Air Sensors International Conference – May 11-13, 2022

Air sensor data management, visualization, and analysis: understanding and meeting the needs of government air quality organizations in the United States

Gayle Hagler[1], Andrea Clements[1], Ryan Brown[2], Daniel Garver[2], Ron Evans[3], Colin Barrette[3], Ethan McMahon[4,*], Dena Vallano[5], Robert Judge[6,*], Sarah Waldo[7], William Wallace[7], Anna Mebust[5], Corey Mocka[3], David Smith[4], Amanda Kaufman[3]

*[1]US EPA Office of Research and Development, [2]US EPA Region 4, [3]US EPA Office of Air Quality Planning and Standards, [4]US EPA Office of Mission Support, [5]US EPA Region 9, [6]US EPA Region 1, [7]US EPA Region 10, *retired or departed from US EPA*

Background

- Air sensor use in government organizations – a data challenge:
  - Example project: A small network of 36 particulate matter (PM) sensors operated offline for one summer, resulted in:
    - 2992 data and sensor status log files
    - Over 1 million rows of sensor data, with two different formats due to firmware changes
    - Regulatory-grade data for nearby stations, at different averaging periods
    - Extensive custom programming in R to process, quality check, and bias-correct the data, then perform analyses
  - Challenging data, but valuable, real-time information on fine PM ($PM_{2.5}$) conditions to answer research questions
- To move the conversation beyond anecdotes, we began an effort in 2019 to dig into the data issue more deeply

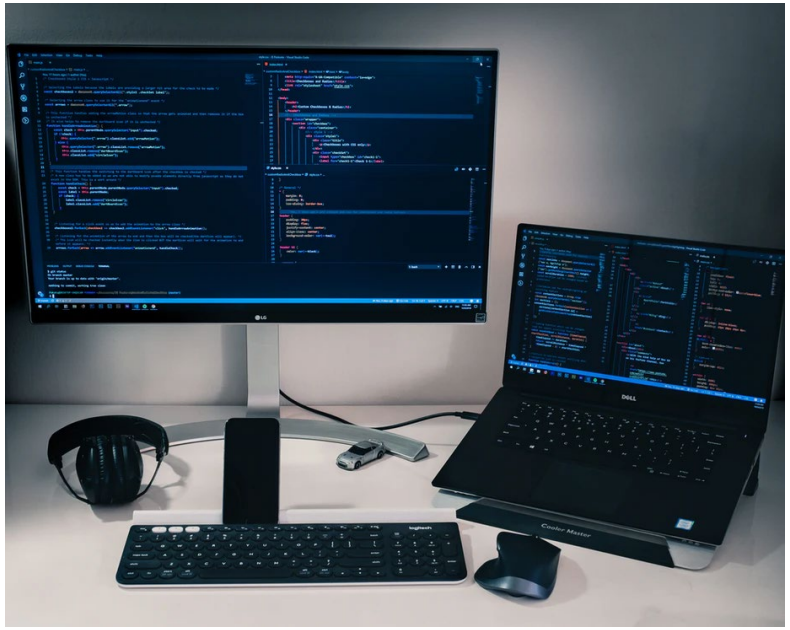
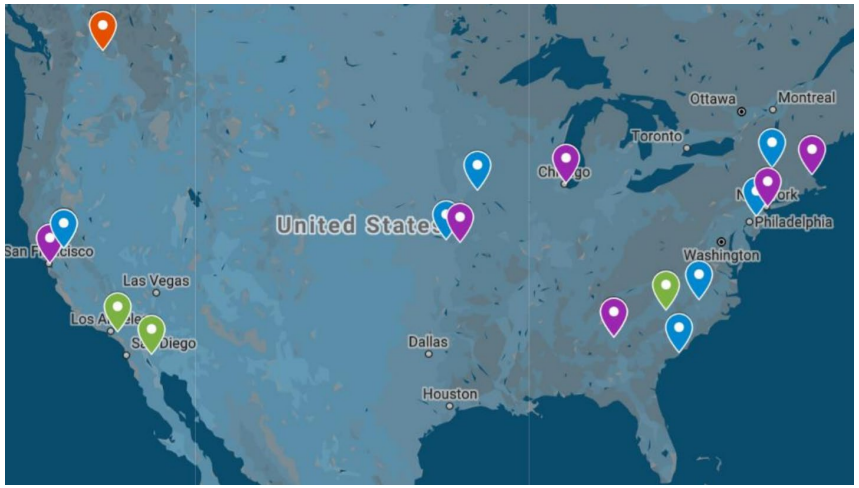


