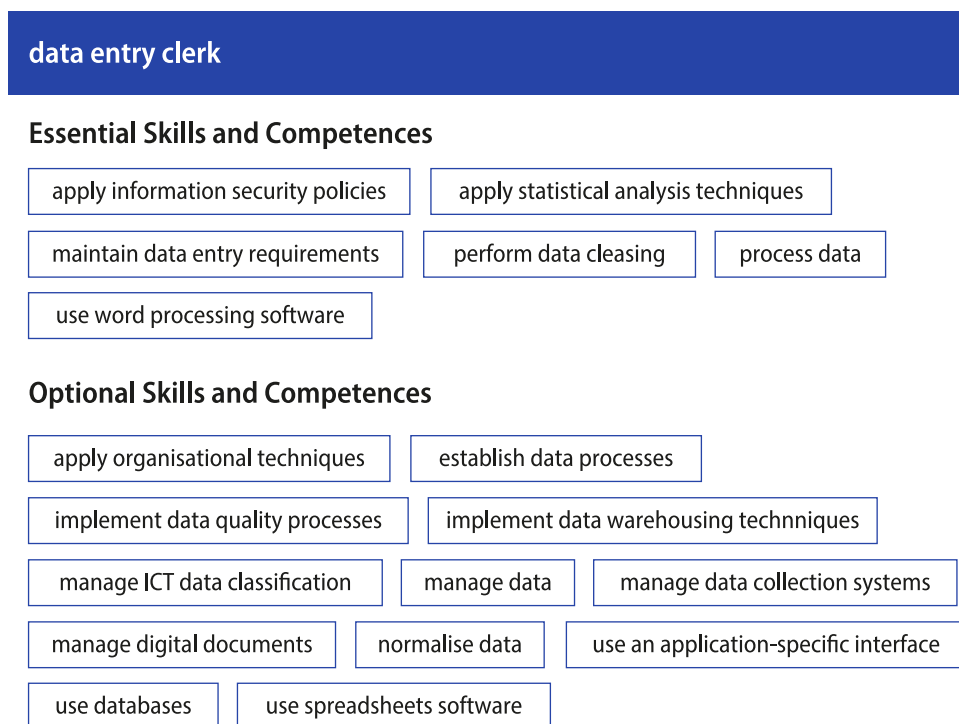
### Step 1: Find Occupational Information System

The first step in applying the CROM is to identify an appropriate occupational information system. Examples include [ESCO](#), [BERUFENET](#), or [O\\*NET](#), which offer comprehensive data on occupations and their associated competences, skills, or tasks. The FSO primarily uses BERUFENET due to its relevance in the German context. However, for this explanation, ESCO is used because of its broader applicability across countries.

ESCO links 13,939 competences to 3,039 occupations. Competences are categorized as essential or optional. [Figure 2.7](#) illustrates the type of occupational information needed to apply the CROM, using the example of a “data entry clerk” from ESCO.

**FIGURE 2.7**

### Competence breakdown for “data entry clerk” in ESCO



### Step 2: Identify Data-Relevant Competences

Out of ESCO’s 13,939 competences, 80 were identified as relevant for data and database activities ([Table 2.1](#)). For example, for the occupation “data entry clerk,” the following data-relevant competences were identified:

- Essential: *maintain data entry requirements, perform data cleansing, process data.*
- Optional: *establish data processes, implement data quality processes, implement data warehousing techniques, manage ICT data classification, manage data, manage data collection systems, normalize data, use databases.*

Other competencies listed, such as ‘*use of spreadsheet software*’ may have some implications for the production of data. However, consistent with what is presented in the handbook recommendation it was considered that only those

competencies that are *likely* in the production of data should be included rather than competencies that are *possibly* involved.

**TABLE 2.1**

## Competences used for selecting data-relevant occupations and calculating involvement rates

Out of ESCO's total of 13,939 competences, the following 80 were identified as relevant for data and database activities:

- Record test data
- manage findable accessible interoperable and reusable data
- obtain financial information
- process data
- collect biological data
- manage data
- keep records on sales
- record survey data
- use data processing techniques
- use databases
- maintain customer records
- analyze big data
- synthesize financial information
- manage database
- migrate existing data
- collect healthcare user's general data
- gather data
- document interviews
- create data models
- maintain credit history of clients
- manage data collection systems
- process collected survey data
- establish data processes
- perform data cleansing
- implement data quality processes
- manage ICT data architecture
- design database scheme
- maintain veterinary clinical records
- perform data mining
- manage cloud data and storage
- collect geological data
- design database in the cloud
- design cloud architecture
- collect mapping data
- analyze data about clients
- integrate ICT data
- normalize data
- record data from biomedical tests
- create data sets
- write database documentation
- balance database resources
- handle data samples
- maintain database security
- collect data using GPS
- collect weather-related data
- collect customer data
- maintain data entry requirements
- maintain warehouse database
- collect financial data
- create database diagrams
- manage standards for data exchange
- operate relational database management system
- collect healthcare user data under supervision
- maintain database performance
- supervise data entry
- manage quantitative data
- compile GIS-data
- implement data warehousing techniques
- manage ICT data classification
- tabulate survey results
- compile statistical data for insurance purposes
- explain interview purposes
- gather data for forensic purposes
- analyze large-scale data in healthcare
- collect ICT data
- record malting cycle data
- use questioning techniques
- conduct public surveys
- design questionnaires
- store digital data and systems
- define database physical structure
- design database backup specifications
- maintain logistics databases
- manage data
- information and digital content
- collect cyber defence data
- compile data for navigation publications
- create freight rate databases
- maintain pricing database

- develop geological databases
- interview focus groups

### Step 3: Select Data-Relevant Occupations

Occupations are considered data-relevant if they include at least one data-relevant competence, whether essential or optional. Using this approach, the FSO identified 641 data-relevant occupations in ESCO.

### Step 4: Calculate Involvement Rates

Involvement rates are computed by translating competences into a numerical value, with double weight given to essential competences. The formula used is:

$$IR_i = \frac{2 \times DE_i + DO_i}{2 \times E_i + O_i},$$

Where:

$IR_i$  = Involvement rate of occupation  $i$ .

$DE_i$  = Number of data-relevant essential competences for occupation  $i$ .

$DO_i$  = Number of data-relevant optional competences for occupation  $i$ .

$E_i$  = Total number of essential competences for occupation  $i$ .

$O_i$  = Total number of optional competences for occupation  $i$ .

For the data entry clerk in [figure 2.7](#), the formula results in an involvement rate of

$$IR_i = \frac{2 \times 3 + 8}{2 \times 6 + 12} = 58\%,$$

meaning they spend approximately 58% of their time on data-related activities. Using this formula, [Table 2.2](#) ranks the top 5 occupations based on the ESCO dataset.

**TABLE 2.2**

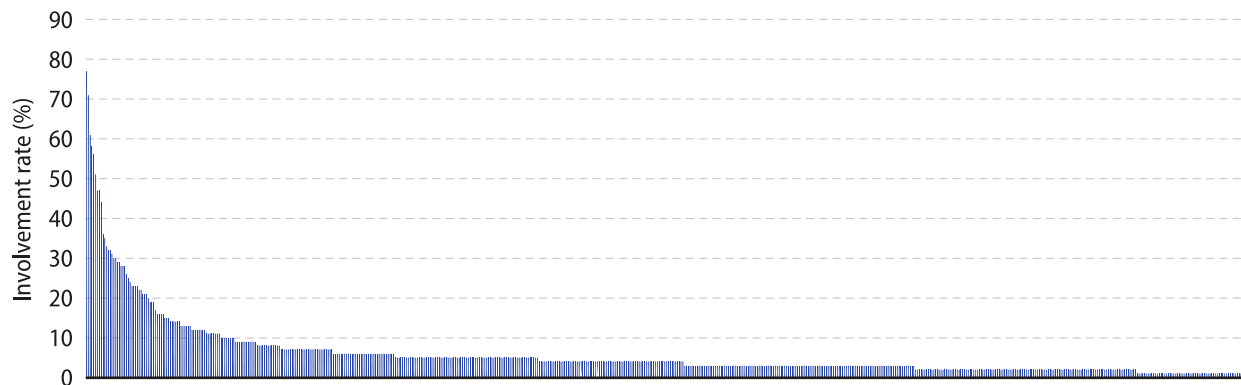
### Top 5 occupations in data-related tasks

ESCO	Occupations	Involvement rate
251120	Data Engineer	77%
25113	Data Analyst	71%
25191	Data Quality Specialist	61%
41321	Data Entry Clerk	58%
34332	Big Data Archive Librarian	56%

The 641 occupations identified as involved in data production ranged from data engineers, with an involvement rate of 77%, to specialized sellers, with a rate of 0.4%. Overall, the distribution of involvement rates across the occupations (see [Figure 2.8](#)) was similar to the results obtained through other compilation methodologies such as machine learning and survey job holders.

**FIGURE 2.8**

## Distribution of involvement rates across data producing occupations



### Step 5: Averaging Involvement Rates Across Occupational Groups

The formula above is applied to narrow occupations, and the rates are then averaged (with equal weighting) across broader ISCO-08 groups. This process helps to determine data relevance at the unit group level, even when detailed statistics at the occupation level are unavailable. For example, as shown in [Table 2.3](#), the involvement rates of ESCO occupations 25211 to 25215 are averaged to calculate the rate for ISCO-08 group 2521. This aggregation results in 146 data-relevant occupational groups at the four-digit ESCO level. [Table 2.4](#) ranks the top 5. This consolidation makes it more likely that the occupations and their respective involvement rates can be applied to occupation-based employment and wage data.

**TABLE 2.3**

### Averaging involvement rates for ISCO-08 group 2521

ESCO	ISCO-08	Occupation (Groups)	Involvement rate
25211		Database Administrator	44%
25212		Database Designer	51%
25213		Database Developer	24%
25214		Database Integrator	47%
25215		Data Warehouse Designer	47%
	2521	<b>Database Designers and Administrators</b>	<b>43%</b>

**TABLE 2.4****Top 5 occupational groups in data-related tasks**

ISCO-08	Occupation Unit Groups	Involvement rate
4132	Data Entry Clerks	58%
2521	Database Designers and Administrators	43%
3511	Information and Communications Technology Operations Technicians	33%
4227	Survey and Market Research Interviewers	29%
3314	Statistical, Mathematical and Related Associate Professionals	25%

**Annex 2.6. Standard list of data producing occupations and involvement rates**

ISCO code	Occupation	Proposed involvement rate	ISCO code	Occupation	Proposed involvement rate
4132	Data entry clerks	64%	3339	Business services agents not elsewhere classified	4%
4227	Survey and market research interviewers	52%	4214	Debt Collectors and Related Workers	4%
2521	Database designers and administrators	32%	4229	Client information workers not elsewhere classified	4%
2120	Mathematicians, actuaries and statisticians	25%	2132	Farming, Forestry and Fisheries Advisers	3%
3314	Statistical, mathematical and related associate professionals	25%	2133	Environmental protection professionals	3%
3511	Information and communications technology operations technicians	25%	4110	General office clerks	3%
2511	Systems analysts	20%	4221	Travel Consultants and Clerks	3%
4131	Typists and word processing operators	18%	4222	Contact Centre Information Clerks	3%
2165	Cartographers and surveyors	16%	4225	Inquiry Clerks	3%
2529	Database and network professionals not elsewhere classified	16%	4411	Library Clerks	3%
3513	Computer network and systems technicians	16%	4415	Filing and Copying Clerks	3%
2512	Software developers	15%	4416	Personnel Clerks	3%
3312	Credit and loans officers	14%	1212	Human Resource Managers	2%
2164	Town and traffic planners	12%	2211	Generalist Medical Practitioners	2%

ISCO code	Occupation	Proposed involvement rate	ISCO code	Occupation	Proposed involvement rate
2514	Applications programmers	12%	2212	Specialist Medical Practitioners	2%
2519	Software and applications developers and analysts not elsewhere classified	12%	2422	Policy administration professionals	2%
2522	Systems administrators	12%	2423	Personnel and Careers Professionals	2%
2523	Computer network professionals	12%	3230	Traditional and Complementary Medicine Associate Professionals	2%
3315	Valuers and loss assessors	12%	3333	Employment Agents and Contractors	2%
3433	Gallery, Museum and Library Technicians	12%	3334	Real Estate Agents and Property Managers	2%
1330	Information and communications technology service managers	10%	3344	Medical Secretaries	2%
2160	Architects, planners, surveyors and designers, nos	10%	3353	Government Social Benefits Officials	2%
2513	Web and multimedia developers	10%	3359	Regulatory government associate professionals not elsewhere classified	2%
3211	Medical imaging and therapeutic equipment technicians	10%	4212	Bookmakers, Croupiers and Related Gaming Workers	2%
3212	Medical and pathology laboratory technicians	10%	4224	Hotel Receptionists	2%
3213	Pharmaceutical Technicians and Assistants	10%	4226	Receptionists (general)	2%
3252	Medical records and health information technicians	10%	7233	Agricultural and Industrial Machinery Mechanics and Repairers	2%
4311	Accounting and bookkeeping clerks	10%	7543	Product Graders and Testers (excluding Foods and Beverages)	2%
4312	Statistical, finance and insurance clerks	10%	1114	Senior Officials of Special-interest Organizations	1%
1211	Finance managers	8%	1223	Research and Development Managers	1%
1346	Financial and insurance services branch managers	8%	1321	Manufacturing Managers	1%
2412	Financial and investment advisers	8%	1411	Hotel Managers	1%
2413	Financial analysts	8%	1431	Sports, Recreation and Cultural Centre Managers	1%

ISCO code	Occupation	Proposed involvement rate	ISCO code	Occupation	Proposed involvement rate
2621	Archivists and curators	8%	1439	Services Managers Not Elsewhere Classified	1%
2622	Librarians and related information professionals	8%	2140	Engineering professionals (excluding electrotechnology), nos	1%
2631	Economists	8%	2141	Industrial and Production Engineers	1%
3114	Electronics engineering technicians	8%	2142	Civil Engineers	1%
3115	Mechanical engineering technicians	8%	2143	Environmental engineers	1%
3116	Chemical engineering technicians	8%	2144	Mechanical Engineers	1%
3117	Mining and metallurgical technicians	8%	2145	Chemical Engineers	1%
3119	Physical and engineering science technicians not elsewhere classified	8%	2146	Mining Engineers, Metallurgists and Related Professionals	1%
3141	Life science technicians (excluding medical)	8%	2149	Engineering professionals not elsewhere classified	1%
3311	Securities and finance dealers and brokers	8%	2166	Graphic and Multimedia Designers	1%
3321	Insurance representatives	8%	2230	Traditional and Complementary Medicine Professionals	1%
3341	Office Supervisors	8%	2250	Veterinarians	1%
3512	Information and communications technology user support technicians	8%	2262	Pharmacists	1%
4211	Bank tellers and related clerks	8%	2264	Physiotherapists	1%
2111	Physicists and astronomers	6%	2269	Health Professionals Not Elsewhere Classified	1%
2112	Meteorologists	6%	2310	University and Higher Education Teachers	1%
2113	Chemists	6%	2351	Education Methods specialists	1%
2114	Geologists and geophysicists	6%	2421	Management and Organization Analysts	1%
2131	Biologists, botanists, zoologists and related professionals	6%	2619	Legal Professionals Not Elsewhere Classified	1%
2411	Accountants	6%	2643	Translators, Interpreters and Other Linguists	1%
2632	Sociologists, anthropologists and related professionals	6%	3118	Draughts persons	1%

ISCO code	Occupation	Proposed involvement rate	ISCO code	Occupation	Proposed involvement rate
2633	Philosophers, historians and political scientists	6%	3122	Manufacturing Supervisors	1%
2634	Psychologists	6%	3123	Construction Supervisors	1%
3342	Legal secretaries	6%	3133	Chemical Processing Plant Controllers	1%
3352	Government tax and excise officials	6%	3139	Process Control Technicians Not Elsewhere Classified	1%
3514	Web Technicians	6%	3142	Agricultural Technicians	1%
4213	Pawnbrokers and Money-lenders	6%	3153	Aircraft Pilots and Related Associate Professionals	1%
4313	Payroll clerks	6%	3154	Air Traffic Controllers	1%
4321	Stock clerks	6%	3240	Veterinary Technicians and Assistants	1%
4322	Production clerks	6%	3255	Physiotherapy Technicians and Assistants	1%
4323	Transport clerks	6%	3257	Environmental and Occupational Health Inspectors and Associates	1%
1213	Policy and planning managers	4%	3259	Health Associate Professionals Not Elsewhere Classified	1%
1219	Business services and administration managers not elsewhere classified	4%	3313	Accounting Associate Professionals	1%
1342	Health Services Managers	4%	3323	Buyers	1%
2152	Electronics Engineers	4%	3324	Trade Brokers	1%
2153	Telecommunications engineers	4%	6221	Aquaculture Workers	1%
2431	Advertising and marketing professionals	4%	7231	Motor Vehicle Mechanics and Repairers	1%
2432	Public relations professionals	4%	7232	Aircraft Engine Mechanics and Repairers	1%
2642	Journalists	4%	7321	Pre-press Technicians	1%
3111	Chemical and Physical Science Technicians	4%	7421	Electronics Mechanics and Servicers	1%
3112	Civil Engineering Technicians	4%	8111	Miners and Quarriers	1%
3113	Electrical Engineering Technicians	4%	9216	Fishery and Aquaculture Laborers	1%
3214	Medical and Dental Prosthetic Technicians	4%			

## Annex 2.7. Numerical example of standard non-labor cost mark up

Due to the absence of specific information on non-labor costs involved in data production, the recommendation in this handbook is for non-labor costs to be incorporated into the final estimate of data production via a single mark-up applied to labor costs. Such a mark-up represents the costs of inputs, taxes less subsidies on production, depreciation used in production, as well as a return to capital (operating surplus).

