IARIW – CIGI 2023

IARIW – CIGI 2023

Thursday, November 2 – Friday, November 3

Estimation of Investment in and Stock of Data and Databases in the Canadian System of National Accounts

Koami Amegble
(Statistics Canada)

Brenda Bugge
(Statistics Canada)

Amanda Sinclair
(Statistics Canada)

Paper prepared for the Conference on The Valuation of Data November 2 – November 3, 2023
Session 1: Methodologies of Data Valuation: National Accounts Approaches
Time: Thursday, November 2, 2023 [10:00-11:30AM EST]

Estimation of Investment in and Stock of Data and Databases in the Canadian System of National Accounts

IARIW–CIGI CONFERENCE: The Valuation of Data

Koami Amegble, Brenda Bugge, Amanda Sinclair

Introduction

Following the 2019 Statistics Canada release of the proposed conceptual framework[1] and experimental estimates[2] of the value of Data, efforts have been focused on updating these estimates using more advance estimation methods. This paper will not delve into the question of whether Data is a produced asset, as there has already been significant work completed by various organizations and there is general agreement that yes, Data is a produced asset – that is it is used repeatedly or continuously in production processes for more than one year.

However there remains significant questions: even though Data is considered a produced, intangible asset that is used in production repeatedly for more than one year, would all Data be considered a produced asset? Or are certain types of Data transitory, they may be kept for extended periods of time, but not really used continually in production? Additionally, what are the boundaries for own-account investment in data, databases, software and research and development to avoid overlap?

This paper will present the results of recent initiatives by Statistics Canada to estimate the flow of investment and the stock of both Data assets and Database assets, as well as a comparison against the initial results from 2019. The results have been explicitly refined to include only those Data that are thought to be used continually in production for more than one-year.

The paper will finish with considerations going forward given the inclusion of Data as a produced asset within the next update to the international standard for national accounting.

Estimation methodology

As noted in the 2019 paper, Data and Databases can be produced and either used by firms on 'own account' or sold on the market. Data and Databases sold on the market are in theory valued at market price (the value of the transaction). Ideally, Statistics Canada would survey Canadian firms and obtain information related to their market sales of Data and Databases, however at this time, Statistics Canada has very little information on these transactions. Data and Databases that are used on 'own account' are valued at the cost of producing the product, including an estimated return on capital. Since Statistics Canada does not have information on the market sales of Data and Databases, the production of these assets, whether for market sale or for own-account use, they have been valued at the cost of producing the product.

[1] Measuring investment in Data, Databases and Data science: Conceptual framework
https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00008-eng.htm

[2] The value of Data in Canada: Experimental estimates
https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm

Investment flows

As previously noted, the estimation methodology employed for the capital investment flow was the sum of costs as this is the recommended method in absence of observable market transactions and for own-account production: the value is determined by how much it costs to produce that capitalized asset. The estimate would include an estimate of labour costs, indirect costs, and capital services. The labour costs are calculated using the wage bill multiplied by the average time spent on these activities. Indirect costs cover additional costs to the entity and include resources such as finance and administration as well as electricity, building maintenance, etc. Capital services would represent the return on capital assets used in this productive activity.

This estimation is completed by industry and by region in Canada (10 provinces and 3 territories).

Within this method, there are significant questions to address:

- What occupations are involved in creating this type of asset?
- What portion of their time do they spend on creating this asset?
- What is a reasonable markup for the indirect costs and return to capital?

The first experimental estimates for Data and Database investment flows and stocks were developed using a set of occupations that were considered Data/Database focused, whereas time spent was an arbitrary ratio based on preconceived notions of types of activities done by those occupations. Given this, the time spent was used as a range, with maximum and minimum shares, in order to calculate a range of investments. The following table provides the occupations and their time use factors.

