CPA CHARTERED PROFESSIONAL ACCOUNTANTS CANADA

Creating high-quality data to reach your digital transformation goals

September 2021

CPA Canada Foresight
REIMAGINING THE PROFESSION.

21-2136

ABOUT CPA CANADA

Chartered Professional Accountants of Canada (CPA Canada) works collaboratively with the provincial, territorial and Bermudian CPA bodies, as it represents the Canadian accounting profession, both nationally and internationally. This collaboration allows the Canadian profession to champion best practices that benefit business and society, as well as prepare its members for an ever-evolving operating environment featuring unprecedented change. Representing more than 220,000 members, CPA Canada is one of the largest national accounting bodies worldwide. cpacanada.ca

Electronic access to this report can be obtained at **cpacanada.ca**

Table of Contents

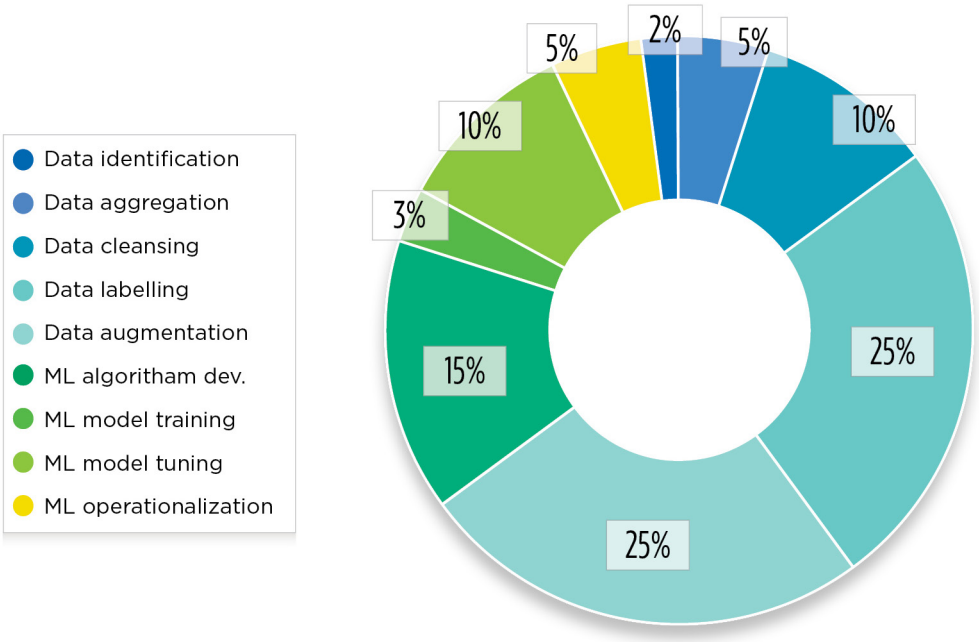
In computing, it has always been clear that garbage in equals garbage out. But this is especially true for machine learning, which powers millions of artificial-intelligence (AI) systems in use around the world. This is why the quest for high-quality data has become top of mind for business leaders. Research from Gartner estimates that AI will create US$2.9 trillion in business value this year and will improve worker productivity by 6.2 billion hours.[1]

In the last decade, organizations made considerable investments aimed at the development and testing of new algorithms and machine-learning tools. Over time, tens of thousands of algorithms have been deployed by data scientists. Now, AI applications are available through open-source platforms, subscriptions to cloud platforms and through licensing agreements. They can be adapted or trained into a wide variety of purposes and applications. Leading organizations managing digital transformation processes have now turned their attention to create (or acquire) high-quality datasets.

As algorithms are set to solve more specialized problems, the cost of generating more granular and nuanced data is getting higher. A recent study by Cognilytica estimated that data preparation represents over 80 per cent of the time consumed in AI and machine-learning projects. This includes time spent on data identification, aggregation, cleaning, labelling and augmentation.

1 Dataiku, 2021. Getting the Most out of AI in 2021: Insights from 10+ Industry Trailblazers. https://pages.dataiku.com/getting-the-most-out-of-ai-2021

Percentage of time allocated to machine learning project tasks



"Data quality is potentially the single most important factor to success," says Jeff McMillan, chief data and analytics officer at Morgan Stanley. "I say that because if you don't have accurate data, nothing else works. A lack of quality data is probably the single biggest reason that organizations fail in their data efforts." Research conducted with early AI adopters by data-labelling firm Appen has found that 93 per cent of companies report that high-quality training data is important to successful AI.[2] Some of the challenges faced by organizations include unlabelled or poorly labelled data, inconsistent or disorganized datasets, too many data sources, a lack of tools to adequately address quality concerns and process bottlenecks.[3]

This primer will help CPAs ensure that datasets that are used in their organizations meet minimal quality requirements for digital transformation. It will also outline a range of approaches to establish and manage data-quality systems and controls.

