Digital Studies / Le champ numérique

Tchoh, Bennett Kuwan, Sara Barnard, Robert Budac, Katrina Ingram, Ryan Chartier, Daniel C. Baumgart, and Geoffrey Rockwell. 2025. "Challenges in Curating Real-Time Data During a Crisis: The Case of the COVID-19 Pandemic in Alberta." *Digital Studies/Le champ numérique* 15(1): 1–27. https://doi.org/10.16995/dscn.15498.

Open Library of Humanities

Challenges in Curating Real-Time Data During a Crisis: The Case of the COVID-19 Pandemic in Alberta

Bennett Kuwan Tchoh, University of Alberta, bennettkuwan@yahoo.co.uk
Sara Barnard, University of Alberta, sbarnard@ualberta.ca
Robert Budac, University of Alberta, rbudac@ualberta.ca
Katrina Ingram, University of Alberta, reganing@ualberta.ca
Ryan Chartier, University of Alberta, recharti@ualberta.ca
Daniel C. Baumgart, University of Alberta, baumgart@ualberta.ca
Geoffrey Rockwell, University of Alberta, grockwel@ualberta.ca

In early 2020, the SARS-CoV-2 virus was recognized as a threat by most countries, and the risk of its spread was already being discussed. After the first cases were identified and the sudden onset of lockdowns, we realized the need to document Albertans' experience of the pandemic. This paper describes our real-time collection, beginning in March 2020, of the discourse surrounding the COVID-19 pandemic in Alberta. We quickly implemented a data collection plan that involved gathering textual data from various sources. We collected the text of public health briefings given by the Chief Medical Officer of Health (CMOH) and the premier of the province, tweets about the pandemic in Alberta, and news articles published in online journals. We conducted preliminary data analysis to assess the suitability of the data for statistical analysis. This included word cloud visualizations, correlation analysis, and Term Frequency–Inverse Document Frequency (TF-IDF) analysis. The results of the analysis were plausible, suggesting that the datasets are suitable for even more advanced data extraction and analysis. The data was deposited in the University of Alberta's Dataverse, an open-source repository, where our datasets have attracted significant interest. Judging by download numbers, the speeches of the premier and the CMOH have each been downloaded more than 2,600 times. We used Otter.ai, an online transcription software, to transcribe the speeches of the premier. The high interest in the speeches of government officials made us realize that more resources should have been allocated to ensure near-perfect transcriptions of the premier's speeches, as the premier was responsible for making the final decisions on pandemic-related policies in the province.

Au début de l'année 2020, le virus SARS-CoV-2 a été reconnu comme une menace par la plupart des pays, et le risque de sa propagation faisait déjà l'objet de discussions. Après l'identification des premiers cas et le déclenchement soudain des confinements, nous avons compris la nécessité de documenter l'expérience des Albertains face à la pandémie. Cet article décrit notre collecte en temps réel, débutée en mars 2020, du discours entourant la pandémie de COVID-19 en Alberta. Nous avons rapidement mis en place un plan de collecte de données textuelles provenant de diverses sources. Nous avons recueilli les transcriptions des points de presse de la médecin hygiéniste en chef (CMOH) et du premier ministre de la province, des tweets concernant la pandémie en Alberta, ainsi que des articles publiés dans des journaux en ligne. Une analyse préliminaire des données a été menée afin d'évaluer leur pertinence pour une analyse statistique. Celle-ci comprenait des visualisations par nuages de mots, des analyses de corrélation et une analyse via Term Frequency–Inverse Document Frequency (TF-IDF). Les résultats obtenus se sont révélés plausibles, suggérant que les ensembles de données sont adaptés à des extractions et analyses plus avancées. Les données ont été déposées dans le Dataverse de l'Université de l'Alberta, un dépôt en libre accès, où elles ont suscité un vif intérêt. D'après les chiffres de téléchargement, les discours du premier ministre et de la médecin hygiéniste en chef ont chacun été téléchargés plus de 2 600 fois. Nous avons utilisé Otter.ai, un logiciel de transcription en ligne, pour transcrire les discours du premier ministre. L'intérêt marqué pour les interventions des responsables gouvernementaux nous a fait réaliser qu'il aurait été nécessaire d'allouer davantage de ressources afin d'assurer des transcriptions quasi parfaites des discours du premier ministre, étant donné qu'il était responsable des décisions finales concernant les politiques liées à la pandémie dans la province.

Introduction

On January 25, 2020, Canada's first case of COVID-19 was recorded in Toronto. Less than two months later, in March 2020, the first lockdowns followed. The effects on Canadians were significant. Canadians witnessed shortages of goods in stores, inability to leave home for school or work, the closure of international borders, and the fear and uncertainty about the health and safety of themselves and their loved ones. Canadians found their lives upended. Therefore, as early as March 2020, it became clear that the societal impacts of COVID-19 on Canadians would be significant.

In response, a team of researchers at the University of Alberta decided to capture the lived experiences of Albertans as they faced their new reality. Aware of the challenges and limitations of retrospective data collection, they began gathering local data about the pandemic in real time. The aim was to create a corpus of textual data about pandemic-related events in Alberta and the societal responses to them as they happened. This corpus would enable discussion and analysis not only of how the pandemic progressed, but also of how the public and other stakeholders' response to the pandemic changed over time. Thus, in March 2020, our team began a project to collect and curate data from online sources about the COVID-19 pandemic in Alberta.

The aim was to create a corpus and make it available online on an open-access public repository such that other scholars interested in the public discourse of the pandemic in Alberta could use the data. To assess the suitability of the data for analysis, we employed exploratory data analysis techniques on the datasets comprising the corpus to evaluate their quality and analytical potential.

This paper outlines the entire process from initiation to completion, describing the rationale behind key decisions, the challenges encountered, and the solutions implemented. It also presents a preliminary analysis to assess the quality of the data. The paper concludes with lessons learned from the project and recommendations for future initiatives.

Literature review

The necessity of curating real-time data during crises and for healthcare purposes was already of concern to healthcare and informatics professionals prior to the onset of the COVID-19 pandemic (Bahk et al. 2016; Kejriwal and Gu 2019; Lee et al. 2018). As a result, various researchers have proposed approaches to data collection and curation during crises or for the purpose of providing healthcare. Some researchers (Bahk et al. 2016) created the Vaccine Sentimeter, a web-based dashboard prototype to detect sentiments about vaccinations online on social media in real time. Another group of researchers (Kejriwal and Gu 2019) developed an end-to-end data pipeline combined with deep learning to preprocess and filter X (formerly Twitter) data immediately after the crisis

caused by the Las Vegas shootings in 2017. More recently, other researchers (Gao et al. 2025) evaluated the use of social media to monitor the spatiotemporal evolution of damage after a typhoon disaster and showed that their methods enabled faster assessment and better emergency response. Their method integrated text and image data analysis through named entity recognition and a deep learning model.

