Indicators for an Open Data Impact Assessment

data.europa.eu
The official portal for European data

This study has been prepared as part of data.europa.eu, an initiative of the European Commission. The Publications Office of the European Union manages data.europa.eu.

European Commission
Directorate-General for Communications Networks, Content and Technology
Unit G.1 Data Policy and Innovation
Email: CNECT-G1@ec.europa.eu

data.europa.eu
Email: info@data.europa.eu

Authors:
Patrick de Bas
Martin Page PhD

The authors would like to thank the following contributors:
Elena Simperl
Oscar Corcho
Hans Graux
Giulia Carsaniga
Jalisa d'Hont
Vincent Schreuder
Cristina Moise

Last update: February 2024
https://data.europa.eu/

ISBN 978-92-78-44051-0 DOI 10.2830/323236 OA OA-05-24-139-EN-N

Contents

Executive summary

The value open data can offer society has been a topic of interest for many years. Several studies have measured the economic impact of open data and public sector information, however, the results of the studies vary, and no consistent picture of the impact of open data has been developed. Structured methods and indicators to measure the impact of open data and the possibilities of basing these indicators on existing datasets still need to be explored and are the subject of this study.

The open data value chain as a framework for understanding how impact is created

A reflection on the open data value chain is necessary to identify potential indicators for measuring open data impact. The open data value chain encompasses the processes applied to data to convert it from a raw form to a valuable use. Figure A presents the four main phases of the value chain model used in this study: data collection, publication, (stimulation of) uptake and application (by reusers).

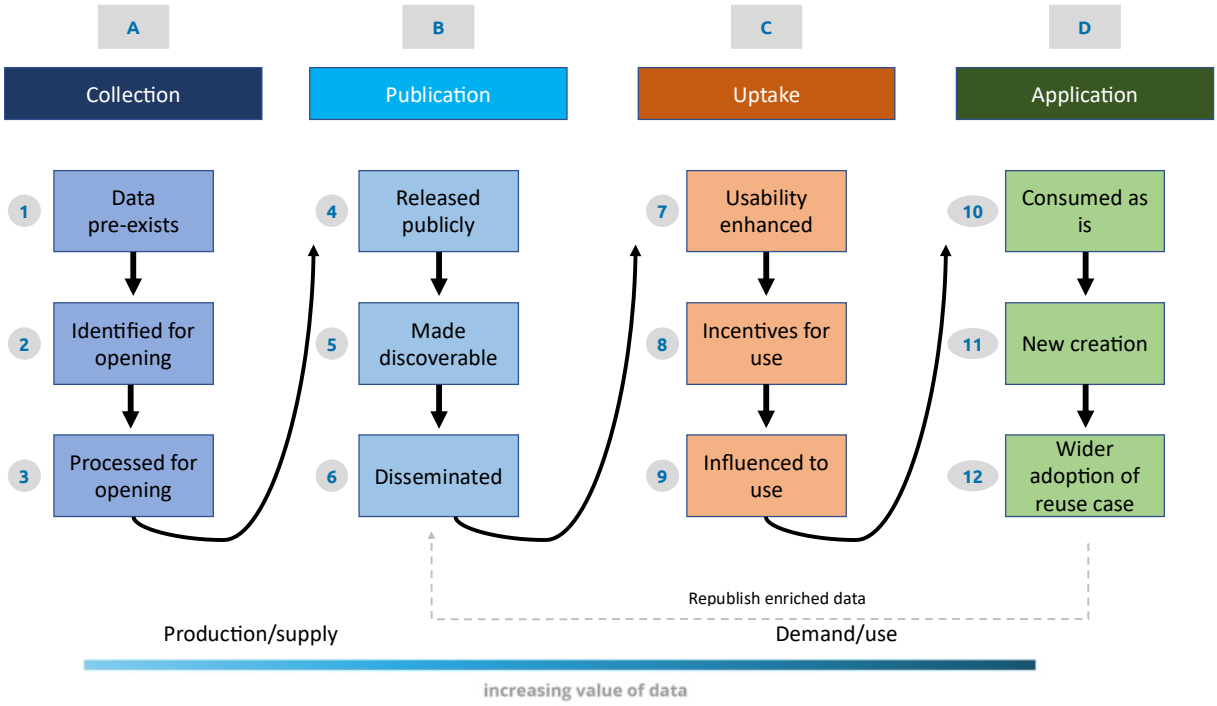

Figure A: Open data value chain, adapted from Open Data Watch ([1])

Using the concepts of output, outcome and impact to assess the effects of open data

Societal effects associated with the introduction of open data can be divided into the following three categories.

- **Outputs** are the direct results of open data availability (for example, a dashboard that integrates data on the built environment at the municipal level into one overview).
- **Outcomes** are the short- and medium-term effects (for example, improved information for monitoring and maintenance).

([1]) https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/.

- **Impact** is a long-term or broad effect (e.g. higher return on public money and increased inhabitant satisfaction).

Identifying potential indicators to measure the output, outcome and impact of open data

At the output level, a wide range of indicators closely linked to the open data value chain are available. Section 4.2 of this report contains an extensive, although non-exhaustive, list of such indicators, accompanied by an assessment of the challenges faced in obtaining data to measure the indicators and the value of the indicators in predicting overall impact. Indicators at the earlier parts of the value chain (i.e. collection and publication), where the focus lies on the supply of datasets , are typically easy to measure but provide a lower level of insight into impact than indicators at the later parts of the value chain. The most promising output indicators, combining relatively low difficulty in data collection and a relatively high extent of insight, are:

- number of reuse cases (success stories) showcased on the national portal;
- explicit references to datasets used;
- portal traffic; and
- application programming interface (API) usage statistics, such as the number of queries.

These indicators are proposed to be further validated and tested in upcoming studies.

At the outcome and impact level, indicators are more closely linked to the reuse cases developed with the support of open data. As the range of reuse cases and the societal challenges they address is broad, the list of potential indicators is similarly extensive. For outcome and impact, the advice is to develop a specific list of indicators depending on the purpose of the impact assessment and the thematic domain – economic, social or environmental – on which the assessment focuses. Section 4.2 contains illustrative categories of outcome and impact with associated examples of indicators.

Operationalising potential indicators in a methodological framework in future studies

Numerous examples of output, outcome and impact indicators are available, each offering a different extent of insight into the impact of open data and having a different difficulty in measuring. A future impact methodology could test the validity of the (most promising) indicators proposed and compile a hierarchy of indicators, giving them different weights based on their relationship to impact. Such a system of weights will allow the creation of a hierarchy of output metrics based on the scale of importance of each indicator.

Insofar as the data for these indicators are obtained from public sources, they may contribute to a system for automated monitoring over time. When making use of non-public data, privacy considerations should be taken into account. This involves meeting the requirements of European privacy and data protection law, mainly Regulation (EU) 2016/679, and Directive 2002/58/EC (as amended). Section 5.2 contains a more detailed list with privacy considerations for specific indicators.

1. Introduction

1.1. Context

Over the years, there has been a keen interest in the value open data can offer society. To exploit this value and its potential impact, more and more countries, regions and municipalities are publishing datasets on their national, regional and local portals, allowing citizens and businesses to reuse them for various purposes. While the number of open government data initiatives has increased considerably over the past decade, the impact of these initiatives still needs to be objectively described. A few studies have measured the economic impact of open data and public sector information ([2]), however, the results of the different studies vary and no consistent picture of the impact of open data has been developed.

To provide evidence of the impact of open data, a workstream on impact assessment has been established within the data.europa.eu. The first report of this workstream, titled *Rethinking the Impact of Open Data* ([3]), reviewed studies about the economic impact of the reuse of public data resources, concluding that a consistent definition of the impact of open data does not exist. As discussed in that report, 'most academic articles that look to explore the impact of open data refer to existing open data frameworks, with the Open Data Maturity (ODM) and Open Data Barometer (ODB) ones most frequently represented'.

The *Rethinking the Impact of Open Data* study also analysed in more detail the different techniques and approaches used by national, regional and local open data portals to measure impact, including the analysis of good practices, the use of forms to get user feedback, the use of usage statistics of open data portals and the measurement of the scope and update frequency of their datasets. It also considered the role of intermediaries (also known as 'infomediaries'), with the Asociación Multisectorial de la Información (Asedie) annual report ([4]) being a reference for this type of analysis. The analysis led to the acknowledgement that impact indicators of open data portals and proposed indicators for open data intermediaries largely rely on proxies for impact measurement. However, the presence of a use case repository or the availability of user statistics is considered insufficient to measure the overall impact of open data. The main challenge ahead is to find ways to improve the genuine insights about measuring the impact of open data.

1.2. Objective of the current study

While the impact of open data has been studied before, typically on an ad hoc basis, structured methods and indicators to measure its impact and the possibilities of basing these indicators on existing datasets still need to be explored and are the subject of this study.

([2]) See Section 3.1 for a more detailed overview of these studies and their findings.

([3]) Publications Office of the European Union, Ooijen, C., Osimo, D., Regeczi, D. et al. (2023), *Rethinking the Impact of Open Data: A first step towards a European impact assessment for open data*, Publications Office, January 2023, https://data.europa.eu/sites/default/files/report/Rethinking %20impact %20of %20open %20data.pdf.

([4]) https://www.asedie.es/en/annual-report.

This report is the second step in defining the methodology for a European-wide assessment of the impact of open data. The study series (see Figure 1) aims to create an interactive monitoring tool for open data impact assessment.