2 Appen, 2020. The State of AI and Machine Learning. Appen, 2020. https://resources.appen.com/wp-content/uploads/2020/06/Whitepaper-State-of-Ai-2020-Final.pdf

3 Dataiku, 2021. 2021 Trends : Where Enterprise AI Is Headed Next. Dataiku, 2021. P. 19. https://www.dataiku.com/stories/2021-trends-where-enterprise-ai-is-headed-next/

The risks of using low-quality data

There are a number of risks associated with using sub-par datasets to run algorithms and machine-learning tools. Of course, the most important one is financial risk. According to International Data Corporation, global spending on AI is slated to double over the next four years, growing from US$50 billion in 2020 to more than $110 billion in 2024.[4] Yet a 2019 business survey estimates that 87 per cent of data-science projects never made it into production. One of the main reasons that projects fail is low data quality. According to Deborah Leff, CTO for data science and AI at IBM, "the problem with data is always that it lives in different formats, structured and unstructured, video files, text, and images, kept in different places with different security and privacy requirements, meaning that projects slow to a crawl right at the start because the data needs to be collected and cleaned."[5]

When trained with low-quality or incomplete data, machine-learning tools can also lead to increased regulatory risk. For example, improperly trained tools can lead to biases or misleading insight, resulting in discrimination by denying under-represented groups for service or treatment. When a prejudice or error is embedded in training data, organizations significantly increase their legal exposure. In a recent Columbia Law Review article, Professor Frank Pasquale looked at the potential for inaccurate or inappropriate data to contaminate machine learning. He argues that "firms relying on faulty data can be required to compensate those harmed by that data use." Emerging doctrinal and regulatory approaches point to the emergence of a duty of care leading to corporate liabilities for failure to maintain adequate safety standards regarding data. This duty of care includes focusing attention on inaccurate, inappropriate or illegally obtained data and "ensuring that the training data for machine learning adequately reflects the domain it governs."[6]

4 International Data Corporation, 2020. Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide https://www.idc.com/getdoc.jsp?containerId=prUS46794720

5 https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

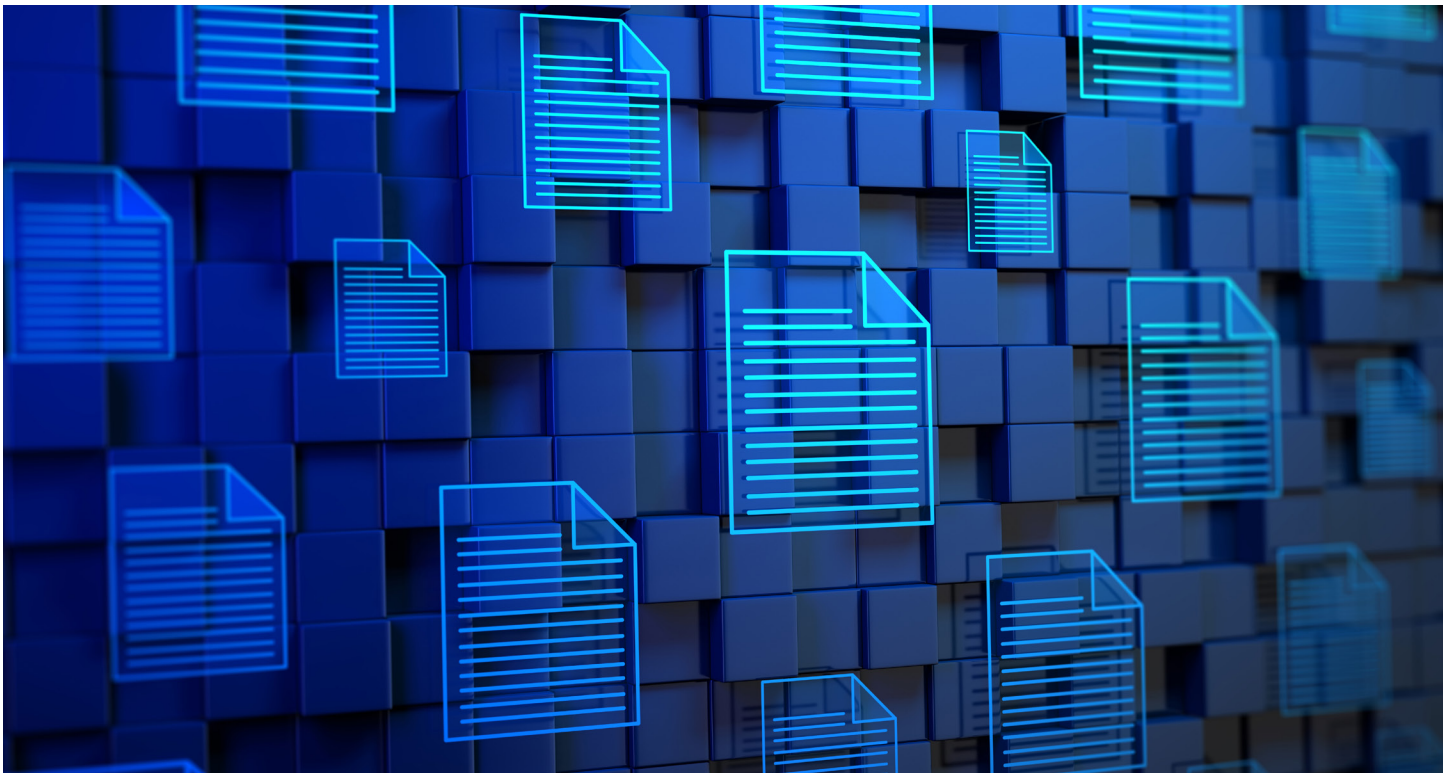
6 Frank Pasquale, 2020, Data Informed Duties in A Development. Columbia Law Review. Vol. 119, pp. 1917-1940. https://columbialawreview.org/content/data-informed-duties-in-ai-development/

