ES CO E ECONOMIC STATISTICS CENTRE OF EXCELLENCE
IARIW-ESCoE Conference
'Measuring Intangible Capitals and their Contribution to Growth'
11-12 NOVEMBER 2021, RSA HOUSE

"Estimating the Value of Data in the Netherlands"

Hugo de Bondt

(Statistics Netherlands)

Nino Mushkudiani

(Statistics Netherlands)

Paper prepared for the IARIW-ESCoE Conference
November 11-12, 2021
Session 4
Time: Friday, November 12, 2021 [9:30-11:00 GMT+1]

Session number: 4: Valuing Data
Time: Friday, November 12, Morning

IARIW-ESCoE Conference “Measuring Intangible Assets and Their Contribution to Growth”

London, UK, November 11-12, 2021

“Estimating the Value of Data in the Netherlands”

Hugo de Bondt and Nino Mushkudiani

For additional information please contact:

Name: Hugo de Bondt

Affiliation: Statistics Netherlands

Email Address: h.debondt@cbs.nl

Estimating the Value of Data in the Netherlands

Hugo de Bondt[1]

Nino Mushkudiani[2]

Abstract:

The value of data has recently been a topic of interest from both micro and macro-economic perspectives. The micro-economic perspective looks at single businesses and tries to establish the value of data (e.g. Li, Nirei and Yamana 2019) for each. The macro-economic perspective mainly focusses on the value of data by either bringing together expenditures and revenues from data (Ker, Spiezia and Weber 2019) or by aiming to estimate investments in data (Statistics Canada 2019a and 2019b). The debate on data as an asset is still ongoing. In strict SNA-terms most data does not conform to the definition of an asset and is placed outside of the asset boundary. As many already have pointed out (ISWGNA sub-group on digitalization 2020; Rassier, Kornfeld and Strassner 2019; Statistics Canada 2019a) there is a case to be made to expand the definition of assets to encompass data, thereby removing the distinction between databases that are sold (includes the value of data) and databases developed on own account. Data then can become a separate asset category, or can be put in a new intellectual property product category together with databases. The research in this paper focusses on the Netherlands and makes an estimate for business investment of data. Specifically, the model developed by Statistics Canada is used (Statistics Canada 2019) and adapted to the data sources available in the Netherlands. Separate estimates are made for each of the stages of the knowledge pyramid (data, databases and data science) as identified by Statistics Canada. As is the case in the original Canadian model, only the own-account expenditure is calculated by using labor input plus a markup for other associated expenditure. Specifically, combining and pooling together labor force survey data with tax data on wages at the personal level provides relatively stable estimates of the cost associated with the production of data assets. The professions selected in the Canadian study were mapped to Dutch (ISCO) ones, with a few alterations. To calculate total labor input, the original weights from the LFS were recalculated by replicating the methodology developed in the Statistics Netherlands paper on free services (Van Elp and Mushkudiani 2019).

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

[1] Corresponding author: Statistics Netherlands, National Accounts, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, e-mail: h.debondt@cbs.nl

[2] Statistics Netherlands, Methodology department, e-mail: n.mushkudiani@cbs.nl

1. Introduction

1.1 Theory and background of valuing data in the Netherlands

Regarding data as an asset has gained a lot of attention in recent years. The economist noted in 2017 the valuation of companies engaged in extracting value from data (Economist 2017). These, largely, big tech companies are able to reap the rewards from economies of scale, almost zero marginal costs and network effects. The media outlet therefore called data the new oil. Research from the statistical community and academics into this topic also started around this time and has continued up until the present (cf. Li et al 2019, Rassier et al. 2019). Not only researchers have taken interest in the topic, also policy makers became more interested in the topic in recent years e.g. the European data strategy (European Commission).

The application of data and it's uses are spurred by the availability of large datasets on all kinds of topics, maintained and connected to other datasets by private enterprises. What's more, companies possess the knowledge to extract useful information from these datasets within the company itself. The anecdotal evidence shows that data can have a large value. This has prompted Statistics Netherlands to conduct an experimental research into the value of data from a statistical and economic perspective. The research was conducted over the course of 2019 and 2020 and commissioned by the Dutch ministry of economic affairs. The goal of the research was to look at the value of data from a broad perspective, also addressing conceptual issues like ownership and classification. The qualitative results from the ICT surveys and in-company interviews with three companies were conducted to gain a broader understanding on the role of data in the economy. The final part of the research an estimate of the value of data was made building on the framework developed by Statistics Canada.

1.2 Data as an asset

Data has had several definitions over the years. Data is often regarded as simply information, in the words of Shapiro and Varian is anything that can be digitized (Shapiro and Varian). A more specific definition was proposed by the ISWGNA, specific for the use in national accounts, which reads: "Data is information content that is produced by collecting, recording, organizing and storing observable phenomena in a digital format, which can be accessed electronically for reference or processing. Data from which its owner(s) derive economic benefits by using it in production for at least one year is an asset." (ISWGNA 2020)

The second definition contains more information than the first one by Shapiro and Varian. Unsurprisingly, since defined by national accountants, it is suited to be implemented in the context of national accounts. Both definitions restrict data to digital data. This does not create much of a problem for recent years, although it must be kept in mind that any data that is non-digital can still hold value. The second sentence of the longer definition is most important for national accounts and this research. The definition complies with the asset definition of national accounts. What's more, not only is it an asset, but also falls within the production boundary according to the first sentence. In practice this means (part of) data production is seen as gross fixed capital formation (investments).

The national accounts definition is not the only definition out there. Business accounting also specifies certain rules under which data can appear as an asset on the accounts. However, the rules are generally such that in practice it would rarely appear on the balance sheet of businesses. Data would most likely fall under the broader category intangible assets. Two conditions have to be met if the asset is to appear on the balance sheet (IFRS rules). Firstly it should be plausible economic benefits accrue to the owner of the asset and secondly the costs associated with the asset have to reliably determined. The first condition is almost identical to the national accounts requirements. The second one however is most likely the reason data do not appear on the balance sheet. Businesses do not record data assets, since they have no reliable way of estimating the cost. Moreover, this requirement also restricts the valuation of internally developed intangible assets. For example internally developed customer databases cannot be valued because they are indistinguishable from other costs made by the business.