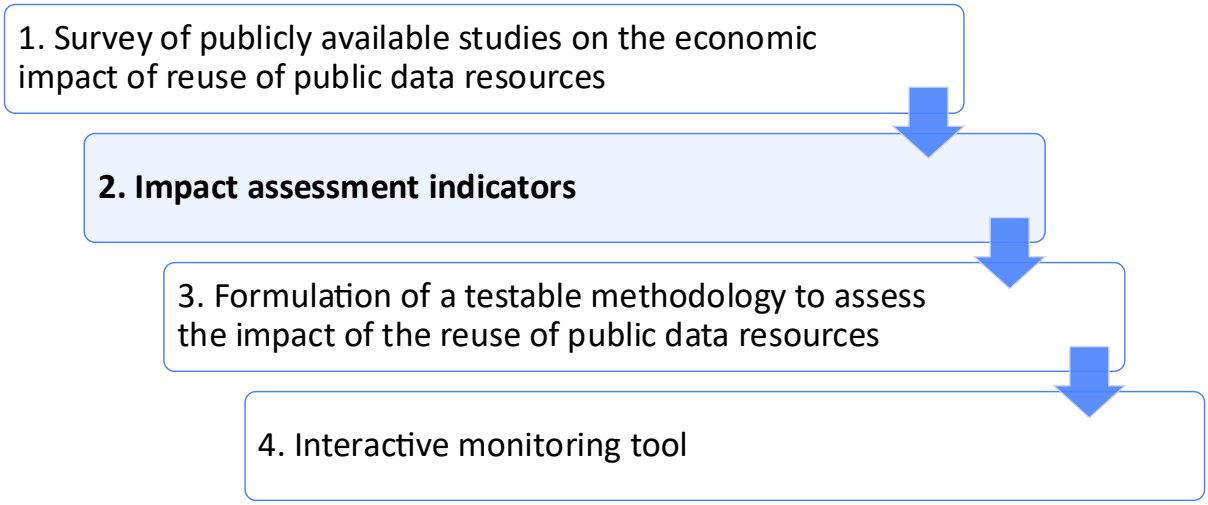

Figure 1: Measuring the impact of open data

This second study aims to provide theoretical input to develop a methodology that analyses the impact of public data resources, focusing on the different impacts to be considered and the indicators that could quantify them. This report uses the results and conclusions presented in the first study on 'rethinking the impact of open data' as a foundation. The results of this study could be incorporated into a third study, where a draft methodology to measure the impact of open data could be developed and piloted.

1.3. Outline

The report is structured as follows.

- **Chapter 2** establishes the analytical framework. This includes defining the concept of open data (Section 2.1) and the open data value chain, allowing for a better understanding of how open data leads to impact (Section 2.2). This chapter defines output, outcome and impact and reviews their possible application in the open data domain (Section 2.3).
- **Chapter 3** investigates existing approaches to measuring impact assessment. After investigating the methodologies of previous studies (Section 3.1), a more detailed review of two studies is conducted: the open data maturity assessment (Section 3.2) and the impact assessment for high-value datasets (Section 3.3).
- **Chapter 4** reviews possible indicators for a future open data impact assessment for output indicators (Section 4.1) and outcome and impact indicators (Section 4.2).
- **Chapter 5** summarises and concludes on the possible improvement of impact assessment for future studies.

2. Establishing the analytical framework

This section aims to explain how the availability of open data can lead to the creation of value and potential impact. First, we define open data (Section 2.1). Second, we propose an open data value chain describing how open data is processed and made more valuable and how this value is realised (Section 2.2). Third, we define the concepts of outputs, outcomes and impacts of open data (Section 2.3.1) and describe how these concepts fit along the open data value chain (Section 2.3.2).

2.1. Open data

Open data refers to data presented in open formats that individuals can use freely and share for any purpose. In the EU, the open data directive ([5]) on the reuse of public sector information provides a common legal framework for a European market for government-held data. The directive addresses materials held by public sector bodies in EU countries, at national, regional and local levels. This includes material held by ministries, state agencies, municipalities, and organisations funded mostly by or under the control of public authorities. It is important to note that not all public sector information is open data; for example, some information is confidential or sensitive. Other legislation, such as the Data Governance Act ([6]), includes measures to stimulate the reuse of public sector data that cannot be made available as open data.

2.2. The open data value chain

A value chain ([7]) disaggregates a process into its strategically relevant activities. Concretely, one can picture a company converting inputs to outputs in a series of steps. Each step of processing adds value to the input, creating a subsequently more valuable output. A value chain comprises the connections between these steps that change a low-value input into a high-value output.

The concept of a value chain can also be applied to intangible goods such as data. In addition to being intangible, another unique feature of data is that using it for a particular purpose does not necessarily make it unavailable for others to use (called 'non-rivalry'; consuming data does not deplete it or exclude others from consuming it). The value chain of data encompasses the processes applied to data to convert it from a raw form to a valuable use (or several uses), such as actionable information or solutions.

Various models of the data value chain are described in the literature, with overlapping elements, for example, Faroukhi et al. (2020) ([8]) and Curry (2016) ([9]) on the big data value chain, as well as models on the lifecycle of data ([10]). In general, data is generated, published, discovered by others, processed and analysed before being exploited to realise value. The early steps of data production and

([5]) https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data and https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024.

([6]) https://digital-strategy.ec.europa.eu/en/policies/data-governance-act and https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX %3A32022R0868.

([7]) Porter, M. E. (1985), *Competitive Advantage: Creating and sustaining superior performance*, Free Press.

([8]) Faroukhi, A. Z., El Alaoui, I., Gahi, Y. et al. (2020), 'Big data monetization throughout big data value chain: A comprehensive review', *Journal of Big Data*, Vol. 7, No 3, Springer Nature.

([9]) Curry, E. (2016), 'The big data value chain: Definitions, concepts, and theoretical approaches', in Cavanillas, J., Curry, E. and Wahlster, W. (eds), *New Horizons for a Data-Driven Economy*, Springer, Cham, pp. 29–37, https://doi.org/10.1007/978-3-319-21569-3_3.

([10]) For example, from the EU's 2013 data value chain strategy, p. 6, https://digital-strategy.ec.europa.eu/en/library/elements-data-value-chain-strategy.

dissemination ensure data is collected and made available in useable formats. However, value is realised from data when it is used.

Nonetheless, data is of value throughout the value chain, not only at the final stage (the headline value). It has different values for different users. For example, open data can be used to monitor the progress of other processes, such as measuring inputs, outputs and outcomes, or it can be used in activities that create awareness about open data. Ultimately, the data value chain relies on reusers finding high-value uses for data and creating the processes to transform the data into an end use. Once data is put to use, it may have an impact.

Figure 2 presents the value chain model used in this study for open data. The model has four main stages. Data is supplied once it is **collected** and **published**. Once available and known to reusers, the data is **taken up** (potentially after promotion and incentives directed at the reuser) and **applied** to some end purpose.

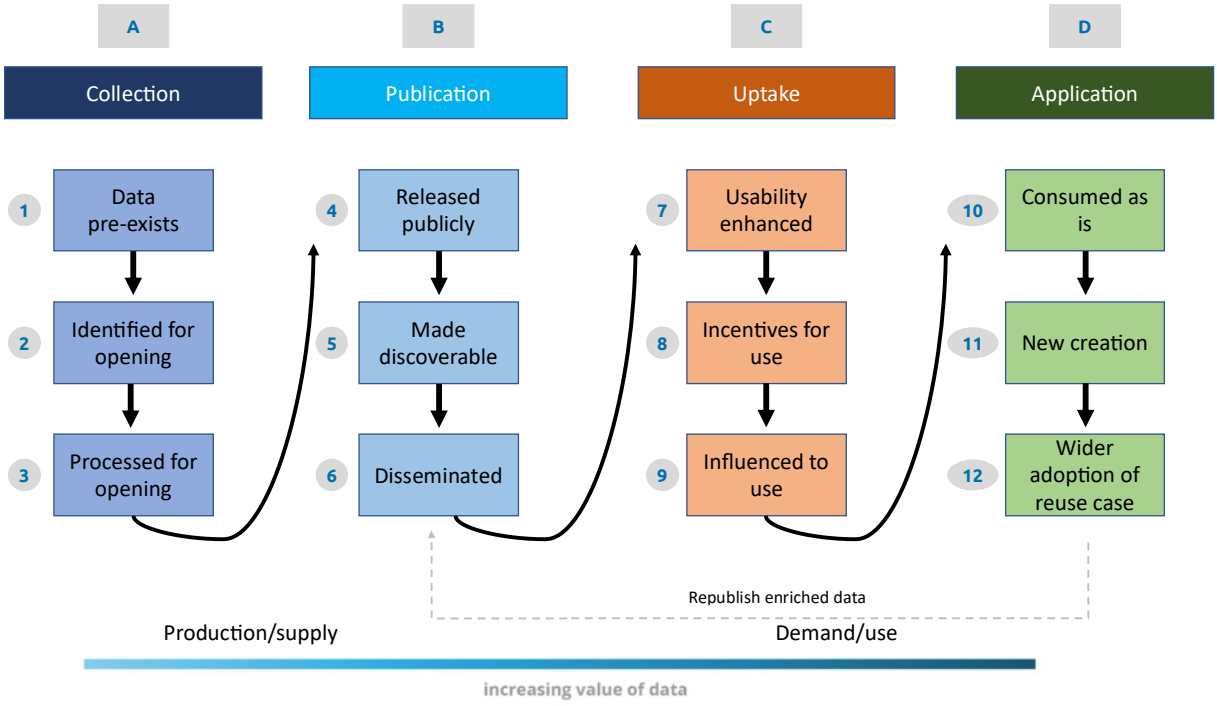

Figure 2: Open data value chain, adapted from Open Data Watch ([11])

The activities within these broader stages are described below.

A. Open data collection

Collecting data is about creating the data that starts the value chain. In the case of open data, the data is already collected because it is needed to support government operations. Therefore, the (1) **data to be made open pre-exists**. This required the relevant data to be identified for collection and for processes to be put in place to collect the data. The public-sector body likely has procedures to process,

([11]) https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/.

clean and format the data for its primary use. There may also be procedures to ensure that the collected data meets a certain level of quality.

The open data value chain starts with (2) **identifying** datasets to be made open. These data (and the accompanying metadata ([12])) must then be (3) **pre-processed** to be ready for publication. For example, data might need to be aggregated, confidential information removed and the dataset stored in formats that allow reuse. The original dataset is now a new data product for the open data value chain.

B. Open data publication

Publishing is about making data available for reusers to discover and access. The data should be (4) **released** in accessible formats with appropriate supporting documentation ([13]). Once publicly available, the data must be made (5) **discoverable** for potential reusers to find, such as through data portals and search engines. Discoverability is enabled by high-quality metadata ([14]) that tags and describes the dataset. To create awareness that the data is available, the data should be (6) **disseminated** ([15]) through channels that can reach the reuser. This can be achieved through activities such as press releases, editorial content, webinars and notifications of new datasets.

C. Open data uptake

Uptake is when reusers start using the available data they have discovered. Various activities foster uptake by making data more usable and encouraging users to search for and put open data to use. However, a user may take up data without such encouragement.

Fostering uptake involves making the available open data more (7) **usable**. For example, portal features to preview tabular and geospatial data help reusers gain a quick understanding of datasets, reducing friction in using the data. In this way, a good user experience can encourage the data uptake from a portal. Open data can also be enriched by combining it with data from other sources, improving the dataset's information and making it more usable.

Data uptake is also stimulated through external (8) **incentives** that encourage reusers to use open data, such as reducing the cost of data use. Uptake is also fostered by (9) **influencing** potential reusers to take up open data. This can be achieved by developing their understanding of the relevance, importance, and potential value of data so that they adopt new habits and learn skills related to data use ([16]). Sharing examples of reuse cases and success stories may also inspire potential reusers to pursue their own ideas using open data.