The relatively sudden onset of the COVID-19 pandemic amplified existing information needs and created a completely new set of information demands. Within healthcare settings such as hospitals and clinics, the most current research about the virus was required to diagnose and treat patients as effectively as possible, especially those belonging to populations with specialized needs, such as obstetrics (Aaronson and Spencer 2021). Similarly, research facilities needed the most up-to-date data available to come up with and communicate more effective treatments (Vahidy et al. 2021). Information needs about the virus extended far beyond traditional healthcare settings, however. Politicians, businesses, schools, and individuals all required accurate and timely information about issues such as the transmissibility of COVID-19 and the effectiveness and availability of personal protective equipment (PPE) to make critical decisions about closures and lockdowns. Subsequently, people needed various types of information to facilitate the monumental task of transitioning to remote work with little to no notice (Aaronson and Spencer 2021). Government agencies, news outlets, researchers, politicians, business owners, etc. communicating on the pandemic resulted in an outpouring of data and information, which created what has been termed an "infodemic." It was noticed that 37,362 papers on the COVID-19 pandemic were listed on the database PubMed alone between December 31, 2019, and August 3, 2020 (Vaghela et al. 2021).

As a result, the need for information transformed into a need for curated information. Curated information allows for the dissemination of information in a manner which is easily accessed by people with a variety of information needs (Aaronson and Spencer 2021; Vaghela et al. 2021). It can also help ensure a level of accuracy, trustworthiness, or truthfulness in the information being disseminated, a necessity in a time when even major health organizations like the World Health Organization were obliged to speak out against the distribution of misinformation (Aaronson and Spencer 2021; Vaghela et al. 2021). Therefore, by curating information, researchers and organizations made it possible for information seekers to easily access accurate and trustworthy information. This need for curated information underlies many of the data curation projects which took place in response to the COVID-19 pandemic.

A group of medical librarians with the support of the Medical Library Association (MLA) created a webpage with curated and verified resources, including lists of open-access societies and publishers, through volunteer work and crowdsourcing. Later,

they were able to add information about PPE, COVID-19 symptoms and treatments, and breaking news and developments to help users quickly access new information (Aaronson and Spencer 2021).

The Houston Methodist System—the medical system serving the Greater Houston area in Texas— also responded to the COVID-19 need for rapid data curation by creating CURATOR, a database which was designed and automated to retrieve new and updated data in near real time. They achieved this by extracting and uploading all recorded data from patients treated within the Houston Methodist system, including information such as pre-existing conditions and presenting symptoms, then integrating this information with a pre-existing virtual ICU telehealth monitoring system, and with available external administrative data sources. Their aim was to gather data in real time to enable early and recurrent analysis to inform treatment decisions.

These projects, while effective in meeting users' needs, appear to have required a large amount of time and manual labour to curate data. However, it is an excellent example of what other researchers have considered to be one of the most important aspects of data curation during the COVID-19 pandemic: collective action (Shankar et al. 2020). According to these researchers, the pandemic highlighted the need for cooperation in curating information, for example, through the joint production and communication of knowledge among researchers. To quickly disseminate information during the pandemic crisis, curators needed to facilitate communication through both human and technical infrastructures, illustrated by how research agencies worldwide shared their data with Nexstrain, an open-source platform to track pathogens (Shankar et al. 2020).

The current project

It was with similar intentions that this project was started. The main aim was to collect data with research value. It differs from some of the studies listed in that the data was not meant for real-time analysis, although it was expected that, at certain points, even while data collection was still going on, with sufficient data already present, the data could be made available to interested researchers internal and external to the project. What drove us to collect data in real time was the following.

- **Capturing data that could be lost**. With the constant generation of data by the different sources already mentioned, retrospective data collection might be difficult or even impossible. Some data might be hosted for a limited time and so no longer available when needed. Some data could even be buried in the hosting system such that access is difficult. Also, as was the case with X (formerly Twitter), changes to the public access policies of data platform owners can put some data behind paywalls after some time.

- **Better data quality and accuracy**. Collecting data in real time reduces the risk of missing critical data or data sources that were relevant at the time the phenomenon happened. Collecting data at the end of a crisis might introduce bias in the data collected if the data collector is influenced either consciously or unconsciously by the current evolved state of the crisis.

Most people were experiencing a pandemic that had such a great impact on their lives for the first time. It affected society in general so much that the government officials, politicians, the media, influencers, business owners, and the public talked about it. Our aim was to collect data that captured the discourse on the COVID-19 pandemic. The data had to be of high quality and high research value such that it would be useful to researchers with diverse research interests. The data was to be processed and made available to those that were part of the project and to every interested researcher. The goal was to deposit the data in open-access repositories after verification that the data has research value.

Methodology

One of the first decisions we had to make at the start of the project was to decide on the scale of the projects, that is, the geographical region from which the discourse would be collected and the sources to include. We had to choose a scale that was manageable, considering the resources available and the effort required for data collection and processing. Considering the political organization of Canada, where provinces are governed semi-independently, we decided to limit our data to the discourse of the COVID-19 pandemic in Alberta.

Throughout the pandemic, information was disseminated in various ways; however, some of the most common methods of dissemination were through digital intermediaries, that is, through digital conduits of information such as social networks, and online news portals (Ćurković et al. 2021). As a result, our team decided that the information we would collect for the purposes of this project would be web-based. When information is disseminated through a specific digital intermediary, the information is automatically curated. The information that a user sees is not necessarily a holistic representation of all available data. Instead, it represents a subset of filtered information, and thus, rather than simply communicating data, a digital intermediary may communicate a specific discourse. Moreover, this discourse will change from platform to platform depending on what its purpose is, who the perceived users are, and who is contributing data to the web host. Thus, by collecting and curating data through digital intermediaries—that is, web-based platforms—we are well-positioned

to study the shifting discourse around COVID-19, both as the pandemic evolved and across different online platforms. Therefore, we decided to collect data from three different online sources: public health briefings on official government platforms, X (formerly Twitter), and news media articles.

From the beginning, our goal was to collect information in real time as the pandemic evolved. For the purposes of this project, real time was defined as within the twenty-four hours to one-week period after online publication. It is necessary to note that from the beginning, our data collection methods were synonymous with data curation. By setting parameters around what data we collected, we automatically curated the data from what was available online. Furthermore, our data collection methods also included the almost simultaneous cleaning and organizing of the data into defined datasets.

