Datasheet for Ontario Drinking Water Quality and Enforcement*

Ontario Drinking Water Quality and Enforcement Dataset (2022-2023)

Kuiyao Qiao

March 11, 2024

The datasheet give a overview of critical data related to the management and safety of drinking water in Ontario, Canada. It document the dataset's motivation, composition, collection process, recommended uses.

Introduction

The datasheet for the 2022-2023 Ontario Drinking Water Quality and Enforcement dataset acts as a detailed manual that helps understand what the data covers, how it was gathered, and how it can be used concerning drinking water quality in Ontario. It summarizes important information about how drinking water is managed and kept safe in Ontario, Canada. This document explains why the dataset was created, what it contains, and how it was collected, and it suggests how it might be used.(Gebru et al. 2021) (R Core Team 2023)

Motivation

The dataset@DrinkingWaterQuality2023 was made so everyone can see and understand how safe and up to standard the drinking water in Ontario, Canada, is. It's all about ensuring the rules for drinking water are followed and that people can trust what comes out of their taps. The Ontario Ministry of the Environment, Conservation and Parks put it together with money from the Ontario government. This was part of their job to look after the environment and our health. They usually keep the exact funding details a secret, as it's part of their regular budget. This dataset is essential for both the government and regular folks to check out how well the drinking water system in Ontario is doing. It's vital for managing the environment and keeping our health in check, giving us the lowdown on the water's safety and quality. "Drinking Water Quality and Enforcement" (2023)

Composition

*Code and data are available at: https://github.com/QPP123/2022-2023-Ontario-Drinking-Water-Quality-and-Enforcement-Dataset

The dataset is about the nitty-gritty of drinking water in Ontario, Canada. It's got a bunch of details on stuff like water test results, problems with water quality, efforts to cut down on lead, checks on water systems, legal actions, and the qualifications of the people running these systems. It doesn't say exactly how many records there are, but it's probably a lot since it covers all of Ontario's water systems and labs for 2022-2023. The goal is to have a complete picture of everything going on with the water systems and labs during that time, not just a snippet. So, in the dataset, each record has various bits of info like test scores, dates, places, types of problems, legal steps taken, and the status of the operators' licenses. There isn't just one main thing this dataset is focusing on; it's got a mix of data that could be used for things like keeping an eye on rules being followed, spotting trends, or helping make decisions about water policies. They don't discuss any information being left out, but there might be gaps here and there because of reporting issues or difficulties in gathering data. The dataset needs to understand how the different records might be connected, but there could be some natural connections based on location or the type of water system. It needs to be set up with specific sections for different uses since it's more about analyzing and keeping tabs on things than machine learning. There aren't any noted mistakes, noisy data, or unnecessary duplications, but that doesn't mean it's perfect – there might be stuff that needs a closer look and cleanup. This dataset is a one-stop shop provided by the MECP, though you might find extra helpful stuff on their website to make sense of the data. They haven't said anything about sensitive or private info being included, but there might be some specifics about the water systems or labs that are a bit sensitive. The dataset is technical and focuses on water quality and system details, so it's likely to be okay with everyone. It doesn't dive into personal details about people; at most, it talks about the professional side of those managing the water systems. Lastly, it doesn't get into personal traits or beliefs. But it touches water quality, which is essential for everyone's health.

Collection process

The data probably came from different ways that Ontario's laws need for checking on drinking water. This could mean regular water tests and check-ups and having to report any water problems to those in charge of water systems and labs. The nitty-gritty of how they got the data isn't spelled out, but it involved a mix of gadgets (like water testing tools), computer programs (for reporting stuff), and hands-on work (like inspections). Ensuring these methods are up to scratch likely falls under the watchful eye of the MECP's quality checks and reviews. This dataset isn't just a random sample; it's meant to be a full roundup of all the relevant data for the period. Gathering the data was a team effort involving MECP staff, the folks running the water systems, and lab workers, who were likely paid as part of their regular jobs or contracts. The info covers 2022 to 2023, showing the data was collected over a year, aiming to give a fresh look at the state of drinking water quality and regulation in Ontario. There must be a word on any ethical review in the dataset description. However, collecting and sharing this data is part of MECP's job to look after public health and the environment, which you'd expect to involve ethical data handling. The data was probably directly gathered from the water system operators and labs through must-do reporting and MECP's checks and monitoring. These operators and labs likely know they must report data to the MECP as it's

part of their rules. However, the dataset description needs to dive into how they were told about the data gathering. Consent for gathering and using the data is probably taken as a given or required by the laws for running water systems and labs in Ontario. However, the dataset needs to get into the exact terms of any consent. There must be a mention of how to get back consent, likely because handing over this data is a must-do under the drinking water laws. The dataset doesn't talk about any analysis of the impact of collecting and publishing this data. However, the MECP probably decided to share this info based on the public's right to know and how transparency can help ensure safe drinking water and adherence to the rules.