Feedback between the steps in the value chain also emerges. As users take up open data, they want assurance about the quality and credibility of the data. Data publishers, in turn, are incentivised to improve the quality of their data or metadata, such as being more transparent about the data collection methods and quality controls. In this way, the good reputation of the data publisher is also a factor that increases the value of the data.

([12]) Metadata refers to data that describes and gives information about other data, such as author, date, etc.
([13]) This activity necessitates that the data is stored somewhere accessible.
([14]) The quality of the data refers to the quality of the data collection and primary processing. The quality of the metadata refers to the quality of the description of the data.
([15]) https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Data_dissemination.
([16]) For example, trainings can be thought to increase the value of data by developing competent users of the data.

D. Open data application

The uptake of open data can lead to a new application of the data to understand a problem or make a decision. The data can be (10) **consumed as is**. For example, the results of a survey might help a reuser make a decision. More likely, the open data must be processed further, giving rise to the (11) **creation** of a new actionable insight or solution to a problem (a reuse case). To support this creation, the data may need to be transformed or integrated with data from other sources to enrich the original dataset. Specialised tools and analytics methods may be required to inspect and manipulate the data to extract value. Depending on the nature of this creation, it may be (12) **adopted** by a broader user base. Finally, this processed data can be **republished for further reuse**, feeding back into the value chain.

At this stage, the data is used and its value is exploited for some purpose, either as direct consumption or by creating a new solution for the reuser or a wider audience. Depending on the purpose of the reuse case and its effect on society, the impact of open data can be realised through the reuse case.

2.3. Understanding the concepts of output, outcome and impact

The value of data can be conceptualised in two ways. One way that data becomes more valuable as it progresses along the value chain is that its **intrinsic value** ([17]) **increases**. For example, machine-readable data accompanied by high-quality metadata is intrinsically more valuable than data without these features. Another way that data becomes more valuable throughout the value chain is through an **increased likelihood of being reused**. For example, data that is machine-readable or is accompanied by high-quality metadata is also more likely to be taken up by reusers. But, for example, promotional activities that increase awareness about a dataset also add value by making the data more likely to be reused, even though the intrinsic value of the data is unchanged.

The concepts of output, outcome and impact are typically used in assessment frameworks that monitor (policy) interventions. An **intervention** ([18]) consists of actions leading to change (for example, activities aimed to improve efficiency in government using open data). **Inputs** are the resources and materials required to implement an action. The inputs can be financial, human, material or (for policy interventions) political decision-making (for example, training and funding to develop a data-driven application to track government spending). This action can lead to the following changes.

- **Outputs** are the direct results of an intervention (for example, a dashboard that integrates data on the built environment at the municipal level into one overview).
- **Outcomes** are an intervention's short- and medium-term effects (for example, improved information for monitoring and maintenance).
- **Impact** is a long-term or broad effect on the intervention (for example, higher return on public money and increased inhabitant satisfaction).

([17]) The intrinsic value could be conceptualised as the cost of having to reproduce the actions applied to a dataset. For example, if data is transformed into a machine-readable format early in the value chain, all reusers can benefit from this action. However, if the data is not made available in a machine-readable format at the point of publishing, each reuser may need to perform this transformation themselves, which takes additional effort.

([18]) https://www.oecd.org/dac/results-development/what-are-results.htm.

3. Measuring the output, outcome and impact of open data

Previous studies have attempted to measure the output, outcome and impact of open data and public sector information, each using a different methodology and indicators.

3.1. Overview of impact assessment studies

The first serious attempt to estimate the economic impact of public sector data was made by Pira ([19]). Using a demand-side estimate through user surveys, the researchers developed a value of public sector information data per user. A total economic impact estimate was developed in combination with an extrapolation to estimate the number of users. Later studies, such as Dekkers et al. (2006) ([20]) and Publications Office of the European Union et al. (2022) ([21]), followed this approach of survey-based demand-side value estimates.

The Open Data Institute used a slightly modified approach ([22]). Based on a model developed by Pollock ([23]) ([24]), the study estimates the gross value added (GVA) of an open-access regime by adjusting the GVA of a cost-recovery regime with various factors, such as ratio of transaction costs and increase in demand.

Another approach for estimating the economic impact of open data relies on a macroeconomic approach. In a European Data Portal ([25]) study, the total open data market size was estimated by averaging the estimated share of gross domestic product (GDP) found in previous studies and applying this percentage to the GPD value at the time of reporting.

WISE (2014) used a bottom-up, macroeconomic model to investigate the impact of big and open data. Estimations are differentiated by country, sector, company size and time. The impact of big and open data is estimated in terms of companies working in a more data-driven way, increased competition, and efficiency improvements.

Finally, the European Data Market Monitoring Tool should be mentioned. The latest report under the tool ([26]) is developed in alignment with previous European data market studies ([27]) ([28]). It gathers data

([19]) European Commission, Directorate-General for the Information Society and Media (2000), *Commercial Exploitation of Europe's Public Sector Information – Executive summary,* Publications Office.

([20]) Dekkers, M. et al. (2006), *MEPSIR Measuring European Public Sector Information Resources – Final report of study on exploitation of public sector information – Benchmarking of EU framework conditions*.

([21]) Publications Office of the European Union, Assen, M., Cecconi, G., Carsaniga, G. et al. (2022), *Open Data Maturity Report 2022*, Publications Office, December 2022, https://data.europa.eu/sites/default/files/data.europa.eu_landscaping_insight_report_n8_2022_1_1.pdf.

([22]) Lateral Economics (2016), *Permission Granted: The economic value of data assets under alternative policy regimes – A Lateral Economics report for the Open Data Institute*, March 2016, https://www.scribd.com/doc/309810679/Permission-granted-The-economic-value-of-data-assets-under-alternative-policy-regimes#.

([23]) Pollock, R. (2010), 'The economics of public sector information', in Fitzgerald, Brian F. (ed.), *Access to Public Sector Information: Law, technology and policy*, Vol. 2, Sydney University Press, July 2010, pp. 14–47. https://ses.library.usyd.edu.au/bitstream/handle/2123/6568/PSI_vol1_chapter3.pdf?sequence=1n.

([24]) Pollock, R. (2008), 'The economics of public sector information', University of Cambridge, Faculty of Economics, 2 December 2008.

([25]) Publications Office of the European Union and Huyer, E. (2020), *The Economic Impact of Open Data – Opportunities for value creation in Europe*, https://op.europa.eu/en/publication-detail/-/publication/1021d8a7-5782-11ea-8b81-01aa75ed71a1.

([26]) IDC and the Lisbon Council (2022), *European Data Market Study 2021–2023 – D2.4 First report on facts and figures*, February 2023, https://ec.europa.eu/newsroom/dae/redirection/document/93914.

([27]) IDC and Open Evidence (2017), *European Data Market SMART 2013/0063 – Final report*, 1 February 2017, https://www.key4biz.it/wp-content/uploads/2018/04/SMART20130063_Final-Report_030417_2.pdf.

([28]) European Commission, Directorate-General for Communications Networks, Content and Technology, Cattaneo, G., Micheletti, G., Glennon, M. et al. (2020), *The European Data Market Monitoring Tool – Key facts & figures, first policy conclusions, data landscape and quantified stories – D2.9 final study report*, https://op.europa.eu/en/publication-detail/-/publication/9fb0599f-c18f-11ea-b3a4-01aa75ed71a1/language-en.

about data professionals, supplier companies, data users and the data market by surveying companies in the ICT and professional services industries and data users in 11 industries. Besides estimating the economic value of the European data economy between 2019–2021, the International Data Corporation (IDC) and the Lisbon Council (2022) also provide forecasts for the years leading up to 2030 based on scenarios that rely on macroeconomic data, along with assumptions on economic growth conditions, policy and regulatory conditions, data market dynamics and global trends affecting all technology markets.

Studies assessing the impact of open data can be summarised in the following three groups.

1. **Quantitative methods studies**, i.e. studies focusing on an economic estimation of the present impact or the forecast value of open data. Sub-categories of these studies include:
    a. **micro- and macroeconomic estimations**, for example, 'Open growth – Stimulating demand for open data in the UK' ([29]); *Permission Granted: The economic value of data assets under alternative policy regimes*; *Independent evaluation of the OFT's 2006 market study into the Commercial Use of Public Information (CUPI)* ([30]); *Open for Business: How open data can help achieve the G20 growth target* ([31]); *Big & Open Data in Europe – A growth engine or a missed opportunity?* ([32]); *Commercial Exploitation of Europe's Public Sector Information – Executive summary*;
    b. **surveys**, for example, *Open Data Maturity Report 2022*; *The Data Economy in the Infomediary Field* ([33]); *Open Data Barometer* ([34]); 'OECD open, useful and reusable data (OURdata) index: 2019' ([35]); 'Review of recent studies on public sector information re-use and related market developments' ([36]);
    c. **a mix**, for example, *Conducting an open data impact study in Cyprus and developing its measurement methodology – Report on the impact of open data in Cyprus*' ([37]); *The Economic Impact of Open Data – Opportunities for value creation in Europe*; *MEPSIR Measuring European Public Sector Information Resources*.

2. **Mixed methods studies**, namely reports using economic estimations through statistics, surveys, and case studies based on interviews, focus groups, and desk research. In this category there are, for instance, European data market studies; *The Economic Impact of Open Data – Opportunities*

([29]) Deloitte (2012), 'Open growth – Stimulating demand for open data in the UK', Deloitte analytics, https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-analytics/open-growth.pdf.

([30]) DotEcon (2015), *Independent evaluation of the OFT's 2006 market study into the commercial use of public information (CUPI) – A report by DotEcon for the CMA March 2015*, 26 March 2015 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/418097/Evaluation_of_CUPI_study.pdf.

([31]) Lateral Economics and Omidyar Network (2014), *Open for Business: How open data can help achieve the G20 growth target – A Lateral Economics report commissioned by Omidyar Network*, June 2014, https://www.academia.edu/35698222/Open_for_Business_How_Open_Data_Can_Help_Achieve_the_G20_Growth_Target.

([32]) DemosEurope and the Warsaw Institute for Economic Studies (WISE) (2014), *Big & Open Data in Europe – A growth engine or a missed opportunity?*, http://www.wise-institute.org.pl/common/files_download.php?fid=4.

([33]) Asedie (2022), *The Data Economy in the Infomediary Field – 10th edition*, https://www.asedie.es/s/Asedie-Report-10-Edition.pdf.