Data collection and curation was carried out primarily by graduate research assistants, some of whom left the project and were replaced by others as contracts began and ended. A permanent team member was assigned to manage the collected data and prepare it for upload onto the open-source repository. This enabled them to manually scan the data for inconsistencies in the collection and curation process; an inevitability in a project with many team members doing the same work. Therefore, the data manager's main task was to re-check the data to ensure it met project standards. They made weekly checks of the data submitted by the collectors to ensure that the correct data was being collected and uploaded in a timely manner. This was critical due to the real-time nature of the collection process, as some data was available for only a short time. For example, the speech transcripts by Premier Jason Kenney and Dr. Hinshaw were only available for about a month on the Government of Alberta website.

Public health briefings

Beginning in March 2020, Alberta's then-Premier Jason Kenney and Chief Medical Officer of Health (CMOH) Dr. Deena Hinshaw regularly held news briefings about the state of the COVID-19 pandemic in Alberta. These briefings included numbers (deaths, infections, hospitalizations, etc.), research findings about the virus, treatments, possible impacts to the public, the steps the government was taking in response to the virus, public health recommendations, etc. The briefings were livestreamed, and recordings were made publicly available online almost immediately. We gathered the transcripts of these speeches, as they not only represent the government's official communications on the pandemic, but also provide a record of key announcements and policy changes, including changing numbers of infections and lockdown as they

were first officially communicated to Albertans. This provided us with data that can be analyzed for questions such as how the discourse of the Albertan leadership changed over time and juxtaposed with the discourse of the other sources.

The news briefings were first gathered in March 2020, and were gathered until the end of March 2022. In total, the corpus contains 200 briefings from Premier Kenney and 249 briefings from Dr. Hinshaw. The text of Dr. Hinshaw's speeches were manually gathered from transcripts provided on the Government of Alberta's website, while audios of Premier Kenney's speeches were downloaded from SoundCloud. The speeches were subsequently transcribed into plain text files with the speech recognition software Otter.ai. A transcription software was used because of how rapidly we could get transcribed data considering the limited human resources we had. Otter.ai has an accuracy of 85–90% but was considered to be one of the best automatic transcription software available. Next, the transcribed texts were manually cleaned by removing other speakers and the question-and-answer session such that only Premier Kenney's introductory speech was left.

Tweets from X (formerly known as Twitter)

The second dataset was curated from X. Previous research has shown that X is by far the most monitored social media platform during crises due to the availability of the published data and its real-time nature (Kejriwal and Gu 2019). Thus, X data can reflect the recorded real-time responses by the public in response to the COVID-19 crisis, possibly providing a very different perspective from that of Alberta's public officials. Tweets were gathered using X's API with a web scraper. This scraper was created using twarc1 (Shaw 2020), a command-line tool and Python library for archiving tweets as JSON data (Summers et al. 2023). At the time of curation, X permitted the free use of its API to gather tweets for academic research purposes. But following the purchase of the social media platform by Elon Musk, and his move to monetize access, by the end of the first quarter of 2023, many researchers lost their access to tweets through the API (Calma 2023).

Beginning on March 25, 2020, data searches were scheduled to take place every twelve hours using a crontab file. The first scrapes searched for the hashtags #abhealth and #albertadoctors. On April 28, 2020, a search for #covid19ab was also added. Tweets were gathered until September 21, 2021. We had intended to continue scraping until October 31, but unfortunately the scraper encountered a problem and stopped prematurely.

The scraper pulled data from X and stored the results as JSON files, with each JSON file representing the results of one search and thus containing multiple tweets. A simple Python script was written to pull the tweets out of the conglomerate JSON files and store them as individual files. The tweet ID was used as a filename. We scraped a total of 1,450,318 tweets. Each tweet was checked for consistency, after which a second Python script sorted them into weekly and monthly CSV files based on the "created_at" timestamp. A third script then extracted the tweet text fields from the CSVs to generate corresponding text files. Retweets were removed before conversion, resulting in 273,368 unique tweets across all weekly and monthly text files.

News articles

The third source of data that we gathered for analysis was online news articles. Throughout the pandemic, the news media had been instrumental in conveying information about the pandemic to the Albertan and international public.

The data collection process was made up of multiple steps. First, alerts were set up on Google so that researchers were fed key stories each week as they were released. This was done using the keywords "abtracetogether," "Cargill alberta," "covid alberta," "first nations alberta covid," "long term care alberta," and "virtual care alberta." The result was about 25–30 alerts each week, which averaged about 1–6 stories gathered per alert. Most of these stories came from provincial news sources such as the *Edmonton Journal*, *Calgary Herald*, CBC, Global TV, and smaller city newspapers' websites.

Each story was saved as a text file and named following an agreed-upon format for easy organization and access. This format was YYYY-MM-DD, followed by ARTICLETITLE.txt. The articles were initially saved locally on the collector's computer, followed by a batch upload once per week to a shared drive. As the articles were saved, relevant metadata was also recorded in a spreadsheet. This includes the formatted file name of the article along with the article's original URL.

Finally, the data was manually cleaned. During this process, extraneous content such as navigation hypertext, sidebars, and ads were removed. Similarly, the titles were double checked to ensure that they conformed to the agreed format. The articles were then organized into weekly and monthly directories and weekly and monthly text files were created from each directory. Thus, weekly and monthly data could be easily accessed and analyzed to create a timeline of events. We stopped gathering news articles on October 31, 2021, with a dataset of 2,959 articles.

Preliminary analysis: Testing the data

We collected electronic text data in real-time on the COVID-19 discourse with the purpose of creating a corpus that is representative of the discourse that happened during the COVID-19 pandemic in Alberta. The goal was to create a corpus that could be used by researchers interested in discourse analysis of the pandemic in Alberta. Before we made the data available to the public, we speculated on and tried out three text analysis methods to assess whether the data was suitable for text analysis. We ran word cloud visualizations, we extracted data from the text and ran a general Pearson correlation, and we ran Term Frequency–Inverse Document Frequency (TF-IDF) analysis. The analyses and brief discussion of the findings are presented next.

Word cloud with Voyant Tools

Voyant (Sinclair and Rockwell 2025) is a web-based suite of tools combining text analysis and visualization tools. In order to explore our data, we uploaded all three of our datasets (the monthly text files of the COVID-19 updates, the X data, and the news articles) to Voyant (Sinclair and Rockwell 2025). This was done to identify the most frequent words used in the different datasets in an attempt to identify differences in the discourse just by visual exploration. To do this, we used the "Cirrus" visualization tool.