Preprocessing/cleaning/labeling

The dataset's info doesn't discuss any steps taken to prep, clean, or label the data before sharing. There might have been some work done to ensure the data is in good shape or to change it somehow, but we need the specifics on that. They need to say if the original, unprocessed data is available. We get to see the final version that's been put out there. Also, there needs to be a mention of what kind of software might have been used to get the data ready, clean it up, or organize it with labels. So, we need to figure out the tools or tech they used to handle the data before releasing it.

Uses

The dataset is mainly for keeping an eye on and reporting the condition of drinking water and how the rules are followed in Ontario. It's likely used by the MECP, public health folks, researchers, and others for different kinds of analysis and to help make decisions. There's no direct mention of specific studies or reports that have used this data in the dataset description, but there may be some in the public or academic circles. This dataset could be handy for several things, like spotting trends in water quality, checking if legal actions help with rule-following, guiding decisions on water management, building models to predict water safety risks, comparing different water system performances, or looking into fairness in water quality and access. However, the dataset has its context and limits. It's only about Ontario from 2022 to 2023, so its insights might not apply elsewhere or at other times. The quality and fullness of the data depend on how consistently and reliably water system operators and labs report their findings. Also, it lacks detailed info on the people using each water system, which could affect its use in exploring issues of fairness or inequality. To make the most of this dataset while being mindful of its boundaries, users should: Thoroughly check the dataset's details and boundaries to understand what it can and can't tell. Use proper statistical techniques and account for other factors when analyzing the data. Bring in extra data or context if needed. Be cautious in how findings are interpreted and shared, noting any limitations. Talk to experts or those involved to confirm what the data suggests. This dataset differs for detailed personal or health info linked to the water systems. Without more context and information, it should not be the sole source from which to judge Ontario's drinking water's overall safety or quality.

Distribution

The dataset is shared via the Government of Ontario's Open Data Catalogue, so it's not just for MECP people but open to others, too. You can find it on Ontario's Open Data Catalogue website, but it doesn't mention a DOI (Digital Object Identifier). The exact release date for the 2022-2023 data has yet to be given, but since they update it yearly, it probably came out in 2023 after the reporting year ended. It's under the Open Government Licence - Ontario, which means anyone can use, change, and share the data freely without paying for it. There's no mention of third-party intellectual property or other restrictions in the description, and with it being under an open license, such limits are unlikely. Also, the description doesn't discuss export controls or legal restrictions affecting how the data can be shared or used.

Maintenance

The dataset is taken care of by the Ministry of the Environment, Conservation and Parks (MECP) in Ontario, Canada, and it's part of their Open Data Catalogue. MECP has provided contact details on their Open Data Catalogue website for any questions or more info. While no typos are mentioned, if any mistakes in the data are found later on, MECP will likely let people know through their Open Data Catalogue site or by contacting those who use the data directly. The dataset gets a refresh every year, covering the fiscal year from April to March, with MECP handling the updates. They share these updates through the Open Data Catalogue, and they might also reach out about these updates through the website or other ways. Since there's no personal info in the dataset, there aren't special restrictions on how long the data can be kept. However, MECP might have its rules or legal requirements about how long they hold onto the data. More specific information is needed on how they handle older dataset versions. They could focus on keeping the latest version up to date and archiving or removing the old ones. The Open Data Catalogue site would probably announce changes regarding handling older dataset versions. The dataset description doesn't discuss how outside people can add to the dataset. Given that a government body runs it, they will likely need a formal way for the public to contribute data. However, people can give feedback or suggestions to the MECP via the contact info on their Open Data Catalogue site.

Acknowledgments

Thanks to Zijun Meng for all the helpful feedback while I was working on this paper.

References

"Drinking Water Quality and Enforcement." 2023. https://data.ontario.ca/dataset/drinking-water-quality-and-enforcement.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.