([34]) World Wide Web Foundation (2017), *Open Data Barometer – Global report – Fourth edition*, May 2017, https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf.

([35]) Rivera Perez, J. A., Emilsson, C. and Ubaldi, B. (2020), 'OECD open, useful and re-usable data (OURdata) index: 2019', *OECD Policy Papers on Public Governance*, No 1, March 2020, http://www.oecd.org/governance/digital-government/ourdata-index-policy-paper-2020.pdf.

([36]) Vickery, G. (2011), 'Review of recent studies on PSI re-use and related market developments', *Information Economics*, Paris, January 2011.

([37]) Deloitte (2021), *Conducting an open data impact study in Cyprus and developing its measurement methodology – Report on the impact of open data in Cyprus*, July 2021.

for value creation in Europe; *The Effect of Free Geodata – Post-measurement* ([38]); *Creating Value through Open Data* ([39]); *Open Data: Unlocking innovation and performance with liquid information* ([40]).

3. **Qualitative methods studies**, meaning research only featuring case studies based on document review, consultations, interviews and peer review, for example, *Open Data in Developing Economies* ([41]).

A significant observation is that no previous study attempted a structured quantitative analysis of open data's social or environmental impacts beyond individual case studies. All quantitative impact studies have entirely focused on the economic dimension of open data. The primary method for identifying the economic impact is developing estimates on the value of open data by users, typically via surveys.

3.2. Deep dive 1 – Mapping the open data maturity assessment to the open data value chain

One of the primary studies on open data impact that is periodically repeated is the open data maturity (ODM) assessment ([42]), conducted by data.europa.eu. In this section, we explore the method in which the ODM plots output, outcome and impact along the open data value chain in more detail.

The ODM is a framework to measure European countries' progress in making public data resources available for reuse. The ODM is informed by the EU's open data policies, primarily the open data directive, which serves as a stimulus of the open data value chain and creates the legal basis for the activities of the open data team. The ODM considers maturity in terms of the following four dimensions.

- **Policy** concerns the national governance models and the measures applied to implement national open data policies and strategies.
- **Portal** concerns the features and functions of national open data portals.
- **Quality** concerns the mechanisms that ensure the quality of the metadata on national portals.
- **Impact** concerns activities to monitor and measure open data reuse and its impact.

These dimensions are evaluated via a questionnaire. The national open data teams of the countries answer the questionnaire. The questions relate to the performance of the countries in supporting the uptake and reuse of open data and can be mapped to the open data value chain (Table 1).

([38]) PwC (2017), Effekten af de frie geodata – Eftermåling – Stryrelsen for Dataforsyning og Effektivisering, 17 March 2017, https://sdfe.dk/media/2916777/de-frie-geodata-eftermaaling.pdf.

([39]) European Commission, Directorate-General for the Information Society and Media, Carrara, W., Fischer, S., Steenbergen, E. et al. (2015), *Creating Value through Open Data: Study on the impact of re-use of public data resources*, Publications Office, November 2015, https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

([40]) McKinsey Global Institute (2013), *Open Data: Unlocking innovation and performance with liquid information*, 1 October 2013, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information.

([41]) Verhulst, S. and Young, A. (2017), *Open Data in Developing Economies: Toward building an evidence base on what works and how*, African Minds, Cape Town, 13 November 2017, https://ssrn.com/abstract=3141435.

([42]) https://data.europa.eu/en/publications/open-data-maturity.

Table 1: The extent to which the questions of the ODM map to activities in the open data value chain

		Stages of the open data value chain			
		Collection	**Publication**	**Uptake**	**Application**
Dimensions of ODM	**Policy**	High	Medium	Low	Low
	Portal	N/A	High	High	Low
	Quality	Medium	High	Low	N/A
	Impact	N/A	N/A	Low	High

A. Open data collection

From a policy perspective, the open data directive encourages EU Member States to make as much information as possible that is collected from public sector bodies and publicly funded organisations available for reuse. This serves as a stimulus for the open data value chain in the EU. It stimulates the data-opening process and guides the pre-processing of datasets through specifications. For example, the implementing regulation on high-value datasets ([43]) identifies categories of existing government data that must be made available for reuse. This implementing act also specifies technical requirements for high-value datasets. For example, these datasets must be made available in machine-readable format via an API and, where relevant, as bulk downloads. More generally, the open data directive also encourages technical standards. For example, it requires national governments to put measures in place that stimulate the publishing of dynamic data and the uptake of APIs. The Inspire directive ([44]) ([45]) for spatial information is another example of EU policy that defines metadata standards.

The ODM assesses the data collection stage of the open data value chain in the policy and quality dimensions. The policy dimension assesses whether national policies, strategies and action plans are in place to stimulate the open data value chain and if these have associated monitoring mechanisms. The questionnaire asks whether there is a policy to maintain a data inventory by public sector bodies, which is important to understand what data exists and can potentially be made open. Furthermore, it assesses whether governance structures are in place to support the data opening process and whether publication plans exist at the public body level. It also evaluates whether measures are in place to incentivise the publication of dynamic and geospatial data (the pre-processing of these datasets falls under the collection stage).

The quality dimension of the ODM assesses whether the open data team provides guidelines and tools to assist publishers in choosing an appropriate licence for their data. The questionnaire also asks if guidelines and tools are provided for data providers to improve the quality of their data publication. It also asks if regular activities are conducted to incentivise data providers to publish data in machine-readable formats and high-quality metadata.

([43]) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2023.019.01.0043.01.ENG.
([44]) https://inspire.ec.europa.eu/inspire-directive/2.
([45]) https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002.

B. Open data publication

EU open data policy guides the preparation of public data resources for publication, such as specifications on metadata and interoperability requirements. Data for dissemination must implement the relevant guidelines and standards to be accessible to reusers.

The ODM evaluates data publication in the questionnaire's policy, portal and quality dimensions. The policy dimension assesses whether a national strategy is in place to foster the discoverability of data on data.europa.eu. The questionnaire also asks whether there are activities to assist data providers with their open data publication process, especially for dynamic and geospatial data.

The portal dimension assesses whether there is a national portal for making open data discoverable and whether the portal offers the ability to make programmatic queries (via APIs or SPARQL ([46]) access points). It also investigates the degree to which public sector data providers contribute to the portal and the concrete actions taken to assist them in their publication process. It asks whether local data sources are made discoverable on the national portal and if measures are taken to optimise the search and discoverability of content. Regarding dissemination activities, the portal dimension includes questions about whether the national strategy includes a description of the portal's target audience, whether the national portal is active on social media and whether actions are taken to promote the national portal's activities and the available open data. The questionnaire asks if users can find information and news on open data topics in the country and receive notifications about new datasets. It also asks whether high-value datasets are promoted on the portal.

The quality dimension assesses whether there is a predefined approach to ensuring that metadata is kept up to date, whether the quality of the metadata on the national portal is monitored and whether the quality of deployment of data is assessed. The questionnaire requests self-reported values about the proportion of metadata that is harvested automatically and the average delay for updates between the source and the national portal. It also asks about the recommended licensing suite and the proportion of open datasets on the national portal that are accompanied by licensing information. The quality dimension also investigates DCAT-AP ([47]) compliance and the proportion of metadata on the national portal that uses mandatory, recommended and optional classes of the DCAT-AP specification. It assesses the proportion of datasets published in machine-readable formats and the proportion that use uniform resource identifiers consistently.

C. Open data uptake

To incentivise the uptake of open data, the open data directive limits exceptions that allow public bodies to charge reusers more than the marginal costs of disseminating the data. Furthermore, high-value datasets must be made free of charge.

The ODM evaluates activities related to fostering the uptake of open data across all its dimensions, with the portal dimension mapping most strongly to this stage of the open data value chain. The policy dimension asks whether processes are in place to assess whether public sector bodies are charging for data above marginal cost. It also evaluates whether there is a professional development training plan

([46]) https://www.w3.org/TR/sparql11-query/

([47]) https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/11.

for civil servants working with data and whether there are activities to promote open data and data literacy among the public.

The portal dimension assesses whether users can rate, give feedback and add supporting material for a given dataset, which can influence other users to take up the open data. The questionnaire asks if users can request datasets and if the status of the request is presented transparently. It also asks if datasets on the portal are linked to available examples of use cases, which can inspire other users with practical examples of how the dataset can be used. Regarding portal features, the questionnaire investigates whether the national portal offers a preview function for tabular and geospatial data. It also assesses whether the portal's traffic and API usage are monitored (e.g. the number of visitors/downloads) and if such statistics and other surveys are used to understand users' needs and improve the portal.

The quality dimension investigates the percentage of datasets linked to other credible sources to provide additional context for the users (linked data), making the datasets more usable. The impact dimension enquires whether public bodies have performed any activities to better understand reusers' needs, which can inform further activities to foster the uptake of open data.

D. Open data application

Ultimately, the purpose of stimulating the reuse of public data resources through the measures in the open data directive is to extract economic and societal benefits from open data. High-value datasets were identified explicitly for their important benefits to society, the environment and the economy. The legislation also foresees that such datasets will serve as key data sources for the development of artificial intelligence.

The ODM evaluates data application primarily through the impact dimension, but the policy and portal dimensions also contain relevant questions. The policy dimension assesses whether a national strategy is in place to support the reuse of open data by the public and private sectors. The portal dimension assesses whether there is a showcase for reuse cases and whether reusers can submit their own reuse case examples.

The impact dimension assesses whether countries have a definition of open data reuse and open data impact and whether they have a methodology in place or studies commissioned to measure open data impact. More generally, the questionnaire asks if there are activities to encourage public bodies to monitor the reuse of their own published data and if the public bodies have performed any activities to map which of and how their open data sets are reused. It evaluates whether countries have processes to monitor the reuse of their open data and if there are systematic ways of gathering and classifying reuse cases. Many questions in the impact dimension investigate the presence of reuse case examples across various impact domains.

While the ODM covers the open data value chain and its associated output, outcome and impact measurements, a rationale exists for further strengthening the approach.

3.3. Deep dive 2 – Impact assessment of the high-value datasets

The implementing regulation that established a list of high-value datasets was supported by an impact assessment ([48]). The study aimed to define concrete high-value datasets in the six thematic areas of the legislation (namely, geospatial, earth observation and environment, meteorological, statistics, companies and mobility). The study included an analysis of the micro- to macro-level impacts of the potential inclusion of given datasets as high-value datasets. It modelled the impact of policy options within economic, social and environmental dimensions. In this section, we describe the method by which the impact assessment on high-value datasets quantified the potential impact of these datasets.

