Public access to fine-grained city air quality data from roving sensors

PAIRQURS is the public data access component of the project LIFE+RESPIRA (www.liferespira.eu). A network of 50 portable air pollution sensor suites are carried by a team of volunteer cyclists during their daily commutes throughout the city of Pamplona, Spain. The sensor suites record at 5 Hz the levels of selected atmospheric pollutants (CO, NOx, airborne particles) as well as auxiliary data (T, HR) and GPS coordinates and transmit processed packets via GPRS messaging to a central database. Data records are fed to a computational fluid dynamics (CFD) model built on the cityscape, and the model in turn enables an online route planner allowing commuters to select low-pollution paths for their cycle, motor, or pedestrian commutes.

Contaminant gasses and particles are recorded at very fine spatial and temporal resolution, and transmitted in near-real time for processing. A huge volume of data is being produced and, after heavy processing, serves to feed an air quality model allowing prediction of best routes for city dwellers. The pilot wants to ensure that citizens and researchers alike can fully access the pre- and post-processed data for any scientific, social, or policy purpose.

The Scientific Challenge

The sensor suites (up to 50) are reading each up to 10 environmental, geo-located, multiple-data parameters at a rate of 5 Hz. Despite heavy internal processing and averaging, we are still producing cumulative, stored data at a high rate. Although these have a limited life for any immediate purpose (e.g. what is NOW the level of this contaminant HERE), air quality models are extremely sensitive to many variables: time, weather, climate, urban structure, winds, etc.; only a large, distributed, nearly-continuous dataset can account for the parametrization of the models—that is, ALL “past” data are useful to understand how air quality is, was, and will be under current and future conditions. Thus, we need to build and store a multi-million-record dataset at a raw resolution better than 10 meters and 10 seconds.

These data may prove invaluable to analyse how air pollution evolves in a city—not only as an overall parameter, but at a human scale. The dataset could thus be used to build models that go beyond the statistical average for an area, down to what the individual can experience during his or her daily walk or ride. Corrective measures could be applied when and where they matter most. Research that we haven’t even figured could be undertaken on the data, and we want to ensure that that research is possible.

Within the LIFE+RESPIRA consortium there are several research subjects that need to filter, select, and group the data according to specific needs—and therefore the project will be the main user of the data at first. But at a larger scale, we want these data to be made available to all: other scientists, officers, technicians, policy makers, and ordinary citizens that may also require selecting and combining the data as they see fit.

Who benefits and how?

The community directly targeted by this pilot, beyond the LIFE+RESPIRA consortium, are the urban air pollution research communities and consortia, and the local, regional and national air quality services that will benefit from the availability of fine-grain air pollution data in near real time through a monitoring model that is expected to extended to European cities.

Another large external consortium targeted is the Long Range Transboundary Air Pollution (LRTAP) working group, in charge of tracking air pollution and airborne pollutant movements, that will experience a dramatic increase in the availability of high-resolution air pollution data across a range of cities throughout Europe if the model is successfully exported.

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

Among the technical communities targeted, the urban services of the cities involved will be the main users of the model and dataflows, enabling them to make plans and take action according to the availability of high-precision, high-resolution, timely data. The beneficiaries of the dataflow, however, will extend to the general population of the cities involved.

Finally, the general population is the ultimate receiver of the dataflow, and the pilot will allow those interested to obtain fine-grain data tailored to their needs. For example, individuals will be able to download data from their immediate neighbourhood, and observe trends, levels, and patterns along time that may allow them to take decisions about how to deal with the air quality around their living or working areas

Technical Implementation

Raw data are pre-processed within the sensor suites prior to upload over GPRS to the provider's (KUNAK) services as individual, time-stamped, tagged tuples. A set of routines (REKAK) have been written to harvest these tuples and ingest them into a master database, together with initial processing.

B2DROP has been implemented as the primary seat of:

- The master database collating all raw and processed records;
- Auxiliary tables containing environmental and calibration data; and
- Processing routines.

The dataflow calls for the master database to be updated exclusively through the harvesting routines. Group members working on further analyses down the path must operate locally during the development phase, but need to do so on either a full local copy of the database while the harvesting routines continue to execute cyclical harvests population the master database, or on a set if exported files being cyclically produced from the master database by processing routines.

In parallel, during the test and implementation phases background tasks run on the B2DROP folders to maintain local and external backups of the master database.

B2ACCESS has also been tested as authentication platform for use within B2DROP or B2SHARE.

Preliminary Results

B2DROP's synchronization capability and lock strategy under file updates were analyzed and found to be appropriate for our dataflow. Therefore, the service is tasked with (1) holding the master database being cyclically updated by automated routines; (2) allowing dissemination across the group's development and production machines of the export files.

The master database is being internally shared as expected through B2DROP, using 55% of the available shared space. The synchronization occurs every four hours and takes about one hour. The harvesting and processing routines pipe the processed data along the corresponding raw data to a series of sequential 2-GB files, each containing approximately 6-month worth of data. A timed routine fires every four hours to create a filtered CSV copy of fit-for-use, valid, non-compromised processed data table that can be directly imported by analytical packages by any team group to which the B2DROP repository synchronizes.

In addition, B2SHARE has been evaluated to act as the public repository of the raw and processed data. Through several workshops it has been determined that in order to function as desired, granular access to the data-record level (as opposed to file-level) should be necessary. While this is not the primary intent of B2SHARE, we could still use it as a file repository for long-term storage and access.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Contacts

- Arturo H. Ariño, University of Navarra, artarip(at)unav.es
- Margareta Hellström, ICOS Carbon Portal, margareta.hellstrom(at)nateko.lu.se

Further Information

Learn more about the LIFE+RESPIRA/EUDAT Collaboration here
Scientific Challenges behind the Pilot
Read the full interview with Arturo H. Ariño

Read more

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.