Curating for Contrarian Communities: Data Practices of Anthropogenic Climate Change Skeptics

Wofford, Morgan F. University of Michigan, United States | mwofford@umich.edu
Thomer, Andrea K. University of Arizona, United States | athomer@arizona.edu

ABSTRACT
The open data movement is often touted as a sweeping strategy to democratize science, promote diverse data reuse, facilitate reproducibility, accelerate innovation, and much more. However, the potential perils of open data are seldom examined and discussed in equal measure to these promises. As we continue to invest in open data, we need to study the full spectrum of what open data facilitates in practice, which can then inform future policy and design decisions. This paper aims to address this gap by presenting an investigative digital ethnography of one contrarian community, anthropogenic climate change (ACC) skeptics, to describe how they process, analyze, preserve, and share data. Skeptics often engage in data reuse similar to conventional data reusers, albeit for unconventional purposes and with varying degrees of trust and expertise. The data practices of ACC skeptics challenge the assumption that open data is universally beneficial. These findings carry implications for data repositories and how they might curate data and design databases with this type of reuse in mind.

KEYWORDS
data practices, data repositories, data curation, data reuse, anthropogenic climate change skeptics

INTRODUCTION
One of the purported promises of the open data movement is its democratizing potential, affording people outside of scholarly communities the ability to access and analyze data (Baack, 2015; Cavalier & Kennedy, 2016; Espinosa et al., 2014; Nielsen, 2011; Ricker et al., 2020; Zuiderwijk & Janssen, 2014). Open data activists typically envision a "democratization of information" (Baack, 2015) as a largely positive development. There are numerous examples of these prosocial outcomes, such as citizen groups like the Anti-Eviction Mapping Project and the Mapping Police Violence Database utilizing open government data to support the needs of marginalized communities (*Anti-Eviction Mapping Project*, 2021; *Mapping Police Violence*, 2021).

However, open data can also be susceptible to misuse or misinterpretation by various communities. For instance, local governments and private companies often misuse climate data when determining financial climate-related risks, resulting in inaccurate projections (Fiedler et al., 2021). Additionally, during the COVID-19 pandemic, anti-mask proponents employed 'orthodox visualization methods' on open government data to bolster their 'unorthodox arguments' against mask mandates (Lee et al., 2021). In another example, DNA samples and 3D facial images were utilized to train machine-learning algorithms used by law enforcement agencies to reconstruct suspects' faces from DNA samples, leading to biased outcomes (Pasquetto, 2018).

Understanding the different ways open data are used and misused is crucial for infrastructural development and supporting appropriate data reuse by scholars and non-scholars alike. These tasks often fall to data repositories managing the discoverability, accessibility, usability, and preservation of datasets housed in their collections. Understanding the diverse data practices of a full range of repository users – even those that may misuse data – is essential in informing design decisions and curatorial activities.

In this paper, we conduct an investigative digital ethnography on a contrarian community (that is, a community that defines itself in opposition to the mainstream and employs misinformation (Ribeiro et al., 2020)) of data reusers -- anthropogenic climate change (ACC) skeptics -- examining their data practices, specifically how they acquire, process, analyze, preserve, and publish/share open climate data. We analyze how ACC skeptics' data practices compare to traditional academic data practices by comparing them to the primary elements of the USGS's data lifecycle model (DLM) (Faundeen et al., 2014). We ask: What are ACC skeptics' data practices and goals? How do these differ from those of academic or traditional data reusers? We find that skeptics' data reuse is similar to that of traditional data reusers, albeit with different research goals, levels of trust, and expertise. We end with a discussion of the implications of these findings for data repositories, including the need for data misuse policies and designing for diverse user groups.

LITERATURE REVIEW
Open Data and Data Re- and Mis-use
Open data refers to datasets that can be accessed and "freely used, modified, and shared by anyone for any purpose" (Open Knowledge Foundation, n.d.). The promises of open data have been espoused for years, with groups touting its potential to facilitate reproducibility (AlQuraishi & Sorger, 2016; Martens & Vizcaíno, 2017; Stodden et al., 2014), accelerate innovation (Knoppers, 2014; Knoppers et al., 2014), democratize science (Cavalier & Kennedy,

86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom.

2016; Nielsen, 2011), promote collaboration (Pasquetto et al., 2019), and derive new insights from old research (Pasquetto et al., 2019; Rung & Brazma, 2013; Zimmerman, 2003, 2008). Sabina Leonelli suggests that the very allure of open data "lies precisely in the impossibility to predict and quantify their potential as evidence in advance," reflecting the reality that it is impossible to thoroughly foresee how data will be reused (2013, p. 9). In the United States, the federal government has issued memorandums (Holdren, 2013), executive orders (Obama, 2013), and more binding legislation (Ryan, 2019) that require all federally-funded research to publish non-sensitive data publicly. In response to these potential benefits and governmental mandates, stakeholders have invested in making the promises of open data a reality via initiatives such as the FAIR guiding principles for scientific data management and stewardship (Wilkinson et al., 2016).

The perils of open data are less frequently explored but not disregarded. Philip Mirowski argues that open data may be more readily utilized by those with greater resources, such as corporate data scientists, which could exacerbate existing power imbalances and fail to truly democratize and diversify science (2014, 2018). There is also a well-established literature on indigenous data sovereignty, which addresses the historical harms caused by data collected about indigenous communities and emphasizes community control and ethics rather than maximum openness (*CARE Principles*, 2022; *OCAP*, 2022; Radin, 2017; Tsosie et al., 2021). Making data public means that contrarian communities, those that oppose mainstream views or attitudes and often create and spread misinformation, such as ACC skeptics, vaccine skeptics, flat earthers, and eugenicists, can also use the data (Ribeiro et al., 2020). Data creators also have concerns about sharing their data due to the potential for misuse or misinterpretation (Cragin et al., 2010; Sayogo & Pardo, 2013; Tenopir et al., 2015, 2018), as evidenced by instances of misuse as defined by subject matter experts in climate science (Fiedler et al., 2021; Hardy & Jamieson, 2017), genomics (Pasquetto, 2018), ecology (Clavero et al., 2022), and medicine (Jourdain, 2021). Furthermore, the very concept of open data can be weaponized in politicized debates (Levy & Johns, 2016), with partisans in legislative bodies using the perceived non-partisan value of scientific transparency and open data to advance their political agendas. At the same time, the open data requirements are worded, perhaps purposefully, to be impossible to satisfy due to technical infeasibility, lack of resources, or missing privacy, security, and confidentiality exceptions.

Open data facilitate data reuse, defined as the "usage of a dataset by someone other than the data creator" (Pasquetto et al., 2017). Various studies examine who searches for open data, why they do so, and how they reuse data. Most studies focus on academic data reuse practices using data access requests (Coady et al., 2017; Federer, 2019), surveys (Curty et al., 2017; Faniel et al., 2016; Gregory et al., 2020; Tenopir et al., 2011; Yoon & Kim, 2017), and published data citation metrics (Fan et al., 2022; Huang et al., 2015; Khan et al., 2019; Lafia et al., 2023; Park et al., 2018; Peters et al., 2017; Piwowar & Vision, 2013; Robinson-Garcia et al., 2017).

Several types of data reuse have been identified, including:

- Comparative data reuse: Data are reused for ground-truthing, calibration, comparison, and confirmation (Pasquetto et al., 2019).
- Integrative data reuse: Data are reused to answer new questions, identify patterns, perform meta-analyses, or develop new statistical methods (Pasquetto et al., 2019; Thomer, 2022).
- Reproducibility data reuse: Researchers reuse the same data and analysis methods to verify a study's findings (Borgman, 2015; Federer, 2019; Pasquetto et al., 2017).
- Infrastructural data reuse: Data are reused to populate a database or repository (Federer, 2019).
- Educational data reuse: Data are reused for educational or instructional purposes (Thomer et al., 2023).