The impact assessment defined a market size for data covered by the open data directive in order to demarcate the scope of the evaluation. The market size was defined as the 'products, services, and content improved or enabled by public data resources'. The study also differentiated between the direct and indirect impact of open data. The direct market size referred to 'the monetised benefits that are realised in market transactions in revenues and gross value added'. The indirect impacts considered the broader benefits of open data, including employment potential in downstream industries, new goods and services, or increased productivity and efficiency.

The economic impact was determined by a top-down analysis (Figure 3). Other studies and the European Data Monitoring Tool, which was extended as a forecast, provided a baseline for the economic value of the data economy. Then, the share attributable to high-value datasets was derived from an existing study and extrapolated. Finally, the study also considered the effect of different policy options based on expert assumptions.

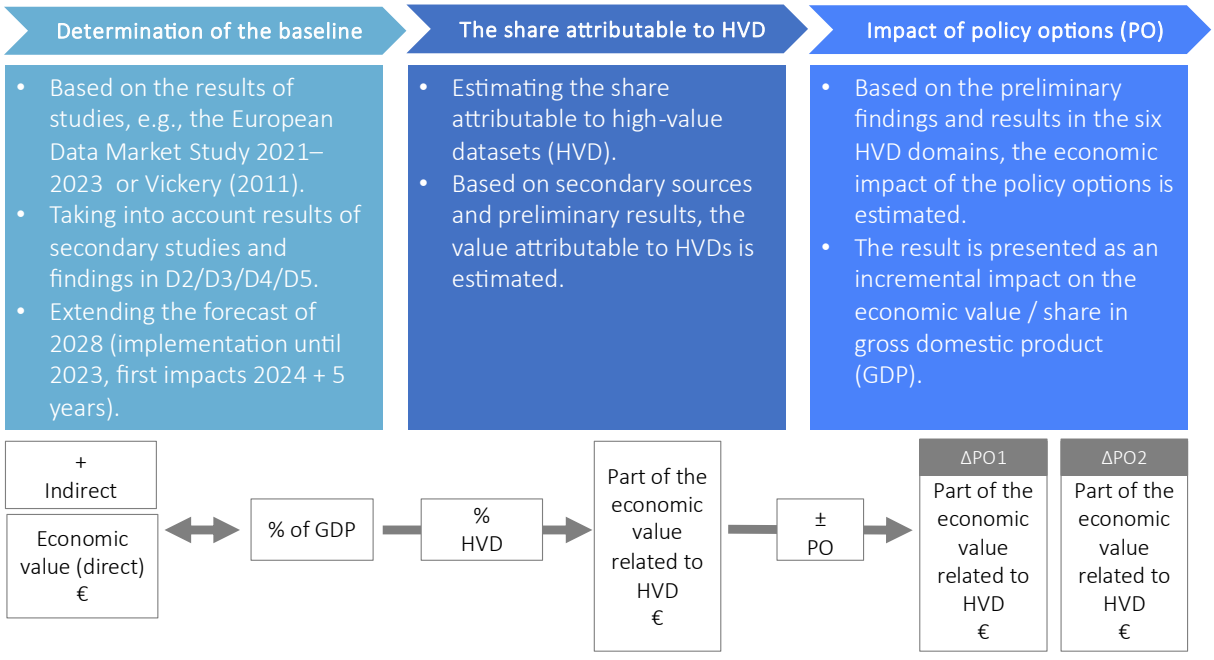

*Figure 3: Top-down approach to measure direct economic impact (*source: *Deloitte, redrawn)*

The indirect economic impacts were calculated based on the estimates of the direct impacts (Figure 4). The main indirect effect the study analysed was the indirect (forward) impact on downstream industries. The study concluded that the indirect impact on the upstream data activities of data

([48]) https://op.europa.eu/en/publication-detail/-/publication/641366a4-8d47-11ec-8c40-01aa75ed71a1/language-en/format-PDF/source-286775476.

providers was expected to be low for the data industry because data providers are usually positioned at an early stage of the value chain. On the other hand, downstream industries that provide products and services based on open data can be used in various other industry sectors. The indirect impacts of open data were then calculated by applying multipliers to the direct estimates based on the results of other studies. Additionally, the impact on employment, government revenues and the number of enterprises were calculated using coefficients in terms of GDP ratios, mainly obtained from official data provided by Eurostat.

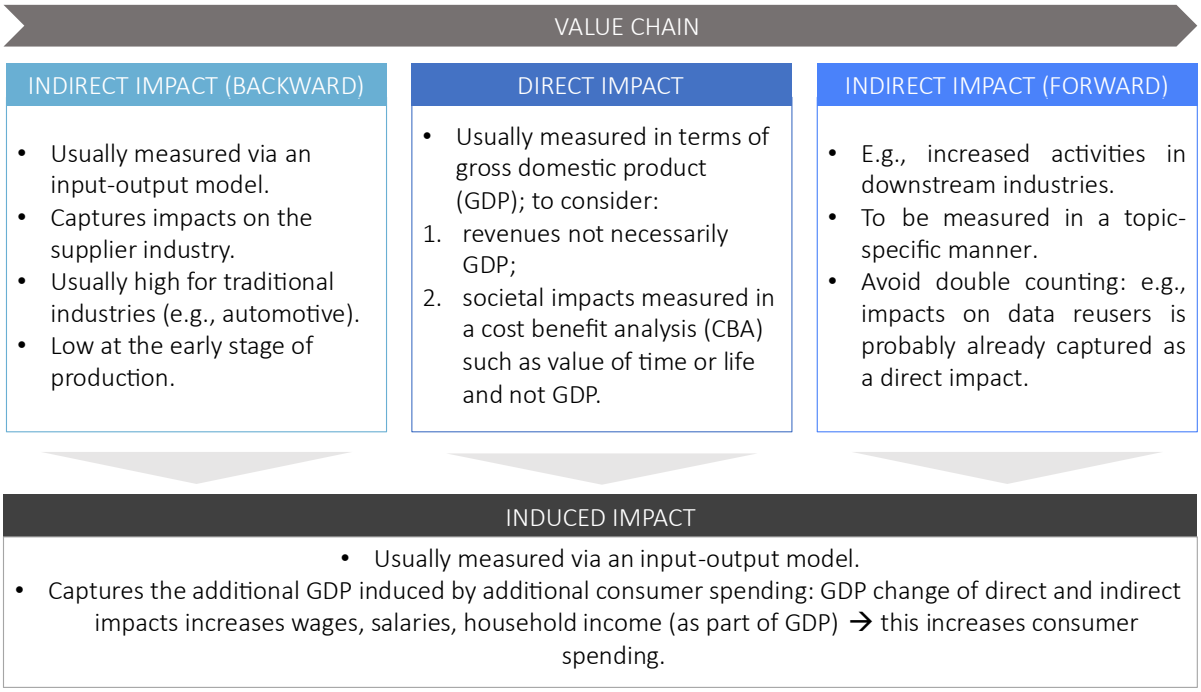

*Figure 4: Overview of direct, indirect and induced impacts (*source: *Deloitte, redrawn)*

To identify categories of data that may carry societal value, the study conducted a literature review to collate categories of value related to the six macro characteristics of potential value derived from open data, as described in the open data directive.

3.4. Overview of output, outcome and impact along the open data value chain

When open data is reused, the impact of the reuse case can be evaluated as a measure of the delivered impact of the open data value chain. This is the headline value where open data is exploited for some purpose, which may have an effect on society. However, when analysing potential output, outcome and impact along the open data value chain, it is important to realise that the phases of the value chain are consequential: some of the later activities can only arise if the preceding steps are performed. For example, data application is only possible if the data is collected and published. Therefore, all activities of the value chain preceding the step where impact is realised contribute to achieving the delivered impact at the end of the value chain. Measuring outputs and outcomes along the value chain gives an indication of how well the value chain is operating. Therefore, such measurements also provide an indication of impact in terms of the potential of the value chain to deliver possible impact.

Output along the value chain

The most straightforward effect to be identified along the value chain is the output, as these are the direct results of the activities conducted. The following output is identified across the stages of the open data value chain (Table 2). Identifying an output does not mean the indicated information is readily available. The availability of information will be discussed in the next chapter.

A. Open data collection

The collection stage revolves around having data available to be published later in the value chain. This means output in this stage is strongly focused on the datasets: pre-existing government data, data identified to be made open and the processing required for the data to be suitable for public consumption. During this stage, data is also prepared for publication. This includes consolidating potentially fragmented data into cohesive datasets and cleaning the data. The **amount and quality of data made ready for open publication** during this process can be considered output for the activities in the collection stage.

B. Open data publication

The publication stage of the open data value chain concerns the activities related to making the data available to the public and making the public aware of the data. The main output associated with this stage is, again, like in the collection stage, the datasets: the **amount and quality of data made publicly available**. In addition to the data-focused output, the available distribution channels may be considered an output. Linked to the availability of channels and data, the general public's (increased) **awareness of open data accessibility** may be regarded as an output measure. In addition, or as an alternative if information on awareness is lacking, the direct results of the dissemination activities may be used as output: the number and quality ([49]) of press releases, webinars, etc.

C. Open data uptake

Upon entering the uptake stage of the value chain, the focus shifts from supply (availability of data) to demand (use of the data). Activities are undertaken to foster this uptake, then reusers can start to consume the data. One key output of this stage is again the general public's (increased) **awareness of the potential uses of open data**, which can be seen in the direct result of influencing activities such as training and sharing examples of success stories. Another key output is the **data being accessed** by reusers to potentially develop a new reuse case. Output indicators include the amount of data downloaded.

D. Open data application

At the application stage, the uptake of **data presents as a new creation or the data is applied to a new purpose (a new reuse case)**. The data is put to actual use, exploiting the value of the open data. Output indicators include the number and type of applications (reuse cases) using the open data.

([49]) Quality should be understood as the effectiveness of achieving the objective of creating awareness among the public.

Table 2: Summary of outputs along the value chain

Open data value chain	Output
Collection	• Data identified to be made open • Data made suitable for public consumption • Data prepared for publication, including consolidation and cleaning of data • Amount and quality of data made ready for open publication
Publication	• Amount and quality of data made publicly available • Available channels for distribution • Direct results of the dissemination activities
Uptake	• Awareness among the public • Amount of data being accessed
Application	• Number and type of applications (reuse cases)

Outcome and impact along the value chain

In addition to output, outcome and impact may also be identified along the value chain. These effects may be observed closer to the value chain's final stages, where open data uptake and application result in activities and products that change people's situations and behaviour ([50]).

A. Open data collection

No outcomes and impacts are identified at this point in the open data value chain.

B. Open data publication

Although making data publicly available is an output of the publication stage, the fact that the data was suitably processed and is of high enough quality to permit publication is an outcome that is realised at the publication stage.