The standard recommendation for calculating such a mark-up is to base it on the ratio of total Remuneration of employees for “Computer programming, consultancy and related activities” (ISIC Rev.5, division 62) and “Computing infrastructure, data processing, hosting, and other information service activities” (ISIC Rev.5, division 63) applied against total gross output for these same industries. The numerical example below shows such a calculation.

The weighted average of the two industries is used to generate a ratio which can be applied as a mark-up. In this example the weighted mark-up of 2.46 is calculated from dividing the sum of gross output (62,677) by the sum of remuneration of employees (25,461) <sup>(30)</sup>. The mark-up is then applied to the data labor costs (15,000), calculated through steps described in this chapter.

This creates a final nominal estimate of data output of 36,925.30

**TABLE 2.5**

### Numerical example of standard non-labor cost mark up

Row #	Economic activity	Computer programming, consultancy and related activities	Information service activities	Sum of both industries
1	Output, Basic prices	53,498.00	9,179.00	62,677.00
2	• Intermediate consumption, Purchasers prices	21,731.00	4,135.00	25,866.00
3	• Value added, gross, Basic prices	31,766.00	5,044.00	36,810.00
4	– Remuneration of employees	22,252.00	3,209.00	25,461.00
5	– Other taxes less other subsidies on production	251.00	-7.00	244.00
6	– Operating surplus and mixed income, gross	9,262.00	1,842.00	11,104.00
7	Non-labour mark-up ratio (Output {row 1} / Remuneration of employees {row 4})	<b>2.40</b>	<b>2.86</b>	
8	Weighted average mark-up	<b>2.46</b>		
9	Value of labor costs of data output	15,000.00		
10	Total data output (labor costs {row 9} * non-labour mark-up ratio {row 8})	36,925.30		

<sup>(30)</sup> The weighted average of the two industries is calculated using the 'remuneration of employees' as the weighting factor.

## Annex 2.8. Summary of recommendations for compiling nominal estimates of data

Measurement step	Broad conceptual recommendation:	Standard recommendation:	Advanced recommendation:	Additional considerations
<b>Valuation approach</b>	Data produced on an own account basis is valued using the already established sum-of-cost method			
<b>Choice of occupation</b>	<p>NSAs should use a list of occupations as the foundation for deriving an estimate of labor costs involved in producing data.</p> <p>Such a list should only include occupations that undertake data producing tasks as an integral part of their role. While occupations will be added and removed over time, reflective of changes in the way that data is produced, in order to ensure that methodological decisions do not create breaks in the time series, this list should remain broadly consistent across periods.</p>	<p>In the absence of other sources, national statistical authorities (NSAs) should use the list of occupations provided in this handbook for the compilation of data output. If the standard list of occupations are used, these must be complimented by the same or very similar involvement rates to those listed in this handbook. The occupation list is presented at the International Standard Classification of Occupations (ISCO – 08) four-digit level for the compilation of data output.</p> <p>If the standard list of occupations is applied, these must be used in unison with the standard involvement rates also applied. As these have been compiled in unison. It is not recommended to apply new or different involvement rates to the standard occupation list.</p>	<p>NSAs are encouraged to derive a list of occupations (and involvement rates) through an objective and systematic approach to better determine which occupations are most likely to be involved in data production. Examples include the use of machine learning or survey information as well as the use of key words when reviewing occupation classifications or job advertisements.</p>	<p>Occupations should be considered for the list if the occupation involves tasks which explicitly contribute to adding value to the production of data <u>and</u> the worker undertakes these tasks as an integral part of their role.</p> <p>It is important to publish the metadata of what occupations have been included (and excluded), so that users have a good understanding of how the results have been derived.</p>

Measurement step	Broad conceptual recommendation:	Standard recommendation:	Advanced recommendation:	Additional considerations
<b>Estimation of Labor costs</b>	<p>The annual nominal estimate of own account data output must reflect non-direct labor costs as well as wages and self-employed income.</p> <p>Since information on average annual wage at an occupational level is unlikely to be available every year, annual movements in wages should be used to move forward total labor costs.</p>	<p>NSAs are recommended to use average annual wage for each of the occupations selected as a starting point in estimating the total labor costs associated with own account data output for a specific year.</p> <p>For annual estimates of labor costs outside of periods when average annual wage is available, the standard recommendation is to use an appropriate indicator to move forward the nominal labor costs at an aggregated level. This assumes that the labor costs of data producers are changing at the same rate as changes in wages for the broader economy.</p>	<p>NSAs are recommended to use average annual wage for each of the occupations selected as a starting point in estimating the total labor costs associated with own account data output for a specific year.</p> <p>Ideally, for annual estimates of labor costs outside of periods when average annual wage is available, total labor costs should be moved forward with an appropriate indicator at the individual occupational level.</p>	<p>The sum of costs methodology includes the labor costs of <u>all</u> specified occupations regardless of the industry or unit they are working in. This may include workers employed by units that predominately produce data for sale, although this fact should be considered when determining final involvement rates.</p>
<b>Involvement rates</b>	<p>Involvement rates, representing the amount of time an employee or self-employed worker actually spends on producing data, are applied at the occupation level so that the actual labor cost associated with producing data is appropriately reflected.</p>	<p>NSAs are recommended to apply the same or very similar involvement rates to those listed in this handbook in order to determine the degree to which each occupation is involved in the creation of data.</p> <p>The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data.</p>	<p>NSAs are encouraged to develop and use involvement rates (and list of occupations) derived through a more systematic and objective means specific to their economy, such as business surveys or results from machine learning.</p>	<p>The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data.</p>

Measurement step	Broad conceptual recommendation:	Standard recommendation:	Advanced recommendation:	Additional considerations
<b>Estimation of non-labor costs</b>	It's recommended that non-labor costs are incorporated into the final estimate via a single mark-up applied to labor costs. Such a mark-up represents the costs of inputs, any adjustment associated with taxes less subsidies on production, the depreciation of fixed capital used in production, as well as a return to capital (operating surplus).	<p>A single mark-up is applied representing the non-labor costs of data production, including a return to capital.</p> <p>The single mark-up, based on the ratio of total gross output applied against remuneration of employees from the "Computer programming, consultancy and related activities" (ISIC 62) and "Information service activities" (ISIC 63) – or similar available aggregate - is applied to total labor costs.</p>	An advanced recommendation is for NSAs to apply multiple mark-ups to the total labor cost estimate separately so that differences in non-labor production costs, depreciation and operating surplus across occupations and industries can be applied more accurately and transparently.	<p>NSAs are encouraged to investigate potential data sources that may provide more detailed information on the non-labor expenses involved in the production of data.</p> <p>A basic alternative, which should be seen as least desirable and only to be used if specific industry level information is not available, is to use the same mark-up as that applied to similar IPP assets compiled in existing compilation via the sum of costs.</p>
<b>Compilation of Quarterly estimates</b>	NSAs are recommended to develop quarterly indicators that can be used for both the interpolation and extrapolation of annual estimates.	<p>NSAs are recommended to use already available quarterly information that displays a correlation with annual data investment estimates. No specific quarterly series is recommended as the choice will depend on data availability.</p> <p>Potential proxies include:</p> <ol style="list-style-type: none"> <li>1. GFCF on assets relating to data production, such as computer hardware and software,</li> <li>2. ROE for data producing occupations,</li> <li>3. ROE for industries heavily involved in data production.</li> </ol> <p>In the interim, NSAs are recommended to extrapolate annual estimates using a trended time series until a more suitable indicator can be identified.</p>	NSAs are recommended to obtain or develop specific quarterly information relating to business expenditure on data production, which can be used to extrapolate the annual estimate.	

Measurement step	Broad conceptual recommendation:	Standard recommendation:	Advanced recommendation:	Additional considerations
<b>Adjustment for short lived data</b>	While acknowledging that it is not just possible, but likely that some expenditure on data production produces an economic benefit for a period less than one year. The handbook agrees with the SNA that “the information needed to separately identify the costs of producing the short-lived data and the costs of producing the long-lived data is often unavailable.” (SNA 2025 §22.31). Due to this it is recommended to capitalise all own account expense on data production.	If information on short lived data is unavailable, the standard recommendation is that all expenditure on production of data on an own account basis is regarded as a capitalised expense and should be classified as GFCF, with no adjustment made to represent short lived data.	When feasible, countries which have obtained statistically appropriate information providing guidance on the proportion of data consumed within one year, are encouraged to make such an adjustment.	

#### Additional recommendations

Expenditure undertaken to update a data asset with newly collected information should be considered as new investment (Gross Fixed Capital Formation - GFCF) rather than repair and maintenance.

The standard list of involvement rates should be applied only when the standard list of occupations is also used to produce estimates of data. NSA should not apply one aspect of the standard methodology (occupation list or involvement rates) and combine them with a list or set of rates obtained separately as this significantly impact the international comparability of the estimates.

#### Treatment of Market transactions

Unless the transaction relates to a purchase of a copy, information on market transactions in data, while useful, should be viewed as a complementary data source which can improve a countries industry and sector splits of data investment, rather than a fundamental component of the compilation process

Data that is purchased and used for less than one year is regarded as being consumed and treated as intermediate consumption.

Data that is purchased and used for more than one year is regarded as an asset.

Data assets that are purchased with exclusive rights are treated as an outright sale of the data asset by the seller to the buyer. However, assuming that the transaction is not a cross-border or inter-sector one, similar to sales of other assets, this transaction would net off and not impact the overall level of GFCF for a specific sector.

Data assets that are purchased without exclusive rights are treated as a purchase of a copy and contributes to the GFCF of the purchaser (and aggregate GFCF) if it satisfies the necessary conditions of GFCF, (i.e. use in production for more than one year).

A copy of a data asset or access to a data asset under a license to use may also be treated as a purchase of a fixed asset if it meets the necessary conditions.

## Annex 2.9. The treatment of transactions in data copies

The use of transaction information on data is considered additional to the requirement to produce accurate estimates of data output and GFCF. However, if NSAs are able to obtain such data, its use can assist in improving the final estimates of data, specifically estimates of GFCF by industry. The following scenario will show in a simplified manner how estimates of data purchases (without exclusive rights) could be treated in the accounts.

In the scenario outlined in Figure 1, there are only two industries, both of whom are producing data on an own account basis. However, Industry A is also producing copies of data that are then sold to Industry B.

Figure 1

Original scenario	Industry A	Industry B
Estimate of Data output (based on sum of cost)	100	50
Own account data production	60	50
Production of copies	40	0
Purchase of copies	0	40

This situation assumes that in addition to industry A's own account production of data (60) the production of copies (40) by industry A is also being picked up in the sum of cost methodology outlined in the handbook. As such, if no transactional information is available, which is likely for many NSAs, the sum of cost methodology on its own would still produce accurate aggregate estimates, that is Total output of 150 and total GFCF of 150 (see [figure 2](#)). However, the GFCF would not be apportioned correctly to the industries.

Figure 2

Based on sum of cost methodology only	Industry A	Industry B
Own account data output	100	50
Total industry data output	100	50
Aggregate Data output (own account + copies)	150	
Data GFCF per industry	100 (should be 60)	50 (should be 90)
Aggregate Data GFCF	150	

Due to the likely absence of transactional information for most NSAs, the outcome outlined in [figure 2](#) is likely to be accepted as suitable for most NSAs. Indeed, the recommendation from this handbook is that countries should view market transactions in data as a complementary data source which can improve a country estimates of data output rather than a fundamental component of the compilation process. As such, countries should focus their resources on obtaining the most accurate estimate of own account data production, as this will provide the largest benefit to the user.

If a value of the purchase of copies is available through business survey or other means, this estimate should be used in two ways. The first is by adjusting up the GFCF of data for the purchaser of the copies (assuming that the purchase satisfies the necessary conditions of GFCF). In the above scenario this would increase GFCF for industry B to 90 (50 of own account and 40 from the purchase of data copies.) However, an equivalent adjustment for the industry producing the copies risks the chance of double counting if the production of the copies has already been included in the sum of cost outlined for own account production of data. In this situation the production of the copies (40) would be counted twice, once as own account production by industry A obtained through the sum of cost methodology, and again when the adjustment is added to the value of output by industry A. This situation is shown in [Figure 3](#).

Figure 3

sum of cost + purchase of copies	Industry A	Industry B
Own account data output	100	50
Total industry data output	140 (100 + 40)	50
Aggregate Data output (own account + copies)	190 (should be 150)	
Data GFCF per industry	100 (should be 60)	90 (sum of cost 50 + copies 40)
Aggregate Data GFCF	190 (should be 150)	

If the value of the copies is added onto the total output for industry A, the own account output of data production for industry A must be reduced. This can be done in two ways.

The first is to reduce the estimate of own account production of data through a reduction in the sum of cost estimate, either through the occupations selected or the involvement rates applied, to systematically lower the estimate of own account data production coming from industry A. (This is shown in [figure 4](#) as a reduction to 60 in the first row). This approach may be favoured if it is well established that the industry is continually producing data explicitly for sale. This would mean that when the value for the copies is added to the value of (the reduced) own account output, it increases the industry's overall estimate of data output but not the industry's estimate of own account data production or GFCF. While conceptually clear, it is acknowledged that when compiling estimates of data production using the sum of costs approach, estimating the separation between production on an own account asset or an original (which should be capitalized) and production on a copy (which should not be current output) is a tricky split to make. As such macro adjustments (the second option) could be seen as more feasible. This macro adjustment would be equal to that made to industry B (the value of the purchase of the copies). This would be made to the original own account data production estimate of industry A, while also adjusting up the overall level of data production. This will correctly identify this output as being produced for sale rather than own account output and thus not GFCF for industry A.