It does not mean at least some part of data assets are not valued. Through merger and acquisition activities, businesses (re)value their goodwill, which can theoretically, capture unvalued data assets. In the experience of Statistics Netherlands, those valuations only occur when M&A activities take place, making this an unreliable or at best anecdotal source for valuing data assets. Moreover, it remains unknown which parts of the goodwill make up data assets.

Both definitions do not cover large or small datasets. Most datasets that are thought to be assets, contain large volumes of information. But any data can have value, even small ones. Under the asset definition small datasets, or even single data cells can be included, but in general large datasets are needed to make use of modern data analytics, e.g. machine learning / AI.

1.3 Ownership and ownership rights on data

The ownership rights of data are not always clear. For instance personal data can be owned by the giver (the person), the receiver (data-business) or the government (for statistical purposes). Many countries, including the Netherlands, limit the use of personal data for business purposes under privacy law. Restrictions to use available data also apply to non-personal data. E.g. data from appliances bought by a third party are gathered and used by the manufacturer. In those cases, each situation has to be judged separately to determine who exercises ownership rights.

In the system of national accounts, two types of owners are recognized, the legal owner and the economic owner. They are not necessarily the same. The legal owner is determined by law, the economic owner is the unit that "accepts risks and rewards" (SNA 2008, para 2.47) and is responsible for the upkeep of the asset. Data assets, like all assets, can be bought and sold. The asset can therefore change hands. Data assets are mostly not bought and sold, but usually part of the sale of an entire business. However some businesses are engaged in producing and directly selling databases (containing data) to other businesses. In those cases the data is sold an recorded (in national accounts terms) as investment in a new asset (gross fixed capital formation).

Because data can be easily reproduced and transported via digital means, the economic owner and place where the asset resides are not immediately clear. If for instance the data asset is produced by one part of a larger company, but used by another part, both can be regarded as the owner. If both units are part of a multi-national enterprise and reside in different countries, the issue becomes even more difficult. The occurrence of payments (license payments) might provide a clue about the owner, but statistics generally lack information on the underlying asset of the license payments, if they in fact do recognize data as an asset (at the enterprise-level). Furthermore, license payments can also be regarded as an investment/asset (license payments for longer than a year) and should not be confused with transfer of the asset itself from one unit to another.

1.4 The value of data from different perspectives

The literature provides three general ways for establishing the economic value of data. One way is to look at supply and use of data related services and products, another one doing case studies on data driven businesses, and the third one regarding data as an investment asset.

1.4.1 Supply and use of data storage and services

The OECD researched the value of data from the perspective of supply and use of data services and data-carriers (data storage) (Ker et al. 2019). The goal was to find the value creation from data services and data carriers in the economy. In essence detailed tables from the US business census were used to create more detail in existing national accounts tables. This study gives an insight into the current share and impact of data services and storage on the economy. It's main advantage is the reliance on current national accounts variables such as gross production, intermediate consumption and international trade in services. By relying on the current framework and currently published figures they are already able to gauge the size of the data economy. A drawback is the need for lots of detail of products and services. The research relied on US data from business surveys that are carried out once every few years, meaning the US supply and use tables do not provide this level of detail on an annual basis. To a lesser extent they were also able to calculate supply and use using Canadian data.

1.4.2 Data from a business perspective

A second perspective is researching the value of data from a single business perspective. Li et al. explore the value of data from several case studies where they take the value of data from the sales, general, and administrative (SG&A) expense of several companies. These costs proxy for organizational capital (Li et al. 2019). Because they specifically target online platform companies, they can estimate the value of data from the organizational capital. In their study they point out the disadvantages of other approaches. The cost based approach (e.g. cost of specific staff) will likely result in underestimation. The other two, market based and income based are less useful or do not result in significant results. Their own approach is most useful for companies that rely almost completely on data to earn income. It is however less useful for more traditional businesses, where the organizational capital cannot be directly related to the value of data. Also, to provide an exhaustive overview of data, their framework is less suitable, because for each business the stock of capital has to be estimated separately. A further disadvantage is the use of company balance sheets, where they use data for the entire company, possibly ignoring residence in several countries (e.g. is Booking Holding a

Dutch or a US company?). This limits the use of their model for use in national accounts statistics and other statistics that rely of the domestic concept.

1.4.3 Data from as an asset in a national accounts framework

The third and final option is regarding data as a produced asset, which is also the concept used in the rest of this paper. Described in paragraph 1.2, produced assets are assets that are used in a production process for at least a year. In contrast to non-produced assets, they are themselves the result of a production process. They therefore appear in the supply and use tables as either gross production or imports and add to total gross fixed capital formation. Statistics Canada was one of the first to make an estimate based on this perspective (Statistics Canada 2019a and 2019b). Their contribution to the discussion consisted of setting up a national accounts framework and making an estimate based on costs. Statistics Canada asserted that data should be classified as a new asset under national accounts (together with data science). The current SNA (SNA 2008) regards databases as an asset, but not the data in it. The only possibility for data to be valued occurs when a database is sold, the transaction recorded should then comprise both the database and the data in it. The ISWGNA have put forward proposals to remedy this inconsistency. In their view all data should be regarded as an asset under current SNA 2008 rules, but leave open the possibility of splitting data into produced data and non-produced observable phenomena (ISWGNA 2020). The advantages of this perspective are the relative ease of implementation, the ability to relate the results to economic variables such as GDP and it does not require very detailed tables. It is however not without drawbacks, such as the underestimation of the value based on costs, and requires a fair amount of assumptions, such as the economic owner, non-existence of cross-border transactions (outright sale of the asset) and the unknown overlap with existing estimates of own account software, databases and R&D.

