Urban Observatories Sensor Report

Philip James, Jennine Jonczyk, Daniel Bell – Newcastle University
Lee Chapman, Nicole Cowell – University of Birmingham
James Evans, David Topping, Thomas Bannan, Ettore Murabito, Emma Tsoneva – The University of Manchester

1.0 About

This report collates the sensor metadata from three Urban Observatory (UO) catalogues – Birmingham, Manchester, and Newcastle. This includes a combined overview of available data, metadata themes per observatory, a discussion of quality metrics, how to access each observatory API endpoints, a discussion on ethics, and questions to consider moving forward.

2.0 Metadata Overview

2.1 Number of Sensors

In total the UO's presently consist of 2146 sensors. However, it must be noted that not all of these are unique hardware devices, but also comprise of software generated detectors. For example, a unit may measure several parameters using a range of sensors or each scan line drawn within a single camera view (e.g., for people counting in different areas of the scene). Each is counted as its own sensor stream despite belonging to the same physical device. This reflects the underlying structure of the data where observations (of a phenomenon in space and time) are the units of record, rather than the physical devices that record them. Collectively we use the nomenclature "Sensor Streams" to differentiate a stream of data from a device (as a device may provide multiple sensor streams).

Urban Observatory	No. UO/University Owned Sensors	No. Third Party Sensors	Total no. of Sensors
Birmingham	132	61	193
Manchester	12	480	492
Newcastle	1178	386	1564
Total	**1357**	**871**	**2246**

Table 1 – Number of Sensors

2.2 Data Availability History

The Urban Observatories have been deploying their own sensors and accessing third party streams for data collection since April 2015. However, available data may go back further as observations were often based on previous sensor networks deployed in cities prior to the project. In addition to this, available datasets were extracted from third party providers (such as Met Office API Wind Speeds) will be obtainable for significantly longer time frames. For an overview, the data timeseries presented in table 2 is taken from first to last sensor reading per theme (from any of its sub-categories) and does not consider any missing data between these periods.

Discipline	Data Timeseries
Meteorology	Jun 2013 – Present
Sound	Nov 2017 – Apr 2019
Atmospheric Chemistry	Aug 2016 – Present
Hydrology	May 2015 – Present
Vehicle Traffic/Parking	Jun 2013 – Present
People Counting	Jun 2018 – Present
Biodiversity Indicators	Jun 2017 – Aug 2020
Carbon Capture	Aug 2017 – Mar 2018
Internal Building Monitoring	May 2015 – Jan 2018
Electrical Power	Jul 2015 – Present

Table 2 – Available Data – Disciplines are ad-hoc collections of sensor streams that are utilised in UO interfaces for collection and aggregation. Disciplines are recorded as metadata within an individual observation.

2.3 Urban Observatory Data Scale

In order to understand the approximate scale of each observatory in terms of live input frequencies and archived data, Table 3 depicts average number of records across all sensors per minute, as well as the number of total records collected since inception.

Urban Observatory	No. Live Sensor streams	Av. no. Records Per Day	Total no. Records Ever
Birmingham	605	147,288	162,000,000
Manchester	4411	998,379	402,576,140
Newcastle	1722	967,000	2,284,270,579
Total	**6738**	-	**2,848,846,719**

Table 3 – Approximate Data Quantity (as of 13th December 2021)

3.0 Metadata Themes

Across the three audited UOs, sensor types can be categorised into ten general themes, each containing a subset of variables and measurement units, as presented in table 4. Currently, sensor types are largely dedicated towards Earth atmosphere observations, such as meteorology and atmospheric chemistry, followed closely by hydrology and vehicle monitoring.

Theme	Variables	Units
Meteorology	Temperature	°C
	Wind Speed	m/s
	Wind Direction	°
	Pressure	Pa // hPa
	RH	%
	Precipitation Depth	mm
	Precipitation Rate	mm/h
Noise	Sound	dB
Atmospheric Chemistry	Temperature	°C
	O3	µg/m^3 // ppb
	CO	ppb
	CO2	ppm
	NO	ppb
	NO2	µg/m^3 // ppb
	NOx	µg/m^3
	PM1	µg/m^3
	PM10	µg/m^3
	PM2.5	µg/m^3
	Relative Humidity	%
	Dew Point	°C
Hydrology	River Level	m
	Relative River Level	m
	Relative Tidal Level	m
	Journey time	m/s
	Sewage Level	m
	Water Temperature	°C
Vehicle Traffic/Parking	Occupied spaces	#
	Plates Count	#
	Matching Plates	#
	Journey Time	HHMMSS
	Traffic Flow	Vehicle/Minute
	Parking Spaces	#
	NO2 Concentration	µg/m^3 // ppb
	Vehicle Speed	mph
	Vehicle Type	Car, bus, van etc.
	Cyclists	#
People Counting	Van Count	#
	Walking X Direction	#
Biodiversity Indicators	Beehive Weight	g
	Beehive Temperature	°C