Table 1 Occupations and time spent ranges, Data and Databases, first method

DATA	**Lower bound**	**Upper bound**
Financial and investment analysts	10%	20%
Customer and information services supervisors	30%	50%
Data entry clerks	100%	100%
Other customer and information services representatives	30%	50%
Survey interviewers and statistical clerks	90%	100%
Mathematicians, statisticians and actuaries	20%	30%
Economists and economic policy researchers and analysts	20%	30%
Social policy researchers, consultants and program officers	20%	30%

DATABASES	**Lower bound**	**Upper bound**
Computer and information systems managers	30%	50%
Database analysts and data administrators	90%	100%
Information systems testing technicians	30%	50%

The mark-up employed in the 2019 method to cover non-direct costs was 0.53% applied to each occupation based on the share that is used for estimating other own-account capital assets (software and research and development).

The second methodology employed by Statistics Canada mirrored the exploratory work undertaken by the United States Bureau of Economic Analysis which used web scraping of a job posting web site and machine learning to identify occupations involved in creating Data and Database assets. Those occupations were then compared to a job tasking website from the United States Department of Labor[3] to determine the importance ranking of Data

[3] O*NET OnLine https://www.onetonline.org/

and Database activities within each occupations described tasks. The following table shows the expanded list of occupations and the estimated time spent using this methodology for Canada.

Table 2 Occupations and time spent, Data and Databases, second method

Occupations: Data and databases	Time spent	Occupations: Data and databases, cont.	Time spent
Financial managers	41%	Computer engineers (except software engineers and designers)	27%
Banking, credit and other investment managers	41%	Mathematicians, statisticians and actuaries	41%
Computer and information systems managers	35%	Database analysts and data administrators	25%
Government managers - health and social policy development …	10%	Information systems testing technicians	25%
Government managers - economic analysis, policy development …	10%	Post-secondary teaching and research assistants	28%
Government managers - education policy development …	10%	Economists and economic policy researchers and analysts	35%
Financial auditors and accountants	37%	Business development officers and marketing researchers	38%
Financial and investment analysts	24%	Social policy researchers, consultants and program officers	30%
Securities agents, investment dealers and brokers	34%	Health policy researchers, consultants and program officers	30%
Health information management occupations	29%	Education policy researchers, consultants and program officers	25%
Records management technicians	23%	Program officers unique to government	30%
Statistical officers and related research support occupations	25%	Customer and information services supervisors	34%
Accounting technicians and bookkeepers	30%	Other customer and information services representatives	33%
Data entry clerks	20%		
Survey interviewers and statistical clerks	26%		

This methodology proved to be quite erratic as it was heavily dependent on what jobs were posted at various times. Given this, further estimation was not completed using this method.

The current approach used by Statistics Canada is based on an update of the US method[4] that uses a databank of job post advertisements rather than web scraping. This method allows for a larger base of job ads to be used in the identification of occupations involved in Data and Database activities.

Although the Canadian method mirrors the US for the most part, an explicit decision was made to exclude those occupations likely to be involved in the creation of other capital assets, such as architectural, engineering, software development and research occupations. Although those occupations can be involved in the production of Data and Database assets, those activities would factor into the creation of other assets, such as architectural services as a cost of production for non-residential structures.

The following table shows the occupations and time spent currently in use for estimating the experimental investment flows of Data and Databases in Canada.

[4] Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings https://conference.nber.org/conf_papers/f159271.pdf

Table 3 Occupations and time spent, Data and Databases, current method (version 3)

Occupations: Data and databases	Time spent
Actuaries	54%
Bookkeeping, Accounting, and Auditing Clerks	17%
Brokerage Clerks	15%
Budget Analysts	46%
Credit Analysts	19%
Data Entry Keyers	90%
Database Administrators	36%
Designers, All Other	33%
Economists	47%
File Clerks	21%
Financial Analysts	67%
Financial Managers	21%
Financial Specialists, All Other	20%
Information Security Analysts	20%
Inspectors, Testers, Sorters, Samplers, and Weighers	50%
Insurance Claims and Policy Processing Clerks	20%

