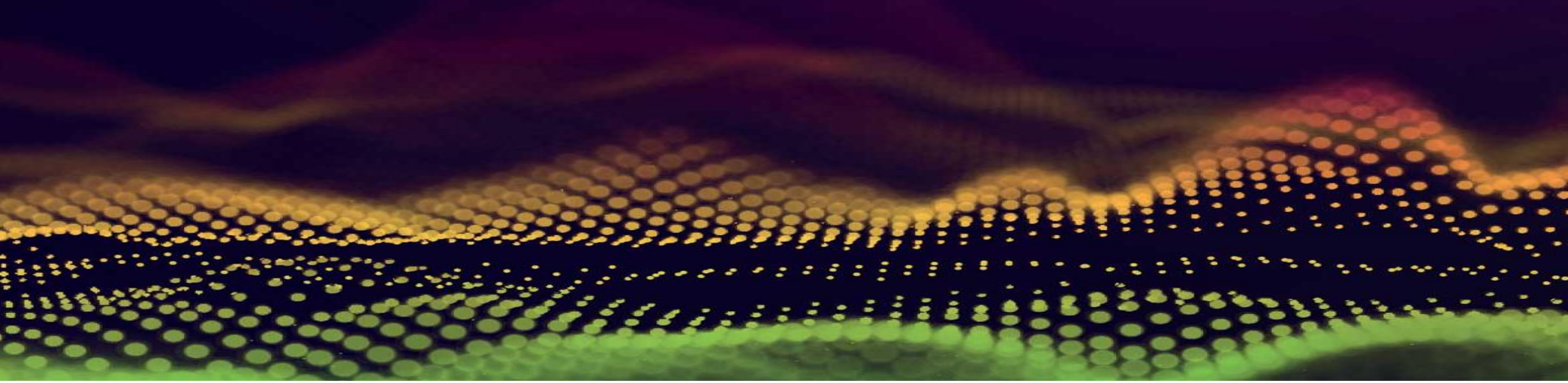


Practical Approaches for Web Scraping

Using AirBnB Data and the UBDC AirBnB Scraping Platform

Urban Big Data Centre

JOINTLY FUNDED BY

UKRI Economic and Social Research Council

University of Glasgow

Programme

- **10:00 - 10:05** Welcome and Housekeeping
- **10:05 – 10:45** Introduction and Background
- **10:45 – 11:45** Technical Walkthrough
- **11:45 – 12:00** Closing Discussion and Wrap-up

Urban Big Data Centre

Welcome and Housekeeping

- Session is being recorded. An edited video will be published on UBDC's YouTube channel in due course (https://www.youtube.com/UrbanBigDataCentre)
- All slides and code will be available to course participants
- Questions can be issued via the Zoom chat – we will have time at the end for these but they will be monitored throughout
- Please keep mics muted throughout the session

About the Urban Big Data Centre

- UK Government (Economic and Social Research Council) & University of Glasgow funded
- Priorities:
  - Data infrastructure and collections
  - Priority research strands: transport & mobility; neighbourhood, housing & environment; education, skills & productivity; big data & urban governance
  - Combining social science research with data analytics and computing science
- Overall aims:
  - Achieve public policy impact
  - Critically evaluate role and value of big data and urban analytics
  - Enhance data and methods

“Promoting innovative research methods and the use of big data to improve social, economic and environmental well-being in cities”

Urban Big Data Centre

Data Service Aims and Objectives

- Overall Aim – to promote innovative research methods and the use of big data to improve social, economic and environmental well-being in cities
- Specifically:
  1. To build ***an outstanding data collection*** which supports high quality, impactful research on UK cities.
  2. To ***maximise the value of the data collection*** by assessing, documenting and enhancing data quality.
  3. To provide a ***high standard, secure, resilient service***, raising demand by ensuring widespread awareness of the service across the UK social science community and beyond.
  4. To ***build capacity*** across the UK social science community for the use of new forms of data through training, the provision of guidance and the development of open source tools.
  5. To play a leading role in debates about the ***policies and practices*** needed to realise the value of new forms of data for urban analysis.

Urban Big Data Centre

Data acquisition – a major challenge

- Getting hold of data to support our research takes up a lot of our time

- There are many reasons why organisation **do not** make their data available
  - Lack of capacity to do so
  - Perceptions of risk (e.g. incompatibilities with privacy legislation)
  - Perceived conflict with existing business models
  - Fears that research results will reflect badly on them, expose them to criticism or be received negatively by other stakeholders

- Conversely, there are several reasons why it **might be a good idea**
  - Establish credibility of data as robust evidence base
  - Facilitate potentially beneficial analysis
  - Often datasets are by definition publicly owned
  - Transparency just might be a good thing

Urban Big Data Centre

Case study – AirBnB (1)

- AirBnB is an online marketplace for arranging or offering lodging, homestays or tourism experiences
- What are the impacts of the rapidly growing sharing economy for private rented sector property market?
  - AirBnB may not be enthusiastic about facilitating such research
  - As an unregulated sector there is little reliable data available

Case study – AirBnB (2)

- Exploring questions such as:
  - What impact does AirBnB have on the availability of private renting stock?
  - Where is that impact greatest?
  - What is the relationship with areas of social deprivation? Does the rise of the sharing economy increase inner-city gentrification and the suburbanisation of poverty?
  - Are AirBnB properties falling below the standards for the PRS e.g in terms of occupancy levels?
  - What impact does AirBnB have on the existing hospitality industry?