Carbon Capture	Soil Temperature	°C
Internal Building Monitoring	Temperature	°C
Electrical Power	Real Power	Real Power +kW

Table 4 – Data Themes, Variables, and Units

4.0 Assessing Sensor Quality and Reliability

There are three principal ways in which data quality is currently validated at the UO:

1. Event triggers – This quality check is largely an automated one. It consists of identifying questionable values, such as those greatly differing from the ordinary, above/below a defined threshold, or impossible values (e.g., a negative NO2 reading), and assigning an invalid flag within the database if it meets any of these constraints. Such information is then prevented from being accessed to users via the API. However, the many unique or edge-case possibilities make defining all of these constraints a difficult task, and therefore must require ample human intervention to stress test the system, take spot tests, update the list of constraints, and relate returned values to real-world contexts for validation.
2. Sensor calibration – At set periods, some sensors (such as those used for air quality measurements) are recalibrated against higher quality precision stations. This involves co-locating both sensors at a common reference point (if possible, at a site that will replace the actual measurement location) and taking continuous measurements for at least 14 days. During this time, slope and offset are calculated to determine linear regression, from which sensor records can be adjusted accordingly. Depending on the application, the measurement period can be extended – for example in capturing seasonal differences in sensor performance. An advantage of this calibration methodology is that the UO's can work alongside sensor manufacturers in order to help improve their algorithms. Some of the best advances in UO sensor quality have been made by this process.
3. External feedback – At larger UO's, such as Newcastle, the final check simply involves listening out for any customer feedback on data quality, through the UO email reporting system, or on platforms such as Twitter. This community of people are consistently accessing Urban Observatory data and are often the first to know of any discrepancies or issues of data retrieval. Furthermore, as those working externally greatly outnumber those who are working internally, they have a much better chance of finding any problems. However, this does assume customers are actively employing their own data quality checks. In the case of smaller UO's, such as

Manchester, where the number of sensors are fewer and there is much less customer feedback, more time and resources can instead be applied to maintaining individual sensors to a higher standard.

There are some considerations to be taken when accessing UO sensor data in terms of quality and reliability:

- The sensors are generally low cost – UO sensors are relatively low cost, but this depends largely on the type of sensor. Widely deployed air quality sensors typically range from £6-9k. Given that the uncertainties from low-cost sensors are not always linear and/or repeatable, and are often poorly defined, it is important that the use of UO data (especially that of air quality sensors) is considered and used in such a manner. For these reasons, data published by the Urban Observatories are intended to be interpreted as a set for analysis on a city-wide scale, rather than results at a single point in location and time. However, this does not mean individual results should be considered completely invalid, as erroneous devices can easily be singled out in the network based upon uncharacteristic results when compared with the many other nearby sensors of the same model. This approach reflects efforts being made in the wider IoT research community to use co-location studies as a reliable approach for making informed decisions.
- Sensor quality information is unavailable within published data – On UO websites, general procedures are described which outline any calibrations or corrections which may be applied to the data by manufacturer or UO alike. Despite this, sensor quality metrics such as precision and accuracy are not readily available through UO API's, leaving it somewhat ambiguous to data users as to the reliability and purpose of each sensor. One solution to this could be to have a UO ontology which is solely dedicated to sensor quality, where sensors may have a score (such as a Key Performance Indicator (KPI)) or expected use context description, based upon hardware specifications and expected longevity in the outside world. Furthermore, in such a case it would also be highly practical to include a note on whether the sensor is a citizen kit, where deployment and maintenance is not necessarily carried out by trained professionals. However, the implication of indicating data quality for a particular sensor is that it may be perceived as passing judgement on the quality of a specific company's product.
- Data frequency – The frequency of retrieved data from any one sensor can largely be attributed to either hardware constraints or sensor use context. In the case of hardware constraints, sensor capabilities are limited to the specifications set by respective manufacturers. However, in some

cases, the rate at which such sensors can return information can be modified in order to optimise for either energy consumption or rate of data return. In terms of sensor use context, the amount of requested data may be limited by software design in order to only receive what is required. For example, extracting only one frame per five minutes from a traffic counting camera still provides enough general information, whilst greatly reducing network bandwidth.