Image: Unsplash.com

Approach: Open-ended dialogues in late 2019

Open-ended dialogues were held in late 2019, focused on:

1. Current sensor data practices
2. Technical pain points
3. Future outlook



EPA Regions (R): 1, 2, 4, 5, 7, 9

Regions were asked to share about their direct use of sensors as well as knowledge of use within their Region by air monitoring organizations

State/Local/Tribal organizations (SLT):

Kansas Department of Health & Environment
Iowa Department of Natural Resources
California Air Resources Board
South Coast Air Quality Management District
Imperial County Air Pollution Control District
Sacramento Metropolitan Air Quality Management District
Georgia Environmental Protection Division
South Carolina Department of Health and Environmental Control
North Carolina Department of Environmental Quality
Mecklenburg County Air Quality Commission
New York State Department of Environmental Conservation
New Jersey Department of Environmental Protection
Confederated Tribes of the Colville Reservation

Dialogue synthesis

Synthesis: Notes were tagged with the following labels, then tagged text was compared across dialogues and interpreted

Current practices

- Level 1 User: limited use (e.g., educational demonstrations)
- Level 2 User: pilot-testing and evaluation of sensors
- Level 3 User: expansive use to meet organization's objectives
- Sensor Type
- Data Use Purpose
- Restricted Data Management
- Unrestricted Data Management
- Showing data to the public
- Data remains private
- Traditional Data Processing and Analysis: e.g., Excel
- Advanced Data Processing and Analysis: e.g., Script-writing (R, Python, etc.)

Future Outlook

- Low Growth Expected
- Moderate Growth Expected
- High Growth Expected

Unmet Needs

- Workforce Needs (i.e., training needs, more staff or staff time needed)
- Sharing of Technical Practices
- Sensor Products - Information and Performance
- New Data Management Solutions Needed
- New Data Analysis Tools or Functionality Needed
- Legal, Data Security or Data Ownership Issues
- Data Standards

Findings

Degree of use varied widely

We categorized the 19 organizations as: **Level I (5 orgs), Level II (11 orgs)** , **Level III (3 orgs)**

Level I: "A quick look at existing data"

- Keeping an eye on the technology developments
- Occasionally viewing public-facing sensor data on a website
- Low amount of time dedicated for analysis, low direct usage of sensors

Level II : **"Increasing use of air sensor data"**

- Using sensors for temporary monitoring to assess a citizen complaint or investigate area of concern
- Testing the sensor data quality by collocating sensors with a reference monitor
- Analyzing sensor data but at a limited level

Level III : **"Full adoption"**

- Sensors integrated into meeting organization's goals
- Sensors used in greater number and for longer periods of time
- Investment in technical staff, data infrastructure to manage large data sets

Findings

Organizational type was a poor predictor of level of sensor use and technical needs

- Level I, II, and III users had commonalities in the purpose of sensor data use and technical needs
- Capacity is another key factor for organizing technical needs
  - Staff time available
  - Staff skill sets for data analysis and programming

All are anticipating sensors to increasingly play a role in their organization's objectives

- Responding to, informing, and loaning equipment to the public
- Use of sensors for screening, temporary monitoring, investigative purposes, and continuous monitoring

Most organizations do not expect additional resources (staff or funds) to support the increasing workload related to sensors