Starting your organization's digital journey

Investing to create high-quality data clearly represents an important step in your organization's journey towards digital transformation. Other important steps to establishing the right framework to succeed include the following:

- adopting a corporate data policy to set data governance rules for data reuse
- approving your digitization strategy and budget
- creating a hybrid team composed of subject matter experts, relevant data experts and IT
- identifying business problems to solve or opportunities to exploit, creating and documenting use cases and selecting appropriate AI solutions or machine-learning tools to train

When these initial steps have been taken, it's time to focus your attention on creating high-quality data. Data needs to be clean, accurate, complete and labelled properly. This can be accomplished through data collection and preparation.



Data collection and preparation

The first steps in creating high-quality data are to identify the relevant data sources across the organization and to create a pathway for your hybrid team to view, access and use the data. This stage requires mapping all relevant datasets from source to repositories. Datasets can include text, numbers, still images, audio, speech, video, objects, time series, sensor data streams, web clicks or product stock keeping units. Review of datasets should include transformation processes (actions required to merge different datasets with different formats into one master dataset), harmonization activities (aligning datasets by using the same name, date, sequence conventions), and final processed output (one master harmonized dataset that is fit for purpose). Variables will differ whether data-collection processes are automated (for example web clicks), originate from sensors (for example Internet of Things devices) or are manually entered into a database (for example a purchase order).

Data cleansing, labelling and annotation

Dirty data is of limited value for feeding algorithms and training machine-learning tools. Data must be checked for compliance with internal policies before being cleaned, labelled and annotated.

Data cleaning involves such steps as ensuring that blank fields in records or databases have been inputted, duplicate records have been deleted, definitions and acronyms used in different datasets have been aligned and time and sequencing conventions have been harmonized. One additional task that may be needed at the data-cleansing stage is to ensure that the datasets are organized in a way that removes personal information identifiers in order to comply with relevant privacy requirements.

Data labelling and annotation are time-consuming tasks. Together, they ensure that the data engineers who are programming machine-learning tools have accurate information about the data they are using. The goal in training machine-learning tools is to show the outcome you want your tool to predict. By feeding enough examples to your model and explaining what to look for and how to find it, it will eventually recognize these features on unlabelled objects on its own and make a decision or take some action as a result.[7] Labelling tasks tend to be specific to the sort of data being used. Primary use cases vary from text tagging and labelling; speech tagging and labelling; object recognition, classification and annotation; as well as image classification, tagging and annotation.