Cirrus is a visualization tool that produces a word cloud with words of different sizes based on the raw frequency of the words in the text. The colours of the words are randomly allocated. It helped us to identify and compare the words used by the CMOH and the premier, as well as discussions on X and in news articles. We made a word cloud of the first 75 highest frequency words for each group of text. The word clouds can be found in **Figures 1** to **4**.

Figure 1: Word cloud from the CMOH.

Figure 2: Word cloud from the premier.

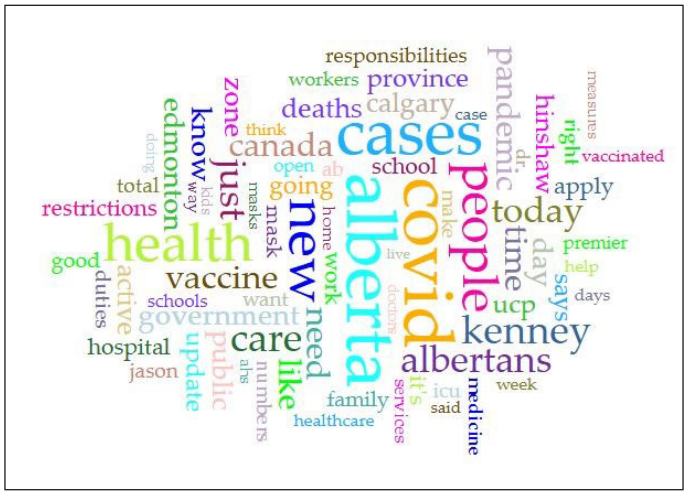

Figure 3: Word cloud from X (formerly Twitter).

Figure 4: Word cloud from news articles.

By comparing the word clouds, we notice that the words in the word cloud from the CMOH and the premier (**Figure 1** and **Figure 2**) reveal immediate differences in their speeches. In the premier's speeches, the words "government," "energy," "economy," and "work" appear frequently, even though they are not words that are thematically related to health. We also notice that words like "health," "covid," "spread," "cases," and "care," which can be grouped under the health theme, are significantly larger in the CMOH's text.

We also notice the word "said" appears in a relatively larger size in the news articles dataset (**Figure 4**). This is not surprising considering that news articles report what government officials and other stakeholders are saying about the pandemic. This suggests that the pandemic generated a lot of talk from different stakeholders and what they said was reported in the news articles. The updates given by the government officials usually end with a Q&A session in which they answer questions from the media and the media report these responses in their articles using the word "said."

Comparing the text in the X word cloud with those in the news articles word cloud (**Figure 3** and **Figure 4**), we notice that the news articles' highest frequency words ("covid," "health," "people," and "cases") are larger than those in the X text, indicating that the larger words in the news articles are used relatively more frequently than in the X corpus. This can be attributed to the fact that news articles report on key issues and give information about the pandemic, while X users, who are in the thousands, have many more issues concerning the pandemic to discuss and use a more varied vocabulary.

There are limits to the use of word clouds. First, the comparison is not fully objective. The word clouds are compared by visual inspections and not by comparing objectively measured values. Secondly, the size of the words is determined by raw frequency, and they depend on the frequency of the other words, thus making visual comparisons between documents difficult. We attempted to address this by using the same image size for all the word clouds and also by comparing the size of the keywords in each word cloud relative to the other words in the same word cloud, rather than only between the different word clouds. Word clouds can be used as a first step to get ideas for more questions to be asked about the data.

Correlation

We ran a general Pearson correlation with data from all our datasets. Before correlations were run, we ensured that our data was fully consistent between the four datasets. The data collection for the different datasets of the corpus began at different times, so we removed weeks that were not present in all the datasets. We also removed weeks in

which the CMOH and the premier did not give any updates from all the datasets. And lastly, weeks in which the size of the text file was less than 3 KB were removed across all the datasets. This left us with 58 weeks of fully consistent data across the datasets for analysis.

To perform this analysis, we used Linguistic Inquiry and Word Count (LIWC) (Boyd et al. 2022) to generate four properties (two groups of opposites) from our weekified data. LIWC is a lexicon-based text analysis tool that is used to analyze word use. The first two properties were the "I" and "They" pronouns, which refer to the use of personal pronouns that refer to self and others, respectively. The second two properties were the "positive emotion" and "negative emotion" of the text, which refers to the use of positive and negative words in the text.

These four properties were generated for each weekly text file in each dataset. Next, we loaded the CSV files generated by LIWC into SPSS 23 (Statistical Package for the Social Sciences) (IBM 2015) and performed a two-tailed Pearson's correlation across all datasets. The results are displayed in **Figure 5**. Only significant correlations (significance level, $\alpha < 0.05$) appear in the figure.

	CmI	CmT	CmP	CmN	PrI	PrT	PrP	PrN	TwI	TwT	TwP	TwN	NwI	NwT	NwP	NwN
CmI	-															
CmT		-														
CmP			-													
CmN				-												
PrI					-											
PrT					-.26 .05	-										
PrP							-									
PrN							-.29 .03	-								
TwI									-							
TwT									.45 .00	-						
TwP		.331 .01	.387 .005						.55 .00		-					
TwN			-.36 .01				-.31 .02		.57 .00	.4 .00		-				
NwI													-			
NwT			.285 .029							-.27 .04			.28 .04	-		
NwP			.33 .01							-.30 .02					-	
NwN												.3 .02	.55 .00			-

Figure 5: Correlation table (Cm = CMOH, Pr = Premier, Tw = X, Nw = News article, I = I pronouns, T = Others pronoun, P = positive emotion, N = Negative emotion). In cells with values, the top value is the correlation coefficient, and the bottom value is the level of significance.

The analysis presents several correlations, of which we will discuss a few of the most interesting. We see that there is a positive relationship between positive emotion in the speeches by the CMOH and positive emotion on X and in the news articles. On

the other hand, there is a negative relationship between the CMOH's use of positive emotion and negative emotion on X. This suggests the possibility that the speeches of the CMOH set the tone for the emotion which is echoed in X and in news articles.

For the premier's speeches, we also notice that there is a negative relationship between the use of pronouns referring to himself and his use of pronouns to refer to others. This relationship between the use of pronouns referring to oneself and to others is positive in both the X and the news articles datasets. Therefore, we might hypothesize that the premier, who is a politician, tends to separate self-reference from reference to others, suggesting a distinct rhetorical or communicative strategy, whereas other datasets, especially X, represent many people freely talking about both themselves and others. We also notice a negative relationship between the positive and negative emotion used in the premier's speeches. As he uses more emotionally positive words, his use of emotionally negative words diminishes.