Overwhelmingly, studies that measure the impact and outcome of open data are done by advocates of that infrastructure, leading to a potential for positive biases (Mayernik et al., 2017). Additionally, by focusing on academic impact, they often fail to consider reuses in educational or policy settings (Mayernik et al., 2017).

Another group of studies examines the data practices within different disciplines (Borgman et al., 2015; Faniel et al., 2013, 2020; Faniel & Yakel, 2017; Weber et al., 2013). The goal is to support repositories as they make design decisions to support these intended users (Faniel et al., 2020; Faniel & Yakel, 2017; Kansa, 2012). Repositories intending to serve broader communities have even greater user requirements to consider. The design choices and material forms of data schemas in repositories profoundly impact how "these systems play in the social settings in which they are used" (Thomer & Wickett, 2020).

Climate Science and the Persistent Presence of Skeptics

Climate science is one of many disciplines embracing open data; the Registry of Research Data Repositories lists 131 repositories associated with the keywords "climate" or "climate change" (*Re3data*, 2014). The majority of climate research focuses on the drivers and effects of ACC, "a statistically significant variation in either the mean state of the climate or in its variability, persisting for an extended period," caused by human activity (VijayaVenkataRaman et al., 2012). Climate scientists use global climate models (GCMs) and simulations to make projections about the future climate (Edwards, 1999, 2010). As a scientific concept, ACC has reached a level of scientific consensus that most theories will never attain, with 99.99% of publishing scientists validating its existence (Powell, 2015). However, despite this relative certainty, ACC skepticism still permeates politics and popular culture, principally in white, conservative, male-dominated circles (McCright & Dunlap, 2011; Sharman, 2014). This skepticism could have disastrous consequences if it delays climate change mitigation and adaptation.

At their most extreme, ACC skeptics use climate science's reliance on models to dismiss the discipline and its findings as imaginary and politically motivated. In less extreme circumstances, they discover and report errors in climate data to spark debate and introduce additional uncertainty (Gramling, 2007), or they use and manipulate open climate data in direct service of their positions and beliefs (Lewandowsky et al., 2016). ACC skeptics use techniques such as data-dredging or p-hacking, both forms of data analysis misuse, to support their claims (Nissen et al., 2016). Lewandowsky et al. performed a blind expert test on multiple contrarian claims about climate data and found them all to be misleading (Lewandowsky et al., 2016). Skeptics present their findings as having the same legitimacy as the research of thousands of climate scientists. Frank Fischer studied how climate science skeptics interpret scientific studies, noting that fact-checking and scientific literacy is insufficient in changing their minds; instead, researchers need to look deeper at how climate skepticism is culturally and historically situated (Fischer, 2019).

METHODS

To gain insights into the data practices of ACC skeptics, we conducted an investigative digital ethnography that combines methods from digital ethnography (Boellstorff, 2012; Coleman, 2010, 2012; Hine, 2000; Krafft & Donovan, 2020; Panofsky & Donovan, 2019; Pink et al., 2015) and investigative journalism (Silverman, 2020). The overall process for conducting an investigative digital ethnography involves identifying relevant topics and artifacts, ascertaining the media ecosystem and influencers, creating a monitoring environment and strategy, auditing assumptions, and analyzing findings (Friedberg, 2020). Following conventional digital ethnography methods, we account for the infrastructures that afford these data practices (Pink et al., 2015). We contextualize these as knowledge infrastructures that are "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010).

Initially, we created social media accounts on Reddit, Twitter, and YouTube and seeded these accounts by following popular ACC organizations and influencers identified from the DeSmog Disinformation Database (Lay, 2022). We then engaged with platform affordances, such as viewing, liking, and reposting content, to influence recommendation algorithms and expand the monitoring environment. The recommendation algorithms suggested new accounts to follow through algorithmic-assisted snowball sampling. The final monitoring environment included 40 Twitter accounts, four Facebook groups, two subreddits, and five YouTube channels.

The data collection period lasted from September 2021 through November 2021, totaling 75 hours. When open data were discussed or reused, we archived the webpage. If posts linked to blogs or organizations that mentioned open data, those websites and subsequent comments were archived. YouTube videos were transcribed. Daily memos were written during the data collection and analysis periods. In total, 189 articles and posts were archived, 17 screenshots were taken, and 19 pages of memos were written.

All the data artifacts were uploaded into NVivo for qualitative coding in January 2022. We synthesized themes across the data using grounded theory (Glaser & Strauss, 1967) adapted for social media analysis (Postill & Pink, 2012), inductively coding the data for emergent themes. While doing this, we compiled lists of data reusers, instances of dataset reuse, analytical tools mentioned by reusers and so on. These themes were then coded into higher-level concepts, constructing theories grounded in the data (Charmaz, 2014). We then organized our findings about ACC skeptics' data practices according to the USGS DLM primary model elements to compare these contrarian practices to normative data practices (Faundeen et al., 2014). We chose this model because it highlights the relationship between data management activities and research project workflows, allowing us to focus on how ACC skeptics interact with data repositories and manage their data. Moreover, the model is designed for types of data that ACC skeptics are likely to reuse (e.g., from the earth and environmental sciences). The primary data life cycle stages include *plan*, *acquire*, *process*, *analyze*, *preserve*, and *publish/share*. Within the *analyze* concept, we additionally coded for the way in which data was reused for analysis, for instance, comparative or integrative as appropriate (Pasquetto et al., 2019; Thomer, 2022). If instances of data reuse did not fit within these categories, new codes were created.

FINDINGS

The ACC Skeptic Community

Across platforms, we observed 77 data reusers posting on their social media accounts, authoring blog posts, and creating websites. Notably, the majority of identifiable influencers in this contrarian community are conservative, white, male, and over 50. Influencers are predominantly located in the United States, Canada, Australia, and England. Some of them have direct ties to the oil, gas, and coal industries as well as the tobacco industry. Three of the individuals were part of the Trump Administration, working in various agencies such as the Environmental Protection Agency (EPA), National Security Council (NSC), National Oceanic and Atmospheric Administration (NOAA), and White House Office of Science and Technology Policy (OSTP).

Generally, reusers are highly educated, many stating they have PhDs. They claim expertise in a wide range of disciplines, including aerospace engineering, atmospheric science, chemistry, climatology, economics, forestry, geography, geophysical science, hydrology, meteorology, modern history, philosophy, physics, and political science. We observed an overlap between ACC skeptics and anti-vaxxers, anti-maskers, and skeptics of the 2020 U.S. presidential election results. Though here we focus on their analysis of climate data, we also found that they apply their data literacy to other public datasets, including 2020 U.S. presidential election data, as well as COVID-19 data.

Plan & Acquire

While the USGS DLM's lifecycle begins with *planning* work, we did not find that our users' planned their projects publicly online. Instead, the majority of early lifecycle discussions focused on the *acquisition* and evaluation of public data and data repositories.

What Data do ACC Skeptics Reuse?

We observed ACC skeptics reuse 61 datasets (Figure 1) from 45 data repositories; of these, ACC skeptics most frequently reuse observational datasets about temperature because the core assumption of ACC rests on the idea that global temperatures are increasing. As a result, skeptics often rely on observational temperature datasets to challenge this assertion and attempt to undermine it in various ways. The top five most reused datasets are as follows:

- *HadCRUT from the Met Office Hadley Centre and the Climatic Research Unit (CRU) (found in 24 documents*) (Brohan et al., 2006; Morice et al., 2012, 2021). HadCRUT combines a sea surface temperature (SST) dataset and a land surface air temperature dataset. Often, skeptics state that these data are less trustworthy because of CRU's involvement with Climategate, a controversy where an ACC skeptic published hacked emails from CRU and cited them as evidence that climate scientists had manipulated data to make global warming appear worse (Leiserowitz et al., 2013).