Case study – AirBnB (3)

- We wanted to automate the **scraping** of AirBnB on a systematic basis to capture and store publicly accessible web content and facilitate research
- Aiming to understand legality of our proposed data collection
- Methods designed to enable us to demonstrate empirically and reliably:
  - Scale and growth of AirBnB
  - Spatial change and focus of short-term lets
  - Occupancy
  - Price
  - Availability

(Introduction to Scraping)

- Scraping means automating the transfer of data from a web site
  - Text pattern matching
  - HTTP programming
  - HTML / DOM parsing
  - Computer vision
- Unsupported by the web content owner, prone to break if the website changes
- Negotiate website controls intended to block nuisance hosts

Existing technical resources

- Various open source projects are out there that we evaluated during the initial stages of our project
  - airbnb-data-collection - https://github.com/tomslee/airbnb-data-collection
  - https-requests-randomiser - https://pypi.org/project/http-request-randomizer

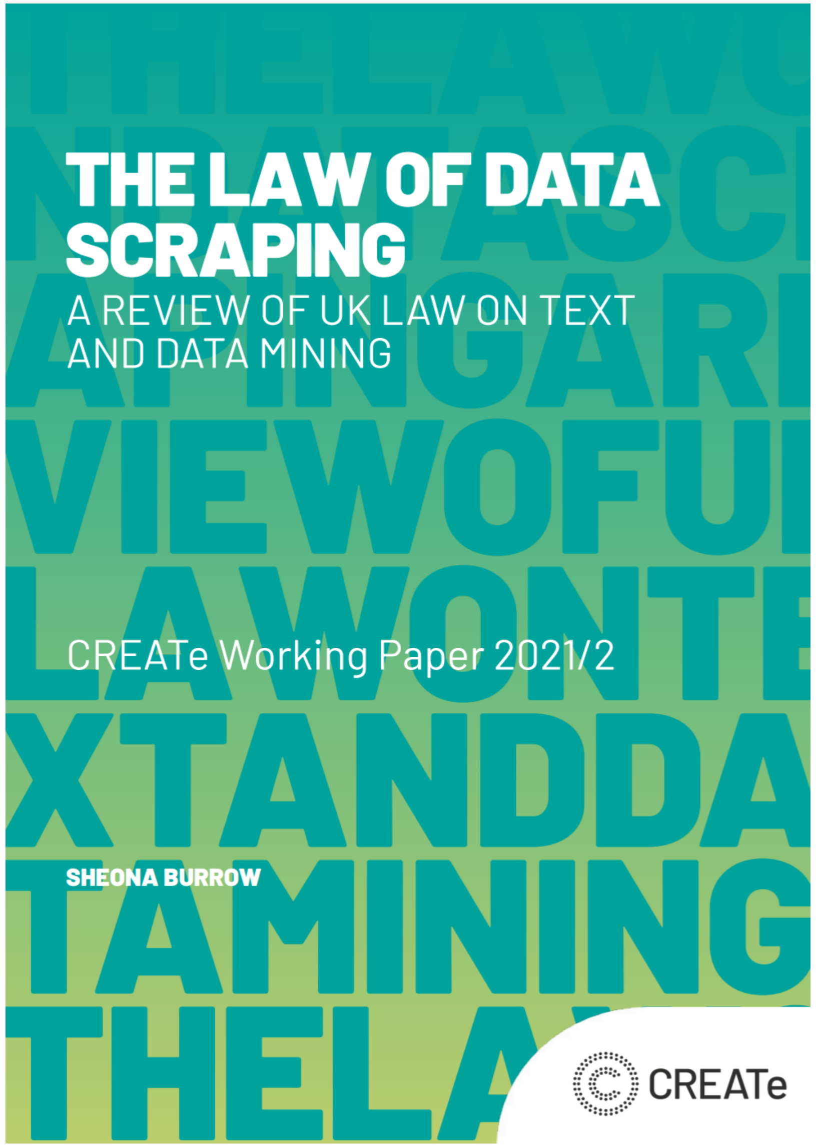
Alternatives to scraping…?