C. Open data uptake

Similarly, the efforts invested at the publication stage to make data discoverable (such as tagging data) manifest in the uptake stage as the ability of users to find relevant datasets. The outputs of the collection and publication stages that put the data in accessible formats manifest in the uptake stage as usable datasets for the user to consume. These are, therefore, outcomes realised at the point of uptake.

([50]) A typical exception to this situation would be negative impact originating from earlier parts of the value chain where materials and other resources are produced. In the case of open data, this effect is quite limited and considered not significant enough to warrant independent analysis.

D. Open data application

The primary outcome of the open data value chain is the use of data, typically in the form of reuse cases developed, at least partially, using open data. Whereas the existence of reuse cases (i.e. their number and type) is an output of the value chain, the purpose that the reuse case achieves is the primary outcome of the value chain. These outcomes contribute to the impact of open data.

As the range of possible use cases creating outcomes and consequentially impacts is broad, the range of possible outcomes of the open data value change is similarly broad. Moreover, as the possible reuse cases are not limited, the possible outcomes and impacts achieved by these reuse cases cannot be exhaustively presented. The next chapter investigates the outcome and impact of open data in more detail.

As for the size of the outcome and impact, two significant factors determine the size of the outcome of open data and, consequently, its impact. These are (1) **the effect of an individual open data reuse** (e.g. does a reuse case help improve administrative efficiency significantly or minimally?) and (2) **the number of reusers** (e.g. is the solution used by a few persons or organisations or by many?). Both factors return in our discussion in the next chapter on the current measurements of outcome and impact and the opportunities to improve them.

As described above, various possible outcomes and impacts have been identified in the literature. Table 3 provides a non-exhaustive overview of possible outcomes and impacts, categorised by impact domain.

Table 3: Examples of outcome and impact in the application phase of the open data value chain

Impact domain	Outcome	Impact

Economic	• Improved productivity
	• Increased innovation
	• Better functioning of markets:
	  ○ reduced transaction costs
		• Increased economic output:
		  ○ GDP
		  ○ investments
		  ○ employment

Social	• Greater social awareness, collaboration and participation
	• More effective and efficient policy interventions:
	  ○ prediction and prevention/reduction of the need for interventions
	• Improved public sector functioning:
	  ○ improved access to public services
	  ○ improved public sector efficiency
	  ○ Increased transparency
		• Improved social services
		• Improved decision-making by policymakers
		• Increased public trust in institutions

Environmental	• Energy management and efficiency • Improved build environment:	• Improving environmental management:

	○ smart cities with improved quality of living • Mobility transition: ○ substitution of less impactful forms of transport ○ reduced traffic due to better planning and monitoring • Citizen engagement in addressing climate change	○ reduced environmental emissions ○ promotion of sustainable development

4. Indicators for measuring output, outcome and impact

Identifying potential impact assessment indicators means finding indicators that can simulate the results of impact assessment studies to a certain extent. Examples of indicators for output along the value chain are presented in Section 4.1, and examples of indicators for outcome and impact are presented in Section 4.2.

4.1. List of output assessment indicators

There are two primary sources of indicators: automated indicators and survey indicators. The literature review conducted in the first report of this series, the *Rethinking the Impact of Open Data* study, summarised that automated indicators can be organised into four main groups, all strongly focused on the data itself and hence closely related to open data output:

1. basic metrics around portal usage and interaction with datasets;
2. metrics for open data APIs;
3. metrics associated with the use of open data in scientific contexts; and
4. metrics associated with mentions of datasets in social networks and other external sources.

In addition to the automated indicators, surveys are another source of potential indicators for an open data impact methodology. As this data is collected in a less automated way, a methodology based on survey indicators may be more labour-intensive but still interesting to explore.

Table *4* summarises examples of indicators that measure output at each stage of the open data value chain. For each output indicator, a description is provided with an indication of how it may relate to assessing the overall impact of open data. Most of the output indicators in

Table *4* are based on previous studies or reports, the references of which are included in the last column. Two additional columns are included: 'difficulty to measure' and 'extent of insights'.

'Difficulty to measure' refers to the expected difficulty of collecting the necessary data for the indicator. Using three categories, a 'low' level of difficulty means data is relatively available, for example, in the form of public data that requires some processing. A 'high' level of difficulty means

that significant efforts would need to be made to obtain the necessary information for that indicator. The 'medium' classification falls between the other two categories.

'Extent of insights' refers to the indicator's expected contribution to the impact of open data. A similar three-level classification as for 'difficulty to measure' is used. A 'low' level of insight means the indicator is expected to provide a poor estimate of the total impact achieved. In contrast, a 'high' level of insight means the indicator is expected to give a fair reflection of the overall impact achieved.

The classification of indicators was based on expert judgement by the project team and forms the foundation for a short list of the most promising indicators to be defined and tested in upcoming studies. Indicators that offer a low level of difficulty to measure and a high extent of insight should be considered the most promising indicators for future use. The assessment of measurability and insightfulness of the indicators should also be validated in the upcoming studies.

Table 4: Indicators for open data output

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
Open data collection				
Availability and extent of published data inventories	A data inventory lists the types of data a public sector body gathers. Data inventories support data publication by cataloguing available data (i.e. they are a proxy for the potential number of publications). Data inventories that are more extensive, such as those covering data across various subject areas or that include more data suppliers, are more likely to address the reuse cases imagined by reusers. It can also be informative to know the number of public organisations in a specific area that are obliged to publish their data. Such publication plans should give a more realistic approximation of how much open data can be published yearly. Data for this indicator could be collected through surveys or an analysis of publicly accessible inventories.	Low	Low	ODM ([51])
Number of data suppliers	The number of data suppliers may reflect the topic coverage of data made available, which in turn is more likely to address the needs of reusers. Collected through surveys, the total number of companies supplying data by specific size and location can be calculated. This indicator can be used as a proxy for measuring the percentage of organisations generating data internally and the share of those publishing open data. More data	Medium	Low	IDC and the Lisbon Council (2022) and Asedie (2022)

(51) See footnote 22, p. 12.

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
	being made available for reuse should result in more data reuse and reuse cases.			
The proportion of public-sector bodies with civil servants made responsible for open data	Having clear processes and responsible persons for making data openly available within public sector bodies makes it more likely that collected data will be processed and published as open data. Data for this indicator could be collected through surveys or an analysis of the governance structures of open data teams published in government documents.	Medium	Low	ODM
Open data publication				
The proportion of eligible datasets (from the public data inventory) published (metric of scope)	This indicator represents the number of datasets published as a total of all datasets that could be published. Publishing a greater volume of data increases the opportunity for reuse. A proportion is more sensible than an absolute value since different public sector bodies collect different types of data and numbers of datasets. Some datasets cannot be made open; therefore, only eligible datasets from the data inventory should be considered. A relative measure is also important when comparing countries of different sizes.	Low	Low	
Completeness of datasets denoted as high-value datasets (coverage of high-value dataset categories and sub-lists)	This indicator represents the number of datasets published as a total of all datasets that could be published. Publishing a greater volume of data increases the opportunity for reuse. The implementing regulation of high-value datasets provides the scope of eligible datasets.	Low	Low	
Number of real-time datasets published (based on a defined list of expected datasets)	This indicator represents the number of datasets published as a total of all datasets that could be published. Publishing a greater volume of data increases the opportunity for reuse. The open data directive specifically mentions stimulating the publishing of dynamic data. The number of datasets must be relative to some expected or minimum list for comparability between countries.	Low	Low	
Number of geospatial datasets	This indicator represents the number of datasets published as a total of all datasets that could be published. Publishing a greater volume of data	Low	Low	

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
published (based on a defined list of expected datasets)	increases the opportunity for reuse. Geodata is covered in the implementing regulation of high-value datasets and previous directives such as Inspire ([52]). The number of datasets must be compared to some expected or minimum list for comparability between countries.			
Percentage of metadata on the national portal that is DCAT-AP compliant in terms of mandatory classes	Better described data that is of higher quality is more likely to be reused. An indicator of metadata quality based on compliance with a metadata standard could be automated. For example, data.europa.eu can analyse the metadata it harvests from its data providers.	Low	Medium	ODM, Metadata Quality Assessment (MQA) ([53])
Percentage of metadata on the national portal that uses DCAT-AP recommended classes	Better described data that is of higher quality is more likely to be reused. An indicator of metadata quality based on compliance with a metadata standard could be automated. For example, data.europa.eu can analyse the metadata it harvests from its data providers.	Low	Medium	ODM, MQA[53]
Percentage of metadata on the national portal that uses DCAT-AP optional classes	Better described data of higher quality is more likely to be reused. An indicator of metadata quality based on compliance with a metadata standard could be automated. For example, data.europa.eu can analyse the metadata it harvests from its data providers.	Low	Medium	ODM, MQA
Percentage of datasets on the national portal that consistently use uniform resource identifiers	Uniform resource identifiers make it easier for search engines and data catalogues to index and discover open data resources, facilitate proper citation and attribution and contribute to the long-term accessibility of open data. Well-catalogued data is more likely to be discovered and be easier for reusers to work with. To measure, data.europa.eu can, for example, analyse the metadata it harvests from its data providers.	Low	Low	ODM, MQA
Percentage of datasets that link to other	Linked data is structured data which is interlinked with other data so that it becomes more useful through semantic queries. A dataset with many links	Medium	Medium	ODM

([52]) https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX %3A32007L0002.
([53]) https://data.europa.eu/mqa/methodology?locale=en.

