Scalable Multi-resolution Spatial Visualization for Anthropogenic Litter Data

Yunfan Kang[1,2], Ziang Zhao[1], Amr Magdy[1,2]
[1]Dept. of Computer Science and Engineering
[2]Center for Geospatial Sciences
Riverside, California
{ykang040,zzhao047}@ucr.edu,amr@cs.ucr.edu

Win Cowger[3], Andrew Gray[3]
[3]Dept. of Environmental Sciences
Riverside, California
{wcowg001,agray}@ucr.edu

Abstract

This paper demonstrates *CleanUpOurWorld*; a research spatial database that is designed and deployed to collect, process, query, and visualize anthropogenic litter data. Such data has a significant importance in the field of environmental sciences due to its important use cases. We make a major on-going effort to collect and maintain such data worldwide from different sources through a community of environmental scientists and partner organizations. With the increasing volume of data, existing software packages, such as GIS software, do not scale to process, query, and visualize such data. To overcome this, *CleanUpOurWorld* digests datasets from different sources, with different formats, in a scalable backend that cleans, integrates, and unifies them in a structured form in a relational spatial database. Frontend applications are built to visualize litter data at multiple spatial resolutions.

ACM Reference Format:
Yunfan Kang[1,2], Ziang Zhao[1], Amr Magdy[1,2] and Win Cowger[3], Andrew Gray[3]. 2019. Scalable Multi-resolution Spatial Visualization for Anthropogenic Litter Data. In *Proceedings of ACM Conference (SIGSPATIAL'19)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large-scale environmental problems involve managing and processing large datasets [1, 2]. A prominent example is anthropogenic litter data, which is data about waste that originates from human activities such as food waste, diapers, construction materials, used motor oil, and hypodermic needles. This waste is generated in a daily basis worldwide as a part of human daily activities. A significant part of this waste ends up in natural dumpsters, such as oceans, which causes several environmental problems, e.g., destroying marine life and increasing the impact of natural hazards like floods. This phenomenon is becoming globally more dangerous to the extent that president of the United States (US) signed *Save Our Seas Act* in October 2018 committing the US to expand efforts to clean up nearly 8 million metric tons of litter polluting the world's oceans [4]. Many environmental scientists and several community and governmental organizations are interested in collecting and visualizing waste data as a preliminary step towards addressing these problems.

Environmental scientists currently use Geographic Information Systems (GIS) software to load and visualize data in the spatial space. However, with the increasing volume of the collected data worldwide, GIS software does not scale to visualize hundreds of thousands, or even millions, of data points. This limits the abilities to analyze sufficiently large datasets. Lack of such scalable visualization hinders several community efforts to clean the world, which recently encouraged interested communities to develop advanced tools to handle this [5]. Moreover, GIS software is currently limited in functionality to solve litter data issues such as cleaning noisy data and data aggregation over different attributes. This complicates integrating new litter reports and datasets, which hinders active monitoring of human litter problems and limits progress in several environmental projects.

This paper demonstrates *CleanUpOurWorld*; a scalable research database that enables environmental scientists and organizations to collect, process, query, and visualize human waste data. *CleanUpOurWorld* overcomes the limitation of visualization scalability through a combination of spatial database technology and data aggregation techniques. In addition, it employs data cleaning and integration modules that take diverse data sources, fix inconsistency problems, and integrate them into a unified logical model. It currently provides six visualization levels: continent, sub-continent, country, sub-country, city, and street levels. The street level shows individual data points while other levels show aggregate number of data points classified based on waste type as shown in our demonstration scenarios (Section 3). Environmental scientists and activists can interactively navigate data through zooming in/out on different spatial levels, panning over different regions, and filtering based on waste type, to visualize and analyze different portions of data with real-time response. In addition, the system allows them to add new litter data sources and download subset of existing data.

SIGSPATIAL'19, Novembe 5-8 2019, Chicago, Illinois, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We demonstrate *CleanUpOurWorld* with an actual system implementation that integrates data from thirty different sources with total of 420K data points collected by environmental scientists and collaborator organizations. Demonstrations attendants will be able to interact with the system through different scenarios as detailed in Section 3.