Figure 4 (recommended approach)

sum of cost + adjustment based on purchase of copies	Industry A	Industry B
Own account data output	60 (100 – 40)	50
Total industry data output	100 (60 + 40)	50
Aggregate Data output (own account + copies)		150
Data GFCF per industry	60	90 (sum of cost 50 + copies 40)
Aggregate Data GFCF		150

# 3

## Creating volume estimates of data

### Introduction

Overall recommendation: **Any price index used to deflate nominal estimates of data must reflect the price change observed in both the labor and non-labor costs involved in data production. In addition, it is recommended that the price index appropriately accounts for the technological and quality improvements that have been observed in the production of digital products over the past several years.**

Standard recommendation: **As a standard position, compilation can be undertaken using an output price index based on an alternative but similar product using chain linking or method consistent with the general practice used for deriving volume estimates.**

Advanced recommendation: **The advanced recommendation is to compile chain volume estimates of data output using a pseudo output price index. This can be created by aggregating appropriate input price indexes and weighted to reflect the actual input costs included in the sum of cost calculation. An adjustment to reflect quality and productivity improvements made to the final output would be added to transform the input price index into a pseudo output price index.**

- 3.1. In the System of National Accounts (SNA), certain high-profile indicators are presented in volume terms as well as nominal terms. In fact, in many cases, such as estimates of production, output and gross fixed capital formation, the movement in the volume series is considered the figure of “principal focus” and importance for users (2025 SNA §18.110). Until now, this handbook has focused only on the production of a nominal estimate of data output and investment.
- 3.2. This chapter will discuss the creation of chain volume estimates of data allowing for data output to be incorporated completely and consistently within the SNA. The chapter will begin from a conceptual perspective, explaining why deflation through a pseudo output price index is considered the most practical option of those presented by the SNA. The chapter will then provide more specific recommendations regarding the choice of price index used when compiling chain volume estimates.

### Deflation options presented in the SNA

- 3.3. When producing outputs in line with the SNA, countries apply several methods to compile volume estimates of production. The choice of which one is usually determined based on the characteristics of the good or service being produced as well as the availability of appropriate price or quantity data associated with the product.

- 3.4. The most popular way for nominal estimates to be represented on a volume basis is to be deflated based on an output price index. Such an index is constructed by recording the market price of the good or service in the current period and the previous period. This allows for an index of the change in the output price to be calculated and applied against the nominal estimate of output produced. Obtaining this price change information is relatively easy for most market transactions, as the collection of prices of good and services on a regular basis is a fundamental data collection in almost all countries.
- 3.5. However, when estimates are calculated as the sum of the costs, as is recommended for data, it is normally reflective of the fact that there is an absence of market prices available or when there are, they are not fully representative of all data production. Therefore, while there is no conceptual concern with deflating a nominal estimate of output compiled using a sum of costs methodology with an index compiled on market prices of output <sup>(31)</sup>, the very fact that a sum of cost methodology is being used usually implies that a true output price index is not able to be easily generated.
- 3.6. The SNA acknowledges this problem when it addresses the compilation of own account volume measures of computer software and databases. Advising two alternatives, an input price index, created by weighting together price indices of the inputs, or a pseudo output price index. (2025 SNA 18.227).
- 3.7. Input price indexes is an established practice in the compilation of the national accounts, for example, they are used heavily in the compilation of volume estimates of non-market output, since this production also exhibits a lack of output prices. While the creation and use of input price indexes is well established, they also have a well-documented weakness. Since the value of the output is continually equal to the value of the inputs, there is no additional value associated with the production itself. In other words, it is not possible to have any value created (or added) by the producer associated with improvements in efficiency or productivity of production reflected in the price index. <sup>(32)</sup> This well-documented weakness means that the SNA explicitly recommends to not use unadjusted input price indexes (2025 SNA §18.227). This recommendation against input price indexes was supported in the follow up manual on compiling capital measures of Intellectual Property Products (IPP) (OECD, 2010).
- 3.8. This leaves a pseudo output price index as the final remaining option. The SNA makes two suggestions on how this can be put together (2025 SNA §18.125).
  - I. By using a similar product(s) where market prices are available to derive an output price index which can then be applied to the output of data.
  - II. by adjusting the calculated input price index to incorporate the observed productivity growth of a related production process.

## Additional consideration in identifying a price index for data measurement

- 3.9. When choosing between the two options, an output price index for a similar product or an adjusted input price index calculated on the actual production of data, there are several considerations to determine which might be more suitable. In fact, the Eurostat handbook on prices and volumes outlines four criteria that a price index must meet to be considered an 'A method.' These include <sup>(33)</sup>,
  - I. it is an index with a coverage of exactly that (group of) product(s).
  - II. it takes proper account of changes in quality of the product(s).
  - III. it is valued in purchasers' prices including non-deductible VAT; and

<sup>(31)</sup> Indeed, it is recommended in the SNA (2025 SNA §18.227) that ideally the nominal sum of cost estimate representing own account capital formation is deflated using an output price index, compiled using market prices.

<sup>(32)</sup> An additional issue, although of slightly less concern is the that the use of input prices often means that the price index is calculating changes in the basic price rather than purchasers' price. As pointed out in the Eurostat manual on price and volume, measurement on this basis assumes that "taxes, transport, installation and the other costs of ownership remain constant in volume terms." The appropriateness of such an assumption can vary depending on the asset and industry, however at this point no adjustment is suggested to account for this assumption (Eurostat, 2016).

<sup>(33)</sup> In the Eurostat manual "A methods" are considered the methods that "approximate the ideal as closely as possible".

IV. the concepts underlying the index correspond to those of the national accounts. (Eurostat, 2016)

- 3.10. Choosing an established price index or Implicit Price Deflator (IPD) would mean that the last three criteria are likely automatically being met including the requirement to properly account the changes in quality. However, by definition, the fact that the established price index or IPD is for a *similar* product rather than the actual product being deflated results in this option not meeting the first criteria.
- 3.11. Conversely, the use of the actual inputs into data production to calculate an adjusted input price index would certainly cover the “exact group of products” but the subsequent adjustment to the index means that the other criteria are met only through the use of assumption or estimation.
- 3.12. Since both options are not perfect NSAs should consider which option comes closest to meeting all four criteria on a regular basis. If the establish price index for the similar product contains inputs that are extremely similar to those used in the production of data, this may prove a superior option. However, if the adjustments made to the input price index are considered accurate and reflective of the quality change occurring, then this option may provide more accurate results. Often, the choice between the two will come down to data availability which may not be consistent across countries. As will be presented later in the chapter, **the use of an output price index based on an alternative but similar product is considered the standard recommendation to produce chain volume measures of data. The advanced recommendation is to compile chain volume estimates of data output using a pseudo output price index. This can be created by aggregating appropriate input price indexes and weighted to reflect the actual input costs included in the sum of cost calculation. An adjustment to reflect quality and productivity improvements made to the final output would be added to transform the input price index into a pseudo output price index.**
- 3.13. Regardless of which option is chosen, **volume estimates of data should not be compiled via an unadjusted input price index.** Advancement in technology used to produce data is on clear display throughout the economy, creating both higher quantities of more granular and accurate data. While the connection between any increase in output and higher quality or value associated with the data is not always linear (see [Box 3.2](#)), the volume estimates of data within the national account should offer some acknowledgement of these developments. Concerns regarding capital going unaccounted for, in particular assets created as part of the digitalisation of the economy was one of the central user concerns that the revised SNA was hoping to address. As such, it is important that the volume estimates of data assets reflect the improved efficiency or quality with which they were produced even if such a consideration is introduced in a generic way and/or are limited to conservative ‘best guesses.
- 3.14. In the preliminary work compiling estimate of data output, NSAs often created volume estimates of data by deflating the nominal estimates of data with the existing price index for other IPP. This is understandable as this initial work was considered experimental and it was deemed that such price indexes would exhibit similar behaviour as those compiled based on inputs to data production. Importantly however, in this initial work, the price indexes chosen represented not simply the market price of other IPP such as Computer Software or Research & Development, but rather the change in price observed for *own account production* of computer software and R & D. This meant that the index contained both a labor and non-labor component. This is an important consideration since as shown by the Australian Bureau of Statistics (and others) the trends exhibited by different price indexes (purchased software vs cost of labor) can be quite different (See [Box 3.1](#)) resulting in the choice of price index having a significant impact on the final chain volume output estimate.
- 3.15. Due to the high amount of data production occurring on an own account basis, the cost of the labor used in the production must be incorporated into the price index used. Therefore, regardless of if the chosen price index is an output price index for a similar IPP or an adjusted input price index associated with the production of data it is recommended that **any price index used to deflate nominal estimates of data reflects the price change observed in both the labor and non-labor costs involved in data production as well as appropriately accounting for the technological and quality improvements that have been observed in the production of digital products over the past several years.**

### BOX 3.1. KEY INPUTS INTO DATA PRODUCTION DISPLAY SIGNIFICANT VARIANCES IN PRICE CHANGE

In their initial work of compiling volume estimates of data investment, the Australian Bureau of Statistics (ABS) tested several different plausible index options that could be used to deflate the nominal estimate.

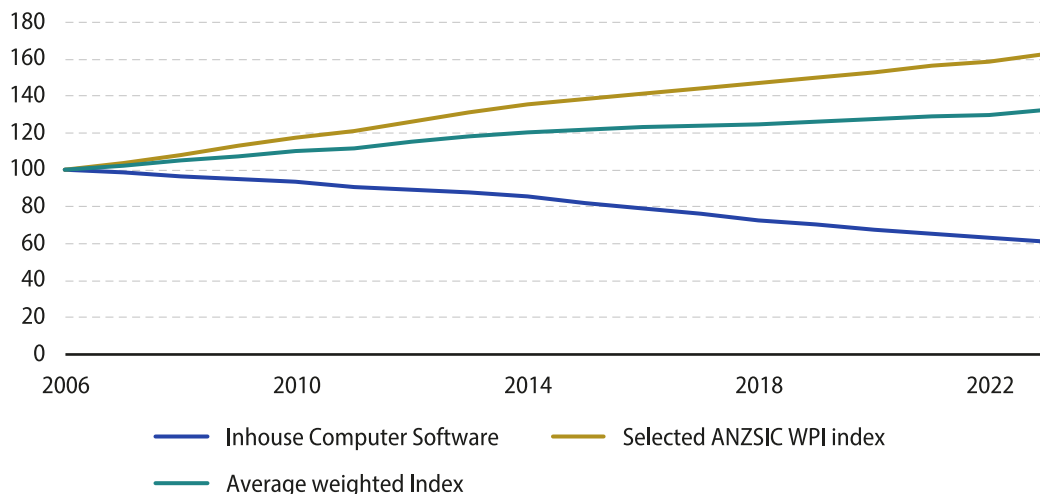
Since the starting point of the nominal estimate of data was labor costs for specific occupations, option 1 involved a wage price index which covered the industries considered as primary data producers. At the same time, aware that such an input price index would not represent any of the productivity or quality improvement that have been so prevalent in the IPP space, Option 2 consisted of a price index representing the cost of in-house computer software was tested.

As displayed in [figure 3.1](#), these two indexes, both considered relatively acceptable options, displayed extremely different trends. The price of labor consistently increased in cost over the time period while the price of software reduced by around 40% over the same time.

With such divergence the ABS ultimately decided the most appropriate option was a weighted mixture of the two, represented by the grey line. Such a decision is in line with the recommendation contained within this handbook.

**FIGURE 3.1**

### Price index for data related activity: Australia, 2006 - 2020



Source: (Smedes, Nguyen, & Tenburren, 2022)

## Use of an established output price index covering similar product(s) or producing industries

3.16. In line with the aforementioned criteria, NSAs, if choosing an established output priced index to deflate data, should look for price indexes that cover production as similar as possible to that observed in the production of data. Understanding which type of output is most closely aligned with the production of data is still a work in progress since

the nominal estimates of data are still under development in most countries. However, once a country has compiled a nominal estimate, it is a relatively straightforward to compare the split of labor and non-labor inputs as well as the specific types of occupations and capital services used as inputs to determine which production process most closely mirrors the production of data. In the initial work in this area, countries have used output price indexes associated with own account production of IPP assets, since the production of these exhibit similar characteristics.

- 3.17. Alternatively, the IPD associated with the output of the relevant data producing industries such as ISIC categories “Computer programming, consultancy and related activities” (ISIC 62), “1 Computing infrastructure, data processing, hosting, and other information service activities” (ISIC 63) or similar available aggregate is also acceptable, however this is considered inferior to the IPD for a specific product since the production of data is unlikely to be concentrated in these few industries.
- 3.18. **As a standard recommendation, compilation can be undertaken using an output price index based on an alternative but similar product. In particular, the Implicit Price Deflator (IPD) associated with own account capital formation of computer software, research and development or broader IPP category, and taken from the compilation of the Supply-Use tables is considered an appropriate standard option.**

## Use of an input price index specific to data production

- 3.19. The production of two assets is never completely alike, therefore while the use of price index based on the production of an alternative product(s) or an aggregate industry is an acceptable standard option, NSAs should strive to create a ‘pseudo output price index’ for data in its own right.
- 3.20. In line with the recommendation in the SNA this could be created by ‘adjusting the calculated input price index to incorporate the observed productivity growth of a related production process’ (2025 SNA §18.125). Such a data specific index, even if based on input prices would provide a better reflection of the change in costs associated with producing data output. Furthermore, such an index could have the weighting updated on a regular basis (i.e. annually) to ensure that it continues to reflect any changes to the proportion of labor and non-labor inputs, or the types of inputs used in the production of data. This is a superior option than the assumption that any compositional input changes occurring in data production are also occurring in the relevant product or industry used as a proxy to deflate data.
- 3.21. The level of detail that such a weighted index may cover largely depends on the information and resources available to the compiler. Ideally, the input price index should reflect as much as possible the full range of costs involved in the production of the data output and importantly, it should be applied at the most detailed level of these underlying costs. For example, rather than applying a single labor cost index representing changes to the cost of labor for the whole economy or industry, the labor cost indexes used should reflect, as much as possible, the wide range of occupations contributing to the labor costs amount as well as the weight that each of these occupations contributes to the overall labor cost amount<sup>(34)</sup>. For example, the USA, when compiling a price index to deflate data assets derives its movement in labor costs from changes occurring at the occupational level (See [annex 3.2](#)).
- 3.22. While the creation of an initial index may involve additional resources, like price indexes used to deflate other components in the national accounts, ongoing work related to the index will likely consist of weighting changes only. These are often undertaken automatically, reflecting the compilation of the nominal estimate. Importantly, such a process will likely bring the deflation of data output in line with the compilation practices already in place for the deflation of other own account IPP.
- 3.23. Therefore, **the advanced recommendation is to compile chain volume estimates of data output using a pseudo output price index. This can be created by aggregating appropriate input price indexes and weighted to reflect the actual input costs included in the sum of cost calculation. An adjustment to reflect**

<sup>(34)</sup> It is considered that the cost of producing data should change relatively consistently across countries, so it should be acknowledged that by applying more detailed occupation-based labor index, this may be exacerbating difference in the change in price between countries. If NSAs have chosen different occupations when compiling their nominal estimate, this may lead to a divergence across countries. Despite this, it is still seen as reflective of the economic situation and therefore preferable to apply a labor cost index at the most detailed level available.

**quality and productivity improvements made to the final output would be added to transform the input price index into a pseudo output price index.**

## Incorporating a quality adjustment to price index

3.24. 3.22 While the quality of some data produced today has declined (i.e., the production of big data sets, requiring a large amount of cleaning) the accuracy, timeliness and granularity of other data is significantly greater than previously produced. Of even less dispute, is the fact that the sheer quantity of data being produced has increased significantly relative to the number of inputs being used. This is almost certainly due to the larger presence of automatic collection tools which have greatly improved the productivity of data collection, and thus data production. Therefore, while the practical implementation and micro measurement of quality and productivity adjustments are not always clear. To ignore these productivity increases when compiling volume estimates of data output would appear to be inconsistent with outcomes experienced in the real economy.

3.25. That said, although the concept of including an adjustment to the price index to represent quality improvements to own account GFCF is often discussed in statistical circles for a range of reasons, there is no definitive advice on its implementation. For example, in the final report of the joint Eurostat – OECD task force on land and other non-financial assets (Eurostat-OECD, 2019), the prospect of including quality adjustments to the price indexes used was largely absent, perhaps an acknowledgement of the conflict that exist between their conceptual reasoning with practical implementation.

3.26. The most prominent of the implementation challenges is the absence of an agreed method on where or how to source this adjustment. For brevity, this handbook will not re-litigate discussions that have already occurred in other guidance on specific technical aspects of hedonic pricing or the best ways to estimate quality and productivity changes when compiling an index. However, for the purpose of an aggregate adjustment, which might be made to the aggregate price index used to deflate nominal estimates of data, some potential starting points for a *simple* aggregated adjustment include:

- I. The calculated difference in growth between the input price index for data and the output price index for a similarly produced product where market prices are available.
- II. The calculated difference in growth between the input price and output price index for similarly produced products where market prices are available.
- III. The total factor productivity estimates for industries that contain a large amount of the occupation identified as data producers.