Occupations: Data and databases, cont.	Time spent
Interviewers, Except Eligibility and Loan	21%
Management Analysts	63%
Market Research Analysts and Marketing Specialists	19%
New Accounts Clerks	31%
Office Clerks, General	11%
Operations Research Analysts	53%
Order Clerks	10%
Procurement Clerks	17%
Social Science Research Assistants	81%
Social Scientists and Related Workers, All Other	58%
Sociologists	31%
Statistical Assistants	30%
Statisticians	69%
Stock Clerks and Order Fillers	4%
Tax Examiners and Collectors, and Revenue Agents	11%
Tax Preparers	11%

The mark-up that is used in the current method was developed based on a composite ratio including consumption of fixed capital, intermediate consumption, and operating surplus, again mirroring the US method. The mark-up factor is developed on an industry space, rather than the occupation focused time spent ratios and is used to estimate a range of investments. The upper and lower bounds were calculated based on production functions where the upper bound would be more data centered.

Given a lack of detail on the prevalence of desktop computers in the workplace, 1981 was chosen as the starting point for investment. The time use factor by occupation in 1981 was only a small portion of that which is used in more recent periods as it was assumed that tasks and activities of occupations would have evolved overtime and gradually became more focused on the activities of Data and Database creation. Therefore, the time use factor increased through time to reflect these changing tasks.

Estimation challenges remain for these methods, including the occupations engaged in creation of Data and Database assets are not always obvious, and those occupations and time spent will change as our ways of working evolve.

Stock of Data and Database assets

The method to calculate the capital stock of Data and Databases in both the preliminary work in 2019 and the current work is the geometric depreciation under the perpetual inventory method (PIM). The value of net capital stock is estimated whereby investment flows are accumulated and depreciated over time, giving rise to a stock of assets. The PIM uses a time series of investment flows, assets services lives and prices, and assumptions regarding methods of depreciation patterns. In the case of the capital assets of Data and Database, depreciation is different from that of other capital assets such as machinery. The depreciation would reflect the obsolescence of these assets.

Within the PIM, depreciation is calculated first based on the existing stock and the new investment in Data and Databases for that reference year. Net stock is then calculated as the addition to the previous year's constant dollar net stock plus the new investment minus the depreciation.

The depreciation and net stock calculation is completed first on a constant dollar basis where the nominal dollar investment flow is depreciated using a price. The resulting constant dollar stock and depreciation are then re-inflated using the same price, to achieve current dollar values. This mirrors the capital stock and depreciation calculation used in the main sequence of Canada's National Accounts.

The price indices used in the first method were constructed from the movement of wage rates for those identified occupations involved in Data and Database activities. This is a similar methodology that Canada employs for deflating own-account capital of software and research and development and reflects the varying prices of the input factors (wage rate). The current method for Data and Database also uses the movement of wages to construct a price index, although proxies from own-account software are used. In the future, price indices based on the movement of wage rates will be developed using the detailed occupations involved in Data and Database creation.

In the 2019 method, the service life for Data was pegged at 25 years to reflect the assumption that Data is useful for a generation. Databases mirrored the software service life of 6[5] years. With the current study, the service life for Data was changed. Given the prevalence of Data in productive activities in the economy, an average service life would need to cover Data used in production for much longer terms, and those in use for shorter terms. There is an assumption that the Data used for a few years in production is most often the case in our economy, therefore a service life of 7 was chosen. Database assets are also deemed to have a service life of 7 which mirrors that used for own-account research and development for Canada.

Preliminary results

Current method

The total real investment in Data and Databases in Canada is estimated to be between $10.4 billion and $17.5 billion in 2021. For comparison purposes, in 2021 the total real investment in capital assets in Canada was $459.7 billion, while for intellectual property products (IPP) it was $62.4 billion. Therefore, based on this expanded methodology, Data and Databases could account for roughly 3% of total capital investment or 22.4% of investment in IPP in Canada[6].