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
renowned sources to provide additional context for the user	to other sources indicates higher use and renown. However, there is no consensus on how to represent this type of information in the DCAT-AP metadata associated with the dataset, and commonly, this is represented as part of the dataset description.			
Quantification of the currency and completeness of metadata (such as based on a defined list of expected datasets that should be updated regularly or with real-time data)	Data that is up to date and time series that are complete and provide an extensive history are more useful to reusers than out-of-date and fragmented datasets. This information may be calculated automatically from the metadata. However, each dataset has a different update frequency and a different historical backlog that is sensible. For example, GDP data may be useful on a quarterly basis, but weather data may be required on an hourly time scale. Decades of historical data may be needed to make accurate weather models. Therefore, definitions of 'current' and 'complete' must be established.	Medium	Medium	ODM
Percentage of datasets with accompanying licenses	Datasets with clearly defined licensing conditions are more likely to be reused. An indicator of metadata quality could be automated. For example, data.europa.eu can analyse the metadata it harvests from its data providers. To be automated, the license needs to be in a structured format.	Low	Low	ODM, MQA
Number of datasets accessible through API/SPARQL endpoint (such as based on a defined list of expected datasets where API distribution is needed for utility) and the responsiveness of API requests (download speed)	APIs make it easier for reusers to access and integrate datasets into their applications, software, or analysis tools, reducing the barrier to entry for utilising the data. However, some datasets may not benefit from API access. Therefore, it is helpful to consider the proportion of datasets with API access where it makes sense to have this functionality. To serve users, API calls must also be processed at high speeds. APIs are queried programmatically and are, therefore, amenable to automation.	Low	Medium	ODM

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
Percentage of datasets available in machine-readable formats	Machine-readable formats facilitate interoperability and enable automated processing of data. Therefore, the format of the data is a measure of its accessibility for reuse. Publication formats may be checked automatically.	Low	Medium	ODM
Number of events, number of attendees or hours of training	Events that promote open data or teach reusers skills to use the data increase the likelihood that reusers can and will develop reuse cases with open data.	Medium	Low	
Number of editorial activities (e.g. number of news pieces) or usage statistics (such as the number of web views, number of reactions, etc.)	Editorial content and social media that promote open data increase the likelihood that open data will be used. In addition to the quantity of such content, engagement with the content provides information on its reach. However, it is more difficult to measure engagement due to a lack of public accessibility to usage statistics.	Medium	Low	
Open data uptake				
Portal traffic (e.g. in terms of the number of unique visitors, visitor profiles, percentage of machine traffic, number of downloads according to the number of datasets)	More visits and downloads reflect more data being used. Given that search terms or datasets may be categorised into topics, it may be possible to automatically identify which ones relate to impact domains. This type of data is commonly provided on many open data portals and must be generated by data publishers and intermediaries. The information on downloads and visits may be collected at the level of individual datasets or at an aggregated level. The difficulty is getting permission to access the statistics if no public dashboard exists.	Low	Medium	ODM
API usage statistics, such as the number of queries	Increased data consumption, as measured by API calls, reflects more data being used. These statistics may commonly be offered as comprehensive logs from the servers that provide the APIs, which would need to be further processed to generate the aggregations that show how many requests were made, with what frequency, for which specific type of data, etc. The difficulty is getting permission to access the statistics if no public dashboard exists.	Low	Medium	ODM

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
Query history for dataset search in open data portals	Increased data searches reflect a potential increase in interest in open data reuse. Similar to the one on API logs, this case would consist of a log of the search queries submitted to the portal.	Medium	Medium	Publications Office of the European Union et al. (2020) ([54])
Usage of standard reference or authority tables in a dataset (e.g. EU vocabularies authority tables)	These tables usually facilitate data integration. Their use may reflect the actual use of the data and, if so, could form an indication of the number of reuse cases. This indicator would be constructed by checking references to authority tables in the descriptions of the datasets (similarly to the case of vocabularies) or by inspecting all datasets and identifying references to values from authority tables there (which is more costly).	Low	Low	
Open data reusers	The total number of reusers should reflect total use. This indicator gives an idea of the number and share of companies using data (see, for example, studies in footnotes 267, 278 and 289). Tracking down which entities specifically use open data can be more complex than for data in general. Studies undertaken by data.europa.eu (footnotes 26 and 41) and Deloitte (footnote 38) have tried to estimate the total number of users based on survey data.	Medium	Medium	
Open data application				
Number of reuse cases (success stories) showcased on the national portal ([55])	Integration of open data into services, products or processes is a clear example of a reuse case. While not wholly representative of the total scope of reuse cases, registration of reuse cases is one of the most direct reflections of the total reuse cases that can be computed automatically. Reuse cases can be categorised in impact domains (e.g. sociopolitical, economic, environmental). Automatically attributing reuse cases to impact domains depends	Medium	High	ODM, PwC (2017)

([54]) Publications Office of the European Union, Ibáñez, L., Kacprzak, E., Koesten, L. et al. (2020), 'Characterising dataset search on the European Data Portal: Analysis of search logs', *Analytical Report*, No 18, Publications Office, 23 September 2020 https://data.europa.eu/sites/default/files/analytical_report_18-characterising_data_search_edp.pdf.

([55]) For stricter measurements, a timeframe could be specified to include only recent reuse cases or there could be a requirement that the reuse case is being actively used. Using the showcase as a source of reuse cases relies on the national portal having a systematic approach for gathering and classifying reuse cases.

Indicators	Description of measurement and its relation to impact	Difficulty to measure	Extent of insights	Reference
	on tags used by open data portals to clarify their reuse cases.			
Explicit references to datasets used	The number of citations of open datasets that reflect use, such as in references in scientific literature or other credible claims ([56]). For example, data.europa.eu assigns persistent identifiers to all datasets and has a data citation button that displays information to cite datasets. Resources like Google Datasets Search offer automation of citation counts. However, some sources, like news media, do not use citations and will not be included automatically in citation statistics.	Low	Medium	European Commission et al. (2023) ([57])
Use of open-source code that uses datasets, processing them for creating other datasets or generating added-value services.	Open-source code using open data may serve as an indication of the total amount of reuse cases. A dataset may be constructed by crawling GitHub or SoftwareHeritage repositories, looking for URLs that reference datasets from open data portals.	High	Medium	European Commission et al. (2023)

As can be observed from the table, indicators at the earlier parts of the value chain (i.e. collection and publication), where the focus lies on the supply of datasets, are typically easy to measure but also provide a low to medium level of insight. The underlying rationale is that impact is generally generated from the uptake of open data, while the supply does not necessarily reflect the uptake. Similarly, the indicators at the later phases of the value chain (i.e. uptake and application) are expected to better reflect the total impact.

Based on

Table *4*, the most promising output indicators, combining relatively low difficulty in measurement and relatively high extent of insight, are:

- number of reuse cases (success stories) showcased on the national portal;
- explicit references to datasets used;
- portal traffic; and

([56]) By credible claims we mean sources such as academic and scientific publications, use cases, mainstream media and other accredited online sources that explicitly attribute certain impacts to open data released by a country's government.

([57]) European Commission, Directorate-General for Digital Services, Corcho, O., Alobaid, A., Amador, E. (2023), *Assessment of the value of data and other types of assets in data.europa.eu*, 11 April 2023, https://data.europa.eu/sites/default/files/report/Assessment_of_the_value_of_data_and_other_types_of_assets_in_dataeuropaeu_final.pdf.

- API usage statistics, such as the number of queries.

These indicators are proposed to be further validated and tested in upcoming studies.

4.2. List of outcome and impact assessment indicators

In addition to the output indicators described in the previous section, indicators related to outcome and impact may be considered. It should be noted that the outcome and impact of open data are typically achieved in reuse cases (see also Section 2.2 on the open data value chain).

One prominent observation when exploring possible indicators for outcome and impact is that open data can be applied for a wide range of reuse cases and, consequently, may impact a wide range of economic, social and environmental domains. Even within domains, many potential impacts may be defined. In addition, an individual open dataset may be used for several reuse cases that realise outcomes and impacts in different domains. This makes it challenging to provide a comprehensive list of outcomes and impact indicators in this report. Instead, any specific study aiming to assess the impact of open data should define the domains it wants to focus on and tailor the impact indicators to the object of investigation.

Given this methodological restriction, it would be an awkward and nearly impossible task to provide a complete list of indicators for each possible aspect of the impact of open data. Instead, this section describes the main categories of impact and potential indicators that may be relevant for each category. The list of possible indicators presented in the tables below – one table for each of the three impact categories: economic, social and environmental – should be considered as inspiration for defining specific indicators for studies, but they are non-limitative.

The examples in the tables below are drafted based on insights from other studies. As open data impact assessments have primarily focused on the economic impact of open data, the inspiration for social and environmental impact indicators has been drawn from studies outside open data impact assessments.

Finally, similar columns on the difficulty to measure and extent of insights, as used for outcome indicators, are lacking here. The measurability of indicators depends on tailoring the indicators to the specific domain on which the impact assessment focuses. A single indicator, such as (impact on) market size, may be either easy to measure or difficult to measure depending on whether the sector subject to investigation has been predefined as one of the sectors in the records in the national statistical offices. Defining the extent of insight is irrelevant for indicators already focused on output and impact.

<u>Economic domain</u>

Nearly all reports that study the impact of open data focus on the economic impact. As such, previous studies have identified a wide range of possible impacts and outputs. See

Table 5 for this overview and possible indicators to quantify the impact and output.

Table 5: Indicators for economic outcome and impact of open data

Outcome/impact	Description of the outcome/impact and data sources for related indicators	Example indicators
Gross domestic product (GDP)	GDP may be considered one of the key elements in the economic domain, reflecting the total value of economic activity achieved within a region. This information may be broken down by sector, with a primary focus dedicated to sectors where open data is expected to have the most significant impact. Information on GDP and related indicators is typically readily available at national statistical offices.	• Economic output from (digital) sectors • Market size • Rate of sector growth • Gross value added
Employment	Employment is closely related to GDP, with an increase in GPD being at least partially reflected in an increase in employment. Open-data-related impact on employment may be expected in high-value jobs like data analysts, with indirect effects observed within the broader employment domain. Employment data is readily available at national statistical offices.	• Total number of jobs (number of full time equivalents [FTE]) • Number of new jobs created • Number of new high-value jobs created • Number of sectors in which jobs are created • Number of new jobs created in high-value sectors
Productivity	Productivity is the other form in which an increase in GDP may be reflected. Productivity gains may be expected via efficiency improvements. Much information on productivity is available at national statistical offices.	• Total-factor productivity / multi-factor productivity • Labour productivity
Innovation	Open data offers a strong foundation for innovation, notably via the reuse cases it facilitates. With the development of new products and services (or improvement of existing ones), higher value output is created, reflecting a higher societal value. National statistical offices often collect information on innovation.	• Number of new products and services • Number of improvements to existing products and services • Number of firms using open data • Growth of firms using open data • Number of registered developers of apps using open data

In addition to the main categories listed in the table above, other, mostly output-related, categories may be considered, such as ‘reduced transaction costs’ and ‘better functioning of markets’. While

these categories do contribute to the realisation of the main impacts listed in the table above, operationalising these categories is challenging due to the lack of clear indicators, especially those attributable to open data. Nonetheless, these categories may be explored regarding their viability for an impact assessment focused on the economic impact of open data.

Social domain

There is limited research into the social impact of open data. Based on the limited studies conducted, the following categories of social impact and example indicators may be developed (see Table 6) ([58]).