- *Global Temperature Report from The University of Alabama in Huntsville Earth System Science Center (found in 13 documents)* (Christy & Spencer, 2022). Two ACC skeptics, John R. Christy, and Roy W. Spencer, created this dataset. This global temperature dataset is derived from NOAA's TIROS-N satellite microwave data and shows less extreme increases in temperature than similar datasets.

- *Goddard Institute for Space Studies Surface Temperature (GISTEMP) Analysis Dataset from NASA (found in nine documents)* (GISTEMP Team, 2022). GISTEMP is comparable to the HadCRUT dataset in combining a land-surface air temperature dataset and an SST dataset to create a global temperature dataset, and the two are often compared.

- *Global Historical Climatology Network monthly (GHCNm) Dataset from NOAA's National Center for Environmental Information (NCEI) (found in eight documents)* (Menne et al., 2018). This dataset provides monthly climate summaries from stations around the world. Climate skeptics critique this dataset saying data are skewed due to urbanization and homogenization.

An additional 15 temperature datasets are found in the sample (see the supplementary material for the complete list).

Two datasets are reused in six documents, tying each other for the fifth most reused dataset. Unlike the previous four datasets, neither are temperature datasets. They include:

- *NOAA's Global Monitoring Laboratory (GML) Earth - System Research Laboratories Monthly Average Mauna Loa CO2* (Tans & Keeling, 2022). This dataset shows monthly mean measurements of atmospheric CO_2 at the Mauna Loa Observatory in Hawaii. Skeptics use this dataset to examine trends in CO_2 and test for correlation with other variables. There is one other CO_2 dataset in the sample.

- *National Snow & Ice Data Center's Multisensor Analyzed Sea Ice Extent – Northern Hemisphere (MASIE-NH)* (U.S. National Ice Center & NSIDC, 2010). MASIE-NH provides a graphical view of sea ice extent. Skeptics use this dataset to explore trends in sea ice over time, suggesting the cause of changes are exaggerated or seasonal. Three other sea ice datasets are in the sample.

A related research topic to sea ice is the exploration of sea level trends, mainly arguing that the rate of sea level rise is not increasing. The sample included four sea-level datasets.

Another popular research topic is natural disaster rates, including fires, hurricanes, droughts, diseases, and associated climate-related death data. In this area of research, ten different datasets were reused. Skeptics commonly try to show that natural disaster rates are not increasing or attribute the observed disaster rate increase to changes in disaster categorization and reporting rates.

Skeptics also suggest alternative causes of climate change. These potential climate change drivers include the sun, the ocean's oscillations, and atmospheric water vapor. The sample contains five solar datasets examining sunspot numbers, total solar irradiance, and outgoing longwave radiation. Six datasets look at the ocean's oscillations. Only one dataset records atmospheric water vapor.

The last common research topic is investigations into the capabilities of fossil fuels and renewable energies. The sample contains seven energy datasets representing energy grid conditions, electricity production data, crude oil production, and fossil-fuel CO2 emissions. Quality of life data, like world development indicators or food production, are often used alongside fossil fuel usage to insinuate that quality of life has increased because of fossil fuels.

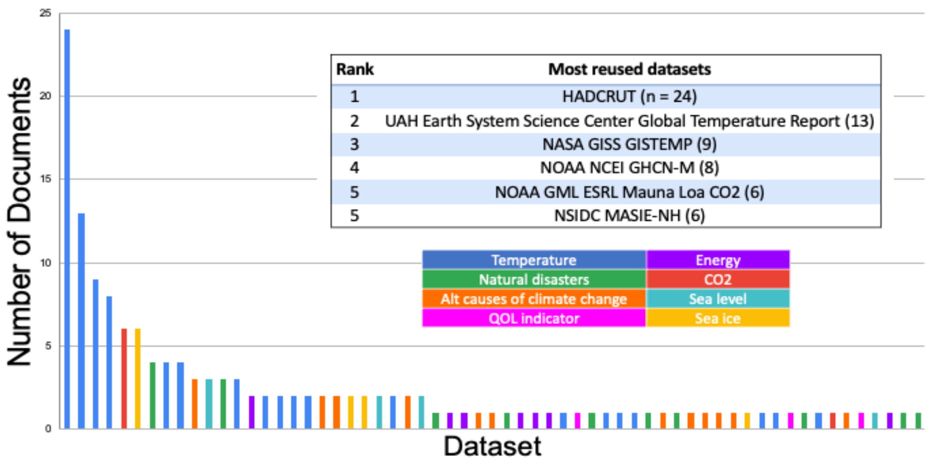

Figure 1: Dataset reuse frequency in collected documents. Due to space constraints, we only list the top 5 most reused datasets; see the supplementary material for the complete list. Frequency is calculated as the number of documents the dataset appears in.

How do ACC Skeptics Evaluate Data and Data Repositories?

In this contrarian community, not all data are created equal; unadjusted or "raw" data are considered superior to all other forms of knowledge because they are the least affected by human subjectivity. By "raw," skeptics refer to data directly collected from instruments with no additional processing, adjustments, homogenization, or manipulation. Overwhelmingly, ACC skeptics favored reusing observational data over other types of data. Climate science, as a discipline, relies on GCMs rather than observational data, which ACC skeptics prefer. Most of the observed ACC skeptics did not engage sophisticatedly with GCMs enough to demonstrate mainstream climate science literacy. Instead, they reject these models outright, considering them inferior to observational data. Observational data are thus often used to critique proxy data and climate model projections.

Consequently, when acquiring data, some ACC skeptics closely examine data repositories, documentation, curatorial decisions, and scholarly products to find processing that has "contaminated" the raw data. For instance, in a Reddit post, a skeptic critiques the GHCNm dataset by directly quoting the dataset's documentation (Menne et al., 2018).

> "… in the introduction to the data (CDRP-ATBD-0859 Rev 1 GHCN-M Mean Temperature-v4.pdf) it states: *'Data are collected from NOAA in situ networks as well as other national and international providers. They are subjected to a series of processes that combine data from various sources, perform quality control, homogeneity corrections, and output the data for customer access and permanent archive.'*
>
> And their readme.txt from their ftp site says in the Introduction: *'. . . The greatest difference from previous version is a greatly expanded set of stations based on the large data holdings in GHCN-Daily as well as data collected as part of the International Surface Temperature Initiative databank (ISTI; Rennie et al. 2014).'*

> So the data you get is different from the raw data version to version. Considering these data have been adjusted in the combined sources with obviously large changes up to Version #4…how would I know they haven't been "adjusting" their data to better align with everyone else?"

ACC skeptics also note when different versions of datasets will be removed or no longer supported actively by repositories and accuse repository staff of changing or removing data to support the theory of ACC. If available, they compare different versions of datasets to note changes using the Wayback Machine.

In the most extreme cases of curatorial oversight, skeptics contact the data repository to question a dataset's limitations and intended uses. For instance, one ACC skeptic discusses repository employees by name from the Centre for Research on Epidemiology of Disasters EM-DATA International Disaster Database, accusing them of endorsing a publication that contradicts the data's recommended use. Within the blog post, the author includes correspondence with repository employees.