2 Framework Overview

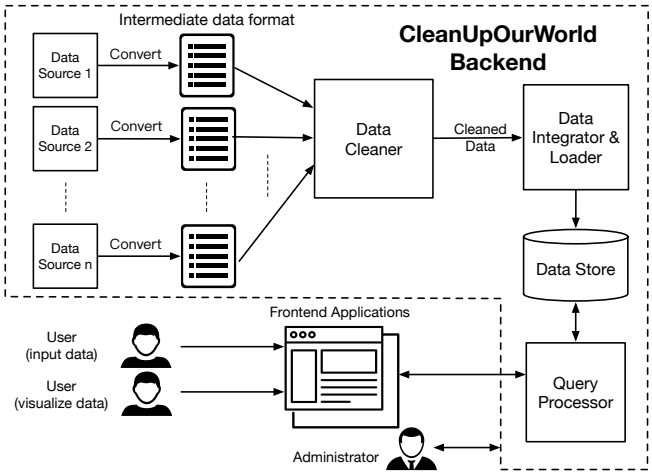


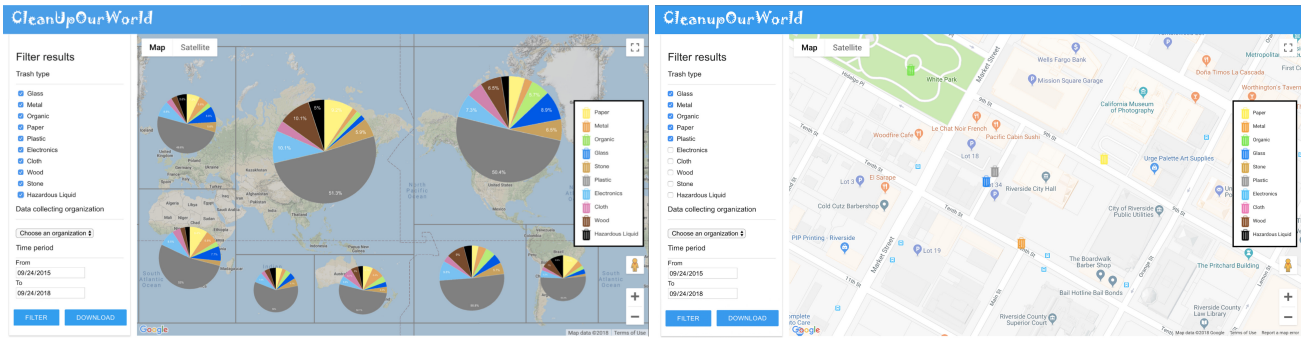
Figure 1. *CleanUpOurWorld* Architecture.

This section gives an overview about *CleanUpOurWorld* components and operations. Figure 1 presents *CleanUpOurWorld* architecture. The backend takes collected datasets and converts them to a common intermediate data format to standardize the input format from different data sources. Then, the converted data is fed to a pipeline of *data cleaner*, *data integrator and loader*, and *data store* that preprocesses the data and store them in a scalable database. Finally, the *query processor* receives queries from end users through frontend applications and access the data through SQL queries. This section briefly outlines backend components and Section 3 outlines frontend applications.

Data collection, formatting, and conversions. In collaboration with *Let's Do it World* [3], a nonprofit organization based in Tallinn, Estonia, *Gray Lab* at the University of California, Riverside started the data collection process by contacting organizations that host the largest datasets such as *Ocean Conservancy*, *Litterati*, *Marine Debris Tracker*, *Alice Ferguson Foundation*, and *US National Oceanic and Atmospheric Administration (NOAA)*. Some organizations, e.g., Ocean Conservancy and Marine Debris Tracker, provide open access to their data while others, e.g., Alice Ferguson Foundation and NOAA, gave us special permissions to access their data. We have collected thirty datasets so far and we are currently working on adding nine more datasets with collaborator organizations and open data sources. In addition, we are running a crowd-sourcing data collection project in Southern California. Datasets are collected in four formats: ESRI Shape files, KML files, XLSX files, and CSV files. Then, all formats are converted into a common CSV format with UTF8 encoding.

Data cleaning. *CleanUpOurWorld* employs a data cleaning process that addresses litter datasets collected from diverse sources. The process includes three operations. First, regulating columns names to avoid empty names, duplicate names, mismatched names, and names that do not comply with SQL standards. The problem of mismatched names arise when a common essential attribute, such as location, appears with multiple names. We detect this for location and date attributes through comparing both data types and names and fill up with a common attribute name that is used in all datasets. Other than mismatched names, empty, duplicate, and non-compliant names are fixed with arbitrary names, through regular expressions, prompting the user asynchronously to alter them with new names if needed. Second, fixing data irregularities where some data values are not compliant with the data type and in other cases values of the same attribute has different formats. Regular expressions are used to discover non-compliant values and unify the format of the same data type. Non-compliant values are filtered out for further manual processing so that the column data type can be assigned the appropriate data type. Third, some attribute irregularities go beyond simple format mismatches to completely different format. For example, the location attribute is an essential attribute in all our datasets and its common format is a pair of latitude/longitude coordinates. However, some datasets have this attribute as a rough textual description, such as street name, city name, etc. To address this, we employ a Python library GeoPy, with *Nominatim* geocoder, that converts this text into precise coordinates.