As shown in the table below, the majority of investments in Data and Databases was in the business sector.

Table 4 Real investment in Data and Databases, by sector, Canada, 2021 ($millions, 2012=100)

Sector	Estimate	Minimum	Maximum
Business	12,363	9,175	15,550
Government	1,570	1,175	1,964
NPISH	19	14	24
Total	**13,952**	**10,365**	**17,539**

[5] Service life of 6 years, with a declining balance rate of 1.65, for a depreciation rate of 27.5%.
[6] These experimental amounts cannot be directly compared to existing estimates of gross domestic product since they overlap to a degree with the published estimates of total gross fixed capital formation. Further work is required to calculate the overlap and refine the estimates.

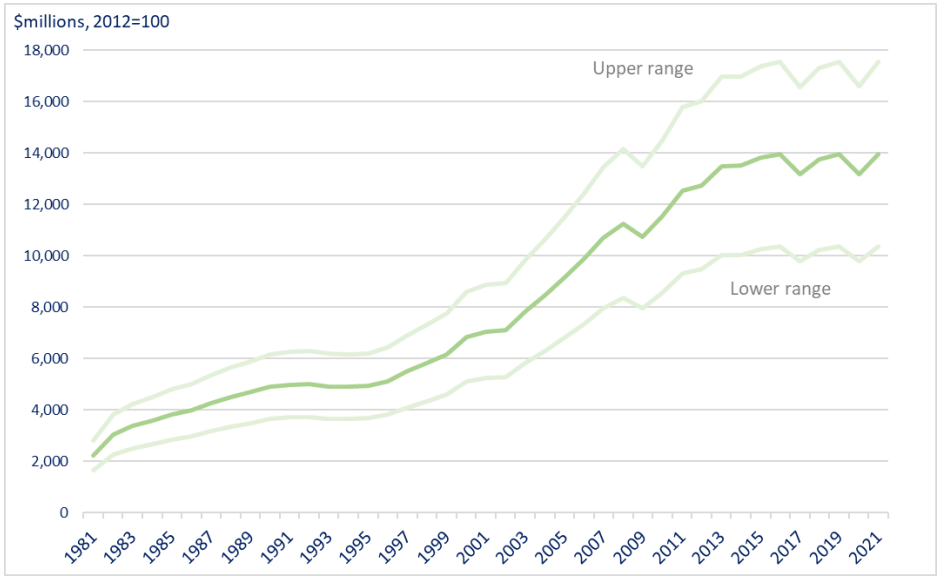
From an industry perspective, finance and insurance, construction and public administration invested the most, accounting for over half of the total estimated investment in Data and Databases in 2021.

Table 5 Industry share of total real investment in data and databases, Canada, 2021

Industry (NAICS)	Share of total investment
Accommodation and food services	0.5%
Administrative and support, waste management …	1.7%
Agriculture, forestry, fishing and hunting	0.3%
Arts, entertainment and recreation	0.5%
Construction	22.9%
Educational services	2.5%
Finance and insurance	23.6%
Health care and social assistance	3.4%
Information and cultural industries	2.8%
Management of companies and enterprises	1.0%
Manufacturing	7.5%
Mining, quarrying, and oil and gas extraction	2.2%
Other services (except public administration)	2.6%
Professional, scientific and technical services	2.9%
Public administration	11.3%
Real estate and rental and leasing	1.3%
Retail trade	2.8%
Transportation and warehousing	4.2%
Utilities	1.1%
Wholesale trade	5.0%

Not surprisingly, there has been significant growth in investment in Data and Databases in Canada since the 1980's, with the largest gains occurring in the early 2000s. Investment in data assets has slowed over the last decade, with a high annual growth of 8.7% in 2011 and lows of -5.7% and -5.5% in 2017 and 2020 respectively.