3.27. None of these options are perfect and countries should research their own compilation methods based on data availability and user feedback. **NSAs must be transparent regarding the source of any adjustments that they make to the price indexes to reflect productivity and quality improvements.**

3.28. Due to the lack of agreed consensus regarding its implementation, no standard recommendation specific to quality adjustment is provided in this handbook. This is one reason why the standard recommendation regarding deflation is to use an established output price index so that no explicit quality adjustment is required. However, if countries compile a data specific price index, then **it is recommended that an aggregate adjustment to reflect quality and productivity improvements is made to the price index used.**

## Final considerations on the compilation of volume estimates

3.29. Due to the impact that any such quality adjustment may have on the final chain volume estimates, some countries have resolved to not make any such adjustment to improve the comparability of results. **The intention of any recommendation in this handbook is not to overrule any such regulations, rather the handbooks' goal is to assist countries compile the most accurate or reliable estimates of data output possible. It is the view of the task team that the introduction of a quality adjustment on top of an input price index is conceptually appropriate and would improve the accuracy of the final estimate. It is accepted that countries will**

**continue to adhere to other frameworks and standards that oversee the compilation of their national accounts.**

- 3.30. While it is not recommended, it should be noted that an alternative to using a price index to generate volume estimates is to undertake a process called quantity revaluation. This process measures the quantity (or volume) of output first and then applies a price index to derive the nominal value second. While rarely used in the compilation of national accounts, it is sometimes considered more accurate for certain agriculture or mineral products where the output exhibits consistent quality and characteristics. However, as pointed out by the SNA, in most cases it is “preferable and more practicable to use price indices to deflate current value [estimates]” (2025 SNA §18.111).
- 3.31. It is conceivable to measure the quantity of data produced and to therefore derive a volume measure of data independently of the compiled nominal estimate. This could in theory be used to derive chain volume measures or use such quantity measure as an output indicator for extrapolating forward the chain volume measure, however, for many reasons the approach of quantity revaluation is not recommended for data (See [Box 3.2](#)).

**BOX 3.2. WHY USING QUANTITY ESTIMATES TO DERIVE CHAIN VOLUME ESTIMATES OF DATA IS NOT RECOMMENDED**

Within the national accounts, volume estimates are occasionally calculated based on an output indicator which often represents a quantity of the good or service produced. This is usually for estimates of production that are relatively homogeneous, and quantity counts are relatively easy to obtain.

In one regard the quantity of data is relatively easy to measure. The bits and bytes that make up data when saved to a computer take up a specific amount of memory. Due to this, it should be, theoretically, possible to measure the additional quantity of data produced each accounting period when compared to the previous period. In fact, this undertaking has already been done by several organizations who estimate that around 2.5 quintillion bytes are created every day with the overall amount of data doubling every two years. However, despite the presence of this estimate there are several reasons why such an estimate of quantity cannot be used for compiling volume estimates of data in the SNA.

The first reason is that this incredible number includes a large amount of data that is *not* data as defined within this handbook and the 2025 SNA. Rather it is closer to an alternative definition of data as Internet Protocol (IP) traffic, or the volume of digitised information stored on servers and other hardware. A large amount of these bits and bytes includes digital activity such as photos, text messages, email and other communications that fail the 2025 SNA data definition. These digital files are usually not produced by accessing and observing phenomena and are not used in productive activities. Rather they reflect the nature of the digital service delivery used by business and consumers alike.

Even if a quantity of data was able to be separated between that used repeatedly in production and that not meeting the SNA definition. There still exists an inconsistent relationship between the quantity of data within data assets and their subsequent value. Much of the data value comes from the content of the information and the context that it has been gathered or could be used. Both these factors are often unrelated to the size of the data. While it is true that data that contains more information is likely to be worth more than data with less information, the relationship is not consistent enough to create any form of reliable value based solely on quantity. Proof of this is the evidence that the huge increase in data production observed over the recent years is driven more by the declining cost and increasing efficiency of data storage than by a positive liner relationship between the amount of data produced and its explicit value.

Overall, while a quantity estimate of data production may be achievable, the use of a direct volume measure within the National Accounts is deemed inappropriate due to the heterogeneous nature

of data as well as the volatility and treatment of prices applying in different markets. Interestingly, data is not the only good that falls into this category, with the 2025 SNA pointing out that the volume estimates of electricity (as well as other utilities) should not be derived through quantity, even though it appears relatively feasible, due to the difficulty in capturing a single representative price. (2025 SNA §18.111).

## Annex 3.1. Volume estimates of data asset: The case of Pakistan

Pakistan Bureau of Statistics (PBS) has compiled both nominal and volume estimates of data asset.

Overall, the PBS compiled five aggregates of data assets using two broad assumptions i.e.,) by using aggregate price changes from industry shares and ii) by deflators. These are discussed below:

### **Industry based shares of nominal estimates:**

1. The first step is to derive nominal estimates of data assets through sum of cost approach including labor and non-labor components following the guidelines presented elsewhere in this handbook.
2. The second step is to deflate these estimates based on the Implicit Price Deflator (IPD) from industry-specific GVA at constant prices and nominal prices. The industries used reflect those contributing to the production of data. These industry-based volume estimates of data assets are then aggregated to have a measure of data asset at the total economy level.

### **Using Input based Deflators:**

PBS specifically compiled four different price indices for data assets based on different aspects of production.

- i. IT equipment index. The only available items in the CPI (2015-16 base) relating to IT equipment in Pakistan are personal computer and laptop having equal weight (See [Table 3.1](#)). It is pertinent to mention that base of price indices and national accounts are the same in Pakistan i.e., 2015-16.
- ii. IT salary index. A wage index in Pakistan is not compiled formally. Therefore, PBS has specifically derived an IT Salary wage index by using data from IT-Salary Surveys conducted by Pakistan Software Houses Association (PASHA). Twelve (12) data related occupations were selected (See [Table 3.1](#)). The weights of these occupations were derived by using employment from Labor Force Survey (2014-15). The fixed weights have been used to compile the IT Salary Index with 2015-16=100.
- iii. Weighted Index: PBS has attempted two specifications of weighted index. In the first specification an equal 50% weight was assigned to both salary index and equipment index. In the second specification, 40% weight was assigned to Salary Index and 60% to equipment. The reason behind using 40% weight for salary is the share of labor component in the nominal estimates of data asset for four labor force surveys for 2014-15, 2017-18, 2018-19 and 2020-21.

**TABLE 3.1**

### Summary of items and weights for data asset in Pakistan

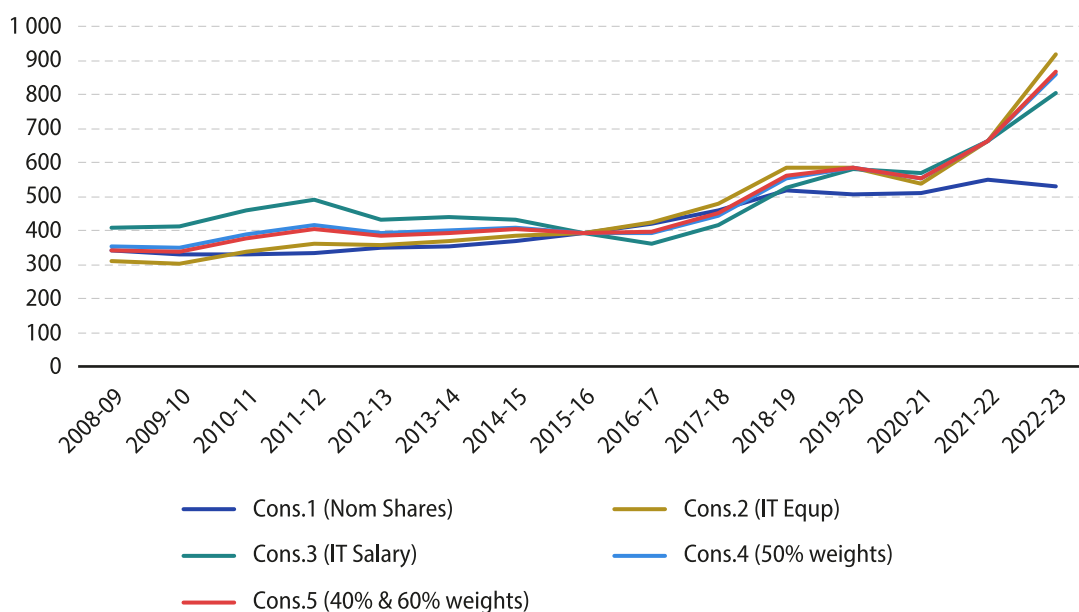
IT equipment items		Weight (CPI)
1	Personal Computer with LED Monitor (17") DELL/HP/ACER CORE i5	0.0489
2	Laptop DELL/HP/ACER Core i5, Display (14'-15')	0.0489
<b>Total</b>		<b>0.0978</b>

IT Salary Occupations		Weight (Labor Force Survey 2014-15)
1	Programmer (IOS)	1.4478
2	Quality Assurance	2.4144
3	Graphics Designer	0.8272
4	Technical Writer	1.5663
5	Product Manager / Business Analyst	10.7004
6	Development Manager	1.5727
7	Project Manager	12.1220
8	Architect	1.5663
9	Database Administrator	3.8942
10	System Administrator	1.2657
11	Sales & Marketing Manager	33.8469
12	Manager/Finance Manager	28.7763
<b>Total</b>		<b>100.000</b>

**FIGURE 3.2**

### Volume estimates of data asset in Pakistan (PKR in billion)



Source: PBS Compilation

## Results and compilation limitations

As shown in [Figure 3.2](#), the different price indexes produce broadly similar growth rates in volume terms regardless of the specific price index being applied when viewed over a 5 – 10-year period however in specific years (2016-17, 2022-23) large divergences are observed between the price change in IT salaries compared to IT equipment. This flows through into different growth rates for the volume estimate of data.

These results are still in their preliminary stage of development as information on IT salaries is not regularly compiled and published, meaning that additional assumptions are required for certain years. Additionally, PBS has not made any quality-based adjustments to the price indexes due to limited data in this area being able to guide decisions on this subject. Despite these limitations, important first steps have been made using the data already available.

## Annex 3.2. U.S. Bureau of Economic Analysis development of adjusted input price indexes for data and database assets

The U.S. Bureau of Economic Analysis developed a productivity adjusted input price index for data and database assets. It is largely inspired by BEA's own account software prices. Both input prices start by considering the change in costs per unit for labor and intermediate inputs.

In terms of the labor costs, BEA used the attributed compensation of employees for gross capital formation in nominal terms per implied unit of labor. In the case of the United States, the implied unit of labor is the employment estimates from the U.S. Bureau of Labor Statistics by occupation multiplied by that occupational group's involvement rate. The annual growth rate in the compensation of employees over effective employment numbers is used as the labor cost rate.

In terms of intermediate inputs, we consider the representative industries Data Processing, Hosting, and Related Services (NAICS: 518) and Computer Systems Design and Related Services (NAICS: 5415). Using the KLEMS published tables, the intermediate inputs are the weighted change in prices for energy-inputs and purchased-services inputs for both industries. The KLEMS tables provide GDP by industry composition, allowing to compute the share for compensation of employees, energy-inputs, and purchased-services for both industries as well as the prices for energy and purchased services for each industry. The change in prices then become the weighted average of the change by the share of each component.

To avoid the output price change equalling the input price change, the BEA also incorporated an adjustment to represent productivity improvement. For this a (simple) average total factor productivity for the two representative industries was used to obtain a 5-years moving average productivity adjustment. This was subtracted from the raw compensation of employees, energy-inputs, and purchased-services computed series derived from changes to the input prices. The resulting adjusted series was used to deflate the nominal estimates of data and database assets.

### Annex 3.3. Summary of recommendations

Subject	Overall Recommendation for deflating data	Standard	Advanced
<p><b>Price index used</b></p>	<p>Any price index used to deflate nominal estimates of data must reflect the price change observed in both the labor and non-labor costs involved in data production. In addition, its recommended that the price index appropriately accounts for the technological and quality improvements that have been observed in the production of digital products over the past several years.</p>	<p>The standard recommendation is to use an output price index based on an alternative but similar product.</p> <p>In particular, the Implicit Price Deflator (IPD) associated with own account capital formation of computer software, research, and development or broader IPP category, and taken from the compilation of the Supply-Use tables is considered an appropriate standard option.</p>	<p>Volume estimates of data output are compiled using a 'pseudo' output price index.</p> <p>This can be created by aggregating appropriate input price indexes and weighted to reflect the actual input costs included in the sum of cost calculation.</p> <p>An adjustment to reflect quality and productivity improvements made to the final output would be added to transform the input price index into a pseudo output price index.</p>
<p><b>Quality adjustment applied to price index</b></p>	<p>It is recommended to include an aggregate adjustment to reflect quality and productivity improvements made to the final output. This is either included in the IPD for the standard method, or explicitly included for the advanced method.</p> <p>Initial options for deriving the aggregate adjustment include:</p> <ul style="list-style-type: none"> <li>• The calculated difference in growth between the input price index for data and the output price index for a similarly produced product where market prices are available.</li> <li>• The growth between the calculated difference in Input price and output price index for similarly produced products where market prices are available.</li> <li>• The total factor productivity estimates for industries that contain a large amount of the occupation identified as data producers.</li> </ul>		

Subject	Overall Recommendation for deflating data	Standard	Advanced
<b>Additional considerations</b>			
<p>While the introduction of a quality adjustment on top of an input price index is conceptually appropriate and would improve the accuracy of the final estimate. It is accepted that countries will continue to adhere to other frameworks and standards that oversee the compilation of their national accounts.</p>			
<p>Volume estimates of data should not be compiled via an unadjusted input price index.</p>			
<p>The use of a direct volume measure within the National Accounts is deemed inappropriate due to the heterogeneous nature of data as well as the volatility and treatment of prices applying in different markets.</p>			
<p>NSAs must be transparent regarding the source of any adjustments that they make to the price indexes to reflect productivity and quality improvements.</p>			

# 4

## Creating Capital Stock estimates

### Introduction

- 4.1. So far, this handbook has focussed on compiling the economic flow of data, that is, the production of data, which usually manifests itself as investment in data, or gross fixed capital formation (GFCF) to use the SNA parlance. These estimates contribute to the compilation of an accurate estimate of GDP, however due to the linkages that exist between the capital services that are an input into production and the subsequent income derived from this production, the SNA expands beyond the production and income account to also include a capital account, other changes in assets account, and balance sheet. The links between the accounts demonstrate the fundamental relationship between the stock of assets and income derived from the economy. As such, an awareness of how much capital stock is available to an economy is a key indicator for forecasting future production and thus income.
- 4.2. This chapter will discuss the transformation of estimates of data output and data investment into an estimate of the capital stock of data, which in turn are used to derive estimates of depreciation.<sup>(35)</sup> These estimates of capital stock and depreciation are usually compiled through the Perpetual Inventory Method (PIM).