Even when ACC skeptics are not accusing repositories of intentionally altering data for political purposes, they often analyze curatorial decisions and documentation for their own reuse. For example, the website Climate4You compares differences in sea surface temperature (SST) using two different datasets, HadSST3 and UAHv6 and notes how changes in curatorial activities in both datasets affect their analysis. However, not all ACC skeptics adhere to appropriate data usage and documentation provided by data repositories in their analyses. For instance, the NSIDC recommends against using the MASIE-NH dataset to compare sea ice trends over time, suggesting instead using the NSIDC Sea Ice Index on a monthly, not daily, basis. Nevertheless, despite the repository's recommendations, Ron Clutz compares MASIE-NH data on specific days in different years in a Science Matters blog post.

Process

In the USGS DLM, the next lifecycle phase is *processing:* this includes data cleaning, data element definition, dataset integration, and calibration to prepare data for analysis. We found limited descriptions of data processing by ACC skeptics for perhaps two compounding reasons. Firstly, many data reusers did not publicly publish or write about their data processing activities. Secondly, as shown in the acquire step, skeptics often disparage any data processing that researchers or repositories carry out on the data, valorizing "raw" data. When ACC skeptics' data processing activities are made public, they often are not extensive, perhaps because of these beliefs.

That said, one common act of processing is converting data files to make them more accessible and computable in spreadsheets. Zoe Phin typically uses more accessible files in her published data reuse blog posts, but does not always convert data files for her own work. For instance, when examining albedo data she uses the netCDF file provided by NASA Earth Data. However, when sharing data with others in her comments, she converts the original netCDF file to a plain text file for ease of access by less technical users. Nowadays, data repositories often provide various file formats, including CSV and TXT files, making the need to convert file formats unnecessary. For example, in a CO2Science blog post, the author uses the HADCRUT CSV version of the dataset over the netCDF version, along with the TXT file of CO2 data from NASA.

Analyze

In the USGS DLM, *analyze* encompasses the activities undertaken by ACC skeptics as they explore and interpret the data they have selected, ultimately drawing their conclusions.

We separate analytical activities into the data reuse typology discussed in the literature review and methods section, which includes comparative or integrative data reuse. We found instances of comparative and integrative reuse (Pasquetto et al., 2019; Thomer, 2022). We additionally observed a new kind of data reuse which we call *quality assessment reuse.* We describe each of these below.

Integrative data reuse entails reusing data for new analyses to "identify patterns, correlations, or causal relationships" (Pasquetto et al., 2019). Integrative data reuse can include meta-analyses or novel statistical analyses. Within the ACC skepticism community, integrative data reuse is widespread and includes identifying patterns in observational data and testing for correlation between variables. ACC skeptics use observational data to identify trends in temperature, CO2 levels, sea ice area, sea-level rise, and natural disasters. Many of their findings directly contradict the scientific consensus: global temperature is cooling, sea ice size is increasing, or natural disaster rates are decreasing. Others assert that there are no identifiable trends in these variables. Others admit to identifiable and even statistically significant trends that support the scientific consensus but assert that they cannot be attributed to anthropogenic CO2 emissions. Finally, some evaluate the dataset and conclude that the data are flawed and unreliable due to changes in measurement practices or other perceived errors. These options lead to the almost inevitable contradiction of the scientific consensus regardless of the data used.

Comparative data reuse consists of reusing data "to assess similarities and differences for purposes such as ground-truthing, calibration, and experimental controls" (Pasquetto et al., 2019). Comparative data reuse within the ACC skepticism community includes using observational data to evaluate model projections and proxy data. It also encompasses the comparison of similar datasets from different organizations or different versions of the same dataset. For instance, observational data are often used to test other types of what ACC skeptics perceive as inferior data, like climate model outputs and proxy measurements for temperature and CO2. One skeptic described observational data as "ground truth data" or "information provided by direct observation as opposed to information provided by inference."

Finally, ACC skeptics reuse data in a way that does not straightforwardly fit into existing typologies, which we call **quality assessment data reuse.** While similar to reuse for reproducibility, it is distinct because they are not trying to reproduce the findings of a publication; rather, they solely analyze the quality of a dataset. For instance, one contrarian site, surfacestations.org, run by Anthony Watts, organized volunteers to survey temperature stations to document "biases and errors through faulty siting, encroachments, or maintenance issues." Even though Watts showed these biases and errors to be statistically insignificant, he continues to critique the GHCN dataset publicly. The ACC skeptic website Climate4You contains perhaps the best example of quality assessment reuse in our sample. The website contains a global temperature page, which analyzes five global temperature datasets in a multitude of ways, including linear and polynomial regression for different periods of time. Here we focus on one example of analysis, which the site calls testing for temporal stability or each dataset's internal degree of stability over time. The site does this by "plotting the net change in their global temperature record" between the May 2008 version of a dataset and the October 2021 version of the same dataset. By comparing the temporal stability of datasets, Climate4You asserts that,

> "…it is not possible to conclude which of the above five databases represents the best estimate on global temperature variations. The answer to this question remains elusive. All five databases are the result of much painstaking work, and they all represent admirable attempts towards establishing an estimate of recent global temperature changes. At the same time it should however be noted, that a temperature record which keeps on changing the past hardly can qualify as being correct."

Tools for Analysis

Most skeptics rely on their own software, like Excel or R, to analyze data and create visualizations. Andy May, who analyzes data using R and Excel, states that "R makes it easy to do the calculation, but it is unsatisfying since we don't get much understanding from running it or from the output." This duplication of data analysis aligns with the skeptical values of teaching others how to reuse data and questioning assumptions behind analysis.

Skeptics also take advantage of interactive data analysis tools provided by data repositories, ACC skeptics' organizations, and agnostic groups (Figure 3). This means that skeptics are not solely reliant on official data analysis tools but have other options that provide additional flexibility and interoperability with various datasets. Wood for Trees is the most popular tool created by an ACC agnostic software engineer, hosting "C++ software tools for analysis and graphing of [historical climate] time series data, and an interactive graph generator where users can play with different ways of analyzing data." The home page asserts that,

> "It's not the place of this Web site (or anyone else) to tell you the answers, even if I could! This is just a tool to help you dig into the data to help you form your own opinions. Whatever you decide the most important thing is that you learned what the issues in analysis are and how to test your ideas against real data."

This statement reflects a mindset common in ACC skeptics that devalues expertise and emphasizes trusting one's own data analysis. However, Clark follows this assertion with the following warning "with sharp tools comes great responsibility…beware of short, cherry-picked trends." This is more guidance than skeptical tools offer.

The main C++ program for Wood for Trees is the analysis tool that performs a variety of processes on time-series data formats before outputting the data to a format suitable for plotting with Gnuplot. The program can read a variety of datasets, including the major global temperature datasets, Mauna Loa CO2 monthly average, PDO index, AMO index, and NSIDC Sea Ice Index. Recently, Wood for Trees added the ability for linear least-squares regression. Still, Clark notes "that it can be fairly dangerous…depending on your preconceptions, by picking your start and end times carefully, you can now 'prove' whatever you want with temperature trends." This affordance, perhaps, contributes to its popularity with ACC skeptics.

RimFrost.no and Sealevel.info are two other tools that provide interactive database visualization tools incorporating numerous open datasets. RimFrost.no, "a unique collection of essential climate info," provides a tool to graphically represent thousands of datasets, including temperature data, ice data "CO2 emissions, CO2-in-the-air measurements, river discharge data, precipitation intensity data, glacier mass balances, ocean energy and ocean water levels, sunspot numbers, and skiing conditions." Sealevel.info is a website created by an ACC skeptic to provide a similar tool as Wood for Trees but for sea-level data. The "spreadsheets…consolidate data from NOAA, PSMSL, and other sources to simplify examination of tide-gauge data for long term sea-level trend analysis."