7 Cloud Factory, 2020. Data Annotation Tools for Machine Learning: Choosing the Best Data Annotation Tool For Your Project. May 2020. https://go.cloudfactory.com/hubfs/02-Contents/2-eBooks/Data%20Annotation%20Tools%20for%20Machine%20Learning%20(Evolving%20Guide).pdf?utm_campaign=RTQD%20%7C%20Data%20Annotation&utm_medium=email&_hsenc=p2ANqtz-8i3RosIvhxfqIn1ZCEspQQNZd_wOxhmw0aYBhTkfcyLfrQV_zuzjje3rETW5tKOHMINA1N0IUJyujsuNRUu2A6Djz5FQ&_hsmi=88137733&utm_content=88137733&utm_source=hs_automation&hsCtaTracking=7bf603ac-2909-467b-b505-0f34b50289a0%7C408970b5-5de5-4328-91b7-7535da77716f

Although it is possible for your hybrid team to undertake data cleansing, labelling and annotation tasks for small datasets, they will need help to manage large or varied datasets that must be aggregated. According to John Singleton, co-founder of data annotation software company Watchful, "A lot of times data annotation is taken by a small and already overworked data-science team who aren't able to focus on their job, which is developing and delivering models that are meaningful."[8]

Starting in 2018, a wave of commercial data-annotation tools became available, offering full-featured, complete-workflow tools for data labelling. Organizations now have many options to choose from. They may opt to download an off-the-shelf data-annotation platform and hire labour to perform annotation tasks manually (although it can be difficult to guarantee quality from a transient labour force unless properly trained and supervised). That being said, a wide variety of self-serve data-labelling tools are available in the marketplace. Well-known tools include Prodigy and Label Studio. Additional open-source and commercial data-annotation tools (such as the hugely popular Dataturk) can be accessed through Git-Hub's Awesome data annotation landing page.

One option gaining traction is to source a self-service data-labelling platform through software-as-a-service (SaaS) arrangements from a vendor. These platforms often embed automation tools and facilitate large-scale collaboration on large datasets. Machine learning can assist in:

- labelling content
- performing quality control checks on human labour
- identifying data types
- pointing to outliers in the structure of a data column and providing guidance to users on how they can clean the data

Organizations like Alegion offer self-service labelling platforms for video annotation and labelling that subject matter experts in your organization can use to label and annotate internally.[9]

Organizations fully engaged in digital transformation may ultimately decide to create full-fledged DataOps systems to support dedicated teams in order to rapidly and repeatedly engineer mission-ready data from all data sources across an enterprise. The term DataOps was coined by IT research firm O'Reilly in 2019 to describe the dedicated systems and teams created by organizations

8 Synced, 2019. Data Annotation: The Billion Dollar Business Behind AI Breakthroughs. https://medium.com/syncedreview/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs-d929b0a50d23

9 https://appen.com/blog/build-or-buy-data-annotation-tool/

aiming to use company data as a strategic asset.[10] With these teams in place, organizations can invest in resident domain expertise to provide highly specialized annotation for sophisticated algorithms.

Companies now offer software to speed up data annotation for deep learning models.[11] A recent report from Cloud Factory provides a comprehensive listing of commercially available annotation tools.

Many organizations opt instead to use third-party organizations that offer a combination of services and advice in data preparation. Large consulting firms like Deloitte, EY, KPMG and PwC offer organizations embarking on digital transformation a range of support that can include advice and guidance on data-preparation activities.[12]