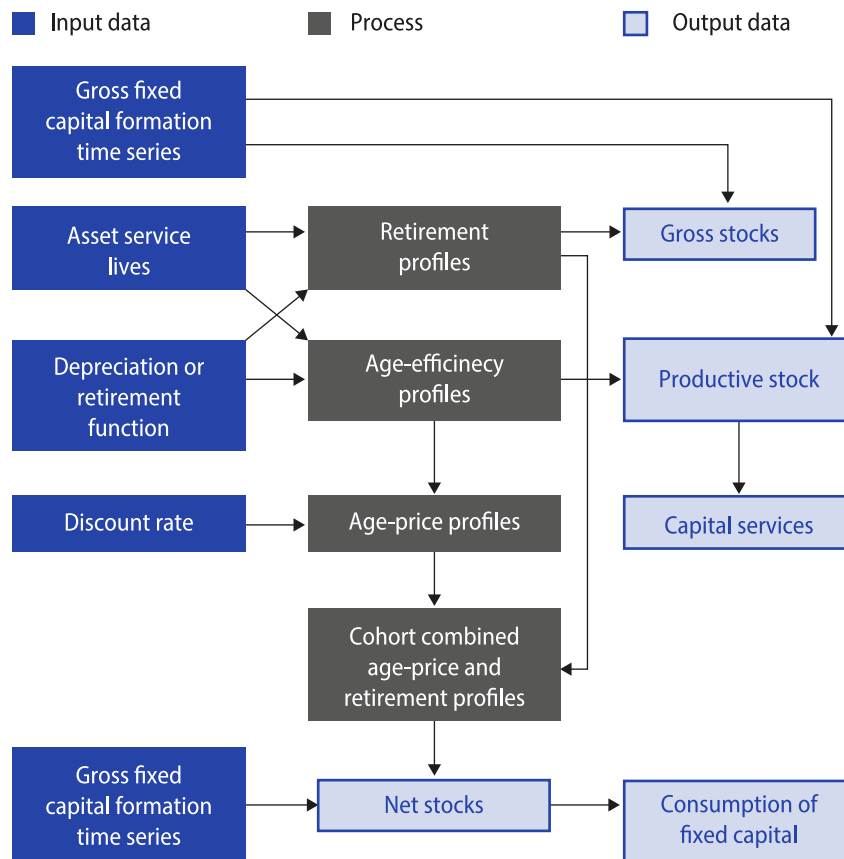
### Background and guidance on the PIM and compiling capital stock

- 4.3. The SNA in its discussion on the compilation of balance sheets includes several key requirements for compiling estimates of capital stock and depreciation of fixed capital. Primarily that “depreciation *must be valued with reference to the same overall set of current prices as that used to value output and intermediate consumption*” (2025 SNA §7.272). In simple terms this suggests that the same prices used to derive output (such as GFCF) should be considered when deriving estimates of depreciation.
- 4.4. The prices used by statistical offices to compile output are often different to those subsequently used to derive depreciation within corporate business accounting. Corporate accounts are usually compiled based on international corporate accounting standards which when combined with respective countries’ taxation rules, create estimates that often have minimal linkages to the actual economic service being provided by the asset. Because of this, the SNA correctly points out that “depreciation as recorded in business accounts may not provide the right kind of information for the calculation of depreciation [in the SNSA].” (2025 SNA §7.273). Furthermore, since data has not yet been recognized as assets according to the current international accounting standards it is likely that NSAs would find it difficult to collect capital stock and depreciation information on data regardless of the established valuation differences.

<sup>(35)</sup> The 2025 SNA has altered the terminology for the amount recorded as declining of the value of an asset over a specific period. Previously this was referred to as Consumption of fixed capital, however the updated SNA now refers to this by the more mainstream terminology of depreciation. Conceptually there has been no change.

- 4.5. In the SNA, estimates of capital stock should reflect the current market price of the asset, which is theoretically associated with its potential future income streams. However, since there is limited information on the market price of second-hand assets, the SNA promotes a more theoretical approach to determining the price of an asset as it ages (2025 SNA §7.274). The theoretical approach subsequently described includes capital stock estimates being *“built up from data on gross fixed capital formation in the past combined with estimates of the rates at which the efficiency of fixed assets declines over their service lives”* (2025 SNA §7.273). This Perpetual Inventory Method (PIM) has now become the standard method used by statistical agencies to estimate the level of capital stock and depreciation in the economy.
- 4.6. A simple diagram representing the PIM is provided in [Figure 4.1](#). This shows the inputs and processes required to calculate estimates of capital stock and depreciation. The inputs into the calculation of capital stock using the PIM are a mixture of compiled outputs (GFCF) as well as assumed parameters. The OECD manual on measuring capital (OECD, 2009) covers these inputs in detail but a summary is provided below.
- 4.7. The primary input into the PIM is a time series of Gross fixed capital formation in volume terms: The level of new investment in the asset occurring each period presented in both nominal and constant price terms. This is the output created in the first 3 chapters of the handbook.
- 4.8. The next four points are assumed parameters required to compile the PIM outputs:
- I. Average service lives – An estimate of the average length of time that the asset is retained in production.
  - II. Retirement profile – an estimate of the level of “retirements” or “discards” each period, which is the number of assets removed from the capital stock each year <sup>(36)</sup>.
  - III. Age – efficiency profile: The age-efficiency function of a single asset reflects losses in productive efficiency due to wear and tear.
  - IV. Age – price profile: The age-profile represents the price of a capital good regarding its age. The age-price profile compares identical capital goods of different age at the same point in time. Typically, this is a declining value, which due to an absence of source data, is often associated with the long-term government bond rate.
- 4.9. These first two points are combined to provide a retirement function, representing an estimate of how much of the initial cohort of assets are still in productive use. The next two points are combined to create a depreciation profile providing an estimate of the current value of a single asset each period which is usually expressed as percentage of the value of a new asset.
- 4.10. The combination of the depreciation profile and the retirement function allows for an estimate of the remaining value of the initial investment based on how many of the assets are still in service and the level of productive services coming from those still in service. This amount is supplemented each year with new investment in the form of gross fixed capital formation, resulting in an ongoing (*or perpetual*) inventory of capital stock.

<sup>(36)</sup> This does not include disposals of non-financial assets to other enterprises or sectors as these are recorded as transactions in nonfinancial assets within the capital account.

**FIGURE 4.1****Simplified diagram of Perpetual Inventory Method (PIM)**

Source: (ONS, 2022)

4.11. This handbook will not cover the specific concepts and practical implementation of each PIM input as this is a considerably detailed and technical endeavour. Furthermore, existing works such as the OECD manual on measuring capital (OECD, 2009) are already in place that cover these topics at length.

4.12. In addition, since the PIM is used by the vast majority of NSAs, including being the required method of countries that adhere to the European System of Accounts (European Commission, 2010), it is seen as redundant to discuss the generic concepts beyond the simple summary provided. Rather this chapter focusses on the specific assumptions that NSAs will make within the PIM when calculating depreciation and capital stock estimates of data.

## Extending countries current approach to PIM parameters to Data assets

**Standard recommendation: It is recommended that except for average service lives, a standard starting position for NSAs is to apply the same parameters in the compilation of depreciation and net capital stock of data as applied currently to other IPPs.**

**Advanced recommendation: Countries should aspire to continually collect additional information on different assumptions and parameters to refine and improve the estimates of depreciation and capital stock being compiled in order to apply assumptions at an industry or occupation level.**

- 4.13. Focusing on the four ‘ingredients’ that make up the assumptions used in the PIM (Age-efficiency profile, age-price profile, average service life and retirement profile) it is unlikely that any two countries apply exactly the same assumptions for any single asset. That said, there is a lot of similarities between countries for some assumptions, while others show only slight differences.
- 4.14. For example, while not absolute, many NSAs apply some version of a bell-shaped retirement profile for the majority of assets. Additionally, many NSAs use similar assumptions regarding the discount rate used to represent the age-price profile, while the age-efficiency profile is usually one of three; hyperbolic, linear, or geometric. Importantly, since retirement functions and the PIM are compiled for a cohort of assets rather than a single one, the final rate of depreciation, which combines the retirement and depreciation function together often resemble a geometric pattern (OECD, 2009), resulting in more conformity between countries, even if slight differences between inputs exist <sup>(37)</sup>. As such it is the average asset life applied to each asset that usually makes the most significant difference to capital stock values. A view that has been reaffirmed by recent testing of data capital stock compilation, using experimental data GFCF estimates (See [Box 4.1](#)).
- 4.15. Fundamentally, the PIM involves certain assumptions for which empirical evidence is difficult to find. Information on the parameters used for the compilation of data are equally difficult to find. The OECD manual on measuring capital recommends various approaches for obtaining information on asset lives and other parameters, this includes those prescribed by tax authorities, company accounts, statistical surveys, administrative records, expert advice, and other countries’ estimates (OECD, 2009). However, not all of these are applicable for data. For example, currently no government or international accounting standard recognises data as an asset for which depreciation can be claimed to reduce a tax liability, therefore immediately ruling out several potential sources.
- 4.16. In the absence of new available source data, NSAs can look to apply the same parameters as already used for other assets. This is useful as recommendation regarding the PIM should not only promote consistency across countries but also, unless clear evidence suggest otherwise, consistency with the parameters used for other assets.
- 4.17. This is not to say the depreciation of data is the same as a dwelling or motor vehicle, they certainly are not in terms of length of service or retirement profiles. However, for the more general, theoretical assumptions covering the age-efficiency and age-price profiles, there is value for both compilers and users in being consistent across different asset classes. In fact, as alluded to previously, due to the absence of data, many NSAs are already holding these two parameters (age-efficiency and age-price profiles) constant within the PIM regardless of the asset being measured (Eurostat, 2024). Even for retirement profiles, the Eurostat report noted that “Only one country uses more than one [type of] retirement profile for the various assets” (Eurostat, 2024) implying that for most NSAs the same retirement pattern is used regardless of the asset type retiring. Overall, it is clear that in a vast majority of countries, for the compilation of depreciation and capital stock NSAs hold most assumptions constant across different asset.
- 4.18. A similar approach has been undertaken in the already compiled experimental estimates of depreciation and capital stock of data. In many cases the average service life for the asset was the only parameter which countries considered changing, preferring to maintain consistency with other assets in the application of age-price, age-efficiency, and retirement profiles rather than trying to determine what was suitable specifically for data.
- 4.19. Overall, the PIM is the standard approach to compiling estimates of depreciation and capital stock estimates across countries. While there are some differences across countries in the parameters used <sup>(38)</sup>, on most occasions NSAs apply a consistent age-price, age-efficiency, and retirement profiles regardless of the asset, with only the average service lives changing across assets (Eurostat, 2024). Although on some occasions, there does appear to be some alternate parameters used in the compilation of depreciation and capital stock of IPP when compared to other fixed assets.

<sup>(37)</sup> It should be noted that the 2025 SNA explicitly recommends a geometric depreciation approach as the standard approach to calculating estimates of depreciation, therefore the standard recommendation in this handbook is consistent with that provided in the 2025 SNA. .

<sup>(38)</sup> For example, the report from the Task Force on fixed assets and estimation of consumption of fixed capital under European System of Accounts 2010 showed that across EU countries, there are at least 10 different retirement functions currently in use (Lognormal, Normal, Weibull, Truncated-normal, Quasi-logistic, Gamma, Linear, Delayed linear, Simultaneous Exit, Geometric).

- 4.20. As such, in the absence of empirical evidence suggesting otherwise **it is therefore recommended that except for average service lives, a standard starting position for NSAs is to apply the same parameters in the compilation of depreciation and net capital stock of data as applied currently to other IPPs.**

#### **BOX 4.1. RECENT TESTING OF DIFFERENT PARAMETERS WHEN COMPILING THE CAPITAL STOCK OF DATA IN GERMANY**

As part of the research and testing of a proposed data measurement methodology, the Federal Statistical Office of Germany (FSO) produced estimates of depreciation and net capital stock for data under several different scenarios. This was undertaken using an experimental set of data investment, compiled via the sum of cost methodology recommended by this handbook.

This work specifically tested different age-efficiency profiles and different average asset lives. Conversely, the bell-shaped density function of the gamma distribution used for the retirement profile and the initial age-price profile were kept consistent. Both were in line with the depreciation and capital stock calculated for other assets within the German wealth accounts.

Three different age-efficiency profiles were tested: linear, and two different geometric curves, one with a declining balance rate of 2 and another with a declining balance rate of 1.65. From a service life perspective, averages service lives of 2, 5 and 10 years were tested.

The results show that the choice of retirement and depreciation functions have only a minor impact of the final estimate, while the assumed average service life has the biggest impact on results. For example, the estimate of depreciation for all three depreciation approaches are similar in size and while the difference between the three are larger for net capital stock, these are minor compared to the increases observed in net capital stock that comes with the increasing length of service life. To this point, the FSO specifically advocated for a standard agreement on the average service life to be applied as this would provide the most significant benefit in attempting to achieve consistency in methodology across countries.

## Estimating average asset lives of data

**Standard recommendation: In the absence of other information, countries should apply as standard, an average service life not longer than 5 years for data assets, this chosen service life reflects the likely inclusion of short-lived data. Any service life chosen should reflect whether an adjustment to exclude short lived data has been made to the nominal estimate.**

**Advanced recommendation: Countries should break up the estimate of data investment by industry in order to allow for different service lives to be applied based on the industry producing the data.**

- 4.21. As explained, of the various parameters, the average service life has the largest impact on the final estimate of depreciation and capital stock. Therefore, several countries have attempted to obtain information on the service life of data. This included Japan, who used the special internet survey to ask firms on how long they intended to use the data asset in production. While the results of the survey would need to be supported with more data, the initial results seem to support the average service life survey proposed in this handbook. More information on the survey included the preliminary results shown in [Annex 4.2](#).
- 4.22. Countries' efforts to obtain information on this is vital, as due to the almost unfathomable range of data being collected, trying to estimate how long the *average* piece of data is used in production for is a near impossible. There is countless anecdotal evidence of collected data being useful for only a few minutes or days. Conversely, most people have provided information to both private and public organizations that have been used by the organization for years. Organizations will frequently make business and production decisions based on data collected over an extended

period of time. While the production of artificial intelligence, including the data sets required to train large language models provides another avenue for certain types of data to be used in production repeatedly.

- 4.23. The acknowledgement that some data is kept for a long time for a specific purpose, need not be dealt with by increasing the average service life, but potentially with a change in the retirement profile applied. Within the task team, it was discussed the possibility that data does not follow a symmetrical bell-shaped retirement profile. It was theorised that a small amount of data collected is likely used for a *very* long time (i.e., static personal information, or production data points that contribute to a time series) which is then offset by a large amount of data collected which is ultimately used for less than the average service life (preferential or variable information with a clear point in time where it becomes obsolete). While there is not empirical evidence supporting this theory, if proven accurate, it would suggest that a traditional bell-shaped retirement profile with a maximum age as double the average age is not suitable for data. Rather a positively skewed bell with an exceptionally long “tail” might better reflect a cohort of data investment.
- 4.24. In the absence of empirical evidence to support this theory, the task team acknowledged that compilers and users’ preference is likely to maintain a connection with the estimation of depreciation and capital stock of other fixed assets rather than applying stand alone experimental assumptions. That said, **NSAs should attempt to continually collect additional information on the average length of time that data is used in production as well as testing different assumptions and parameters to refine and improve the estimates of depreciation and capital stock being compiled.**
- 4.25. Due to the absence of empirical evidence suggesting otherwise, most countries in their initial estimates of data decided to implement an average service life like other IPP such as computer software. This is seen in [table 4.1](#) which lists the average service life applied for data in a range of countries.

**TABLE 4.1**

### Average service life applied in preliminary estimates of capital stock and depreciation of Data

Country	Asset life applied
Australia	Average 3 years (max 5 years)
Canada	7 years
Germany	2 / 5 years
USA	5 years

Source: (Smedes, Nguyen, & Tenburren, 2022) (Amegble, Bugge, & Sinclair, 2023) (Calderón & Rassier, 2022)

- 4.26. Overall, it is recommended that **in the absence of other information, the standard recommendation is that NSAs apply an average service life for data assets of not longer than 5 years. Any service life chosen should reflect whether an adjustment to exclude short lived data has been made to the nominal estimate.** Such a recommendation takes into consideration feedback from the task team, the small amount of empirical information available as well as a desire to maintain some consistency with the compilation of depreciation and capital stock estimates of other assets (especially IPP).
- 4.27. If relevant source data is made available, suggesting a longer or shorter service life, countries should look to implement a revised average service life. Furthermore, if allowed by the source data, **NSAs should break up the estimate of data investment by industry in order to allow for different service lives to be applied based on the industry producing the data.** (See [Box 4.2](#))

#### BOX 4.2. CREATING MORE DETAILED ASSET COHORTS TO IMPROVE DATA OUTPUTS FROM THE PIM

The PIM calculates depreciation and capital stock estimates for all assets used in the economy, however it is not feasible to measure the depreciation and capital value of each asset individually. Rather, the PIM groups together cohorts of assets based on their characteristics and use in production. Therefore, GFCF is grouped together and placed into the PIM in cohorts, representing a set of similar kind of assets, which have entered production at the same time.