In the X dataset, the use of first-person pronouns has a positive relationship with both negative and positive emotion, while the use of pronouns to refer to others has a positive relationship with negative emotion only. This suggests that X is used to express personal opinions, whether negative or positive, but when more pronouns referring to others are used, more negative words are also used. This suggests that more negative words are used when referring to others than when referring to oneself. This is consistent with studies that have found that othering language online is often accompanied by negative opinions (Burnap and Williams 2016).

In the news articles, there is a positive correlation between the use of personal pronouns referring to oneself and negative emotion, but not with positive emotion. This suggests that as the article writers refer more to themselves, they also use more negative words. It might also suggest that they explicitly state their opinions, especially when they are giving negative critique.

Correlation helps us see important relationships that could be explored further. However, it must be noted that correlation does not imply causation. The assumptions made were not verified by close reading of the text or other methods and are made to raise ideas for further research. This analysis was done on weeks of text, which is similar to data obtained from individuals completing several questionnaires, but here the individuals are the weekly text files. Future studies could examine whether these correlations change at different time lags between the updates given by the government officials and the response on X and in the news media.

TF-IDF

We ran TF–IDF analysis on the monthly text files for each dataset to check if the analysis will reveal keywords that inform on topics of the discussions that were characteristic of that month. TF–IDF is used to identify words that are uniquely significant to a document in a collection of documents. It is calculated for each word by taking the product of the word's relative frequency in the document and the inverse of its frequency across the documents in the collection. We ranked the words by their TF–IDF score in descending order and selected the first five. **Table 1** shows the first five words for each month in the four datasets.

Month	Dr. Hinshaw	Pr. Kenney	X (Twitter)	News Articles
March 2020	self-isolation	sait	becoz	
	self-isolate	cent	saudis	
	self-isolating	deferral	babylon	
	rosedale	sands	tele-health	
	suspect	downturn	saudi-backed	
April 2020	plant	ventilators	jbs	
	cargill	elevated	connector	
	camp	scenario	cargill	
	kearl	probable	towne	
	jbs	models	mckenzie	
May 2020	relaunch	trafficking	médecin	smokers
	brooks	gouging	hygiéniste	vitamin
	app	flooding	jbs	hellomd
	cargill	jams	relaunch	smoking
	stage	arrivals	chef	plants
June 2020	relaunch	speaker	médecin	jun
	stage	firearms	hygiéniste	plants
	treaty	racism	chef	arguelles
	territory	parole	community-specific	smith-fraser
	demonstrations	recommendation	affecte	popeyes
July 2020	nhl	amendment	ama	jul
	beach	bill	ndp-era	hutterite
	hutterite	addiction	ophthalmologists	colonies

(Contd.)

Month	Dr. Hinshaw	Pr. Kenney	X (Twitter)	News Articles
	samaritan	thousands	good-quality	beekeepers
	heritage	enoch	american-style	legion
August 2020	august	cameras	samaritan	oele
	samaritan	lakeland	southgate	shoesmith
	teachers	hinton	re-entry	kamaran
	southgate	college	Aug2	ifr
	sherene	vermillion	aimga	southgate
September 2020	cohorts	fish	wilfrid	sansom
	copies	throne	recitation	sep
	foothills	speech	festering	bison
	map	creek	kooky	bonds
	harms	park	culprits	yammine
October 2020	halloween	airdrie	wildcat	morales
	candy	veterans	oct	thome
	flu	suncor	thanksgiving	firewood
	trick	hydrogen	halloween	thanksgiving
	voluntary	ring	housekeeping	oct
November 2020	beds	gatherings	sailed	aalak
	elected	app	recordings	lovestrom
	Sits	emissions	sondage	remdesivir
	backlog	limit	fooked	grauman
	diwali	sports	nov	vigueras
December 2020	holiday	opioid	nuh	christmas
	holidays	addiction	christmas	mst
	hanukkah	substance	malfeasance	dec
	season	vaccine	xmas	agovideo
	december	pfizer	santa	chinatown
January 2021	vial	ally	allard	jan
	terminology	easing	variant	vandelannoite
	syringes	biden	hawaii	redman

(Contd.)

Month	Dr. Hinshaw	Pr. Kenney	X (Twitter)	News Articles
	resumed	vaccine	variants	variant
	december	presidential	tranmissions	torkelson
February 2021	border	gymnastics	variants	olymel
	pilot	club	variant	variant
	variants	jerry	positioned	variants
	variant	boreal	metamorphosis	mchugh
	administered	constituti	rousers	coates
March 2021	astrazeneca	diagnostic	variant	mar
	born	scans	variants	variant
	clots	imaging	astrazeneca	variants
	o	ct	doses	papenfuss
	birth	voters	misogynists	astrazeneca
April 2021	astrazeneca	astrazeneca	variants	gracelife
	variants	vaccine	variant	coates
	clots	doses	sentencing	welsh-rollo
	p1	vaccines	p1	variants
	variant	rockyview	astrazeneca	variant
May 2021	rural	vaccine	rodeo	rodeo
	dose	vaccines	bowden	vitt
	vaccinated	vaccinated	doses	franchise
	doses	montana	variants	douma
	target	doses	vaccinated	blackfeet
June 2021	dose	dose	delta	variant
	astrazeneca	lottery	doses	prizes
	doses	hydrogen	variant	delta
	mrna	muslim	dose	eddie
	vaccinated	vaccinated	pfizer	jun
July 2021	august	whip	delta	jul
	mid-august	associate	variant	variant
	universally	augustana	abandons	delta
	wastewater	digital	chickenpox	lambda
	standpoint	cameras	unvaccinated	srp

(Contd.)

Month	Dr. Hinshaw	Pr. Kenney	X (Twitter)	News Articles
August 2021	sustainable	film	unvaccinated	una
	covid-related	productions	delta	delta
	stood	hbo	vaccinated	unvaccinated
	expectations	livestock	syphilis	variant
	self-referral	filming	variant	neudorf
September 2021	vaccinated	unvaccinated	unvaccinated	passport
	health-care	vaccinated	passport	unvaccinated
	pregnant	icu	vaxxed	vaccine-eligible
	models	dose	vaccinated	passports
	delta	vaccination	unvaxxed	vaccination
October 2021	pregnant	rtp		ivermectin
	halloween	qr		qr
	maternal	exemption		nagase
	fertility	thanksgiving		passport
	vaccinated	unvaccinated		unvaccinated

Table 1: Monthly top keywords identified by TF-IDF in four data sources during the COVID-19 pandemic (March 2020–October 2021).