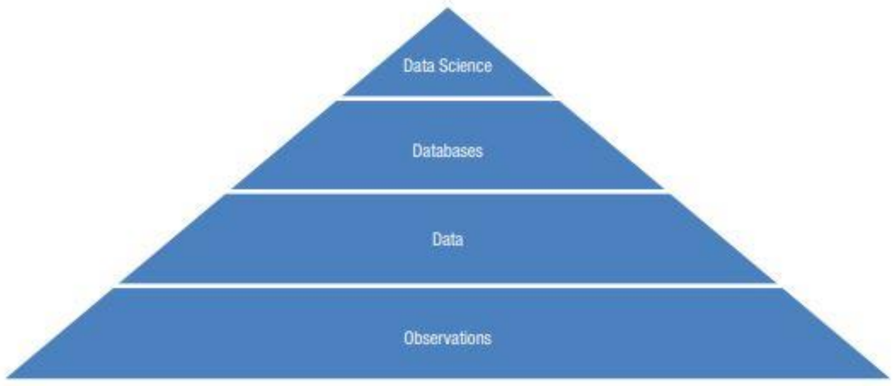
2. The application of a value of data framework

2.1 The data value chain

Statistics Canada (2019a and 2019b) propose a data value. The first tier of value chains consists of observations (or observable phenomena (ISWGNA 2020)). This tier is a stage before digitization. Examples are the outside temperature, aspects of human behavior, results of a sports match. When the observations are converted into a digital format, they have become data. As usually the case, data needs to be structured in a database in order to be used in a productive manner. The database is usually constructed alongside the data, making it harder to separate the two. Also databases are constructed using software, or software can create databases without much human interference (Nijmeijer 2018). This limits the practical separation, but nonetheless we continue along the lines of the method.

Data science is the final stage of the data value chain. This stage consists of extracting valuable insights from the data that is in the database. In this stage artificial intelligence and machine learning techniques are applied to aid the researcher. Nguyen and Paczos give this stage far reaching implications, noting the rise of new business models and complete transformation of existing businesses (Nguyen and Paczos 2019). Generally it is expected this stage of the value chain is most valuable (Li et al. 2019). The role of this stage is also to produce insights that feed into new and additional data collection, thereby creating a feedback loop into the value chain.

Figure 1 Data value chain



Source: Statistics Canada.

2.2 Produced and non-produced assets

Statistics Canada and the ISWGNA make the distinction between the non-produced observations and produced data assets. Because observations exist without human interference they are considered non produced. However in practice no value estimate is made, it does not mean they can’t hold any value. Produced assets are the result of human effort. The national accounts framework so far does not have an explicit category for the value of data (stage 2 of the value chain). Stage 3, the value of databases, is already part of the core national accounts.

For the Netherlands, databases are part of an asset class together with software. In practice it proved not possible to make separate estimates for both software and databases, however some countries do compile statistics for each separately. The final stage, data science, is most likely a type of R&D. Although a more recent phenomenon, Statistics Canada asserts that "this part of the information chain does not signify a deviation from the 2008 SNA standard"(Statistics Canada 2019a). The goal of data science is to gain knowledge on a systemic basis and use it to devise new applications, which is also the general description of R&D in the SNA 2008.

2.3 A framework for the value of data in the Netherlands

2.3.1 Restrictions and assumptions

To compile an estimate of the value of data in the Netherlands, some concepts and assumptions need to be put in place. Following the approach by Statistics Canada, only the own account production (and own account investments) are calculated. This practical limitation allows us to disregard all data originals prepared and sold on the market. Generally this does not need to create a large omission in the statistics, since it is expected businesses engaged in data activities create their own datasets and use them in-house.

A second restriction concerns government. Governments do not produce services at market prices. This means that the value of any data investment is difficult to establish, even for cost based approaches. Even if we can get a figure for cost associated with the production of data originals, the government does not necessarily want to recoup all costs, since data are used to perform the duties of government. For businesses we do assume their costs in producing a data asset will be covered by the income generated by the data asset. Cost based estimates of government investment are not unusual. Own account software and R&D are a part of current national accounts tables, and the value of these investments is usually determined by calculating costs. It means this issue can be regarded as a smaller problem when calculating the value of data. Of greater importance is the issue of ownership. Government data are not entirely owned by governments, e.g. data about its citizens is clearly owned by the citizens. Furthermore, a lot of government administered data is freely available, e.g. through NSIs. For something to qualify as an asset, economic ownership needs to be asserted. In absence of these characteristics, it is difficult to regard government data as a government asset.

In practice this has led us to exclude the economic activities government and education from this research (SIC 2008 sections O and P). It can be argued that this does not fully exclude all non-market activities, because nonprofit institutions serving households and government units outside the mentioned economic activities remain part of the population. Also, some market activities are part of the education sector, and are therefore excluded. The current data sources do not allow us to exclude/include these units. The net effect of this restriction remains unknown.

A further assumption needed to come up with an estimate is to also disregard the trade in already existing datasets, both inside the Netherlands and cross border. Trade in second hand assets inside a single country does not make a large difference to the set-up since we only distinguish total non-government activities. That would in theory omit the trade between government and market, which we expect is small. Cross border trade in second hand data originals is very difficult to establish with currently available statistics (if it exists at all). It is generally known that multinational enterprises shift their IPP assets with relative ease between jurisdictions. Creating an asset in the Netherlands and moving it somewhere else will therefore not be detected by the current setup. This assumption most likely distorts the figures more than trade within country borders, but it is not possible to make an estimate of its size. Li et al. suggest one of the reasons for acquisitions of other businesses are found in the value of data. The outright sales and purchases of data originals might be limited, this is much less the case for acquisitions of entire businesses whose main assets are data (Li et al. 2019).
To be sure, the framework does not leave out cross border data services such as advertisement services, as long as it does not involve the actual transfer of a data original across borders.
The result is the assumption that the producer of the asset is also the owner. In our case it will only lead to a bias if producers and users of the asset are located in different countries.

A final word of caution on calculating government own account production of data. As shown by De Haan and Haynes (2018) on R&D, including past depreciated R&D into the sum of costs calculation of new own account R&D will result in an ever increasing output of R&D. This mechanism could also apply to data. If we consider data a produced asset, it will inevitably depreciate (through obsolescence). If this depreciation is used as an input in creating new data (the cost of existing data, i.e. depreciation), is used as a part of the sum of cost calculation of own account output, the same mechanism will apply.

Finally the framework we adopt uses the definition of assets, but does not have an explicit boundary for short lived data (i.e. under a year) and data that is longer in use. The information that allows us to distinguish between the two is not present. It is not of great importance, given the continued use of data. Data is usually not discarded after a short period of time, but instead is part of an ever growing data set. Data has a large option value, where it is hard to predict the value changes of current data. Businesses may want to keep data stored, rather than deleting it (Mitchell et al. 2021). The PIM-model of Statistics Canada also takes into account this notion by assigning data an average service of 25 years (Statistics Canada 2019b).