There are no set rules regarding the size or make up of these cohorts, theoretically, the more cohorts there are the more accurate the outputs of the PIM should be. The methodology of the PIM lends itself to cohorts being made up of assets with similar expected service lives. In some case this means that two identical assets may be placed into different cohorts if they are used differently in production. For example, in Italian national accounts, a building or industrial structure in the mining and quarrying industry is expected to have an average life of 35 years, considerably less than the building or industrial structure used within the wholesale and retail trade industry which is expected to be used in production for 65 years (OECD, 2009). In practice due to the source data and methodologies that exist in most countries, this means that cohorts are usually broken up by the type of asset that they are (transport equipment, machinery, dwellings, buildings) and by the industry in which they operate.

On occasions, countries have decided to break down assets even further than what is typically published in the break down of expenditure on GFCF. An example of this is dwellings. While some countries will simply include all residential property into a single cohort, some countries will break it down further based on material or dwelling type. In the United States, the BEA applies different average service lives to the following categories, all of which contribute to a single dwelling GFCF number (OECD, 2009).

- 1-to-4-unit structures–new
- 1-to-4-unit structures–additions and alterations
- 1-to-4-unit structures–major replacements
- 5-or-more-unit structures– new
- 5-or-more-unit structures–additions and alterations
- 5-or-more-unit structures–major replacements
- Manufactured homes

Similarly, the Australian Bureau of Statistics breaks up its dwelling investment number into the following categories, each with a different average service life but while maintaining the other parameters (ABS, 2021).

- Private brick homes
- Private timber, fibro, and other houses
- Private non-house dwellings (units, flats, etc.)
- Private alterations and additions
- Public

Since it is considered that each category has a slightly different asset life, it is seen as advantageous to try and separate them and apply a separate asset life in the calculation. Alternatively, countries can attempt to apply an average life to the total that covers the different composition of the diverse types of assets.

Data assets could potentially benefit from a similar more detailed breakup. It is well established that some data will be used repeatedly for many years while other data may be more time sensitive, either because the information it contains is constantly being updated or because more detailed

data has become available. The way that differing types of data is collected by different industries would suggest that they are likely to be used for differing lengths of time. Certain industries leverage time sensitive and continually changing data, such as consumer preferences, location dynamics etc. Conversely, other industries tend to rely on the collection of data that remains unchanged or that gets modified only sporadically over the course of a lifetime. As such these two distinct types of data may necessitate two different service lives.

Such an additional break up for data is conceptually possible and would likely create a more accurate estimate. There already exists numerous typologies of data, however none that have become internationally authoritative. Furthermore, since the depreciation of data represents obsolescence rather than physical wear and tear, the “type” of data matters less compared to the information that the data holds, information which is likely highly dependent on the industry creating the data. Finally, since an industry split of data investment is likely already being compiled as part of the overall industry estimates of GFCF, it is likely that these input series are already being produced by countries. Due to this a breakdown based on industry that is producing the data would appear an attractive and obtainable option for deconstructing data to a more granular level to improve the accuracy of depreciation and capital stock estimates.

## Annex 4.1. Capital stock estimates of data asset: Case of Pakistan

Pakistan Bureau of Statistics (PBS) has compiled capital stock estimates of data asset using the Perpetual Inventory Method (PIM) following linear and geometric models. While the application of linear model requires fewer assumption i.e., nominal and volume estimates of data asset, asset life, price indices, the application of geometric model require additional assumptions i.e., retirement pattern (cut-off values), age-efficiency and age-profiles of assets. PBS has valued the capital stock of data assets at prices of base year i.e. 2015-16, the reference year of national accounts. PBS has also used capital stock estimates of data assets to derive estimates of consumption of fixed capital (CFC) under both linear and geometric methods.

### Assumptions

Both CFC and capital stock have been estimated at constant prices from 2015-16. The nominal estimates have been derived by using weighted index for data asset assuming 40% weight for salary index and 60% for equipment index, while applying a service life of 5 years.

A simplified form of a geometric model has been used as given below:

$$X \times p^t = X \times C$$

$$p^t = C$$

$$p = C^{1/L}$$

where

$X$  = Gross Fixed Capital Formation at constant prices of the base year

$p$  = geometric parameter ( $0 < p < 1$ )

$L$  = service life

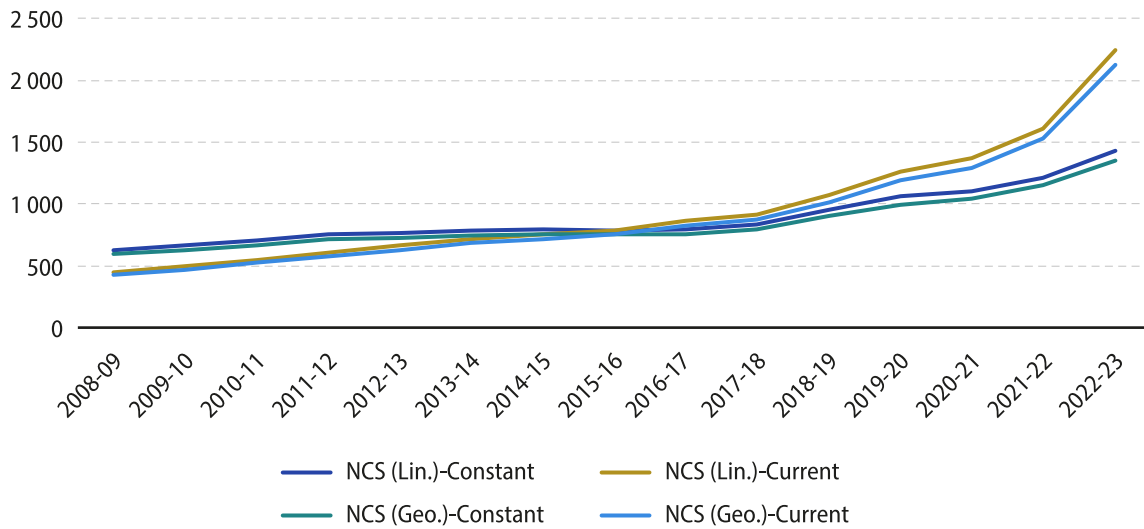
$C$  = Cut-off point (Residual value),  $0 < C < 1$ , e.g., if 15 %,  $C = 0.15$ )

The exact cut-off-point of data asset was not known. Therefore, a sensitivity analysis for geometric estimates of data assets has been conducted for three values i.e. 10%, 15% and 20%. Stock of data asset has been valued at the prices of base year of national accounts of Pakistan i.e., 2015-16.

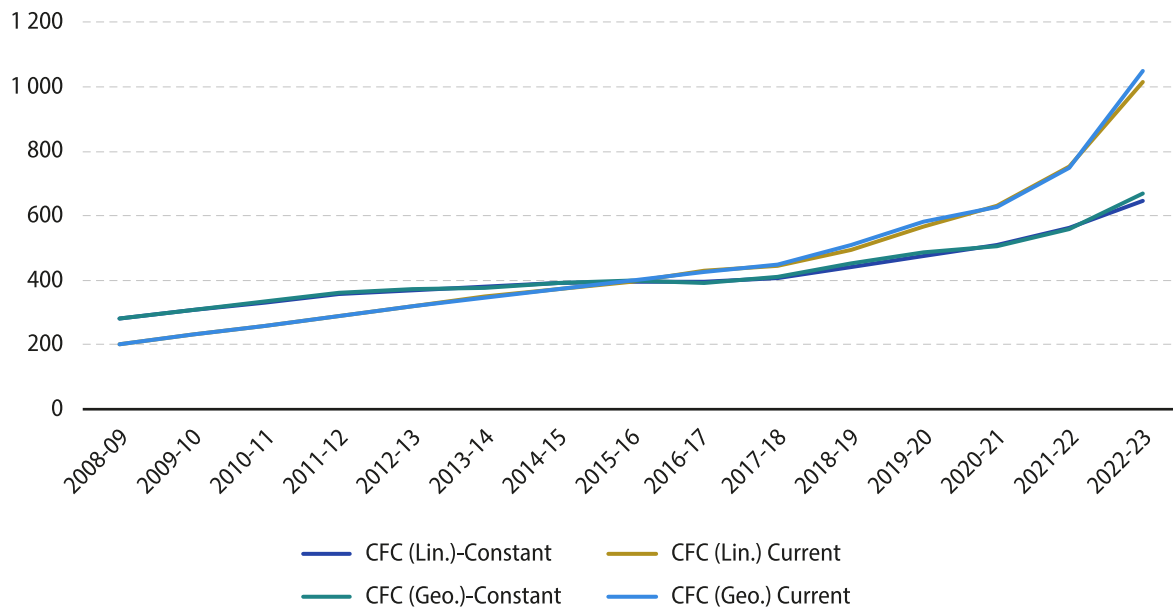
The estimates of net capital stock at current and constant prices under linear and geometric models are presented in [Figure 4.2](#).

**FIGURE 4.2**

### Estimates of net capital stock of data asset in Pakistan (PKR in billion)



The estimates of CFC at current and constant prices under linear and geometric models are presented in [Figure 4.3](#).

**FIGURE 4.3****Estimates of CFC of data asset in Pakistan (PKR in billion) (2015-16=100)**

Results of the testing show that estimates of both capital stock and CFC remain similar regardless of which linear or geometric models are used. Further confirming that the choice of service life remains the most impactful assumption of those made when using a PIM. Since there is an absence of information on the service life of data in Pakistan, the 5-year service life was applied, like the service life used for other IPP. Similarly, no new information on other assumptions required (retirement distribution) on data is available so the same assumptions as used in other IPP capital stock compilations have been used.

## Annex 4.2. Obtaining information on asset life via surveys: the experience in Japan

In 2022 and 2023 the Economic and Social Research Institute (ESRI) within the Japanese Cabinet Office undertook surveys of individuals focussed on their production of data in their employment. Questions were asked regarding the type of data they produce, and the amount of time spent producing data, this information is extremely useful for assumptions dealing with the compilation of a nominal estimate of data output. In addition, the survey also asked responders to estimate how long the data they produced was used by the company or organisation.

Specifically, the respondent was offered the following choices to estimate the “usage period” for the data they produced, the options are presented along with the percentage of respondents who answered that option.

Total		
1	Most usage periods are 1 year or more but less than 3 years	35.9%
2	Most usage periods are 3 year or more but less than 5 years	20.4%
3	Most usage periods are 5 year or more but less than 7 years	8.7%
4	Most usage periods are 7 year or more but less than 10 years	7.2%
5	Most usage periods are 10 year or more but less than 15 years	5.0%
6	Most usage periods are 15 year or more	13.5%
7	Don't know	9.4%

This question followed one undertaken in the previous years survey where respondents could choose simply between producing data that was used for less than or more than 1 year. When these two information points are combined, an average asset life of data was able to be estimated. Both for all data produced as well as an average life of data only expected to be used for more than one year, which is data considered data assets. The results were:

Type of data	Average service life
For all data output	3.63
For data output expected to be used for more than 1 year	7.05

While the survey sample size was relatively large, there is a few caveats attached to this work beyond the standard statistical noise. As may be expected when obtaining information on a new subject, the respondents' interpretation of the question and subsequent answer may be quite subjective or based on imperfect knowledge that a respondent might have about how their company or organisation manages and uses the data.

Despite this, in an area where source data is severely lacking, this information is extremely useful. For instance at first glance it seems to support the theory that while a majority of data is used relatively soon, the decline is not linear and there is a non trivial amount that is kept for a long time, in other words the retirement profile is not evenly distributed and likely to have a long tail, both of which are different to the retirement profile normally used in the PIM.

### Annex 4.3. Summary of recommendations

Subject	Standard	Advanced
<b>Parameters used in PIM (Excluding Average service life)</b>	The standard recommendation for NSAs is to apply the same parameters in the compilation of depreciation and net capital stock of data as applied currently to other IPPs.	Countries should aspire to continually collect additional information on different assumptions and parameters to refine and improve the estimates of depreciation and capital stock being compiled in order to apply assumptions at an industry or occupation level.
<b>Average service life</b>	In the absence of other information, the standard recommendation is that NSAs apply an average service life for data assets of not longer than 5 years.	If relevant source data is made available, suggesting a longer or shorter service life, countries should look to implement a revised average service life.  NSAs should break up the estimate of data investment by industry in order to allow for different service lives to be applied based on the industry producing the data.

# 5

## Incorporating estimates of data production into existing national account outputs, including backcasting

### Introduction

- 5.1. This chapter discusses additional compilation aspects not addressed in the previous chapters 1-4. This includes the presentation of data within the accounts, backcasting estimates of data output and investment, and a further clarification on the conceptual split between data and other assets.
- 5.2. Several items in this chapter discuss how estimates of data production and investment are incorporated into the existing national account statistical outputs. While perhaps not scrutinised as much as the compilation of the estimate itself, countries should carefully consider how this is done and be pro-active in explaining it to users.
- 5.3. The introduction of data into a country's national accounts output reflects the altering of the production boundary in the System of National Accounts (SNA), which is just one of many changes being introduced in the revised version of the 2025 System of National Accounts (2025 SNA). Therefore, while the introduction of data into a country's official economic statistics will likely be done at the same time as other changes brought on by the revised SNA, the revisions, and changes to headline indicators due to the introduction of data should not be concealed amongst other changes. On the contrary, countries should be transparent regarding both the compilation of the estimates of data production and their impacts (regardless of the size) on national account aggregates such as total Intellectual property products, total gross fixed capital formation (GFCF), and Gross domestic product (GDP).
- 5.4. Official economic statistics are vitally important and changes to them need to be done appropriately to maintain their value as a significant knowledge asset. The trade-off between the options discussed in this chapter regarding how estimates of data are disseminated and back casted needs to be managed carefully to maintain the trust that users have in the set of accounts. Doing so will ensure that the value of the statistical output to the community at large is conserved.

### Reporting of data within the national accounts

- 5.5. From a production point of view, it has already been established that data can be produced by all industries and sectors within the economy. Therefore, estimates of data output should simply be incorporated into existing output series within the SNA production and income accounts. There will be no specific series identifying data within these tables as the tables are usually presented on an aggregate industry or sector basis.

- 5.6. When the output of data is capitalised (as will often be the case) data will be separately identifiable within the SNA capital account <sup>(39)</sup>. As noted in the 2025 SNA, “despite their conceptual difference, *data and databases are difficult to measure separately because they are produced with similar inputs and because transaction prices generally reflect the combined value of the database and the data. Therefore, it is recommended that data and databases are combined and reported in a specific single detailed intellectual property product (IPP) called data and databases.*” (2025 SNA § 11.115)
- 5.7. This handbook advocates this approach, meaning that any published series would incorporate values associated with the production of data covered in this handbook and capitalised for the first time as a produced asset, alongside the existing values associated with the production of databases already recorded as a produced asset and published in an existing series (most likely with computer software).
- 5.8. The 2025 SNA further suggests that this detailed product is then combined with “software including artificial intelligence to form a higher-level class of IP product”.(2025 SNA § 22.25). This statement should be viewed as a suggestion, as the actual dissemination and reporting of asset classes is a decision for each NSA based on data availability, user demand or any reporting requirements that they are aligned to, such as transmission to an international organization.
- 5.9. While undeniably intertwined and often used in correlation, ‘software’ undertakes many tasks and roles beyond creating data, likewise a single software asset can be used repeatedly in the production of multiple data assets. Additionally, data and computer software exhibit many distinct characteristics, in both the way they are produced and used. As has been discussed in previous chapters, it is these types of characteristics that determine some of the assumptions and methodology used for deflating and depreciating nominal estimates of IPP. If different assets are exhibiting distinct characteristics, ideally compilation should be undertaken separately to improve the accuracy of the final estimate. The accuracy of chain volume and capital stock estimates of data and databases investment will be improved if the nominal estimates are separated from computer software. Likewise, decisions on the compilation of capital stock and chain volume estimates of software will benefit from not needing to take into consideration the life cycle and input costs of data and databases.
- 5.10. From a user perspective, or when considering the extended time series of both assets, the evolution of investment in software and data are likely to deviate at certain points reflecting the various innovations at specific points in time. The benefits and use of software in production were established significantly earlier than the incorporation of data as a fundamental input into production. By delineating the two series, the rise of each asset can be better observed from a user perspective while also helping to alleviate concerns regarding double counting or missing values (See [Box 5.1](#)).
- 5.11. Therefore, for this handbook, **the standard recommendation is that for the purposes of the SNA capital accounts, investments (GFCF) in data and databases should be reported separately from computer software.** This is strongly recommended so that the conceptual changes introduced into the 2025 SNA can be fully realised and observed by users. This is regarding both data and databases but also the estimate of software which will now contain a greater focus on software associated with artificial intelligence. If such a deviation is simply not possible **then a less desirable basic proposition is to publish estimates of data and database GFCF alongside computer software in a combined higher-level category of IP.** While the dissemination of data and databases together with computer software as a single asset category may be considered a preliminary step in the publication of data output, NSAs should strive to separate the two types of output as soon as practically possible.