Table 6: Indicators for social outcome and impact of open data

Outcome/impact	Description of the outcome/impact and data sources for related indicators	Example indicators
Improved public sector functioning, including: • improved access to public and social services • improved public sector efficiency	Similar to how open data supports commercial innovation (see Table 5), open data may be used by (suppliers to) governments to develop products and services that improve government social services. While there is no comprehensive overview of the quality levels of social services provided, there are periodic monitors for (some of the) services provided ([59]).	• Ease of access by citizens to healthcare facilities, educational facilities and other public facilities • Public sector procurement • Effectiveness of public service provision
Improved decision-making by policymakers	Improved decision-making includes more effective and efficient policy interventions and prediction and the prevention/reduction of the need for interventions. With open data more readily available, decision-makers will have a broader base of information on which to base their decisions, which should lead to improved decision-making. Assessing the quality of a decision-making process is much more complex than testing the quality of products or services. Obtaining objective measures on the quality level is expected to be difficult. As an alternative, subjective impressions on the quality of decision-making may be used.	• Perception of quality of decision-making

([58]) See, for example, Keserű, J. and Chan, J. K-S. (2015), *The Social Impact of Open Data*, May 2015, https://www.researchgate.net/publication/298646716_The_Social_Impact_of_Open_Data.

([59]) See, for example, Capgemini's annual eGovernment benchmark.

Increased public trust in institutions	Public trust is closely associated with the population's perception regarding the decisions taken by institutions. As such, it is related to the previous category on improved decision-making by policymakers. However, decision-making may focus more on operational aspects, and public trust is strongly connected to more abstract concepts like fairness and equity. Since objective measurement of such concepts is nearly impossible due to a lack of assessment criteria, using a subjective assessment may be the most viable approach. Hence, a public survey is considered the primary source of this information.	• Level of trust in institutions

As the social domain covers the full range of interaction between citizens and governments, the range of impacts realised may be correspondingly large.

Environmental domain

The environmental domain covers different aspects, such as the climate, the natural environment, available energy and raw materials, and flora and fauna. Significant challenges related to the environmental domain include the transition to green energy and sustainable mobility. Table 7 summarises examples of indicators that may be developed.

Table 7: Indicators for environmental outcome and impact of open data

Outcome/impact	**Description of the outcome/impact and data sources for related indicators**	**Example indicators**
Energy management and efficiency	Energy management concerns the planning and operation of energy production and energy consumption, while energy efficiency relates to the development of energy usage. For both, sufficient data should be available at statistical offices or in publications of regulators in the energy sector.	• Levels of energy use • Frequency and size of energy market mismatches
Reduced environmental emissions	With the current attention to global warming and the role of greenhouse gasses, such as CO_2 and nitrogen oxides (NO_x), reducing emissions of these gasses is considered highly impactful to reduce the strength of climate changes. Due to the current attention, ample information about emissions is available in the public domain.	• Level of CO_2 emissions • Level of NO_x emissions

Improved built environment	In the built environment, quality of life is one of the major concerns, especially in densely populated areas. Challenges like heat stress, floods and air quality may be addressed by cities, for example, by transforming into a smart city with the application of various open data-supported reuse cases. Data in the domain of the built environment is often publicly available or (going to be) measured by cities.	• Heat stress levels • Floodings • Air quality
Environmental resilience	Linked to the challenges in (and outside) the built environment is environmental resilience, meaning how those challenges are addressed, either beforehand in the form of prevention schemes or afterwards in the form of response to incidents. Although information on these topics should be available, the accessibility of that information may be a challenge. Surveys among selected target groups, such as local and regional government, may be a valuable source of this information.	• Number of flood prevention schemes and other natural disaster prevention schemes • Responsiveness to environmental emergencies / effectiveness of disaster prevention schemes • Number of citizen-led environmental initiatives
Mobility transition	One of the major drivers behind the mobility transition is the positive environmental impact that can be achieved by altering transport modes, particularly from private to public transport. Public data on transportation is readily available.	• Substitution of less impactful forms of transport • Reduced traffic due to better planning and monitoring

Especially with the environmental challenges identified in the past decades, the range of indicators to be included in any environmental impact assessment may be subject to change.

5. Reflections and conclusion on indicators for open data impact

This section concludes with the outlook of using indicators (such as the examples provided in this study) to improve the measurement of open data impact, to be validated and tested in upcoming impact assessment studies or pilots. In addition, some reflections on the future testing of these indicators are provided.

5.1. Priority of the different automated indicators

Numerous examples of output, outcome and impact indicators are available, each offering a different extent of insight into the impact of open data and having a different difficulty in measuring.

A future impact methodology could compile a hierarchy of indicators, giving them different weights based on their relationship to impact. The aim would be to calculate a total impact score from a

hierarchy of output metrics based on the scale of importance of each indicator. Such a scheme must first be established. Indicators that are easier to measure and offer the most insight into impact are the most ideal for a future impact methodology.

When selecting indicators, a distinction can and should be made between output indicators on the one hand and outcome and impact indicators on the other hand. Output indicators are closest connected to the value chain (described in Chapter 2), but the link with impact is less direct. For outcome and impact indicators, the situation is precisely the opposite. The link with impact is strong, but the contribution of the open data to the selected indicators becomes more difficult to measure.

Of the identified **output indicators**, the indicators linked to open data reuse or reuse cases are the strongest proxy for assessing the added value provided by open data reusers. The next best indicators are those that reflect actual data reuse, for example, the types and amount of data downloaded and the number of downloaders. This actual data reuse might reflect the actual number of applications developed with the open data. The subsequent best indicators are those related to the need for open data, such as search requests, which estimate the potential demand for open data. Finally, supply-side indicators, such as the quality of metadata and the quantity of datasets made available, are weaker predictors for demand-side developments. Among the most promising output indicators in terms of data availability and expected weight in explaining impact are:

- number of reuse cases (success stories) showcased on the national portal;
- explicit references to datasets used;
- portal traffic; and
- API usage statistics, such as the number of queries.

Several of these and other output indicators can be automated. For example, the number of datasets released annually could be checked automatically. While the number of released datasets may have a low influence on impact, it is still an important output. Measuring how many high-value datasets are published is even more critical. Furthermore, if published datasets have good quality metadata, are linked to the vocabularies, have APIs, open licenses, CC BY 4.0 or alike, they have a chance to create a more significant impact. Automated measurements regarding metadata quality could be related to data.europa.eu's current metadata quality assessment methodology (see footnote 55).

Outcome and impact indicators are wide-ranging and should be defined based on the specific needs of the organisation conducting the assessment. To operationalise this, a future impact assessment methodology should start by defining which outcomes and impacts are of interest to the organisation requesting the assessment. For example, an EU-wide methodology from the perspective of the Commission should be flexible enough to meet the needs of multiple Directorates-General, which monitor several Commission priorities and targets. Detailed indicators can be developed once the priorities and targets have been set. Chapter 4 contains example indicators for three domains: economic, social and environmental.

Once the complete list of indicators for a specific impact assessment has been developed, a further review of the data sources and opportunities for the automation of data retrieval and analysis can be explored as a first step toward creating an automated monitoring dashboard. For many of the indicators, public data sources have been identified. However, non-public data sources are considered necessary for some indicators, most notably surveying a target population.

For the (automated) analysis of impact, it should be noted that causality between open data and impact needs to be established. The effect generated by open data should be isolated from other exogenous effects.

5.2. Privacy considerations

Most of the indicators identified above would be collected and evaluated using cookies or comparable techniques. In this context of monitoring and evaluating the access and use of open datasets, the central question will be whether the use of the indicators referenced above will satisfy the requirements of European privacy and data protection law, mainly Regulation (EU) 2016/679 ([60]), and Directive 2002/58/EC (as amended) ([61]). For example, indicators can be deployed and collected but may require the active consent of users. It is worth underlining that all indicators could conceptually be used if prior active consent is sought. In addition, there are low-risk cases where prior consent can be skipped in the case of first-party cookies used for anonymous, aggregate statistics used under specific assumptions and safeguards. In general, indicators selected in a future impact methodology would need to be assessed for their privacy impacts. Some key privacy impacts of using automated indicators are summarised in Table 8.

Table 8: Privacy impact assessment of automatic indicators

Indicator	Privacy impact assessment
Statistics of consumption of API calls	**Low impact**. Can be measured with first-party cookies; no substantial risks.
Number of downloads and visits for each dataset	**Low impact**. Can be measured with first-party cookies; no substantial risks.
Aggregated number of downloads and visits for each dataset on the different portals	**Medium impact**. Requires cross-domain linking. Can be done without, however, using tracking technologies or re-linking findings to individual users.
Query history for dataset search in open data portals	**Medium impact**. Arguably, it no longer falls under statistical/aggregate analysis but qualifies as user profiling, requiring consent. The concern is removed if only statistical and fully anonymised information on searches is retained – i.e. no links to an individual user, only aggregate search statistics across **all** users.
Explicit links among datasets and with existing vocabularies	**Low impact**. Can be measured without any reference to users – tracks datasets, not users.
Availability of datasets in other general-purpose or community-specific services	**High impact** if implemented through user monitoring. **Low impact** if implemented merely by tracking datasets themselves rather than users.
Explicit links among datasets and with existing vocabularies	**Low impact**. Can be measured without any reference to users – tracks datasets, not users.

([60]) See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX %3A32016R0679.
([61]) See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX %3A02002L0058-20091219.

Generation of persistent identifiers and times that a dataset is referenced in scientific literature	**High impact**. Conceptually, it would appear to track datasets rather than users (which would make it low impact). However, the description also mentions tracking its usage by individuals outside of the portal (literature, social media, conferences), thus implying extensive monitoring at the individual level that seems hard to justify from a data protection perspective.
Number of additional reuse cases for political impact, the difference in the number of reuse cases with open data publishers, and concrete impact or revenues generated by reusers of open data intermediaries	**High impact**, since this again implies extensive monitoring at the individual level outside of the portal itself, which is hard to justify from a data protection perspective.

The privacy considerations listed should be considered when drafting a specific approach for the automated monitoring of impact indicators.

5.3. Conclusion

Drawing on the insights from the previous chapters and sections, the conclusion is that various indicators for measuring the impact of open data have been identified, with many of these indicators having potential for automation. It should be noted that the options presented offer guidance and attempt to be realistic. However, this does not mean everything may be achieved immediately. In some cases, cooperation from third parties may be necessary, while constraints on time and resources may exist for other options.

Moreover, the suitability and relative importance of each indicator still need to be tested in practice. Once tested, the relative importance of each indicator can be determined, with the quantified results providing the basis for a hierarchical scheme of (automated) indicators related to impact in varying degrees. These tests, to be conducted in a future study or pilot, will also offer insight into the viability of an automated monitoring tool.

Publications Office
of the European Union

ISBN 978-92-78-44051-0