2.3.2 Determining costs

Cost based calculations of own account production/investment are not unusual in national accounts, in contrast to income based and market based approaches. Own account software and R&D are quite often, if not always, calculated based on costs. mainly because in absence of other direct methods, the SNA recommends using sum of costs. In terms of available data, R&D makes use of the Frascati survey results, and software the number of IT-staff, in accordance with the OECD manual (OECD intellectual property 2009??). Software own account is based on the number of (specializes) IT-staff. Own account data resembles the method for software more closely than R&D. The framework of Statistics Canada rests on selecting a number of occupations of occupation groups that are deemed to contribute to data assets. For each of three subparts of (data, databases, and data science) a percentage of their labor input is assigned to the production of data. The lack of empirical research on these time-factors or labor-share factors are a main reason for providing high and low range values. The values used in this research are equal to the ones used by Statistics Canada and not based on empirical sources.

Table 1. Mapping of occupations from Statistics Canada to Statistics Netherlands

Statistics Canada	Statistics Netherlands		Percentage dataproduction					
occupation description	**occupation code (ISCO 08)**	**occupation description**	**data**		**databases**		**datascience**	
			min	**max**	**min**	**max**	**min**	**max**
Customer and information services supervisors	1221	Sales and marketing managers	30	50				
Data entry clerks	4132	Data entry clerks	100	100				
Other customer and information services representatives	243	Sales, marketing and public relations professionals	30	50				
Survey interviewers and statistical clerks	4312	Statistical, finance and insurance clerks	90	100				
Mathematicians, statisticians and actuaries	212	Mathematicians, actuaries and statisticians	20	30			50	60
Economists and economic policy researchers and analysts	2631	Economists	20	30			50	60
Financial and investment analysts	2412+2413	Financial and investment advisers, Financial analysts	20	30			50	60
Social policy researchers, consultants and program officers	2632	Sociologists, anthropologists and related professionals	20	30			50	60
Information systems testing technicians	3513	Computer network and systems technicians			30	50		
Database analysts and data administrators	252	Database and network professionals			90	100		
Computer and information systems managers	252	Database and network professionals						
Statistical officers and related research support occupations	3314	Statistical, mathematical and related associate professionals					90	100

Source: Statistics Canada and Statistics Netherlands

Table 1 shows the mapping of the occupations selected by Statistics Canada to the occupational classification used by Statistics Netherlands. The matching has been done by matching the description of the occupations. A matching based on codes was not possible, because Statistics Canada uses its own national occupational classification (NOC) codes instead of ISCO-codes.

All percentages are then applied to the labor costs (compensation) per year. The other costs are determined using the R&D (Frascati) survey. We expect the survey to give a reasonable estimate of the additional costs, due to the related nature of data and R&D. We found that additional costs are around 60 percent of labor costs over the most recent years of the survey, which is 10 percentage points higher than the percentage applied by Statistics Canada. Finally, we also apply a markup of 3 percent for capital.

2.3.3 Determining prices

The price changes of data are needed to calculate volume changes[3]. By definition discerning prices for assets that are not traded on a market is not possible. The second best option is to use the price changes of inputs. We have three components, labor, intermediate consumption (for other costs) and capital, and for each we assign the price index from the supply and use tables. Labor deflators are taken from the compensation of employees of SIC 62 support activities and of IT and 63 information services activities for data and databases. Deflators of labor costs for data science are based on SIC 72 Research and Development. Other costs are deflated by total intermediate consumption and capital by the deflator of total gross fixed capital formation, excluding sales of existing assets. Weighting of the components is based on shares in total costs. The aggregated deflator is then corrected for expected productivity increases of 1 percent annually. An input based deflator does not keep track of these increases, while data is expected to increase in productivity. These productivity increases should ideally be based on empirical evidence and not just by assumption. The assumed increases do however have some merit. For instance, connecting different datasets, as done by many businesses, increases the usefulness and productivity of the data itself. We consider this modest yearly productivity increase a reasonable assumption.

2.3.4 Estimating labor cost per occupation

The data used to calculate the labor costs uses two sources, first the LFS (labor force survey) and second the integral tax register on wages of employees[4]. The LFS is a rotating panel survey, with 5 separate instances where the survey is conducted throughout a year. The administrative record is called Polis. It is an administrative dataset that combined information from different sources, mainly tax data, but also the government labor agency (CWI) and the benefits agency (UWV). Polis contains information on persons, households, jobs and benefits and pensions. It covers the entire Dutch population, including persons living abroad, but working in the Netherlands or enjoying a pension or receiving benefits from a Dutch institution.

For this research we take normal hours worked plus hours worked for paid overtime. As labor compensation we take labor compensation without pension contributions.

The combination of LFS and register results based on the combination of three annual LFS results and one register. E.g. the register of 2014 is combined with LFS results from 2013,

[3] Price changes are also needed for creating a PIM model, but is kept out of scope for this paper.
[4] The method presented here draws strongly on the method developed by Van Elp and Mushkudiani (2019)

2014 and 2015. It is assumed people generally do not switch occupation that often, thereby allowing the pooling of consecutive LFS years. The combination does require a recalculation of the original LFS results. Each year of the time series is calculated the same way. Therefore only the estimation and correction of year 2014 is presented to illustrate the method. For both estimates of hours worked and compensation this method can be used, since both variables are drawn from the register.