Data integration, loading, storage, and visualization. After each dataset is cleaned, it is forwarded to a data integrator and loader module. This module goes over three steps. The first step loads the dataset in a separate SQL table, which contains all input data attributes, in a PostGIS database that represents the main data store in Figure 1. The loading process is performed through a dynamic SQL function, using *plpgsql* language. Then, it creates a SQL table, populates the data from the file to this table, and adds an auto increment primary key field that guarantees a unique identification for each record in the system. The second step loads only essential attributes of each record from the newly loaded dataset in a centralized table, called *maintable*, that integrates data from all existing datasets in a snowflake-like fashion. This table is created to scale up query processing as we will elaborate later. The *maintable* have three essential attributes per record: a location as latitude/longitude pair, a date, and a collecting organization, in addition to the record id and the dataset name to link the record back to its original dataset table with full attribute set. The *maintable* has one record corresponds to each record in each dataset.



(a) Litter Data on Continent Level

(b) Litter Data on Street Level

Figure 2. Visualizing Litter Data on Multiple Spatial Levels.

The third step aggregates data from the *maintable* in the *aggregatetable*. When environmental scientists visualize data in large regions, e.g., city, country, continent, or ocean, the large number of points in the region will limit the existing visualization frontends, such as GoogleMaps and GIS software, to show all points while still being interactive to users. To overcome this problem, *CleanUpOurWorld* visualizes individual data points only at the street level, while aggregate data is visualized on higher levels. To speed up such aggregate visualization, *aggregatetable* maintains aggregate counts from *maintable* at five different spatial levels that corresponds to the frontend visualization levels. At each level, the whole world is divided into a set of spatial tiles. Then, the *aggregatetable* maintains the number of data points in each spatial tile classified by the litter type. When a visualization query comes, this table is queried to retrieve data right away and visualize them to users.

Query processing. The query processor receives three types of queries from frontend applications and accesses the database to return their results. First, a spatial query that finds all data points in a certain spatial range. This query is answered from *maintable* with a single spatial range query. Second, a spatial query that finds aggregate counts in a certain spatial range. This query is answered from *aggregatetable* with a single spatial range query as well. Third, a query that finds all attributes of a single data record. This query is answered from the corresponding dataset table with a single query given the record id. Using *maintable* and *aggregatetable* enables each incoming query to be translated into a single SQL query, which allow scalable management and visualization of litter data.

3 Demonstration Scenarios

CleanUpOurWorld functionalities and internals are demonstrated using the first batch of real datasets collected from thirty different sources with total of 420K data points. New data batches with hundreds of thousands of points are being collected as described earlier. Our demo attendees would be able to interact with *CleanUpOurWorld* through one or more of the following scenarios.

3.1 Scenario 1: Hypothesising and Locating Litter Through Interactive Visualization

The main objective of *CleanUpOurWorld* is to enable environmental scientists and community organizations to address human litter problems through analysis, proposing hypotheses, or locating data to clean up certain places. Thus, demo attendees will be able to interactively visualize litter data based on multiple dimensions: different spatial levels, arbitrary time periods, categories of different litter types, and collecting organizations. Figure 2 depicts the main visualization screen of *CleanUpOurWorld* demonstration. Figure 2(a) depicts the highest spatial level of visualization. At this level, the whole world is divided into ten large regions based on continents and their adjacent oceans, eight of them are shown in the figure. The figure shows the spatial boundaries of each region. In addition, each region has a single pie chart that shows percentages of each litter type, out of ten types depicted in the legend. On hovering over the pie chart, the demo attendees will be able to know the actual number of points of each litter type. Zooming on the map to deeper levels shows a pie chart for each sub-area, either in oceans or in continents. This enables scientists and activist to narrow down areas of interest based on certain litter types, e.g., determining that the northern part of the Pacific ocean witnesses the highest plastic pollution.

The environmental scientist can zoom in and out on the map view to show finer granular data on six different levels: continent, sub-continent, country, sub-country, city, and street level. By zooming in/out on the map view, the data is automatically divided or merged to show data that corresponds to the current level. The sub-continent level shows data from multiple countries in one spatial tile and divides the whole world into 10x10 tiles, and so on up to the street level. Figure 2(b) depicts the street level visualization in Riverside, California. At this level, only individual data points are