Perspective in 2019: current practices and pain points

User Level	Data Storage	Data Analysis / Processing	Current Pain Points
I	None – sensor use limited	No in-depth analysis – quick viewing on a website or instrument integrated screen	• No easy way to visually compare sensors with nearby reference monitors • Feel they are not up to speed on the latest information on sensors, projects, and best practices
II	***Mixed:*** On-board storage and manual download OR Dependent on manufacturer server	***Mixed:*** Excel or other spreadsheet-based analysis (e.g., JMP) OR Custom scripts developed by the user (R most common followed by Python)	• Limited staff time available • Time burden of data management and analysis • Learning curve and time burden to do custom-scripting (“everyone is an amateur coder”) • Lack of data standards • Insufficient information from sensor manufacturers on data
III	Custom-built data management system; cloud-based or transitioning to cloud; some use manufacturer servers	Custom-developed scripts and user interfaces (R, Python)	• Burdensome cost of cloud-based data management • Lack of data provenance / data standards • Uncertain data security • Time / effort needed to make data understandable to the public

Level II users: pain points with expanding use

1) **Data management methods –** *especially meeting data security needs and IT policies of government organizations*
2) **Code sharing and code development support –** *especially open source languages*
3) **Quick air sensor data analysis and visualization, integrated with other data**
4) **Data standards**
5) **Increased technical information sharing**

"The volume of data is pretty overwhelming." – Kris Ray, Confederated Tribes of the Colville Reservation

Data pain points

Level II and Level III users are grappling with several sensor data-related issues throughout implementation

Online sensors (cell, wifi)

Offline sensors (on-board data storage)

Multiple types of platforms, data online/offline

Cost of equipment and data management

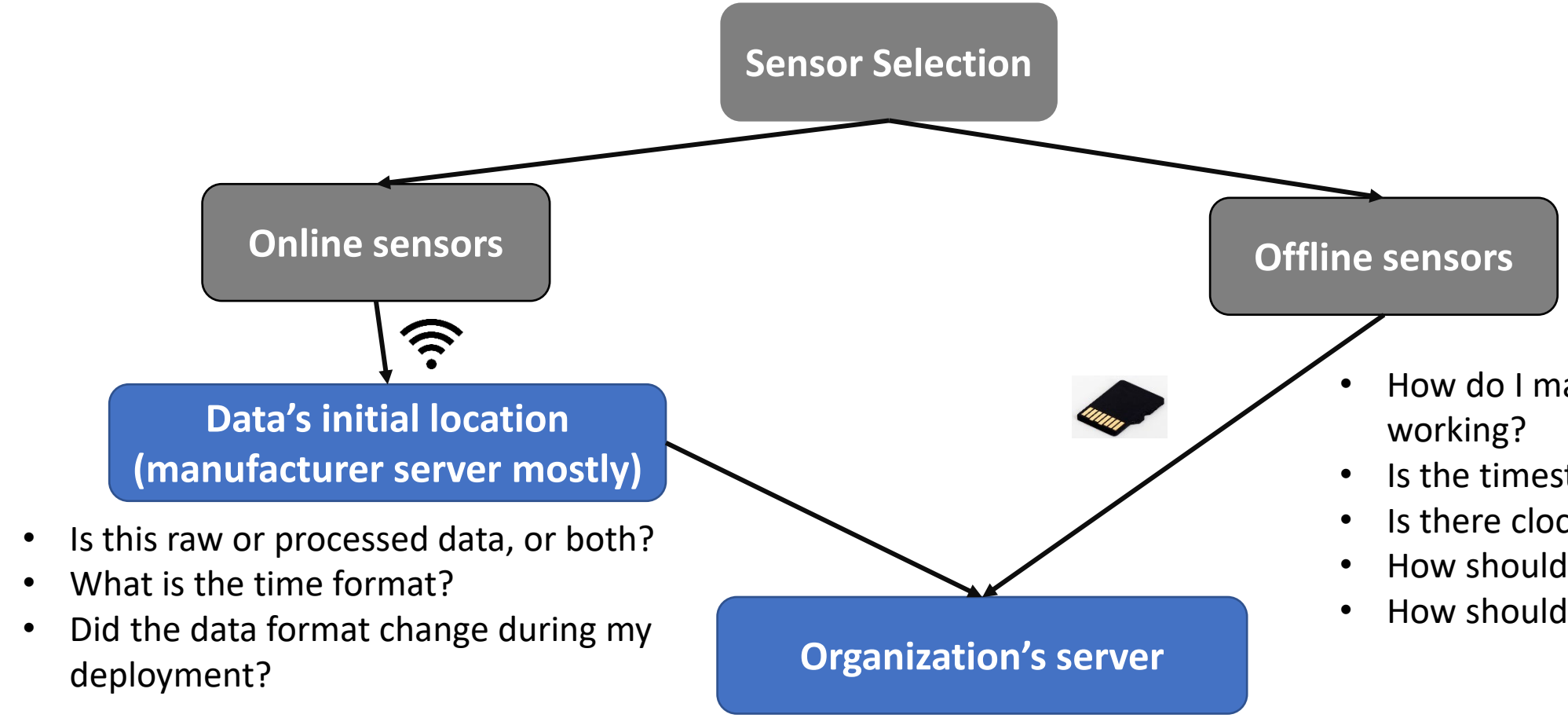
- Can this equipment interface easily with my existing data management system?
- Can I handle the workload of manual data downloads?