In year 2014 there are n_{2014}= 2735 respondents (unique annual observations) of all selected occupations. Using the LFS annual weights we can estimate the total of occupations within the Netherlands in 2014, which is 316791. The annual weight represents the total amount of people for which the response in the survey has been made. The total of 2735 is then matched and joined to the Polis population, in order to estimate hours and compensation. The LFS is unfortunately based on addresses and although fairly large, will give a biased estimate of hours and compensation. To enlarge the number of annual observations, the LFS data from 2013 and 2015 are also added. We assume respondents of the LFS do not switch occupations within a year, and people from the 2013 and 2015 survey also have the same occupation as they have in 2014. The total sample then triples in size.
In 2013 we have n_{2013}= 2822 respondents with the selected occupation, in 2015 we have n_{2015}= 2812 respondents. This corresponds to a total 316720 of and 320839 respectively when using the LFS weights. N_{2014} denotes the estimate of the total of selected occupations in the Netherlands:

$$N_{2014} = \sum_{occupations} w_{i,2014}^{LFS} \qquad (1)$$

Here $w_{i,2014}^{LFS}$ denotes the annual weight of LFS respondent i. In the same vein we define N_{2013} for annual totals of 2013 and N_{2015} for 2015. If we take up the respondents from 2013 and 2015 into the 2014 sample, the weights have to be adjusted:

$$w_{i,2013}^{1} = \frac{w_{i,2013}^{LFS}}{N_{2013}+N_{2014}+N_{2015}} \cdot N_{2014} \qquad (2)$$

$$w_{i,2014}^{1} = \frac{w_{i,2014}^{LFS}}{N_{2013}+N_{2014}+N_{2015}} \cdot N_{2014} \qquad (3)$$

$$w_{i,2015}^{1} = \frac{w_{i,2015}^{LFS}}{N_{2013}+N_{2014}+N_{2015}} \cdot N_{2014} \qquad (4)$$

If we add up all new weights we arrive at a total of N_{2014} people with the selected occupations, the exact number we would like, 291433.

These unique respondents are matched and linked to the Polis register of 2014. Some could not be linked, and some double records had to be removed. A total of 7451 people then remain for which the variables of interest are available.

The next step is then to correct for the loss of observations.

$$w_{i,2013}^{2} = \frac{w_{i,2013}^{1}}{N_{2013}^{polis}} \cdot N_{2013} \qquad (5)$$

$$w_{i,2014}^{2} = \frac{w_{i,2014}^{1}}{N_{2014}^{polis}} * N_{2014} \qquad (6)$$

$$w_{i,2015}^{2} = \frac{w_{i,2015}^{1}}{N_{2015}^{polis}} * N_{2015} \qquad (7)$$

HereN_{2013}^{polis}, N_{2013}^{polis} and N_{2013}^{polis} denote the LFS respondents that can be linked to the Polis register.

The linked data makes it possible to estimate hours worked and compensation per industry. The final step is to multiply the hours or compensation with the adjusted weights. However, some respondents have more than one job, i.e. they can work in different industries. For these people the weights have to be divided.

Suppose person X worked in 2014 in both industry A and B, for which it worked H_A and H_B hours and received S_A and S_B compensation. The weight for the hours worked for industry A then becomes:

$$w_{X,2014}^{3H} = \frac{w_{X,2014}^{2}}{H_A + H_B} * H_A \qquad (8)$$

And for the compensation:

$$w_{X,2014}^{3S} = \frac{w_{X,2014}^{2}}{S_A + S_B} * S_A \qquad (9)$$

The same applies to industry B.

2.4 Classification of data

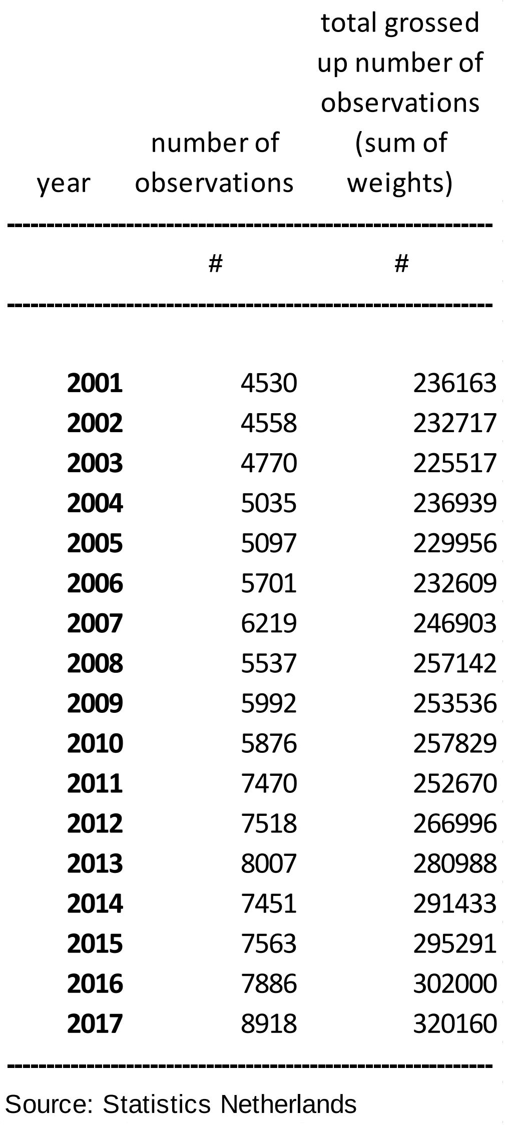
Not a lot of attention in recent literature has been given to classification of different kinds of data. Distinguishing between different types will help in creating more understanding of the data economy and also allows the construction of different price indices for each type, to be used in supply and use tables and PIM models.

A contribution to the discussion is made by Nguyen and Paczos (2019). They distinguish between types of data along the dimensions of funding of the collection process, ownership, identifiability (whom or what is it about), data source and method of collection. The dimensions do not exclude each other, i.e. a dataset can be classified according to each dimension, but some dimensions are more useful than other, depending on the asset. In the case of the research presented in this paper, ownership is of main importance. We only make an estimate on the investments of the non-government part of the economy.

2.5 Data description

The period studied ranges from 2001 up until 2017. Data for the period 2001-2005 are based on a different administrative data source than later years. This results in a series that is not entirely comparable over the years 2001-2017. The number of observations vary, but increase over the years. Of main interest is the sum of weights, suggesting between 230 thousand and 320 thousand people are engaged in data producing activities over the years 2001-2017.