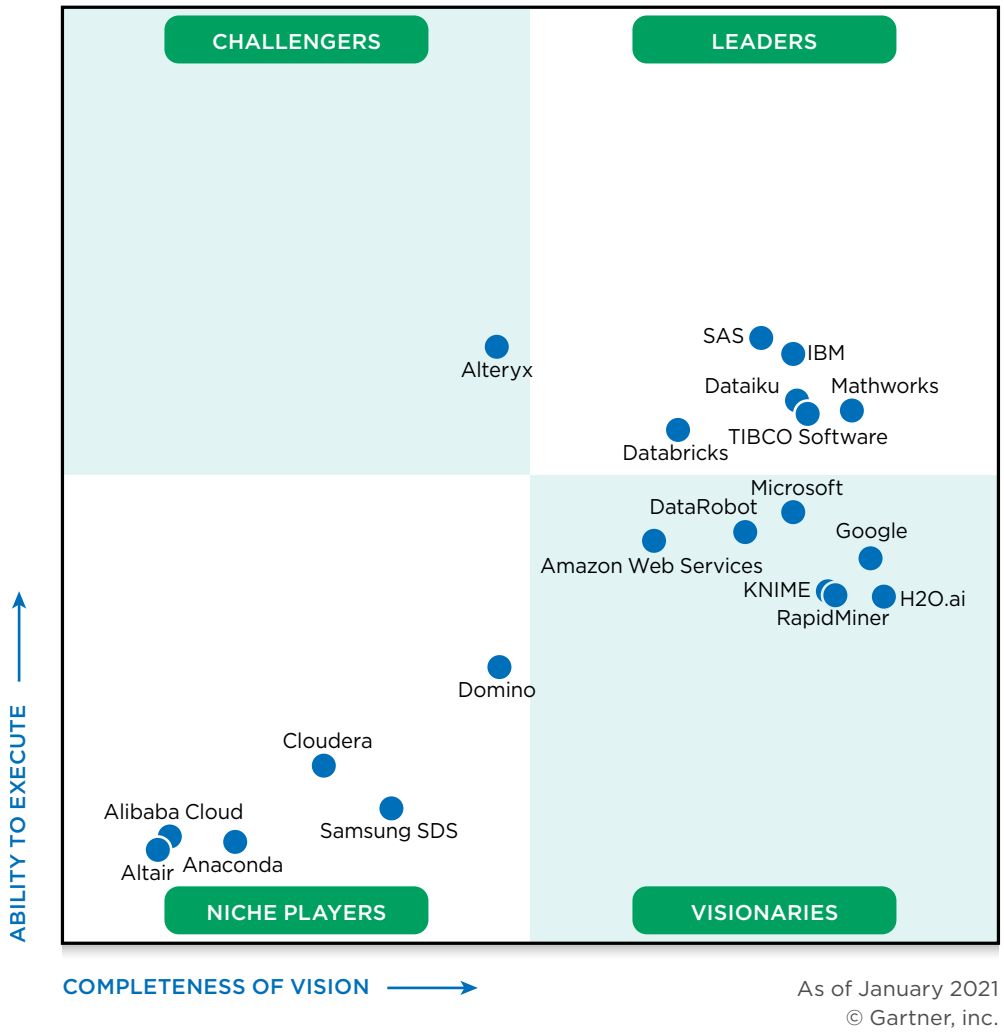
Platform-based solutions are also growing in importance. In 2021, Gartner published a comprehensive report comparing offerings from 20 data science and machine-learning platforms (DSML) that organizations can use to source data, build models and operationalize machine learnings. According to the report, all listed DSML service offerings include software, tools and guidance to manage data preparation and exploration features. Although some of these service providers are focusing on niche markets covering specific sectors, many large platform service providers have developed approaches that can be adapted to a variety of specific-use cases. As illustrated below, established global leaders include SAS, IBM, Dataiku, MathWorks, Databricks and TIBCO Software.[13]

10 Andy Palmer, Michael Stonebreaker, Nik Bates-Haus, Liam Cleary and Mark Marinelli. 2019. Getting DataOps Right. O'Reilly, 2019. 58 pages. https://get.oreilly.com/ind_getting-dataops-right.html

11 For example, see service offering by Supervisely, used by more than 25,000 companies in a growing number of sectors: https://supervise.ly/

12 http://canadian-accountant.com/content/business/consulting-in-canada-where-the-big-four-firms-are-making-money

13 https://www.gartner.com/doc/reprints?id=1-24M0T9F3&ct=201117&st=sb



As of January 2021
© Gartner, inc.

You may also want to consider entrusting a third party to perform data-preparation functions on your behalf. Vendors have developed a series of aides, often powered by machine-learning tools, to improve the quality and increase the speed of labelling and annotation functions. These tasks are becoming increasingly specialized and domain specific. For example, innovative image medical technology requires machine-learning models that can identify a wide range of pathologies within medical images such as clots, fractures, tumours and obstructions. Datasets need to be consistently labelled with the specific pathologies. This requires knowledge and domain expertise from legions of data-labelling and annotation operators.