Information security limitations

- Am I allowed to use wifi at the deployment location?
- Do I have full ownership and control of the data?
- Am I allowed to have my organization's data stored on the manufacturer's cloud / server?
- Do I trust how a manufacturer would treat the data? (data processing algorithms, data ownership)

Data flow pain points

Level II and Level III users are grappling with several sensor data-related issues throughout implementation

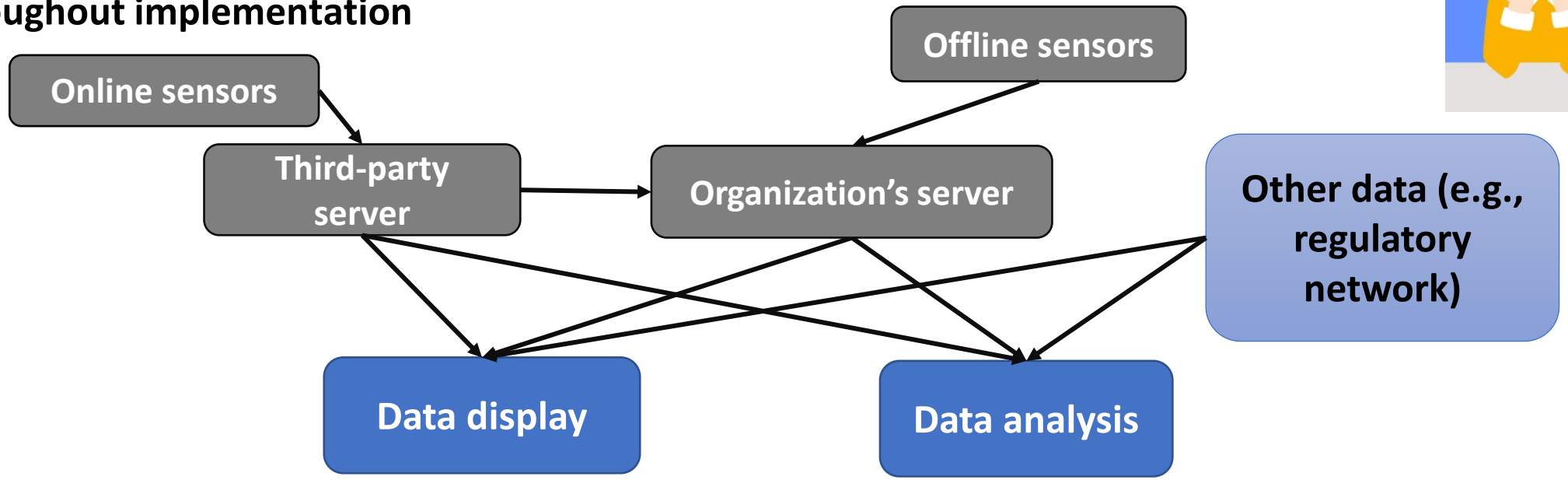


- Is this raw or processed data, or both?
- What is the time format?
- Did the data format change during my deployment?