The seven other data analysis tools are associated with specific data repositories or governmental agencies. Generally, these have more limitations than the tools above, only including data from their own repository. The most used of these is NSIDC's Cha*rctic* Interactive Sea Ice Graph, which visualizes the Arctic and Antarctic yearly seasonal cycles of sea ice extent beginning in 1979, which is the start of the satellite record.

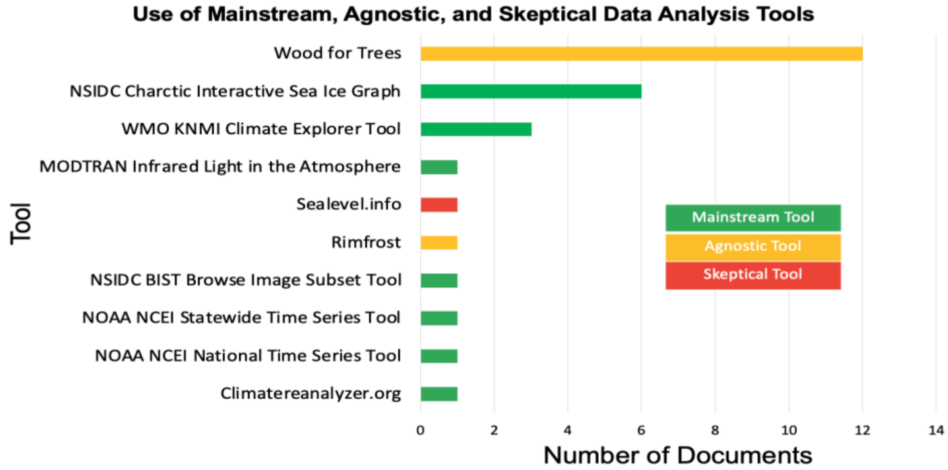

Figure 3: Frequency of interactive database visualization tool use in collected documents. Frequency is calculated as the number of documents the tool is used in.

Preserve

The *preserve* stage of the USGS DLM involves activities related to storing data for long-term use and accessibility. This includes creating documentation and metadata, using accessible storage file formats, and creating supplementary products for preservation.

Many skeptical organizations utilize data to populate their websites and databases, providing convenient access to data in one location. When datasets are shared, application-neutral storage formats such as CSV files are commonly used. For example, Wood for Trees provides "raw" data in tabular plain-text files accompanied by metadata. The metadata in the file includes information on who processed the data (e.g., www.woodfortrees.org), the original source of the data (e.g., NASA Goddard Institute for Space Studies), a link to the original data (e.g., data.giss.nasa.gov/gistemp), the file name (e.g., GLB.Ts.txt), and a brief description of the data (e.g., Time series (gistemp) from 1880 to 2023.08). Other websites' file formats are more diverse. For instance, SeaLevel.info provides JSON files with extensive metadata, copies of NOAA's data in CSV files, spreadsheets of tabulated data in Microsoft Excel HTML format, and documentation for metadata in plain-text files.

Other websites and groups, such as Climate4You and Going Green Canada, rely on data repositories only providing links to publicly-available datasets without creating their own databases. They often choose to provide links to the documentation of these datasets to promote further understanding of the data. However, relying on these data repositories for data preservation means that any changes made by the official organization can affect the availability of data. Data and documentation may be moved to different sites, such as when many NOAA datasets were moved from NCDC to NCEI. In other cases, data repositories may remove certain datasets due to errors, newer versions, or administrative changes. As a result, links to datasets and documentation on these sites may occasionally be dead as site managers fail to update the links.

Publish/Share

In the USGS DLM, activities in the *publish/share* phase conventionally include the publication of findings in peer-reviewed journals as well as the distribution of data online. However, within the sample, ACC skeptics, with one exception, never published in peer-reviewed journals. Instead, they post their findings in blog posts, skeptical conferences, and on social media.

Frequently blog posts with data reuse cases have extremely long comment sections, reaching up to hundreds of pages. Commenters include ACC skeptics and some non-skeptics, although their specific stance on the issue is often difficult to determine. When errors are pointed out that the data reuser agrees with, the work is frequently changed. For instance, one skeptic responds to an error in his figure by stating, "Thank you very much for finding my labeling error in Figure 10 that inadvertently switched the labels for land and deep ocean, ending up disagreeing with my own calculations. I updated the PDF above to correct this error." In a way, the comments section functions as peer review. Commenters provide constructive critique, and reusers revise their work accordingly.

When discussing data sharing or publishing efforts, ACC skeptics emphasize the importance of involving more people in reusing climate data. If a commenter wants to reproduce work, the reuser will provide data, instructions, guidance, and in some cases, the code used. Very occasionally, data reusers refuse to provide data and code when asked for by commenters. One data reuser, Willis Eschenbach, states that the size of the data, along with their unwieldiness and lack of documentation, make them difficult to reuse. Instead, he would "MUCH rather do new and interesting research than spend time beating my code into usable shape." While Eschenbach's refusal may be due to flawed code or deliberate misuse, it seems just as likely that it results from the time commitment needed to make code understandable and reusable. However, he does provide the commenter with a folder with the basic functions he used. As another data reuser mentions in the comments of her blog post, "making small mistakes gets people involved. I think I'll do it more often. I just made 2 more experts."

DISCUSSION

ACC Skeptics' Data Practices in Existing Lifecycle Models and Reuse Typologies

ACC skeptics' data practices align with those described by the USGS DLM in the sense that skeptics acquire, process, analyze, preserve, and publish/share data, often in that order. However, skeptics diverge from traditional researchers in their data evaluation, research goals, criteria for publishable findings, and publication outlets. First, very rarely do ACC skeptics collect their own data and thus are entirely reliant on data they find in open access repositories. Second, the end research goal of many ACC skeptics is to discredit data by finding alleged errors. Additionally, skeptics do not process their data to the extent that traditional researchers would, perhaps, in part because they view mainstream processing activities as contaminating "raw" data, potentially leading to biases. Furthermore, ACC skeptics rarely publish in peer-reviewed journals; instead, they often post or present their findings in blog posts, skeptical conferences, and on social media.

Several studies have identified and categorized different types of scientific data reuse (Coady et al., 2017; Federer, 2019; Gregory et al., 2020; Pasquetto et al., 2019). These typologies are based on the data reuse practices of typical data reusers, such as academic researchers. Do the data reuse types found in the ACC skepticism community fit within these typologies? Largely, the answer is yes, with some caveats. Within the ACC skepticism community, data reuse types can be broadly classified into integrative data reuse (Pasquetto et al., 2019), comparative data reuse (Pasquetto et al., 2019), quality assessment data reuse, and infrastructural data reuse (Federer, 2019).

Integrative data reuse, as defined by Pasquetto et al. (2019), refers to the use of datasets for new analyses with the aim of identifying patterns, correlations, or causal relationships. Integrative reuse is widely observed within the ACC skepticism community, often involving the identification of patterns in observational data and testing for correlations between variables. However, it is a less common practice in the academic communities studied in the Pasquetto et al. paper. Pasquetto et al. argued that significant amounts of tacit knowledge are needed to facilitate *effective* integrative reuse; we speculate that because ACC skeptics are not performing integrative data reuse in the normative epistemological context of climate science, they do not share mainstream scientists' qualms about integrating other people's data. They do not consider tacit knowledge valid and do not require it for integrative reuse. Thus, integrative reuse is more common in ACC skeptic communities than in mainstream.