Examining the words from the different datasets in the table, it can be noticed that the words from Dr. Hinshaw's speeches are all obvious references to the COVID-19 pandemic. This is unlike the words from Premier Kenney's speeches, in which, during several months, the most prominent terms are not explicitly pandemic-related. This simply shows that, as premier of Alberta, Kenney addressed many other issues during the pandemic. Obvious references to the pandemic appear in his speeches in April 2020, and then in December 2020, and become irregular towards the end of the dataset. Throughout the period covered in the dataset, premier Kenney talked about the pandemic, but the mentions were not revealed by the TF-IDF analysis because they were consistently present.

It is also interesting to notice that the analysis suggests that in the month of April 2020, beyond the regular talks about the pandemic, what stood out most in Dr. Hinshaw's speeches was the outbreaks in the meat-processing facilities. In contrast, Premier Kenney's speeches during the same period focused on ventilators (which were almost running out) and prediction models and possible future scenarios. The

analysis also shows that Dr. Hinshaw's speeches addressed the major events and feasts during the pandemic. We see mentions of the NHL, Halloween, Diwali, and Hanukkah. Premier Kenney in his speeches used to extended greetings to communities celebrating these occasions, but the analysis shows that he did not dwell on the topic in a way that related words had high TF-IDF values. The feasts and public events usually involved public gatherings and so had to be experienced differently. Dr. Hinshaw, in her COVID-19 briefings, consistently addressed these occasions and provided related public health guidance.

X data collection began in March 2020 using general health-related hashtags (#abhealth and #albertadoctors) specific to Alberta. A hash tag specific to the COVID-19 pandemic (#covid19ab) was added in April 2020. This is reflected in the results, where the keywords such as "saudi," "babylon," and "tele-health" point the public concern related to the use of Babylon Health a medical service provider with financial ties to Saudi Arabia. In April and May 2020, the COVID-19 outbreaks in meat-processing facilities—particularly JBS and Cargill—stand out. In the later months, which were still early in the pandemic, the discussions were so broad that few core pandemic-related terms appeared. Then starting in early 2021, discussions about variants and vaccinations become so popular that they dominate the discussion on X. A similar trend is observed in news articles, where, from early 2021 onward, coverage became dominated by topics related to variants and vaccination. New variants, which were more contagious, emerged, making the push for vaccine availability and increased vaccination rates. This, in turn, was met with resistance from individuals concerned about vaccine safety, resulting in a surge in highly polarized discussions online.

TF-IDF analysis made it possible to observe local topics of discussion in the different months of the pandemic by highlighting month-specific document keywords. The results suggest that the datasets differ in terms of the topics covered although they are all related to the pandemic. Dr. Hinshaw's speeches, for example, were observed to remain focused on the pandemic.

A limitation of TF-IDF is that it suppresses common terms hence common themes across the dataset, which might represent central points of interest in the datasets are underrepresented. This preliminary TF-IDF analysis considered only the top five keywords with the highest TF-IDF values, and this undoubtedly limited the number of highlighted topics. This limitation is even more significant considering that the text was not lemmatized. Future studies should consider lemmatizing the text and including more keywords per document to gain a clearer understanding of the differentiating themes across the documents.

These analyses are different in nature and suggest the versatility of use of the datasets we curated. Word clouds are frequency based and are serving primarily as visualizations, and their analysis is largely subjective. But they have an element of aesthetics which makes them pleasant to look at. Literature has widely discussed the appeal of visualizations and how it could influence perception (Prantl, Moeller, and Koesten 2023). The data extraction and correlation show that the data could be used for more empirical analysis. Researchers would be able to extract the property in the data that interests them and pass the data through the analysis and/or visualization they are interested in. TF-IDF analysis provided a quick overview of the topics that stood out across different months in the dataset and enabled comparison between the datasets.

These preliminary analyses and the fact that during the data collection process we were contacted by scholars requesting part of our data showed us that the data being collected was valuable and would be used by scholars and so should be made easily available to the public.

Depositing the data: Dataverse

We had as goal to make our data freely accessible online in order to contribute to the advancement of research. Our datasets were uploaded to the University of Alberta's Dataverse on Borealis (the Canadian Dataverse repository), a data repository for scholars at the University of Alberta.

In addition to ensuring the quality and consistency of the data collected, the data manager had to ensure that uploading the data online did not go against any use policy by entities that held the copyright or management rights of the data. According to X's then terms of service, it was not permitted to share tweet datasets, but tweet identifier (or Tweet IDs) datasets could be shared. This dehydrated dataset can then be rehydrated (Tweet text added to the identifiers) using a service provided by X. This preserves users' rights as content creators, allowing them to remove public access to their tweets when they wanted. The tweets will no longer be available to anyone after the rehydration process even if they have the tweets' identifiers. We faced a similar issue with the news media articles gathered from many different news publishers. The text of the news articles are in most cases copyrighted material and so could not be published. But we wanted to publish data that is easily accessible and directly usable to the potential researcher. We decided to publish aggregate statistics of the data we had. The data was processed as follows.

Media briefings

The text from the media briefings given by Premier Kenney and CMOH Dr. Deena Hinshaw was grouped into weekly and monthly folders, and weekly and monthly text

files were created from these folders. The created text files were then passed through sentiment analysis using VADER (Hutto 2014) and through word frequency analysis using scikit-learn. The output was CSV files. Since there were no legal restrictions on obtaining and using text from official news updates, the raw text of the media briefings was also included in the dataset.

X data: Tweets

To process the collected tweets, we used VADER (Valence Aware Dictionary for sEntiment Reasoning), a sentiment and rule-based sentiment analysis library and scikit-learn, a machine learning library. Both libraries run in the Python programming language. The tweets were grouped into weekly and monthly CSV files and retweets were filtered out. Hashtags, X handles, and URLs were removed from each tweet. VADER was used to get sentiment scores for each tweet. Weekly and monthly text files were created by combining the text from all the cleaned tweets from the weekly and monthly CSV files. The weekly and monthly texts were then passed as a whole through sentiment analysis using VADER and CSV files with the positive, negative, and neutral scores for each week and month were created. scikit-learn was used on the weekly and monthly text files to get the raw frequency, relative frequency, and TF-IDF scores of the first 10,000 high-frequency words of the texts. The output files were also saved as CSV files. We also produced shuffled text data of the monthly and weekly text files. Shuffling the text ensured that no tweet could be identified and used to identify the creator, hence protecting the creator's identity. We also shuffled the text to allow its publication without violating X's terms of use.