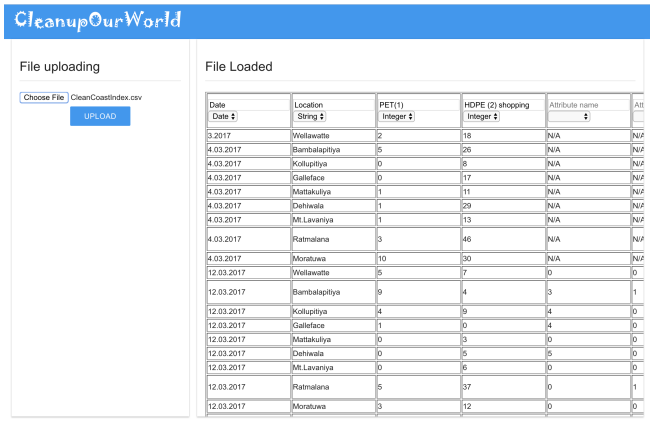


Figure 3. Adding a New Litter Data Source.

shown without any aggregation. The viewed data subset is reflecting the applied filters of litter type, time period, and collecting organizations.

3.2 Scenario 2: Adding a New Litter Data Source

The on-going data collection process and the natural continuity of human litter necessitates adding new datasets to the database. A simple way is adding new data manually through the administrative tools, e.g., PostGIS admin tools. However, this will limit both number of collaborations and amount of data that contribute to our repository. To make it a comprehensive research database that serves as many environmental scientists worldwide as possible, we are enabling external data entry so that scientists and organizations can contribute their data to our repository easily. Figure 3 depicts a data entry screen that enables uploading external data to *CleanUpOurWorld*. The screen allows the user to upload a file of a certain format. Currently supported formats are ESRI Shape files, KML files, XLSX files, and CSV files. By default, the first data row is considered attributes names. After the file is successfully parsed and loaded, the user can rename attribute names and determine their data types. In addition, she can edit any data value in any row so she can possibly add any missing values or correct any incorrect values. Once all corrections are made, the user can submits the uploaded data to our backend database.

The newly uploaded data is not directly integrated with our existing data. Instead, the new data is loaded in a new separate database table and the system administrators are automatically notified with a new dataset addition request. Then, the request is reviewed for data adequacy in terms of completeness of necessary attributes, matching attribute names, and lack of any data problems such as SQL injection data, severe noise data, etc. In several cases, it is needed to follow up with the data owners to fix data problems before allowing the new data to be integrated with the existing data. Once the uploaded data is put into a good shape to be integrated, the cleaning, integration, and loading process of *CleanUpOurWorld* is executed so the new data is processed as part of our database. Although this process is lengthy and might include several cycles of interactions between database administrators and data owners, it still allows much faster process than individual collaborations as uploading and refining the data is much faster.

3.3 Scenario 3: Downloading Litter Data

CleanUpOurWorld will have its prominent value with enabling environmental scientists to use it as data repository where they can *add* and *get* data in addition to being a scalable data manager and visualizer. This open culture encourages the environmental scientists and community organizations to invest time, efforts, and contribute data to this repository, which significantly enrich our collaboration network. For this, we will enable users to download subsets of data based on terms and conditions of each dataset. Figure 2 shows a download button that enables demo attedees to download the subset of data that is currently visualized on the map view with all the applied filters. The downloaded data will include any aggregate data that is shown on the map view, in addition to any individual data points, contributing to these aggregates, that are permitted by the data owners to be downloaded.

3.4 Scenario 4: Administrating Existing Litter Data

Despite the automated tasks provided by *CleanUpOurWorld* in maintaining litter data, database administrators may still need to manually fix data issues, e.g. clean unknown formatted attributes or tweak any dataset-specific problem. Thus, demo attendees will be able to interact directly with the backend database via the administration console through either command line or GUI interfaces. This includes executing SQL queries of different types. Examples are selection queries to view existing data records, update statements to modify certain data values, alter table statements to modify data scheme properties (changing attributes names, data types, adding attributes, etc). In addition, the demo attendees will be able to interact with dynamic SQL functions that are used in several operations such as data loading, part of the cleaning process, etc.

References

[1] C. Guo, Y. Ma, B. Yang, C. S. Jensen, and M. Kaul. EcoMark: Evaluating Models of Vehicular Environmental Impact. In *SIGSPATIAL*, pages 269–278, 2012.
[2] C. Guo, B. Yang, O. Andersen, C. S. Jensen, and K. Torp. EcoSky: Reducing Vehicular Environmental Impact Through Eco-routing. In *ICDE*, pages 1412–1415, 2015.
[3] Let's do it! World. https://www.letsdoitworld.org/, 2018.
[4] Lawmakers reauthorize NOAA Marine Debris Program. https://www.americangeosciences.org/policy/news-brief/lawmakers-reauthorize-noaa-marine-debris-program, 2018.
[5] World Waste Platform. http://opendata.letsdoitworld.org, 2019.