- Live stream vs Archived streams- Whilst the observatories host several live streams, sensors will be retired over time either due to reaching their lifespan or project constraints. Currently on the observatories, some data from sensors which are no longer live is available, but this may be archived over time. Although live streamed data is key in the context of digital twins, historic data may be useful for setting context and creating background datasets for models.
- Sensor lifespan – As with any type of hardware, sensors have an expected lifespan based on their components, build quality, battery life, and external conditions they are exposed to. These largely differ by sensor type – for example, road traffic cameras which have a continual power supply and are generally mounted highly above the ground, typically last many more years and months than air quality sensors, which rely on being more exposed to the elements and on frequent maintenance and calibration for sensitive components. The UO's therefore have critical operations structures in place in order to manage the day-to-day running and maintenance of sensors, as well as longer term funding, upkeep, and expansion.

5.0 API Information

5.1 Standards

Urban Observatory API standards are a work in progress. They can be viewed on the following sites:

- https://urbanobservatory.github.io/standards/#struct-geography
- https://urbanobservatory.stoplight.io/

5.2 Development Potential

Previous conversations between the Observatories have aimed to develop a unified set of API standards, where a shared ontology can make accessing data from any source easier for users resulting in a partially completed set of standards. The Observatories all differ in terms of how much they follow the existing standards, with Birmingham leading the way out of the three discussed in this report. In moving forward internally, and externally with partners such as DfT, the following considerations should be taken in developing API architecture:

- A standardised vocabulary – An issue which is regularly brought up in conversation is the lack of consistency when it comes to vocabulary, in terms of hierarchical structures, themes, and sensor categorisation. As the use of understood terminology is a critical mechanism when it comes to accessing information via an API, this concern should be prioritised when developing a unified set of standards. In many cases, this may simply mean changing item names to match an agreed upon term. However, it has been suggested that a formal hierarchical system could be established, whereby within each theme there are defined platforms (used for keeping track of fixed/changed sensor locations), followed by individual sensors (including variables and measurement units), and finally individual observations at the lowest tier. Ultimate decisions on what modifications should be made must continue to prioritise expected end users and the intention behind the Urban Observatory open-data model.
- Versioned APIs for AGILE development – In order to create a continuous cycle in which API architecture can be constructively designed, developed, tested and evaluated, all Observatories should continue to implement versioned APIs in future updates.

5.3 Access

API access along with instructions can each be found on respective UO websites:

- https://newcastle.urbanobservatory.ac.uk/api_docs/
- https://manchester-i.com/
- https://data.birminghamurbanobservatory.com/map/platforms

5.4 Data Download Formats and Post Processing

Data download formats offered by the UOs come mostly in JavaScript Object Notation for Linked Data (JSON-LD) and Comma-Separated Values (CSV). While JSON-type formats are often expected and preferred by technical users who are frequently streaming/requesting data packets, this is often not the case for any non-technical user who would likely prefer working with the spreadsheet style of a CSV. In order to improve open data accessibility for all users, work is underway to process all forms of archived records into easily downloadable hourly, weekly, monthly, and yearly CSV aggregates.

6.0 Ethics

The UOs go through ethical approval through their respective research ethics systems. In addition, Newcastle has conducted a full Data Protection Impact Assessment (DPIA) (specifically around potentially personal data such as CCTV). No personally identifiable data is stored in UO systems. For instance, CCTV pedestrian counts are aggregated to 5-minute totals (total number of pedestrians crossing a line every 5 minutes). AI based image-analytics are run automatically without human intervention and video is deleted after processing. Site selection is based on discussions with local stakeholders (e.g., the local authorities) or to support other research activities. We are working with the Alan Turing Institute to understand issues around sensor placement inequality[1] and are developing a tool to assess inequality of sensor placements using ONS data.

7.0 Questions Moving Forward

1. Potential use cases of real-time data (in DfT and beyond) impact metadata, API and data structures
    a. Eg. Pull requests for historical or near past data extraction v's streaming APIs for integration into real-time monitoring/analysis
2. Quality metrics and how to convey these are often raised by users and data managers – challenges include
    a. Using data inappropriately
    b. Assumptions that data is always correct
    c. Quality is associated with purpose not the observation itself
    d. What Level of Service/QA needed for intended use eg. By DfT

    Solution may include:

    1. Purpose / task-based metrics
    1. Suggested methods of analysis

[1] https://www.turing.ac.uk/research/research-projects/spatial-inequality-and-smart-city

3. Required formats and download structures (eg. JSON, CSV, NETCDF etc.)
4. Additional data needed for real-time sensor streams (in DfT metadata catalogue)
    a. Compatibility with existing API and UO approaches (and vice versa)
    b. Sustainability issues (I.e. duplicate data holdings)
5. What additional post-processing or services might be required
6. Management of external data streams and the stakeholders who own them
7. Alignment with emerging and new standards for metadata, data query and response
8. Additional services that can be applied to the data to improve the user experience e.g. temporal averaging and aggregation, geographical aggregation etc.