Table 2. Number of observations

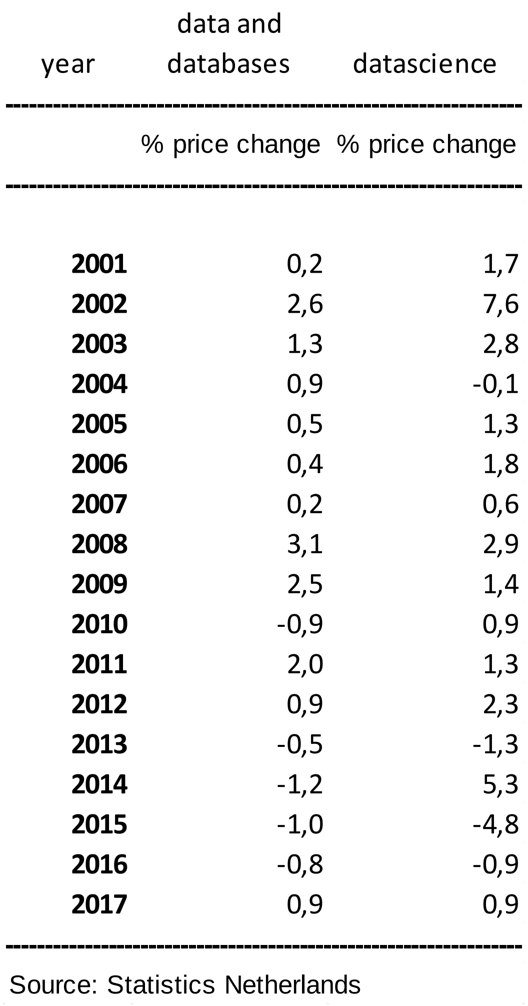


year	number of observations	total grossed up number of observations (sum of weights)
	#	#
2001	4530	236163
2002	4558	232717
2003	4770	225517
2004	5035	236939
2005	5097	229956
2006	5701	232609
2007	6219	246903
2008	5537	257142
2009	5992	253536
2010	5876	257829
2011	7470	252670
2012	7518	266996
2013	8007	280988
2014	7451	291433
2015	7563	295291
2016	7886	302000
2017	8918	320160

Source: Statistics Netherlands

Price changes are based on the method described in 2.3.3. The price changes for data and databases are different from data science, because they use SIC 62 & 63 opposed to SIC 72 in the supply and use tables for the deflation of labour costs. As a whole the prices of data science are more volatile, owing to the smaller industry size and overall volatility of the R&D industry.

Table 3. Price changes of data



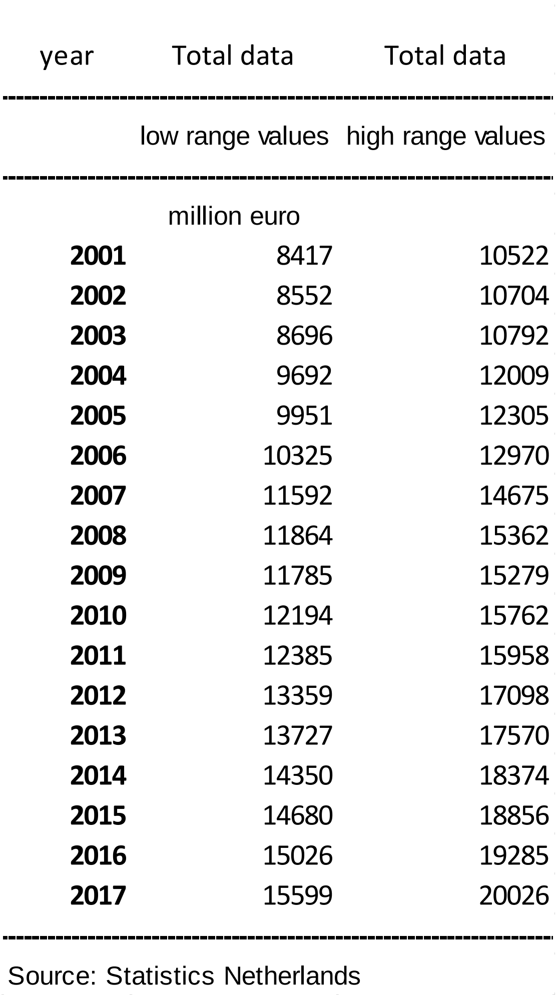
year	data and databases	datascience
	% price change	% price change
2001	0,2	1,7
2002	2,6	7,6
2003	1,3	2,8
2004	0,9	-0,1
2005	0,5	1,3
2006	0,4	1,8
2007	0,2	0,6
2008	3,1	2,9
2009	2,5	1,4
2010	-0,9	0,9
2011	2,0	1,3
2012	0,9	2,3
2013	-0,5	-1,3
2014	-1,2	5,3
2015	-1,0	-4,8
2016	-0,8	-0,9
2017	0,9	0,9

Source: Statistics Netherlands

3. Results

Due to the use of different sources, only results from 2001 – 2005 and 2006 - 2017 are entirely comparable. However, the results do not show a significant break in the series between 2005 and 2006. Still we chose to represent the volume growth for the period 2001-2005 separately from the later period. The growth is fairly constant over time for the values. Depending on the high range or low range values, production of data (and investment) grew from 8.4 billion euro in 2001 to 15.6 billion in 2017 for the low range values, and 10.5 to 20 billion euro in 2017.