The HathiTrust Research Centre (HTRC) does something similar with books in its collection and calls the data output "non-consumptive book data." It publishes data both for books that are in the public domain and for books that are copyright protected (HathiTrust 2025).

News articles

Weekly and monthly text files were created from the text of the news articles that were gathered. These files were passed through sentiment analysis with VADER and word frequency analysis with scikit-learn. scikit-learn was used to get the raw frequency, relative frequency and TF-IDF scores of the first 10,000 high frequency words of the texts. The output was a series of CSV files that made up the dataset. We also generated shuffled versions of the monthly and weekly text files as a way to publish the content without infringing on copyright.

A readme file was created for each dataset and the datasets were uploaded to the COVID-19 Dataverse (see Appendix). As of May 2025, the tweets dataset has been

downloaded 213 times; the media briefings dataset by then CMOH Deena Hinshaw has been downloaded 2,733 times and that of then-Premier Jason Kenney 2,636 times; and the news articles have been downloaded 630 times.

Reflections and lessons learned

The conception and execution of this project required the creation and implementation of a methodological framework within a short time frame. Indeed, the entire concept of a project to collect and curate data for research on the COVID-19 discourse was a reactive, rather than a proactive, decision. We recognized that the sudden and extreme societal changes resulting from the widespread effects of the COVID-19 virus would have a significant impact and thus would be an important area to understand. However, to collect real-time data about the pandemic as it unfolded meant that we needed to begin our data collection process almost immediately, without a buffer period to deeply research the ways that others were collecting and curating real-time data, or to create in-depth data infrastructures, such as data pipelines. Thus, the immediacy and real-time aspects of our project significantly impacted the framework and methodology of the collection and curation process.

If we had not been subject to these time constraints in beginning the process, we would likely have attempted to increase the level of automation used in the data collection process. Our project faced the same challenge that many other data curation projects regularly face: the need for significant human resources and substantial investments of time and labour to collect and curate large amounts of data (Lee et al. 2018). This was especially true of the web-based news articles we collected. While the alerts that we set up were automated, we did not have a scraper set up to collect the text of the web articles that were returned by the alerts. Instead, a researcher would manually copy the text of the article into a plain text document, and then manually clean the data by removing extraneous information, such as ads. While this method works well for retrieving materials and is easy to implement on short notice, to save hours of manual work, we would recommend implementing a basic web scraper or AI tools to collect and clean the articles. Another benefit from using AI-assisted automation processes would be the reductions of human error, bias, and subjectivity in the data collection. The news alerts received from Google had to be reviewed by the data collectors and decisions had to be made on whether to include the webpages in the alerts in the dataset. With an automated AI-assisted process, human labour could be limited to supervising the process and monitoring the accuracy of the returned results.

We observed a wide variety of COVID-19-related hashtags used on X throughout the pandemic, with many related issues emerging at specific points in time. Although some of this discourse was captured in our dataset—as indicated by the presence of the

hashtags in the raw data collected—it is likely that some tweets contained only those temporally popular hashtags and not the general hashtags we used. Implementing an automated process that identifies and uses popular related hashtags to scrape X might have been a better method. This snowball technique in which new hashtags are used to discover and use other hashtags might have produced a richer and more representative sample of the discourse on the COVID-19 pandemic in Alberta.

With limited human resources, we had to rely on Otter.ai, an online transcription software to transcribe the speeches given by Kenney. With an accuracy of 85–90%, this implies that 10–15% of Kenney's speech may be either misrepresented or omitted. It is important that the speech of the premier of the province be correctly represented because of the sensitive issues and important decisions that were made during the pandemic. As shown by the number of downloads, the public was much more interested in the speeches of the premier and the CMOH, hence more effort should have been made to ensure that the premier's speeches were accurately transcribed. More human resources should have been allocated to the collection of this dataset. The dataset would have been much more accurate, valuable, and useful if the transcripts produced by Otter.ai were reviewed by a human.

Digital tools can be of significant value when collecting and curating datasets. However, like other data collectors and curators, we found that our greatest strength lay in the collective action of our team members and the people and groups working adjacent to the project. From the beginning, our team consisted of researchers from a wide range of backgrounds, which included diverse specialties in the digital humanities, library studies, philosophy, medical sciences, and computer science. As a result, our team members were able to contribute diverse ideas to the project, for example, methods for data collection, locations from which to collect data, ideas on how to manage the data, and ideas of how to avoid copyright infringement.

Digital tools enabled us to conduct three analyses on the curated datasets using data visualization and text analysis. These analyses were intended to assess the data's suitability for empirical research. These analyses demonstrated that the data were adequate and valuable for statistical use. It is advisable to begin preliminary analysis early in the data collection process, as insights from the results can help refine data collection to better meet original or evolving research goals.

Conclusion

In Alberta, as in the rest of the world, the COVID-19 pandemic began rapidly and caused significant societal changes, including an overwhelmed healthcare system, panic buying, and lockdowns. Thus, our team decided to collect and curate data on

COVID-19 pandemic in Alberta as it happened. We achieved this by collecting data from the speeches given by our premier and CMOH, X data, and online news articles. We met our goal, which was to create a representative sample of the discourse of the COVID-19 pandemic in Alberta. Our preliminary analysis indicated that the data was suitable for research purposes. Our corpus was deposited in the University of Alberta's Dataverse, an open-source repository. The corpus has attracted significant interest, as indicated by the high number of downloads. In making the datasets available to the public for future research, we made sure that we did not violate any use policies. We did so by publishing detailed statistics of the word use in the datasets instead of the raw text, and we also published some datasets in a non-consumptive format.

Although the corpus we created is representative and valuable, it represents only a fraction of the COVID-19 pandemic discourse in Alberta. The premier and the CMOH made numerous other comments on the pandemic, such as in town halls and through their official social media accounts. X is only one of the many other social networking sites like Facebook, YouTube, Snapchat, and Telegram from which data was not curated in this project. In hindsight, we recognize that the 85–90% accuracy of the premier's speech transcripts was insufficient. More resources should have been allocated to ensure near-perfect transcription, as the premier was the primary decision maker during the pandemic in Alberta, and his communication should be represented with the highest accuracy.