Labelling firms are rapidly evolving to provide domain-specific capabilities. According to a recent report from Cognilytica, more than 35 companies are currently engaged in providing human labour to add labels and annotation to data. Some of these firms use general crowdsourcing approaches to data labelling, while others rely on their own managed and trained labour pools to address domain-specific data labelling needs. Rising stars in this field include Scale, which leverages AI-powered annotation tools and 30,000 contractors for labelling text, audio, pictures and video. The demand for data-labelling services from third parties is expected to grow from US$1.7 billion in 2019 to more than $4.1 billion in 2024.[14]

Finally, another possible approach is to acquire data from a third party to train algorithms and machine-learning tools. Organizations like Thompson Reuters, Orbital Insights, Appen and Collibra Marketplace have created new business lines where curated datasets can be leased or purchased. There are also a growing number of open-source data sets made available by governments, not-for-profit organizations and universities.[15] Gartner estimates that by 2022, 35 per cent of large organizations will be either sellers or buyers of data via formal online data marketplaces, up from 25 per cent in 2020.

Data marketplaces and exchanges provide platforms to consolidate third-party data offerings. These marketplaces and exchanges provide centralized availability and access that create economies of scale to reduce costs for third-party data. However, a recent Garnter report from Gartner advises that: "to monetize data assets through data marketplaces, data and analytics leaders should establish a fair and transparent methodology by defining a data governance principle that ecosystems partners can rely on."[16]

14 https://orbitalinsight.com/

15 https://appen.com/open-source-datasets/; https://www.thomsonreuters.com/en/artificial-intelligence/machine-learning.html; https://orbitalinsight.com/

16 https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/

Accuracy, quality control and grading

Checks and balances are required to ensure that the data has been appropriately cleaned prior to analysis. Combined or aggregated datasets should be reviewed and deemed complete, accurate and reliable.

Grading is about determining whether data is appropriate for the purpose for which it will be used. Not all data is equal and different decisions require different reliability of data. Using data without understanding its reliability limitations will likely result in a poor outcome if the data is assumed to be more reliable than it is. When it comes to AI and machine learning, CPAs need to be on the lookout for: a) high quality data in order to avoid prejudices from incomplete or faulty datasets, and b) well-designed algorithms and machine-learning tools in order to avoid prejudices embedded from faulty assumptions. At the same time, it takes time and money to collect highly reliable data, so requiring highly reliable data for all decisions may result in missed opportunities.

To illustrate the idea of fit-for-purpose, data underpinning external reporting is of high reliability and is typically certified as such. Conversely, data used for internal purposes can be of varying reliability. Professional accountants need to understand the differences and be able to inform decision-makers of the reliability of the data underpinning a decision, including an assessment as to whether the data aligns with the data-reliability requirements of a decision point. For example, an "options analysis" data-reliability requirement is less stringent then a "go/no-go" decision point. Calibrating data reliability to decision points can improve timeliness in decision-making, risk disclosure and transparency. It's essential to have formalized check-in processes to ensure that decisions are validated as data is updated. Otherwise, previously sound decisions may be overtaken by subsequent events.[17]

CPAs are well positioned to manage data-quality control functions and systems in organizations engaged in digital transformation. As explained in CPA Canada's article, Making Sense of Data Value Chains, new professional classes are emerging to perform a series of new tasks and functions. Data scientists are playing a critical role in the development of new algorithms. We are seeing the emergence of data engineers, machine-learning engineers and software engineers to manage data-collection activities, data-management systems and interfaces and machine-learning programming. Managing data access and sharing platforms, organizing and maintaining data dashboards, keeping tabs on data queries and ensuring compliance with data-sharing contracts as well as applicable laws and regulations will create a strong demand for data controllers. If operated strictly as a public service, this new professional category could combine competencies ranging from library sciences, contract law, privacy law, cybersecurity as well as compliance monitoring and reporting.

Although assessing, testing and reporting on data quality has become top of mind, no new professional class has emerged with the right combination of competencies and skills to manage data-quality systems and controls. It is a role that is well suited to CPAs. Looking forward, CPAs would benefit from the development and maintenance of data-quality standards to properly frame new systems and controls. CPA Canada is looking forward to supporting the development of appropriate guidance that will contribute to the collection and use of high-quality data.

17 CPA and IFAC, 2021. Professional Accountants' Role in Data – Discussion Paper. Joint publication, CPA and IFAC. January 2021. 23 pages.

Where to find further information

CPA Canada Mastering Data series

- Corporate data policy and its elements
- Building a digitization strategy for your company
- Making sense of data value chains

Data annotation and labelling tools

- Prodigy
- Label Studio
- Dataturk
- Git-Hub
- Cloud Factory
- Scale

CHARTERED
PROFESSIONAL
ACCOUNTANTS
CANADA

277 WELLINGTON STREET WEST
TORONTO, ON CANADA M5V 3H2
T. 416 977.3222 F. 416 977.8585
CPACANADA.CA