Table 4. Data investment in million euro current prices 2001-2017

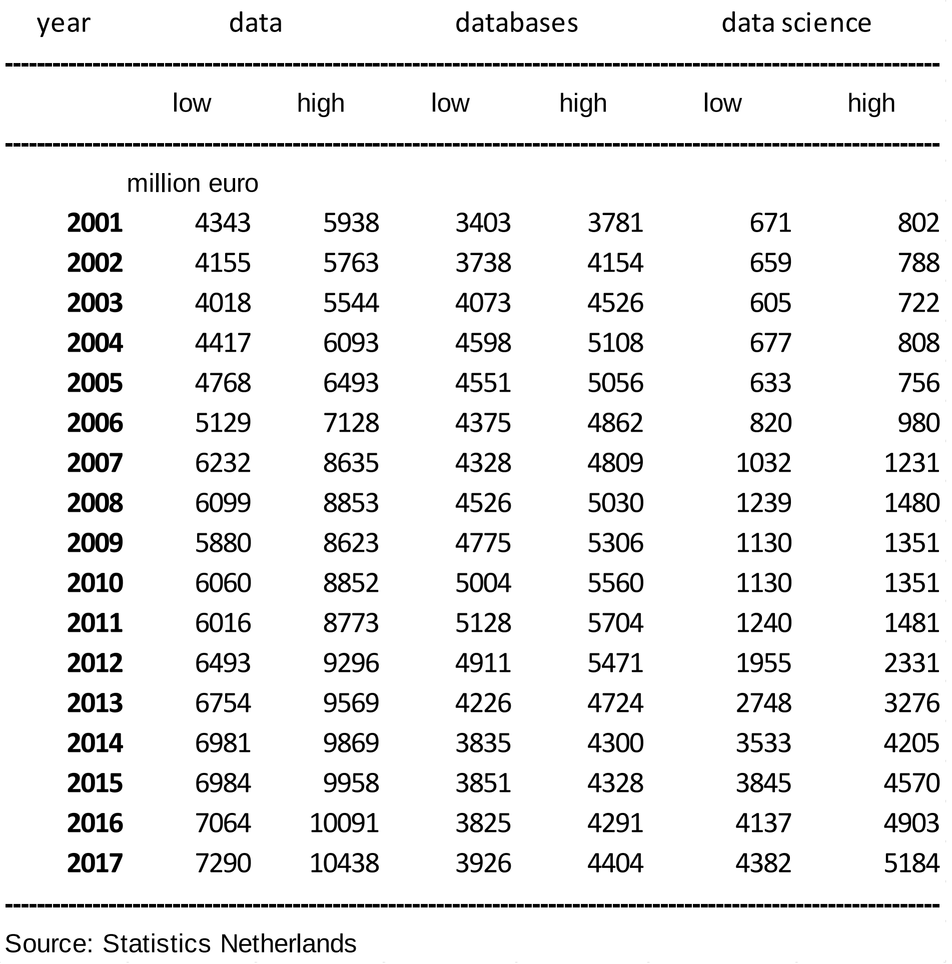


year	Total data	Total data
	low range values	high range values
	million euro	
2001	8417	10522
2002	8552	10704
2003	8696	10792
2004	9692	12009
2005	9951	12305
2006	10325	12970
2007	11592	14675
2008	11864	15362
2009	11785	15279
2010	12194	15762
2011	12385	15958
2012	13359	17098
2013	13727	17570
2014	14350	18374
2015	14680	18856
2016	15026	19285
2017	15599	20026

Source: Statistics Netherlands

The subtypes show more heterogeneous growth. Considering both the high and low range values, the strong growth of data science is striking, especially in contrast to the decline of databases investment in absolute and relative terms for a couple of years after 2012. The subtype data is growing over time, but at a much smaller rate than data science.

Table 5. Subtypes of data

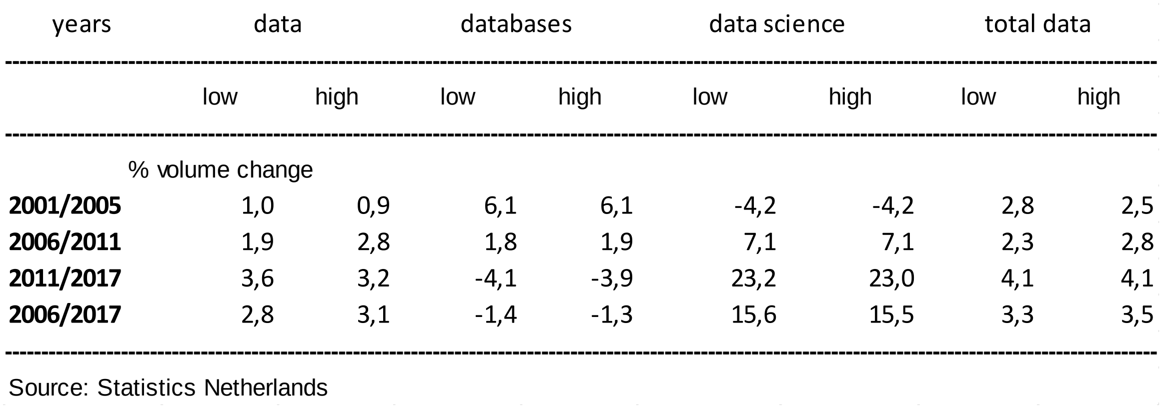


year	data		databases		data science	
	low	high	low	high	low	high
	million euro					
2001	4343	5938	3403	3781	671	802
2002	4155	5763	3738	4154	659	788
2003	4018	5544	4073	4526	605	722
2004	4417	6093	4598	5108	677	808
2005	4768	6493	4551	5056	633	756
2006	5129	7128	4375	4862	820	980
2007	6232	8635	4328	4809	1032	1231
2008	6099	8853	4526	5030	1239	1480
2009	5880	8623	4775	5306	1130	1351
2010	6060	8852	5004	5560	1130	1351
2011	6016	8773	5128	5704	1240	1481
2012	6493	9296	4911	5471	1955	2331
2013	6754	9569	4226	4724	2748	3276
2014	6981	9869	3835	4300	3533	4205
2015	6984	9958	3851	4328	3845	4570
2016	7064	10091	3825	4291	4137	4903
2017	7290	10438	3926	4404	4382	5184

Source: Statistics Netherlands

Volume growth for total data in 2001-2005 showed a growth comparable to the following periods. The contribution of subtypes however differed strongly. Consecutive periods have shown very large volume growth of data science and more moderate growth of data. Databases showed decreases in volume growth in the period 2011-2017.

Table 6 Volume growth of data, databases and datascience, average yearly growth



years	data		databases		data science		total data	
	low	high	low	high	low	high	low	high
	% volume change							
2001/2005	1,0	0,9	6,1	6,1	-4,2	-4,2	2,8	2,5
2006/2011	1,9	2,8	1,8	1,9	7,1	7,1	2,3	2,8
2011/2017	3,6	3,2	-4,1	-3,9	23,2	23,0	4,1	4,1
2006/2017	2,8	3,1	-1,4	-1,3	15,6	15,5	3,3	3,5

Source: Statistics Netherlands

4. Discussion

The results show a strong growth of data investments in the Netherlands. These findings support the general impression that investment in data have experienced strong growth in recent years, especially the subtype defined as data science. Research suggests these business activities to be most profitable (Li et al. 2019). Also, the results from the ICT survey support this finding. In the Netherlands the share of businesses carrying out analysis on their data sources increased from 19 to 27 percent between 2015 and 2019. What is more puzzling is the decline of databases. Other research by Statistics Netherlands suggests offshoring of ICT functions, including database management, to be an influencing factor (Statistics Netherland 2018). Dutch businesses located support activities both inside and outside Europe, especially eastern and central Europe and India. Between 2001 and 2006 the percentage of Dutch companies engaged in offshoring support activities was around 67 percent, between 2009 and 2011 that percentage increased to 70 percent. For the years 2014-2016 similar percentages were found as in 2009-2011. Another possibility for the decline can be the increased productivity of database management, requiring less staff to administer a database. The deflators would then require a more substantial correction for productivity increases to eliminate decreases in volume growth.