Figure 1 Total real investment in Data and Databases, Canada, 1981-2021 ($millions, 2012=100)



The total year-end net stock of Data and Databases in Canada was between $31.6 billion and $53.5 billion in 2021 – amounting to roughly 1% of the total capital stock and 19.4% of IPP. As a result of investment concentrated in the business sector, the same holds for the net stock, with the same industrial concentration. Finance and insurance, construction, and public administration were the dominant industries in terms of net stock of Data and Database assets.

Table 6 Real year-end net stock in Data and Databases, by sector, Canada, 2021 ($millions, 2012=100)

Sector	Estimate	Minimum	Maximum
Business	37,875	28,113	47,636
Government	4,592	3,438	5,746
NPISH	65	49	81
Total	**42,532**	**31,600**	**53,463**

Table 7 Industry share of total net stock of data and databases, Canada, 2021

Industry (NAICS)	Share of total investment
Accommodation and food services	0.6%
Administrative and support, waste management …	1.8%
Agriculture, forestry, fishing and hunting	0.3%
Arts, entertainment and recreation	0.6%
Construction	22.7%
Educational services	2.7%
Finance and insurance	23.0%
Health care and social assistance	3.5%
Information and cultural industries	2.9%
Management of companies and enterprises	1.4%
Manufacturing	7.7%
Mining, quarrying, and oil and gas extraction	2.4%
Other services (except public administration)	2.7%
Professional, scientific and technical services	2.7%
Public administration	10.8%
Real estate and rental and leasing	1.4%
Retail trade	2.8%
Transportation and warehousing	3.9%
Utilities	1.1%
Wholesale trade	5.1%

Comparing current and previous estimates

The current method used to estimate investment and net stock in Data and Database assets yields significantly smaller results in comparison to the previous method used in the 2019 paper. Comparing estimates for 2018, the most current reference year available in both studies, the new method produces results for investment between 33% and 42% lower than the previous method, while the net stock is between 70% and 75% lower.

Table 8 Comparing current and previous estimates for 2018, Canada, ($millions, 2012=100)

Real investment	Minimum	Maximum
Previous method	17,464	25,841
Current method	10,229	17,306
Current method reduction	**-41.4%**	**-33.0%**

Net stock	Minimum	Maximum
Previous method	123,516	178,043
Current method	31,023	52,499
Current method reduction	**-74.9%**	**-70.5%**

The results from the current method are notably smaller in comparison to the first estimation method due to a variety of factors. First, the current method employed a more refined selection of occupations and applied time use factors so that only the proportion of time spent on the creation of Data and Database assets, based on performed tasks, would be included. For example, even though an occupation like an economist would be involved in the creation of Data, not all their tasks should be capitalized as such. Second, the current method applies adjustments to the mark-up to remove transitory data, that is data that does not last or is used for more than a year. Finally, the net stock results are also impacted by the shorter duration of service life used in the current method (6 years versus 25 years), thereby leading to a considerably smaller stock estimate.

The significant reduction in both investment and net stock with the current method indicates that estimates for own-account Data and Databases are extremely sensitive to the methods and assumptions used in the sum of costs approach, and for this reason continued research will be undertaken. Specifically, more research is warranted on the mark-up factors used for capital and intermediate inputs, as well as methods for avoiding duplication with other intangible assets, such as software and research and development.

Conclusion

Significant progress has been made on estimating the asset Data in many national statistics offices on the road to integrating this asset within the production boundary as a component in the updated System of National Accounts, SNA 2025. While Canada is amongst these efforts, further research is required to determine more precise occupations and time use, as well as prices and depreciation profiles. The significance of overlap with already capitalized intangibles will also need review.

Considerations

- It will be difficult to determine which portion of Data is used continually in production processes for more than one-year. It is expected that most Data are transitory or temporary. Determining this component will be difficult.
- A service life for Data and Databases will need to be set. If these appear within the same asset class, they will require the same service life. The service life should reflect the expected length of time the asset can be used in production, this would not be the same as the retention period.