Comparative data reuse entails the practice of utilizing data to assess similarities and differences for various purposes, such as ground-truthing, calibration, and experimental controls (Pasquetto et al., 2019). Pasquetto et al. found comparative reuse to be common in the academic communities they studied; we similarly found that comparative data reuse is a prevalent and routine practice within the ACC skepticism community. Skeptics use observational data to evaluate model projections and proxy data, as well as to compare similar datasets from different organizations or different versions of the same dataset.

We observed on type of data reuse not found in existing typologies. In **quality assessment data reuse**, skeptics analyze the quality of a dataset to identify flaws or weaknesses, as shown in tests for temporal stability or surfacestations.org surveys. This differs from quality assessment done by the mainstream climate science community, where the quality of their selected data is assessed *prior* to reuse. For mainstream climate scientists, this assessment is rarely the final public output of the reuse. Additionally, ACC skeptics often conduct these quality assessments solely to discredit the data and climate science in its entirety.

Finally, we additionally observed instances of infrastructural data reuse, which involves reusing one or more datasets to populate a database or repository (Federer, 2019). ACC skeptics reuse datasets to populate databases on multiple sites, including Climate4You, Sealevel.info, and CO2Science. They bring together datasets from various agencies and universities that are often siloed from one another into one accessible database. This serves a larger goal of creating an alternative knowledge infrastructure that mimics and takes advantage of climate science's mainstream knowledge infrastructure in an attempt to weaken it. We term this infrastructure a **parasitic knowledge infrastructure** that builds on the back of the mainstream climate science knowledge infrastructure (Wofford, 2022). A parasitic knowledge infrastructure generates, shares, and maintains its knowledge using components of another knowledge infrastructure while simultaneously weakening that infrastructure it relies upon. The "hypertransparency [of] open data, open code, commodity software tools, and alternative publication venues" allows skeptics to selectively use these artifacts (Edwards, 2019, p. 21).

Implications for Data Repositories

Our findings show that ACC skeptics frequently evaluate and interact with data repositories. Data will inevitably be reused in ways repositories do not anticipate and, perhaps, do not agree with. This will be difficult, if not impossible, to avoid through any act of design or curation. Nonetheless, our study has implications for repositories in dealing with contrarian reusers. While these implications primarily apply to scientific data repositories, they may also hold relevance for other contexts, disciplines, or data types. Future research is needed to investigate the interactions between contrarian communities and different types of data repositories.

First, repositories must consider how they might handle the misuse of data and develop policies accordingly. Throughout our work, we have struggled with how to characterize ACC skeptics' use of data. The reusers we observed would not consider their work misuse of data, but certainly, many (if not all) mainstream scientists and data curators see ACC skeptic's practices as inappropriate or flawed. Data repositories and curators are likely not in a position to act as arbiters for truth, but it is essential to recognize that repositories, like all technologies, are not neutral. Curatorial policies have the power to shape not just data reuse but public perceptions of scholarship.

Second, it is crucial to recognize that trust in a repository, its curators, and its data are not guaranteed. In fact, certain communities of reusers will come to a repository with high degrees of distrust. This distrust is not limited to contrarian communities. It may also be seen in other contexts, such as communities in cities expressing distrust in data collection and representation (Yoon & Copeland, 2020) or environmental justice groups distrusting federal data repositories due to perceived ties with polluting companies. Previous studies have linked more traditional data reusers' trust in repositories with "organizational attributes, user communities (recommendations and frequent use), past experiences, repository processes (documentation, data cleaning, and quality checking), and users' perception of the repository roles" (Yoon, 2014, p. 17). ACC skeptics bring up these factors when discussing their level of trust in repositories. Often data reusers lose trust in repositories due to validity errors or missing variables. Thorough and transparent documentation thus enhances trust and helps avoid violations in the first place (Yoon, 2017). However, even with increased documentation and decreased errors, it is unlikely that ACC skeptics would stop questioning and critiquing repositories. At the same time, these steps would give ACC skeptics less fodder for skepticism.

Furthermore, data repositories should be designed with various potential user groups in mind, even if all reuse cases cannot be predicted. Different user groups will come to the repository with various levels of expertise and goals; subsequently, they will interact with the repository and data differently. Data and tools, even if not publicized, will be found by non-experts. Because of this, repositories should design tools that do not easily afford misuse, even, in some cases, at the expense of flexibility. Yoon et al. suggest using data intermediaries to empower prosocial communities with limited resources to build data literacy and capacity (2018). Data intermediaries are agents positioned between two other agents in a data supply-demand chain that facilitate the reuse of data (Schalkwyk et al., 2015; Yoon et al., 2018). In this case, however, data intermediaries could be used to understand how data are being reused and, with contrarian communities, to decrease the likelihood of data misuse.

CONCLUSION AND FUTURE WORK

This paper examines the data practices of one contrarian community of data reusers, ACC skeptics. These practices are made possible through the parasitic relationship that exists between the skeptics' knowledge infrastructure and the mainstream climate science knowledge infrastructure and its increased openness. In future work, we will further develop the concept of parasitic knowledge infrastructures and how contrarian knowledge infrastructures can simultaneously rely on and weaken consensus-based knowledge infrastructures through this case study. Other research, soon to be published by our team, looks at how three different contrarian communities, pro-life activists, Young Earth Creations, and ACC skeptics, engage with scientific values and norms (e.g., objectivity), standards (e.g., uncertainty), validation processes (e.g., consensus), and artifacts (Pasquetto et al., 2023). Finally, we hope to see whether this parasitic relationship distinguishes skeptics' knowledge practices from other forms of public-led epistemic practices, such as those seen in environmental justice groups.

ACKNOWLEDGMENTS

Thanks to Irene Pasquetto and Libby Hemphill, whose feedback helped shape this paper.

REFERENCES

AlQuraishi, M., & Sorger, P. K. (2016). Reproducibility will only come with data liberation. *Science Translational Medicine*, *8*(339), 339ed7-339ed7. https://doi.org/10.1126/scitranslmed.aaf0968

Anti-Eviction Mapping Project. (2021). Anti-Eviction Mapping Project. https://antievictionmap.com

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*, *2*(2), 2053951715594634. https://doi.org/10.1177/2053951715594634

Boellstorff, T. (2012). *Ethnography and virtual worlds: A handbook of method*. Princeton University Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. The MIT Press.

Borgman, C. L., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who Uses the Digital Data Archive? An Exploratory Study of DANS. *Proceedings of the Association for Information Science and Technology*, *52*. https://doi.org/10.1002/pra2.2015.145052010096

Cavalier, D., & Kennedy, E. B. (2016). *The rightful place of science: Citizen science*. Consortium for Science, Policy & Outcomes.

Charmaz, K. (2014). *Constructing Grounded Theory*. SAGE.

Christy, J. R., & Spencer, R. W. (2022). *The University of Alabama in Huntsville Earth System Science Center Global Temperature Report*. *32*(10). https://www.nsstc.uah.edu/climate/archives.html

Clavero, M., García-Reyes, A., Fernández-Gil, A., Revilla, E., & Fernández, N. (2022). On the misuse of historical data to set conservation baselines: Wolves in Spain as an example. *Biological Conservation*, *276*, 109810. https://doi.org/10.1016/j.biocon.2022.109810

Coady, S. A., Mensah, G. A., Wagner, E. L., Goldfarb, M. E., Hitchcock, D. M., & Giffen, C. A. (2017). Use of the National Heart, Lung, and Blood Institute Data Repository. *New England Journal of Medicine*, *376*(19), 1849–1858. https://doi.org/10.1056/NEJMsa1603542

Coleman, E. G. (2010). Ethnographic Approaches to Digital Media. *Annual Review of Anthropology*, *39*(1), 487–505. https://doi.org/10.1146/annurev.anthro.012809.104945

Coleman, E. G. (2012). *Coding Freedom: The Ethics and Aesthetics of Hacking*. https://press.princeton.edu/books/paperback/9780691144610/coding-freedom

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data Sharing, Small Science, and Institutional Repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLOS ONE*, *12*(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Edwards, P. N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. *Science as Culture*, *8*(4), 437–472. https://doi.org/10.1080/09505439909526558

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Edwards, P. N. (2019). Knowledge infrastructures under siege: Climate data as memory, truce, and target. In D. Bigo, E. F. Isin, & E. S. Ruppert (Eds.), *Data politics: Worlds, subjects, rights* (pp. 21–42). Routledge, Taylor & Francis Group.