<sup>(39)</sup> Countries will disseminate GFCF from many different perspectives. By industry, by asset by sector, or even a combination of both. It is not expected that data will be reported any differently from the manner which existing assets are reported.

### BOX 5.1. SEPARATING COMPUTER SOFTWARE, DATA AND DATABASES, WHEN REPORTING GROSS FIXED CAPITAL FORMATION

Data in the SNA refers solely to digital data. This results in almost all (currently produced) data being constructed in unison with software and subsequently stored on a database. At face value this makes the three asset categories inextricably intertwined.

The small but clear difference between data and databases is best represented in [Figure 1.1](#) shown in [chapter 1](#). This outlines that the production of data focuses on the act of accessing and digitally recording information elements of observable phenomena. Conversely, the production of databases focuses on the act of organizing this data into a structured format, making it possible to analyse, draw conclusions and use in the production of other goods and services.

The obvious link between data and databases is explicitly noted in the SNA that states that *“Despite their conceptual difference, data and databases are difficult to measure separately because they are produced with similar inputs and because transactions prices generally reflect the combined value of the database and the data. For reporting purposes, data and databases are therefore combined into a single detailed intellectual property product (IPP) called data and databases”* (2025 SNA §11.115). This passage outlines the fundamental relationship between data and databases in the SNA. That while they do represent two distinct assets, they will most likely always be measured and reported as a single item.

The SNA also acknowledges the link between software and databases by noting that *“a computerized database, including the relevant data, cannot be developed independently of a database management system (DBMS), which is itself computer software”* (2025 SNA §11.111). However, the SNA still advocates that the expenditure on the two distinct assets should be separated if possible, suggesting that the cost of the database management system (DBMS) used should *“not be included in the costs of creating a database, but be treated as a computer software asset.”* (2025 SNA §11.116) In this way the SNA is outlining an advanced goal of separately identifying the expenditure on the different assets, in line with their conceptual differences, while acknowledging that this may not always be achievable.

This convergence of multiple assets becoming more intertwined leading to the production of multiple outputs has become more common in recent times, such as when software is added to vehicles and machinery. Likewise, from the perspective of data, the development of new production tools, such as online platforms and applications, which while usually developed to assist with other facets of production, are also designed with data production very much in mind.

For example, social media, online commerce and on-demand service platforms very rarely derive revenue directly from selling data, however the data that these applications and platforms capture is fundamental to their owners’ overall production strategy, and in fact their ability to harvest data from their software can often provide a competitive advantage regardless of the underlying service being offered. At face value this would then appear that the development of these platforms and applications is an extension of the production of data and should be included in the value of the asset. This is not the case for the SNA.

Despite the consideration on data capturing when developing the platforms, the platforms are first and foremost a computer program and from an SNA perspective any expenditure on their production should be considered the production of computer software rather than data. Therefore, even if the software is integral to the production of data (which in many cases it is), this does not mean expenditure on it should be considered GFCF of data asset. Many other assets are involved in the production of data (computer hardware, sensor equipment, etc.) however from the SNA perspective, these are individual capital assets used as an input into the production of another asset (Data).

In much the same way, cranes, cement trucks and construction equipment are all used to construct a high-rise building. Although the building could not have been produced without these inputs, and certain construction equipment cannot be used for anything other than constructing a building it is not suggested that the value of these should be added to the value of the building. Rather when measured via the sum of cost, a value associated with using these capital assets (depreciation) is included, but the overall value of the input remains separate from the final output.

Such an approach should be followed for measuring data. Ideally, countries would be able to separate the expenditure on producing data and databases from the expenditure on the assets used to produce data including expenditure on software. Doing so would allow for a more accurate recording of the evolution of the data economy, whereby to produce data assets to use as an input into their production, companies and organization must initially invest in other assets, which in turn allow for the production of data to occur. Separately reporting these different assets will provide users with the tools to better understand this evolution.

## Backcasting

### The need for backcasting

- 5.12. Chapters 1-4 have focussed on producing an estimate of data output and data investment in both current price and chain volume terms. These have covered both the flow of data investment and the capital stock of data assets; however, they have focused on producing an estimate for the current period only.
- 5.13. While most users will have upmost interest in the most recent period, the movement observed in the current period or current business cycle must be presented within the context of a full time series. Those responsible for forecasting the economy or to evaluate any appropriate policy reactions are dependent on the use of long time series to estimate and project the dynamics of the economy. As such, most NSAs have national account outputs that extend back many years. Therefore, for data to be appropriately incorporated into the System of National Account framework, estimates for an extended time series must be compiled.
- 5.14. The inputs and methodology described in previous chapters should allow NSO's to compile more than a single year's worth of estimates. However, it is unlikely that countries will have the required information for an extended period, or at least for the period of which digitalised data was considered a productive asset in the economy. As such it is likely that some form of backcasting will be required prior to officially publishing estimates.

### How long should digital data be backcast for?

- 5.15. A fundamental requirement in the backcasting of any statistical output is a decision on how far back the series should go. Often when NSO's introduce a new concept or classification, it is agreed that the change should be observable for the entire time series. While it depends on the length of the existing series, conceptually from the perspective of the SNA, data may not have been produced for the duration of the entire time series. Although numerous examples exist of non-digital data being collected and organized to be a fundamental input into production throughout history<sup>(40)</sup>, these do not meet the criteria of data used in the SNA. As outlined in [chapter 1](#), from the perspective of the SNA, data refers to *"Information content that is produced by accessing and observing phenomena; recording and storing information elements from these phenomena in a digital format, which provide an economic benefit when used in productive activities."* (2025 SNA 22.22) With this in mind, the output of this type of "SNA data" should certainly not predate the emergence of digitisation in the economy.

<sup>(40)</sup> Examples include the formation of the first modern insurance fund in 1740's (Ferguson, 2009) to the Feist publications court decision in 1991 (Supreme Court of the United States, 1991).

- 5.16. Using a strict interpretation of digitisation, there are clear examples of rudimentary digitisation assisting with economic production as far back as population census's in the 1890's (Columbia University, 2023; United States Census Bureau, 2023) and even simple digitalisation<sup>(41)</sup> within the 1950 airline industry (Sabre, 2024). However, these and other preliminary examples should generally be considered as not representative of the level of data being produced by the economy at large. Rather the creation of systematic digital data, produced to provide an economic benefit, is generally considered a more recent phenomenon. Beginning first with data that business and organizations produced about themselves and their own business practices before turning their attention to producing data based on the (incredibly significant) increase in user generated information elements following the widespread take up of the internet and social media.
- 5.17. While it is broadly agreed that the take up in data production has been exponential, there is no clear agreement as to the specific point in time when digital data first began being regularly produced, and thus should be included as a separate asset in the system of national accounts.
- 5.18. As such this handbook does not recommend a specific year when data should be back casted to. Rather this decision should be made on an individual basis based on the economy in question, user demands for the data, available information, as well as the existing start data for other comparable series, such as other IPP asset classes. However, to promote some form of consistency across countries, and **in the absence of additional information to the contrary, this handbook recommends incorporating a time series to at least the period covering 1995 – 2005**. It's considered that this would appropriately reflect the emergence of data production for many countries.
- 5.19. If length of the back series currently published begins after 2005, then it is likely that the backcasting of data need not be reduced to zero and estimates of data production will exist for the entire time series. In this case, to accurately estimate the capital stock of the data within the economy using the PIM, estimates of GFCF will need to extend beyond those published. Put simply, if the service life of data is T and capital stocks need to be published from date t onwards, the corresponding GFCF series should start at date t-T (at least) (Eurostat, 2023).

## Types of backcasting potentially suitable for estimates of data output and investment

- 5.20. Several different methods are available for backcasting national account outputs. The choice of which is most suitable will likely depend on the series being back casted, the availability of source data and how the back cast series interacts with other aggregates in the accounts. The methods listed below are included in the preliminary draft of the UN handbook on backcasting (UNSD, 2018). They are not an exhaustive list but are those most likely to be used by compilers of data output and data GFCF.
- 5.21. **Bottom-up estimation.** This method refers to compiling a back casted series by building up source data components, in the same or similar method to what is done for the current period estimates, at least for periodic benchmark levels. Since it is somewhat a continuation of the method used for the current period, the bottom-up approaches usually produce the most robust results. However, bottom-up estimation is also the most data intensive, requiring the same or similar source data for the whole period required by the backcasting.
- 5.22. **Retrapolation.** This approach uses growth rates from the currently compiled series to project backwards. There is no minimum or maximum number of periods required to determine the applied growth rate; however, it would usually match the number of years for which information is available for. The overriding assumption of this method is that growth observed in the recent period is consistent with previous periods. As such, the implication is that the periods used to determine the back cast growth rate is representative of the growth rate for the entire series. Such an assumption can be misleading if the recent period picks up only one portion of the business cycle or the specific series is heavily impacted by innovation or structural changes in the economy which is likely to produce data. However, retrapolation is arguably the easiest backcasting methodology to implement as it requires no new data sources. Rather the already compiled growth rate is applied backwards to 'guide' the series to a previously decided upon point in time.

<sup>(41)</sup> As outlined by the OECD, (OECD, 2019) Digitisation is the conversion of analogue data and processes into a machine-readable format. Digitalisation is the use of digital technologies and data as well as interconnection that results in new or changes to existing activities.

- 5.23. **Modelling the back series using a proxy indicator.** This method relies on the assumption that the relationship of an indicator or indicators to the estimate are stable over time and can be applied to the back period. Backcasting using this method does not need to mean that movements in the newly compiled back series are identical to the proxy indicator. Rather the indicator can be used to “guide” the back series to a previously decided point in time or estimate, which may or may not be zero. Such a method has already been recommended for compiling data in the periods (both annual and quarterly) where the source data contributing to the sum of cost approach is unavailable (See [chapter 2](#)).

## Applying backcasting methods to data series

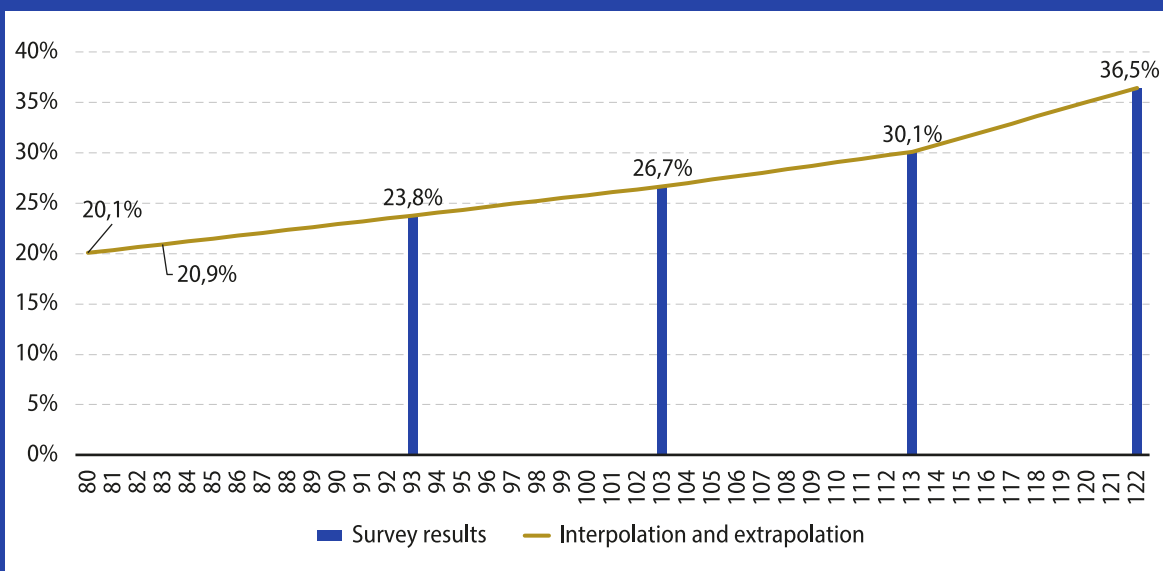
- 5.24. In the case of data, it is likely that NSAs can use a mixture of the above methods, depending on the length of time series desired. Some of the fundamental data sources used in the compilation of data production such as average wage and number of workers in certain occupations likely have relatively long availability due to them often being sourced from population census data. As such, it’s likely that the bottom-up approach is a viable solution for many NSAs, even if estimates are only produced for the years aligned with the population census.
- 5.25. An important consideration if applying the bottoms up approach is that the standard occupations and involvement rates provided in [chapter 2](#) are based for the compilation of data in the current period. They are based on current data production and likely do not reflect the occupations and involvement rates of data producers in historical periods. This is due to both the fundamental shift across the economy, whereby businesses and organizations are devoting more resources to the production of data but also due to the micro changes occurring at the occupational level. For example, data focused occupations such as ‘Computer systems technicians’ or ‘Medical records and health information technicians’ have likely always been involved in producing data. However, for other occupations, the production of data as a standard task within the job is a more recent phenomenon. This assumption is not only a logical conclusion based on the innovation of data over recent years, but also reflects the evidence obtained by the Japanese statistical office in their special internet survey which obtained some information on historical production of data (See [Box 5.2](#)).
- 5.26. As such, **the standard recommendation is to populate the back series with an already available series that displays a correlation with the growth observed in the compiled annual estimates of data output and GFCF.** This is similar to the standard recommendation used to populate current quarterly and annual periods where data is not available (See [chapter 2](#)). No specific proxy indicator is recommended as the choice will depend on data availability. Potential options include, 1) Remuneration of employees for data producing occupations, 2) Remuneration of employees for industries heavily involved in data production, or 3) GFCF on assets relating to data production, such as computer hardware and software. Importantly it is not recommended for the newly constructed series to mirror the indicator series exactly as the end points are likely different between the two. Rather a decision should be made as to when the production and GFCF of data should commence and the indicator series used to ‘guide’ the back series to this point.
- 5.27. If required data sources are available, the advanced recommendation is for NSAs to **compile a back series via the bottom-up approach. However, it is not recommended to replicate the current period compilation method exactly using earlier iterations of the same data source. Rather adjustments should be made to the chosen occupations and involvement rates over time to reflect the evolution of data production over the previous periods.**

**BOX 5.2. JAPANESE SURVEY ON HISTORICAL DATA PRODUCTION**

It is unusual for NSO's to use survey data to estimate back series within the national accounts. However, as part of their 'special internet survey' (Japanese Cabinet Office, 2024), the Economic and Social Research Institute (ESRI) within the Japanese Cabinet Office obtained information that would assist them to do exactly that. Based on a sample from a previous internet survey, respondents were asked to retrospectively provide information on the proportion of people and time engaged in data-related work in their past organizations. This was done for four-time points; 1 year, 10 years, 20 years, and 30 years. Estimates for the years not covered by the survey were estimated via linear extrapolation between the obtained values.