- How do I make sure the sensor is working?
- Is the timestamp accurate?
- Is there clock drift?
- How should I process the raw data?
- How should I organize the data?

Data visualization and analysis

Level II and Level III users are grappling with several sensor data-related issues throughout implementation



Data display mind set:

- If publicly viewable through manufacturer – are the data quality-checked / processed to be reliable source of information to the public?
- Could I incorporate the data into my public data website?

Data analytics mind set:

- Are data analysis tools provided by the manufacturer?
- Do we have staff with time/skills to do advanced data analysis?
- Do we have resources to develop custom solutions?

Many desired attributes for data visualization and analytics

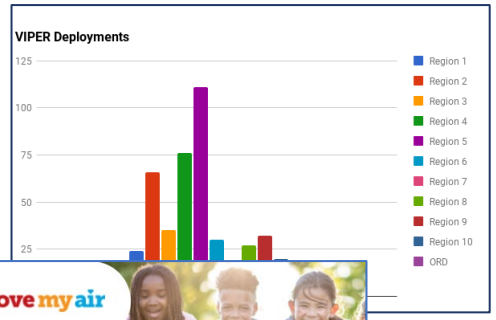
- Low technical barrier to use (non-programmer friendly)
- Allows offline data to be manually added
- Supports data screening
  - Outlier detection / flagging
  - Application of data correction schemes if available for a particular sensor
- Supports geospatial exploratory analysis
- Supports combined air pollution and meteorology analysis
  - Wind flow patterns
  - Pollution-wind comparisons (e.g., pollution roses)
  - Measurement artifacts (e.g., PM sensor data vs. relative humidity)
- Avoids implication of endorsement (of the data, of the sensor types)
- Training support provided (videos)

After the dialogues

Second phase of effort – over 2021-2022 four EPA staff teams explored the current landscape of options, with an updated focus after deliberation on:

- Data hosting options
- Code sharing and code development support
- Data Quality Objectives and Data Quality Indicators
  - *A transition and broadening of focus from data standards*
- Quick air sensor data analysis and visualization options

Data hosting team



AirNow

South Coast AQMD

Agilaire LLC

openaq

PurpleAir

aeroqual

Review of a spectrum of options to understand the variety of strategies for sensor data management

Questions we asked:

- Who owns the data hosting system?
- What is the primary purpose of the data hosting system?
- Are there standard data import / upload options?
- Does the system support off-line / batch data uploads?
- What data time-averaging intervals are supported?
- Who has access to use the system?
- What is the cost to use and who owns the data?
- What are future development plans?

QREST: QUALITY REVIEW AND EXCHANGE SYSTEM FOR TRIBES

Code sharing and support team

Focused exploration on options compatible with current EPA IT:

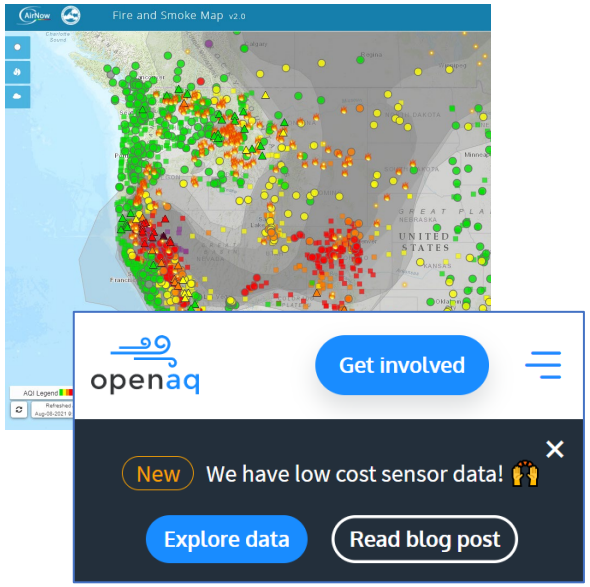
Questions we asked:

- What are options for technical exchange of information?
- What are options for code co-development?
- Where could a code repository be housed?
- How can code developers / users find and receive expert troubleshooting advice?
  - Note: R identified as most widely used open source language, followed by Python

Exploration was limited to options EPA could help facilitate, with our current IT restrictions

Data analysis and visualization team

AirNow Fire and Smoke Map

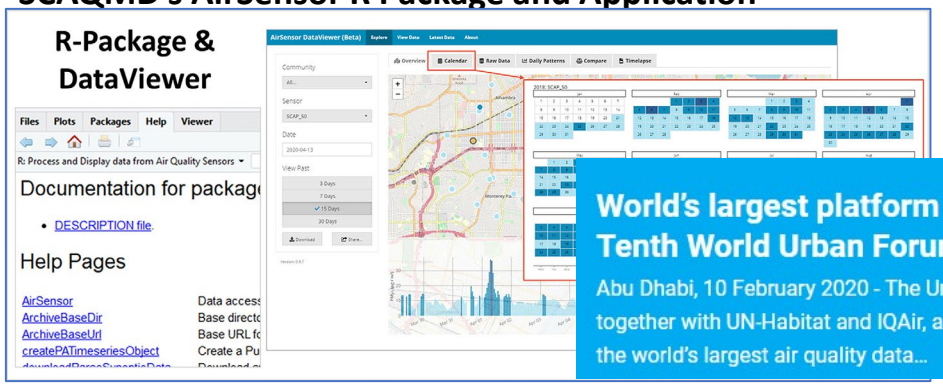


Discussion with organizations currently building or sustaining air sensor data analysis and visualization applications

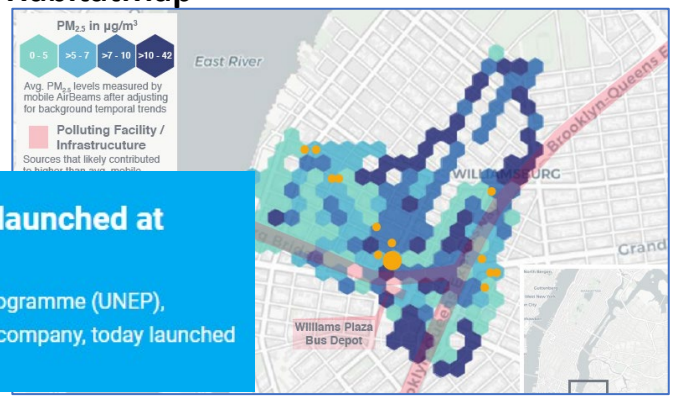
- How do the dialogue participant needs line-up with the application?
- What are the future development plans?
- What could be opportunities to foster further development to meet the identified needs?
- Given multiple applications under development in parallel, can EPA add benefit as a convener, community-builder, connector of application developers while representing end user needs?

EPA's RETIGO

SCAQMD's AirSensor R Package and Application



HabitatMap



DQOs and DQIs team

Focused exploration of past or ongoing non-regulatory air monitoring projects conducted by government organizations in the USA – including use of sensors and other non-regulatory monitors.

- What were the monitoring objectives?
- What were the measured parameters and technologies used?
- What data quality objectives (DQOs) were set?
- What data quality indicators (DQIs) were utilized during data collection?
- Were any data standards used?