Espinosa, R., Garriga, L., Zubcoff, J. J., & Mazón, J.-N. (2014). Linked Open Data mining for democratization of big data. *2014 IEEE International Conference on Big Data (Big Data)*, 17–19. https://doi.org/10.1109/BigData.2014.7004479

Fan, L., Lafia, S., Bleckley, D., Moss, E., Thomer, A., & Hemphill, L. (2022). Librarian-in-the-Loop: A Natural Language Processing Paradigm for Detecting Informal Mentions of Research Data in Academic Literature. In *ArXiv*. http://arxiv.org/abs/2203.05112

Faniel, I. M., Frank, R. D., & Yakel, E. (2020, May 12). *Context from the Data Reuser's Point of View*. OCLC. https://www.oclc.org/research/publications/2019/context-from-data-reuser-point-of-view.html

Faniel, I. M., Kansa, E. C., Kansa, S. W., Barrera-Gomez, J., & Yakel, E. (2013). The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 295–304. https://doi.org/10.1145/2467696.2467712

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

Faniel, I. M., & Yakel, E. (2017). Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation. In L. R. Johnston (Ed.), *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* (pp. 103–126). Association of College and Research Libraries. http://www.oclc.org/research/publications/2017/practices-do-not-make-perfect.html

Faundeen, J. L., Burkey, T. E., Carlino, J. A., Govoni, D. L., Henkel, H. S., Holl, S. L., Hutchison, V. B., Martin, E., Montgomery, E. T., Ladino, C. C., Tessler, S., & Zolly, L. S. (2014). *The United States Geological Survey Science Data Lifecycle Model* (Open-File Report No. 2013–1265; Open-File Report, p. 12). U.S. Geological Survey.

Federer, L. M. (2019). *Who, what, when, where, and why? Quantifying and understanding biomedical data reuse*. https://doi.org/10.13016/60jd-9hux

Fiedler, T., Pitman, A. J., Mackenzie, K., Wood, N., Jakob, C., & Perkins-Kirkpatrick, S. E. (2021). Business risk and the emergence of climate analytics. *Nature Climate Change*, *11*(2), 87–94. https://doi.org/10.1038/s41558-020-00984-6

Fischer, F. (2019). Knowledge politics and post-truth in climate denial: On the social construction of alternative facts. *Critical Policy Studies*, *13*(2), 133–152. https://doi.org/10.1080/19460171.2019.1602067

Friedberg, B. (2020). Investigative Digital Ethnography: Methods for Environmental Modeling. In *Media Manipulation Casebook*. https://mediamanipulation.org/research/investigative-digital-ethnography-methods-environmental-modeling

GISTEMP Team. (2022). *GISS Surface Temperature Analysis (GISTEMP), version 4*. NASA Goddard Institute for Space Studies. https://data.giss.nasa.gov/gistemp/references.html

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Publishing.

Gramling, C. (2007, October). *Geotimes—October 2007—Error in NASA climate data sparks debate*. http://www.geotimes.org/oct07/article.html?id=nn_climate.html

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

Hardy, B. W., & Jamieson, K. H. (2017). Overcoming Endpoint Bias in Climate Change Communication: The Case of Arctic Sea Ice Trends. *Environmental Communication*, *11*(2), 205–217. https://doi.org/10.1080/17524032.2016.1241814

Hine, C. (2000). *Virtual Ethnography*. SAGE Publications Ltd. http://methods.sagepub.com/book/virtual-ethnography

Holdren, J. P. (2013). *Increasing Access to the Results of Federally Funded Scientific Research*. Executive Office of the President, Office of Science and Technology Policy. https://www.usaid.gov/sites/default/files/documents/1865/NW2-CCBY-HO2-Public_Access_Memo_2013.pdf

Huang, Y.-H., Rose, P. W., & Hsu, C.-N. (2015). Citing a Data Repository: A Case Study of the Protein Data Bank. *PLOS ONE*, *10*(8), e0136631. https://doi.org/10.1371/journal.pone.0136631

Jourdain, O. (2021). Resistance to vaccination in France: History and the influence of social media. *Journal of Gynecology Obstetrics and Human Reproduction*, *50*(7), 101997. https://doi.org/10.1016/j.jogoh.2020.101997

Kansa, E. C. (2012). Openness and archaeology's information ecosystem. *World Archaeology*, *44*(4). https://doi.org/10.1080/00438243.2012.737575

Khan, N., Thelwall, M., & Kousha, K. (2019). Data citation and reuse practice in biodiversity—Challenges of adopting a standard citation model. In *1*. International Society for Scientometrics and Informetrics. https://wlv.openrepository.com/handle/2436/623005

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal*, *8*(1), 3. https://doi.org/10.1186/s11568-014-0003-1

Knoppers, B. M., Harris, J. R., Budin-Ljøsne, I., & Dove, E. S. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Human Genetics*, *133*(7), 895–903. https://doi.org/10.1007/s00439-014-1432-6

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. *Political Communication*, *37*(2), 194–214. https://doi.org/10.1080/10584609.2019.1686094

Lafia, S., Thomer, A., Moss, E., Bleckley, D., & Hemphill, L. (2023). *How and Why do Researchers Reference Data? A Study of Rhetorical Features and Functions of Data References in Academic Articles* (arXiv:2302.08477). arXiv. https://doi.org/10.48550/arXiv.2302.08477

Lay, J. (2022). Climate Disinformation Database. *DeSmog*. https://www.desmog.com/climate-disinformation-database/

Lee, C., Yang, T., Inchoco, G. D., Jones, G. M., & Satyanarayan, A. (2021). Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3411764.3445211

Leiserowitz, A. A., Maibach, E. W., Roser-Renouf, C., Smith, N., & Dawson, E. (2013). Climategate, Public Opinion, and the Loss of Trust. *American Behavioral Scientist*, *57*(6), 818–837. https://doi.org/10.1177/0002764212458272

Leonelli, S. (2013). Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. *Bulletin of Science, Technology & Society*, *33*(1–2), 6–11. https://doi.org/10.1177/0270467613496768

Levy, K. E., & Johns, D. M. (2016). When open data is a Trojan Horse: The weaponization of transparency in science and governance. *Big Data & Society*, *3*(1), 2053951715621568. https://doi.org/10.1177/2053951715621568

Lewandowsky, S., Ballard, T., Oberauer, K., & Benestad, R. (2016). A blind expert test of contrarian claims about climate data. *Global Environmental Change*, *39*, 91–97. https://doi.org/10.1016/j.gloenvcha.2016.04.013

Mapping Police Violence. (2021). Mapping Police Violence. https://mappingpoliceviolence.org

Martens, L., & Vizcaíno, J. A. (2017). A Golden Age for Working with Public Proteomics Data. *Trends in Biochemical Sciences*, *42*(5), 333–341. https://doi.org/10.1016/j.tibs.2017.01.001