Appendix

URLs of datasets

DataVerse: https://borealisdata.ca/dataverse/covid-discourse.
Hinshaw: https://voyant-tools.org/?corpus=61c75f0e097308c01412b92a53f02d71.
Kenney: https://voyant-tools.org/?corpus=8e7dc183e06c2df22348657e3bb65bf2.
X: https://voyant-tools.org/?corpus=ed07a3c032256f89821931230d748652.
News articles: https://voyant-tools.org/?corpus=30d2e9e5b2767fba48b704b83ffd4260.

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial

Authorship in the byline is by magnitude of contribution, with the last author being the overall project supervisor. Author contributions, described using the NISO (National Information Standards Organization) CrediT taxonomy, are as follows:

Author names and initials:

Bennett Kuwan Tchoh (BKT)
Sara Barnard (SB)
Robert Budac (RB)
Katrina Ingram (KI)
Ryan Chartier (RC)
Daniel C. Baumgart (DCB)
Geoffrey Rockwell (GR)

Authors are listed in descending order by the significance of their contribution. The corresponding author is BKT.

Conceptualization: GR
Data Curation: SB, RB, KI, BKT
Formal Analysis: RC, BKT
Funding Acquisition: DCB, GR
Methodology: GR, RC, BKT
Project Administration: GR, DCB
Software: RC, RC, BKT
Supervision: GR, DCB
Validation: GR, DCB
Visualization: RC, BKT
Writing – Original Draft: SB, BKT, RC
Writing – Review & Editing: GR

Editorial

Section Editor

Davide Pafumi, The Journal Incubator, University of Lethbridge, Canada

Copy Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

References

Aaronson, Ellen, and Angela Spencer. 2021. "Informing Information Professionals: A Case Report on Creating a Shared Site of Pandemic Resources." *Journal of Hospital Librarianship* 21 (2): 198–201. Accessed August 2, 2025. https://doi.org/10.1080/15323269.2021.1899789.

Bahk, Chi Y., Melissa Cumming, Louisa Paushter, Lawrence C. Madoff, Angus Thomson, and John S. Brownstein. 2016. "Publicly Available Online Tool Facilitates Real-Time Monitoring of Vaccine Conversations and Sentiments." *Health Affairs* 35 (2): 341–347. Accessed August 2, 2025. https://doi.org/10.1377/HLTHAFF.2015.1092.

Boyd, Ryan L., Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdfpdf.

Burnap, Pete, and Matthew L. Williams. 2016. "Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics." *EPJ Data Science* 5: 1–15. Accessed August 2, 2025. https://doi.org/10.1140/epjds/s13688-016-0072-6.

Calma, Justine. 2023. "Twitter Just Closed the Book on Academic Research." *The Verge*, May 31. Accessed August 2, 2025. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research.

Ćurković, Marko, Andro Košec, Marina Roje Bedeković, and Vladimir Bedeković. 2021. "Epistemic Responsibilities in the COVID-19 Pandemic: Is a Digital Infosphere a Friend or a Foe?" *Journal of Biomedical Informatics* 115: 103709. Accessed August 2, 2025. https://doi.org/10.1016/j.jbi.2021.103709.

Gao, Songfeng, Tengfei Yang, Yuning Xu, Naixia Mou, Xiaodong Wang, and Hao Huang. 2025. "Enhancing Disaster Situation Awareness through Multimodal Social Media Data: Evidence from Typhoon Haikui." *Applied Sciences* 15 (1): 465. Accessed August 2. https://doi.org/10.3390/app15010465.

HathiTrust. 2025. "Where to Start: Understanding the Data." HTRC Analytics. Accessed September 8. https://analytics.hathitrust.org/

Hutto, Clayton J. 2014. "vaderSentiment." GitHub. Accessed August 2, 2025. https://github.com/cjhutto/vaderSentiment.

IBM (International Business Machines Corporation). 2015. "IBM SPSS Statistics for Windows, Version 23.0." IBM Corp.

Kejriwal, Mayank, and Yao Gu. 2019. "A Pipeline for Rapid Post-Crisis Twitter Data Acquisition, Filtering and Visualization." *Technologies* 7 (2). Accessed August 2, 2025. https://doi.org/10.3390/technologies7020033.

Lee, Kyubum, Maria Livia Famiglietti, Aoife McMahon, Chih-Hsuan Wei, Jacqueline Ann Langdon MacArthur, Sylvain Poux, et al. 2018. "Scaling Up Data Curation Using Deep Learning: An Application to Literature Triage in Genomic Variation Resources." *PLOS Computational Biology* 14 (8): e1006390. Accessed August 2, 2025. https://doi.org/10.1371/JOURNAL.PCBI.1006390.

Prantl, Verena Ingrid, Torsten Moeller, and Laura Koesten. 2023. "Untangling Rhetoric, Pathos, and Aesthetics in Data Visualization [arXiv preprint]." arXiv:2304.10540. Accessed August 2, 2025. https://doi.org/10.48550/arXiv.2304.10540.

Shankar, Kalpana, Wei Jeng, Andrea Thomer, Nicholas Weber, and Ayoung Yoon. 2020. "Data Curation as Collective Action during COVID-19." *Journal of the Association for Information Science and Technology* 72 (3): 280–284. Accessed August 2, 2025. https://doi.org/10.1002/asi.24406.

Shaw, Ed. 2020. "twarc." GitHub. https://github.com/DocNow/twarc.

Sinclair, Stéfan, and Geoffrey Rockwell. 2025. "Voyant Tools." Accessed August 2, 2025. https://voyant-tools.org/.

Summers, Ed, Igor Brigadir, Sam Hames, Hugo van Kemenade, Peter Binkley, Tina Figueroa, et al. 2023. "DocNow/Twarc: v2.14.0." Zenodo. August 2, 2025. https://doi.org/10.5281/zenodo.593575.

Vaghela, Uddhav, Simon Rabinowicz, Paris Bratsos, Guy Martin, Epameinondas Fritzilas, Sheraz Markar, et al. 2021. "Using a Secure, Continually Updating, Web Source Processing Pipeline to Support the Real-Time Data Synthesis and Analysis of Scientific Literature: Development and Validation Study." *Journal of Medical Internet Research* 23 (5): e25714. Accessed August 2, 2025. https://doi.org/10.2196/25714.

Vahidy, Farhaan, Stephen L. Jones, Mauricio E. Tano, Juan Carlos Nicolas, Osman A. Khan, Jennifer R. Meeks, et al. 2021. "Rapid Response to Drive COVID-19 Research in a Learning Health Care System: Rationale and Design of the Houston Methodist COVID-19 Surveillance and Outcomes Registry (CURATOR)." *JMIR Medical Informatics* 9 (2): e26773. Accessed August 2, 2025. https://doi.org/10.2196/26773.