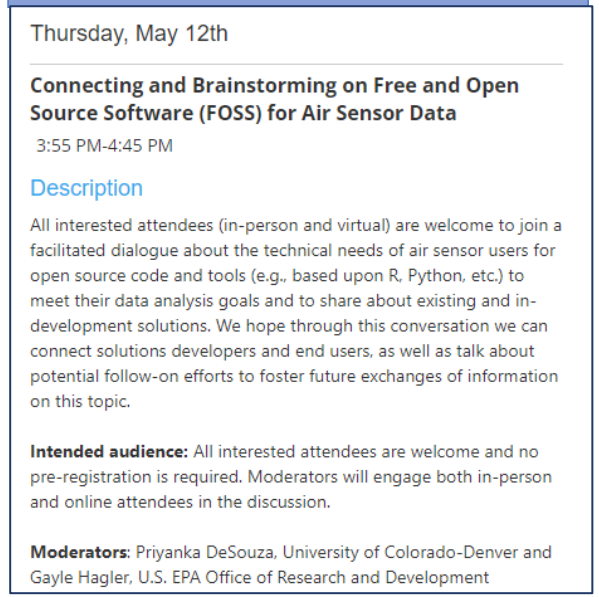
We cross-walked commonalities across projects with similar objectives and discussed further information or guidance EPA could provide to complement or update what exists.

Image: EPA's Village Green Project location in Durham, NC

Summary and next steps

- As sensor use increases, government organizations in the “Level 2” category feel the data pinch – continued innovation is needed in data management, visualization, and analysis solutions.
- We plan to release a number of our past research project quality assurance documents and undertake new research related to air sensor network quality assurance.
- Exploring initiating a developer and end user community for free and open source code for air sensors; building and releasing code from EPA in-house research efforts.
- EPA plans to continue efforts to build, support, and cheerlead solutions for this challenge.

Related: All are welcome to join the conversation!



Thursday, May 12th

Connecting and Brainstorming on Free and Open Source Software (FOSS) for Air Sensor Data

3:55 PM-4:45 PM

Description

All interested attendees (in-person and virtual) are welcome to join a facilitated dialogue about the technical needs of air sensor users for open source code and tools (e.g., based upon R, Python, etc.) to meet their data analysis goals and to share about existing and in-development solutions. We hope through this conversation we can connect solutions developers and end users, as well as talk about potential follow-on efforts to foster future exchanges of information on this topic.

Intended audience: All interested attendees are welcome and no pre-registration is required. Moderators will engage both in-person and online attendees in the discussion.

Moderators: Priyanka DeSouza, University of Colorado-Denver and Gayle Hagler, U.S. EPA Office of Research and Development

Related: Poster #69 – “*sensortoolkit* - A Python Code Library for Standardizing the Ingestion, Analysis, and Reporting of Air Sensor Data for Performance Evaluations”

Related:
GitHub: https://github.com/USEPA/sensortoolkit
PyPI: https://pypi.org/project/sensortoolkit/

Thank you for your time!

Points of contact: hagler.gayle@epa.gov and clements.andrea@epa.gov

Acknowledgements:
EPA: David Smith, John Masters (retired), Marta Fuoco, Stephen Krabbe, Mary Clare Hano, Joe Schaefer, Alysha Thompson, Catherine Taylor, Rachael Leta-Graham, Anahita Williamson, Leland Grooms, David Bright, Andy Hawkins, Gwen Yoshimura

SLTs: Representatives from Kansas Department of Health & Environment, Iowa Department of Natural Resources, California Air Resources Board, South Coast Air Quality Management District, Imperial County Air Pollution Control District, Sacramento Metropolitan Air Quality Management District, Georgia Environmental Protection Division, South Carolina Department of Health and Environmental Control, North Carolina Department of Environmental Quality, Mecklenburg County Air Quality Commission, New York State Department of Environmental Conservation, New Jersey Department of Environmental Protection, Confederated Tribes of the Colville Reservation, City & County of Denver, Puget Sound Clean Air Agency, QREST