Results from Canada and the US can be compared to the results from this research. The framework in this paper is identical to the Canadian research, but differs in practical application. The main difference is the inclusion of government in the Canadian data. Overall the results are comparable, measured as average annual value growth divided. In the first period (2005/6 – 2010) growth in the Netherlands was somewhat higher, but in later periods (2010-2015) the growth in Canada was higher. The largest differences can be found in investments in databases, where Canada did not experience a decline. Data science did however grow more strongly in the Netherlands, but both countries showed a slower growth in the final period compared to earlier ones.

The BEA conducted a somewhat comparable study, where they also added up costs from data-related activities. They found a growth of 7 percent annually, which is quite a bit more than recorded in this research. The methods are however less comparable, because BEA selected different occupations and applied a different way of calculating additional costs (Rassier et al. 2019).

Relating the investments data to other national accounts variables is not without difficulties. Both Statistics Canada and BEA are cautious about calculating data investments as percentage of GDP and total investment (GFCF). The main reasons are the choice of occupations and the percentages of cost attributable to own account production. The unknown overlap with own account software and R&D, which also rely on sum-of-cost calculations makes the exercise even more difficult. For example, in the view of Statistics Canada data science is part of R&D and the explanatory texts on survey forms used in the Netherlands for the R&D survey support this view. However, when asked during interviews as part of this research, businesses were reluctant to confirm this view, shedding some doubt on the overlap of own account data science and own account R&D in practice.

Determining the share of GDP and investment (GFCF) requires therefore some assumptions. One advantage of our framework is the exclusion of government. The output of government is based on costs. Past government data investment would change consumption of fixed capital, and therefore also output. To calculate shares of GDP and shares of investment more assumptions have to be put in place, besides the ones listed in paragraph 2.3.1. The first one is overlap between the regular estimate of (software and) databases and the one presented in this research. We assume the regular estimate already captures the database investment figures presented in this paper. On the other hand, we assume no overlap between the regular R&D investment figures and data science. Investment share, measured in levels of current prices, then ranges between 8 percent for the low range values and 10 percent for the high range values. This number increases to 9.7 percent in 2017 for the low range values and more than 12 percent for the high range values. Measured as a share of GDP level the share is 1.7 percent in 2006 for the low range values and 2.2 for the high range values. In 2017 the share had grown to 2.1 percent for the low range values and 2.7 percent for the high range values.

5. Conclusion

This experimental research has shown that it is possible to compile an estimate of the value of data for the Netherlands. The framework developed by Statistics Canada proved to be a practical solution. The main strengths of this approach are using national accounts concepts, such as defining data as assets, own-account investments/production and sum-of-costs. Assigning production of the assets to several selected occupations provided a way of compiling the investment figures with data available at Statistics Netherlands. The method of pooling LFS data and merging it with register data provided the necessary data for calculating the sum-of-cost by occupation. Just as was the case in the Canadian study, we made an estimate for low and high range values. The total value of data investments, both data, databases and data science, amounted to 15.6 billion euro for the low range values and 20 billion for the high range values.

The number of assumption required make the framework still not ready for official statistics, even if data and its subtypes are fully included in core national accounts. Further research can cover, but is not restricted to, different topics such as trade in second-hand assets, cross-border trade (sale of data originals cross border), use of data within multi-national enterprises, and deflation. Also the inclusion of government data assets and overlap between data science and R&D estimates. The research presented in this paper did not cover a PIM estimate of (net) stocks. These types of estimation require even more assumptions on the parameters, due to the lack of research on this topic. A broader issue involves the delineation between produced and non-produced assets. A consensus is arising that regards part of the value of data as non-produced. These observable phenomena hold value even separate from digitized information. The method of assigning value for each part however is yet unclear (cf. ISWGNA 2020).

References

CBS (2018), Uitbesteden van werk aan het buitenland door bedrijven in Nederland, Internationaliseringsmonitor 2018-2. (https://longreads.cbs.nl/im2018-2/uitbesteden-van-werk-aan-het-buitenland-door-bedrijven-in-nederland/)

Elp, M. van and N. Mushkudiani (2019), ‘Free services’. CBS paper.

European Commission, European data strategy. (https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en)

Haan, M. de, and J. Haynes (2018), ‘R&D capitalisation: where did we go wrong?’, EURONA 2018-1.

ISWGNA sub-group on digitalization (2020), ‘Recording and Valuation of Data in National Accounts’.

Ker, D., V. Spiezia and A. Weber (2019), ‘Perspectives on the value of data and data flows’. Working Party on Measurement and Analysis of the Digital Economy.

Li, W.C.Y., M. Nirei and K. Yamana (2019), ‘Value of Data: There’s No Such Thing as a Free Lunch in the Digital Economy’.

Nguyen, D. en M. Paczos (2019), ‘Measuring the Economic Value of Data and Data Flows’. OECD Working Paper.

Mitchell, J., M. Lesher and M. Barberis (2021), ‘Going Digital Toolkit Note: Measuring the economic value of data’, DSTI paper published for official use (DSTI/CDEP/GD(2021)2).

Nijmeijer, H. (2018), ‘Issue paper on Databases’, Joint Eurostat-OECD Task Force on Land and Other Non-Financial Assets.

Rassier, D.G., R. J. Kornfeld and E.H. Strassner (2019), ‘Treatment of Data in National Accounts’. Paper prepared for the BEA advisory committee.

Shapiro, C. and H. Varian, (2000), ‘De nieuwe economie’. Nieuwezijds, Amsterdam.

Statistics Canada (2019a) , ‘Measuring investment in data, databases and data science: Conceptual framework’.

Statistics Canada (2019b), ‘The value of data in Canada: Experimental estimates’.

United Nations (2009), ‘System of National Accounts 2008’.