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2017). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, *68*(6), 1341–1359. https://doi.org/10.1002/asi.23721

McCright, A. M., & Dunlap, R. E. (2011). Cool dudes: The denial of climate change among conservative white males in the United States. *Global Environmental Change*, *21*(4), 1163–1172. https://doi.org/10.1016/j.gloenvcha.2011.06.003

Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J., & Lawrimore, J. H. (2018). The Global Historical Climatology Network Monthly Temperature Dataset, Version 4. *Journal of Climate*, *31*(24), 9835–9854. https://doi.org/10.1175/JCLI-D-18-0094.1

Mirowski, P. (2014). What is "Open Science" open to? In *The Electronic Condition*. https://www.academia.edu/11571042/What_is_Open_Science

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, *48*(2), 171–203. https://doi.org/10.1177/0306312718772086

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set: THE HADCRUT4 DATASET. *Journal of Geophysical Research: Atmospheres*, *117*(D8), n/a-n/a. https://doi.org/10.1029/2011JD017187

Nielsen, M. A. (2011). *Reinventing discovery: The new era of networked science*. Princeton University Press.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *ELife*, *5*, e21451. https://doi.org/10.7554/eLife.21451

Obama, B. (2013). *Executive Order—Making Open and Machine Readable the New Default for Government Information*. https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Open Knowledge Foundation. (n.d.). *Open Definition 2.1—Open Definition—Defining Open in Open Data, Open Content and Open Knowledge*. Retrieved April 18, 2023, from https://opendefinition.org/od/2.1/en/

Panofsky, A., & Donovan, J. (2019). Genetic ancestry testing among white nationalists: From identity repair to citizen science. *Social Studies of Science*, *49*(5), 653–681. https://doi.org/10.1177/0306312719861434

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, *69*(11), 1346–1354. https://doi.org/10.1002/asi.24049

Pasquetto, I. V. (2018). *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses* [Ph.D. Dissertation, UCLA]. https://escholarship.org/uc/item/1sx7v77r

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-008

Pasquetto, I. V., Thomer, A. K., Wofford, M. F., Jordan, C., & Order TBD. (2023). *On Post-Truth Science Opposition: From Current Understandings to New Research Directions [Manuscript in preparation]*.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. I. (2017). Zenodo in the Spotlight of Traditional and New Metrics. *Frontiers in Research Metrics and Analytics*, *2*. https://doi.org/10.3389/frma.2017.00013

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & Tacchi, J. (2015). *Digital Ethnography: Principles and Practice*. SAGE.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Postill, J., & Pink, S. (2012). Social Media Ethnography: The Digital Researcher in a Messy Web. *Media International Australia*, *145*(1), 123–134. https://doi.org/10.1177/1329878X1214500114

Powell, J. L. (2015). Climate Scientists Virtually Unanimous: Anthropogenic Global Warming Is True. *Bulletin of Science, Technology & Society*, *35*(5–6), 121–124. https://doi.org/10.1177/0270467616634958

Re3data.org | Registry of Research Data Repositories. (2014). http://www.re3data.org/

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

Ricker, B., Cinnamon, J., & Dierwechter, Y. (2020). When open data and data activism meet: An analysis of civic participation in Cape Town, South Africa. *The Canadian Geographer / Le Géographe Canadien*, *64*(3), 359–373. https://doi.org/10.1111/cag.12608

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, *11*(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*(2), 89–99. https://doi.org/10.1038/nrg3394

Ryan, P. D. (2019, January 14). *H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018* (2017/2018) [Legislation]. http://www.congress.gov/

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19–S31. https://doi.org/10.1016/j.giq.2012.06.011

Schalkwyk, F. V., Canares, M., Chattapadhyay, S., & Andrason, A. (2015). *Open Data Intermediaries in Developing Countries*. 640993 Bytes. https://doi.org/10.6084/M9.FIGSHARE.1449222

Sharman, A. (2014). Mapping the climate sceptical blogosphere. *Global Environmental Change*, *26*, 159–170. https://doi.org/10.1016/j.gloenvcha.2014.03.003

Silverman, C. (2020). Verification handbook for disinformation and media manipulation. *European Journalism Centre*. https://datajournalism.com/read/handbook/verification-3

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing Reproducible Research*. CRC Press.

Tans, P., & Keeling, R. (2022). *NOAA Global Monitoring Laboratory (GML) Earth System Research Laboratories—Monthly Average Mauna Loa CO2*. https://gml.noaa.gov/ccgg/trends/data.html

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Christian, L., Allard, S., & Borycz, J. (2018). Research Data Sharing: Practices and Attitudes of Geophysicists. *Earth and Space Science*, *5*(12), 891–902. https://doi.org/10.1029/2018EA000461

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thomer, A. K. (2022). Integrative data reuse at scientifically significant sites: Case studies at Yellowstone National Park and the La Brea Tar Pits. *Journal of the Association for Information Science and Technology*, *73*(8), 1155–1170. https://doi.org/10.1002/asi.24620

Thomer, A. K., & Wickett, K. M. (2020). Relational data paradigms: What do we learn by taking the materiality of databases seriously? *Big Data & Society*, *7*(1), 2053951720934838. https://doi.org/10.1177/2053951720934838

Thomer, A. K., Wofford, M. F., Lenard, M. C., Dominguez Vidana, S., & Goring, S. J. (2023). Revealing Earth science code and data-use practices using the Throughput Graph Database. In X. Ma, M. Mookerjee, L. Hsu, & D. Hills (Eds.), *Recent Advancement in Geoinformatics and Data Science*. Geological Society of America. https://doi.org/10.1130/2022.2558(10)

U.S. National Ice Center, & NSIDC. (2010). *Multisensor Analyzed Sea Ice Extent—Northern Hemisphere (MASIE-NH), Version 1*. https://doi.org/10.7265/N5GT5K3K

VijayaVenkataRaman, S., Iniyan, S., & Goic, R. (2012). A review of climate change, mitigation and adaptation. *Renewable and Sustainable Energy Reviews*, *16*(1), 878–897. https://doi.org/10.1016/j.rser.2011.09.009

Weber, N. M., Thomer, A. K., Mayernik, M. S., Dattore, B., Ji, Z., & Worley, S. (2013). The Product and System Specificities of Measuring Curation Impact. *International Journal of Digital Curation*, *8*(2), Article 2. https://doi.org/10.2218/ijdc.v8i2.286

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wofford, M. (2022). Parasitic Knowledge Infrastructures: Data Reuse by Anthropogenic Climate Change Skeptics. *Proceedings of the Association for Information Science and Technology*, *59*(1), 837–839. https://doi.org/10.1002/pra2.743

Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. *Archival Science*, *14*(1), 17–34.

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730

Yoon, A., & Copeland, A. (2020). Toward community-inclusive data ecosystems: Challenges and opportunities of open data for community-based organizations. *Journal of the Association for Information Science and Technology*, *71*(12), 1439–1454. https://doi.org/10.1002/asi.24346

Yoon, A., Copeland, A., & McNally, P. J. (2018). Empowering communities with data: Role of data intermediaries for communities' data utilization. *Proceedings of the Association for Information Science and Technology*, *55*(1), 583–592. https://doi.org/10.1002/pra2.2018.14505501063

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, *39*(3), 224–233. https://doi.org/10.1016/j.lisr.2017.07.008

Zimmerman, A. S. (2003). *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists* [Ph.D Dissertation]. University of Michigan.

Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology, & Human Values*, *33*(5), 631–652. https://doi.org/10.1177/0162243907306704

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17–29. https://doi.org/10.1016/j.giq.2013.04.003