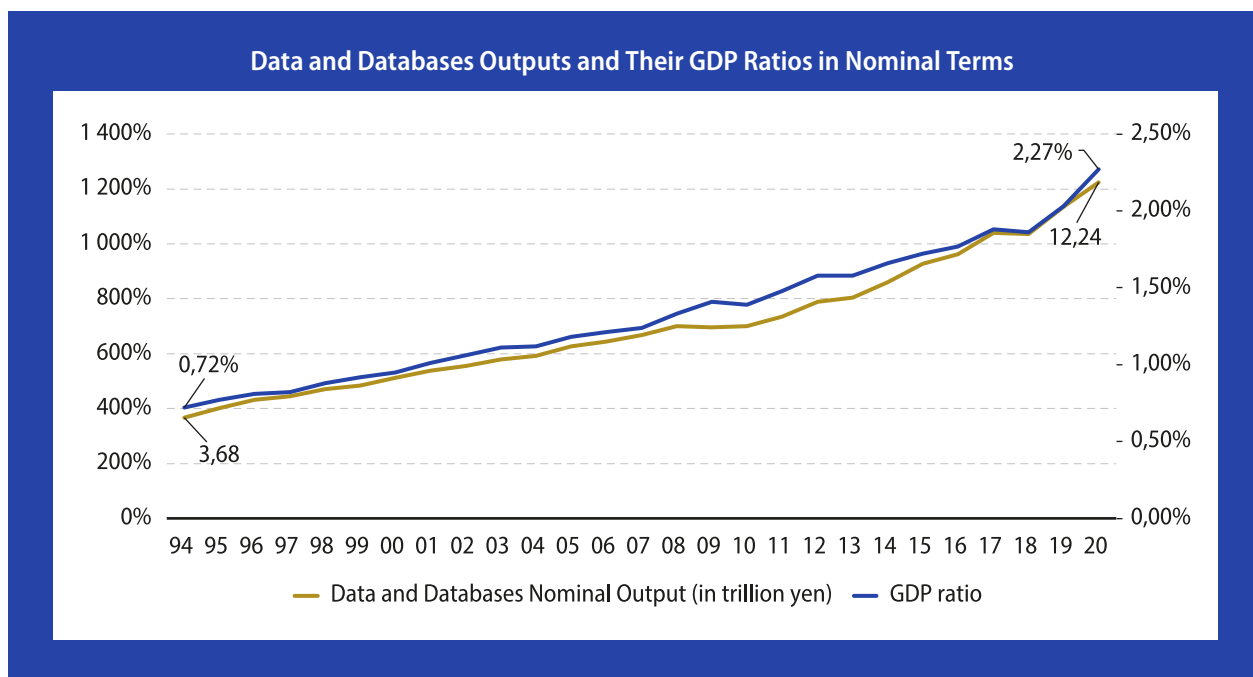
The proportion of people involved in data production has increased over time, as might be expected. In 1980 around one fifth of people were involved in data production in some form. This increased to just above one third in 2022 (see figure below). Similarly, the proportion of their working day that these people spent on data production (a similar concept to the involvement rate used in this handbook) also increased over time.

Proportion of people engaged in data-related work



Importantly, as well as asking about their role and time spent on data production, the survey also asked what methods were used to store and use the data created. In response to this question, a considerable number of respondents reported storing data on paper. Based on this, it was identified that several respondents had answered the question in relation to all data including the production of non-digital data, which is excluded from the SNA definition of data. Such a revelation allowed for an adjustment to be made to the backcasting estimates to remove the production of non-digital data.

Based on the results from the survey, the Japanese cabinet office was able to compile a back series for data production in the Japanese economy which showed that data production grew from 0.7% of GDP in 1994 to 2.27% in 2022 (See below). This novel way of obtaining information on historical data production has not been further reviewed to determine its accuracy or comparability but provides an alternative solution to backcasting estimates of data that can be independent of other IPP assets.



Source: (Japanese Cabinet Office, 2024)

## Annex 5.1. Summary of recommendations for incorporating estimates of data production into existing national account outputs

Issue	Advanced recommendation	Standard recommendation	Additional considerations
<b>Reporting of data in capital account</b>		For the purposes of the SNA capital accounts, investments (GFCF) in data should be reported together with databases as a single IP product, but that this should be reported separately from computer software.	If data and databases are unable to be separated a final basic option is to publish estimates of data and database GFCF alongside computer software.
<b>Backcasting – length of series</b>		In the absence of additional information to the contrary, this handbook recommends incorporating a time series to at least the period covering 1995 – 2005.	If the back series for currently published IPP estimates does not extend to 2005, then data should be back cast until the beginning of published estimates.

Issue	Advanced recommendation	Standard recommendation	Additional considerations
<p><b>Backcasting – choice of methodology</b></p>	<p>If required data sources are available, the advanced recommendation is for NSAs to compile a back series via the bottom-up approach.</p>	<p>The standard recommendation is to populate the back series with an already available series that displays a correlation with the growth observed in the compiled annual estimates of data output and GFCF.</p> <p>This is similar to the standard recommendation used to populate current quarterly and annual periods where data is not available.</p>	<p>It is not recommended to replicate the current period compilation method exactly using earlier iterations of the same data source. Rather adjustments should be made to the chosen occupations and involvement rates over time to reflect the evolution of data production over the previous periods.</p>

# References

- ABS. (2021). *Concepts, Sources and Methods, Australian National Accounts*. <https://www.abs.gov.au/statistics/detailed-methodology-information/concepts-sources-methods/australian-system-national-accounts-concepts-sources-and-methods/2020-21/chapter-14-capital-account/consumption-fixed-capital/sources-and-methods-annual#mean-asset>.
- Ahmad, N. (2005). The measurement of databases in the National Accounts. <https://unstats.un.org/unsd/nationalaccount/aeg/papers/m3Databases.PDF>.
- Ahmad, N., & van de Ven, P. (2018). Recording and measuring data in the System of National Accounts. [https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12\\_3c1\\_Data\\_SNA\\_asset\\_boundary.pdf](https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12_3c1_Data_SNA_asset_boundary.pdf).
- Amegble, K., Bugge, B., & Sinclair, A. (2023). *Estimation of Investment in and Stock of Data and Databases in the Canadian System of National Accounts*. <https://iariw.org/wp-content/uploads/2023/10/Amegble-Bugge-Sinclair.pdf>.
- Calderón, J., & Rassier, D. (2022). *Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings*. <https://www.bea.gov/system/files/papers/BEA-WP2022-13.pdf>.
- Columbia University. (2023). *Columbia University Computing History: Herman Hollerith*. New York.
- Corrado, C., Haskel, J., Iommi, M., & Jona-Lasinio, C. (2022). *The value of data in digital-based business models: Measurement and economic policy implications*.
- Corrado, C., Haskel, J., Iommi, M., Jona-Lasinio, C., & Bontadini, F. (2023). *Data, Intangible Capital, and Productivity*.
- Coyle, D., & Manley, A. (2022). *What is the Value of Data? A review of empirical methods*. [https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/07/policy-brief\\_what-is-the-value-of-data.pdf](https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/07/policy-brief_what-is-the-value-of-data.pdf).
- Destatis. (2024). *Feasibility Project on the Recording of Data and databases in the national accounts*.
- EC et al. (2009). *2008 System of National Accounts*.
- European Commission. (2010). *European System of Accounts*. <https://ec.europa.eu/eurostat/documents/3859598/5925693/KS-02-13-269-EN.PDF>.
- Eurostat. (2016). *Handbook on prices and volume measures in the national accounts*. <https://ec.europa.eu/eurostat/documents/3859598/7152852/KS-GQ-14-005-EN-N.pdf/839297d1-3456-487b-8788-24e47b7d98b2>.
- Eurostat. (2023). *DMES task force on fixed assets and estimation of consumption of fixed capital under ESA 2010*. <https://ec.europa.eu/eurostat/documents/24987/19760111/task-force-fixcap-final-report.pdf/6ae7756a-509e-f4b8-4516-beb3c3328e8b?t=1723539577790>.
- Eurostat. (2024). *Directors of Macroeconomic Statistics Task Force on fixed assets and estimation of consumption of fixed capital under European System of Accounts 2010. UNECE: Conference of European Statisticians*. [https://unece.org/sites/default/files/2024-03/6\\_DMES%20TF%20on%20fixed%20assets%2C%20Eurostat.pdf](https://unece.org/sites/default/files/2024-03/6_DMES%20TF%20on%20fixed%20assets%2C%20Eurostat.pdf).
- Eurostat, IMF, OECD, UNSD, World Bank. (2025). *2025 System of National Accounts*.
- Eurostat-OECD. (2019). *Final report from task force on Land and other Non-Financial assets - Intellectual Property Products*. <https://ec.europa.eu/eurostat/documents/24987/725066/Eurostat-OECD+Report+on+Intellectual+Property+Products.pdf>.

- Ferguson, N. (2009). *The ascent of money : a financial history of the world*. New York: Penguin Books.
- ISWGNA. (2023). *SNA Guidance note, DZ.6: Recording of data in the National Accounts*. [https://unstats.un.org/unsd/nationalaccount/RADOCS/ENDORSED\\_DZ6\\_Recording\\_of\\_Data\\_in\\_NA.pdf](https://unstats.un.org/unsd/nationalaccount/RADOCS/ENDORSED_DZ6_Recording_of_Data_in_NA.pdf).
- ISWGNA. (2023). SNA revision Guidance note; DZ. 9 Improving the visibility of Artificial Intelligence in the national accounts. [https://unstats.un.org/unsd/nationalaccount/RAdocs/ENDORSED\\_DZ7\\_AI.pdf](https://unstats.un.org/unsd/nationalaccount/RAdocs/ENDORSED_DZ7_AI.pdf).
- Japanese Cabinet Office. (2022). *Special internet survey*. <https://www.esri.cao.go.jp/en/esri/prj/menu-e.html>.
- Japanese Cabinet Office. (2024). *eport of Research on the measurement of the Digital Economy toward 2025SNA : Recording Methods for Data as Capital*. <https://www.esri.cao.go.jp/jp/esri/prj/hou/hou088/hou088.html>: (in Japanese).
- Mitchell, J., Ker, D., & Leshner, M. (2022). *Measuring the economic value of data*. [https://www.oecd-ilibrary.org/science-and-technology/measuring-the-economic-value-of-data\\_f46b3691-en](https://www.oecd-ilibrary.org/science-and-technology/measuring-the-economic-value-of-data_f46b3691-en): OECD publishing.
- Nguyen, D., & Paczosi, M. (2020). *Measuring the economic value of data and cross-border data flows: A business perspective*. <https://doi.org/10.1787/6345995e-en>.
- OECD. (2009). *Measuring Capital, OECD Manual*. <https://unstats.un.org/unsd/nationalaccount/docs/oecd-capital-e.pdf>.
- OECD. (2010). *Handbook on deriving capital measures of intellectual property products*. <https://unstats.un.org/unsd/nationalaccount/docs/oecd-ipp.pdf>.
- OECD. (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Paris: OECD publishing. doi:<https://doi.org/10.1787/9789264239012-en>
- OECD. (2019). *Going Digital: Shaping Policies, Improving Lives*. <https://doi.org/10.1787/9789264312012-en>: OECD publishing.
- OECD. (2022). *Measuring the value of data and data flows*. <https://www.oecd-ilibrary.org/docserver/923230a6-en.pdf?expires=1695240160&id=id&accname=guest&checksum=A23446F17CC73C2BAA1DEEAADB8AE8F3>.
- OECD, based on Eurostat. (2021). *Measuring the value of data and data flows*.
- ONS. (2022). *Capital stock user guide*. <https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/methodologies/capitalstockuserguideuk>.
- Sabre. (2024). *History of Sabre*. Retrieved from <https://www.sabre.com/files/Sabre-History.pdf>
- Schmidt, J., Pilgrim, G., & Mourougane, A. (2023). "What is the role of data in jobs in the United Kingdom, Canada, and the United States?: A natural language processing approach". <https://doi.org/10.1787/fa65d29e-en>.
- Smedes, M., Nguyen, T., & Tenburren, B. (2022). "Valuing data as an asset, implications for economic measurement". <https://www.abs.gov.au/system/files/documents/7bfccb4ddb8aded818330bebe6b76b14/Smedes%20-%20Valuing%20data%20as%20an%20asset.pdf>.
- Statistics Canada. (2019). *The value of data in Canada: Experimental estimates*. Retrieved from <https://www150.statcan.gc.ca/n1/en/pub/13-605-x/2019001/article/00009-eng.pdf?st=ifEOEPUK>
- Supreme Court of the United States. (1991). *Feist Publications, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340. <https://supreme.justia.com/cases/federal/us/499/340/>.
- Task team on measuring data in the National Accounts. (2023). Terms of Reference. [https://europea.eu.sharepoint.com/:w:/r/teams/GRP-PRO-Eurostat-IMF-Task-Team-Data-as-an-asset/\\_layouts/15/Doc.aspx?sourcedoc=%7BE329D06D-E2CD-415E-A811-2E89BBOFF7D7%7D&file=ESTAT-IMF\\_TF\\_Data-as-an-Asset--ToR.docx&action=default&mobileredirect=true](https://europea.eu.sharepoint.com/:w:/r/teams/GRP-PRO-Eurostat-IMF-Task-Team-Data-as-an-asset/_layouts/15/Doc.aspx?sourcedoc=%7BE329D06D-E2CD-415E-A811-2E89BBOFF7D7%7D&file=ESTAT-IMF_TF_Data-as-an-Asset--ToR.docx&action=default&mobileredirect=true).
- United Nations Statistical Commission. (2024). Report on the fifty-fifth session. [https://unstats.un.org/UNSDWebsite/statcom/session\\_55/documents/2024-36-FinalReport-E.pdf](https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/2024-36-FinalReport-E.pdf).
- United States Census Bureau. (2023). *History of the Census: 1890*. Retrieved from [https://www.census.gov/history/www/through\\_the\\_decades/overview/1890.html](https://www.census.gov/history/www/through_the_decades/overview/1890.html)

UNSC, Task Team on ISIC. (2024). Explanatory Notes of the Standard Industrial Classification of All Economic Activities, Revision 5 (ISIC Rev. 5). *United Nations Statistical Commission 55th session*. [https://unstats.un.org/UNSDWebsite/statcom/session\\_55/documents/BG-4e-ISIC5\\_Exp\\_Notes-E.pdf](https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/BG-4e-ISIC5_Exp_Notes-E.pdf).

UNSC; Task Team on ISIC. (2024). Introduction of the Standard Industrial Classification of All Economic Activities, Revision 5 (ISIC Rev.5). *United Nations Statistical Commission; 55th session*. [https://unstats.un.org/UNSDWebsite/statcom/session\\_55/documents/BG-4e-ISIC5-Introduction-E.pdf](https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/BG-4e-ISIC5-Introduction-E.pdf).

UNSD. (2018). Draft handbook on backcasting. *12th Meeting of the Advisory Expert Group on National Accounts*. [https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12\\_8iiiB\\_Backcasting.pdf](https://unstats.un.org/unsd/nationalaccount/aeg/2018/M12_8iiiB_Backcasting.pdf).

UNSD et al. (2009). System of National Accounts 2008. <https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf>.

UNSD et al. (2025). System of National Accounts 2025.

UNSD, OECD, Eurostat, IMF, WTO, UNWTO. (2010). *Manual on Statistics of international trade in services*.



## GETTING IN TOUCH WITH THE EU

### ***In person***

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### ***On the phone or in writing***

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls);
- at the following standard number: +32 22999696;
- via the following form: [european-union.europa.eu/contact-eu/write-us\\_en](https://european-union.europa.eu/contact-eu/write-us_en).

## FINDING INFORMATION ABOUT THE EU

### ***Online***

Information about the European Union in all the official languages of the EU is available on the Europa website ([european-union.europa.eu](https://european-union.europa.eu)).

### ***EU publications***

You can view or order EU publications at [op.europa.eu/en/publications](https://op.europa.eu/en/publications). Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### ***EU law and related documents***

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex ([eur-lex.europa.eu](https://eur-lex.europa.eu)).

### ***EU open data***

The portal [data.europa.eu](https://data.europa.eu) provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

# Handbook on measuring data in the System of National Accounts

The inclusion of “Data” as a new class of produced assets represents one of the most significant innovations in the 2025 System of National Accounts (SNA). This handbook, developed by an international task team jointly led by Eurostat and the International Monetary Fund, is designed to support compilers in incorporating this important new feature into their national accounts systems.

Additionally, the handbook serves as an informational resource for users who wish to understand the representation of data output and data capital formation in the estimates.

---

**For more information**

**<https://ec.europa.eu/eurostat/>**



Publications Office  
of the European Union

ISBN 978-92-68-29884-8