- Some online resources are already available:
  - Inside AirBnB – http://insideairbnb.com/
  - AirDNA – https://www.airdna.co/
- These tend to suffer from limited sampling, data quality concerns and use of black box processing
- They obstruct in particular our understanding of occupancy, price and availability which are key to our interests
- Various open source projects are out there that we evaluated during the initial stages of our project
  - airbnb-data-collection - https://github.com/tomslee/airbnb-data-collection
  - https-requests-randomiser - https://pypi.org/project/http-request-randomizer

Urban Big Data Centre

Legal and ethical issues (1)

- We have worked closely with CREATe at the University of Glasgow to understand the legal landscape related to online scraping
- Unfortunately the picture is not wholly clear
- We refer to the "text and data mining" exception within copyright law as our lawful basis for undertaking scraping
- However, the applicability of this law, and the effects of other legal regimes (privacy, contracts, confidentiality) which appear at times contradictory mean we have no definitive right to scrape.
- Until case law emerges this is a risk based policy decision


THE LAW OF DATA SCRAPING

A REVIEW OF UK LAW ON TEXT AND DATA MINING

CREATe Working Paper 2021/2

SHEONA BURROW

CREATe

DOI 10.5281/zenodo.4635759

Urban Big Data Centre

Legal and ethical issues (2)

- Our collection includes:
  - Pictures of hosts
  - Further information about hosts (including first name and personal descriptions)
  - Property locations
- AirBnB employs web controls intended to limit effects of nuisance hosts issuing multiple requests within a short time period – mainly blacklisting
- We circumvent this by distributing requests across proxy host network. This is to bypass controls intended to safeguard the stability and functionality of the AirBnB website
- There may be legitimate questions about this approach – specifically related to the costs of serving content for consumption at scale

Urban Big Data Centre

Challenges

- Data-Chaos: Data that we can only guess what they represent.
- Evolving needs of our Clients
    - Capture X attribute. Later: I wish we knew about Y attribute too.
- Evolution of Airbnb Data model:
    - They used to mention about that, where is it now?
- Airbnb Service orientating attitude:
    - 300 results per search. Throttling.
- Scalability
    - Daily harvests, should finish within one day
- Historical Data.

Introducing the UBDC Platform

- https://github.com/urbanbigdatacentre/ubdc-airbnb
- Our platform is available under the Apache License, Version 2.0
- This license allows users to:
  - Distribute
  - Modify
  - Use
- Requires original attribution notices to be preserved

Urban Big Data Centre

Urban Big Data Centre

Walkthrough / Demo

JOINTLY FUNDED BY

Economic and Social Research Council

University of Glasgow

Initial Approach

The initial approach that was tried was to get data from the website directly:

```
*for each* page in search-result,
          extract the *listings-links*,
                    and for each link in all the *listing-links*:
                              extract data from *listing-details* pages
                                        using html-text scraping techniques.
                              store in a Database
```

Initial Approach

Nothing bad for this approach, except:

- Airbnb’s limiting the search results to 300 per query.
    - So high level queries that produce more than 300 results (e.g. ‘Edinburgh’ with over 10k listings) do no capture everything.
- Fault tolerantly is low. As, as each search-page is unique if the *while* operation fails you can’t easily resume.
- Linear operation.
- Constantly monitor and react to any Airbnb’s html changes.
- No retain of historic data; once you have parsed the initial html you’re discarding it.

AirBnB Website

- If you look closely the Airbnb website is designed as a responsive web application.
  - The website is a combination of HTML + JS.
- The html code provides a starting template, while the JS code enriches the html, with data coming dynamically based on parameters (i.e the calendar of a Listing ID).
  - The data come from so called ‘endpoints’; URLs that instead of html pages, give back data using a JSON notation.

Instead…

- Instead of extracting the information from html nodes, which is complicated, why not just use directly their API?

- The API generates a structured JSON (key=value), out of each *request* which is more resilient to schema updates, and use that instead to create tabulated data of interest?

- Or even better, just store the returned *response,* and then transform the JSON to tabulated data for further processing.

AirBnB API Explained

- Instead of constructing http requests manually, you can use your favourite programming language to do it for you. We use a python.
- https://github.com/urbanbigdatacentre/airbnb-python
  - It’s a python library, that internally generates API calls to known end-points
  - Forked from https://github.com/nderkach/airbnb-python
  - The actual endpoints can be seen in the code. There are too many to list here, and the code is not complicated.
- Allows you to communicate with Airbnb API in a straightforward manner:

```
  import airbnb
  api = airbnb.Api()
  data = api.get_homes(north=55,south=50,west=-4,east=-3)  # get homes in this
bounding box
```

```
"pdp_listing_detail": {
  "id": 4487641,
  "lat": 52.28909,
  "lng": -6.95914,
  "city": "New Ross",
  "state": "Co Wexford",
  "id_str": "4487641",
  "photos": [
   {
    "id": 33865340,
    "large":
"https://a0.muscache.com/im/pictures/5635
5477/4f40ce5e_original.jpg?aki_policy=large"
,
    "caption": "Unique Country House in a
tranquil garden setting",
    "scrim_color": "#1C271F",
    "aspect_ratio": 1.3333333333333333,
    "is_professional": false,
    "picture_orientation": "LANDSCAPE",
    "preview_encoded_png": “….",},
  ….
}
```

Software Requirements

- Required:
  - Docker - https://www.docker.com/get-started
  - QGIS - https://qgis.org/en/site/forusers/download.html
  - Git - https://git-scm.com/downloads
- Optional but recommended
  - A PostgreSQL database client (we use DBeaver - https://dbeaver.io/)
  - A decent text editor capable of formatting JSON strings (we like sublime - https://www.sublimetext.com/ or Notepad++ - https://notepad-plus-plus.org/)

Inside the Docker Containers

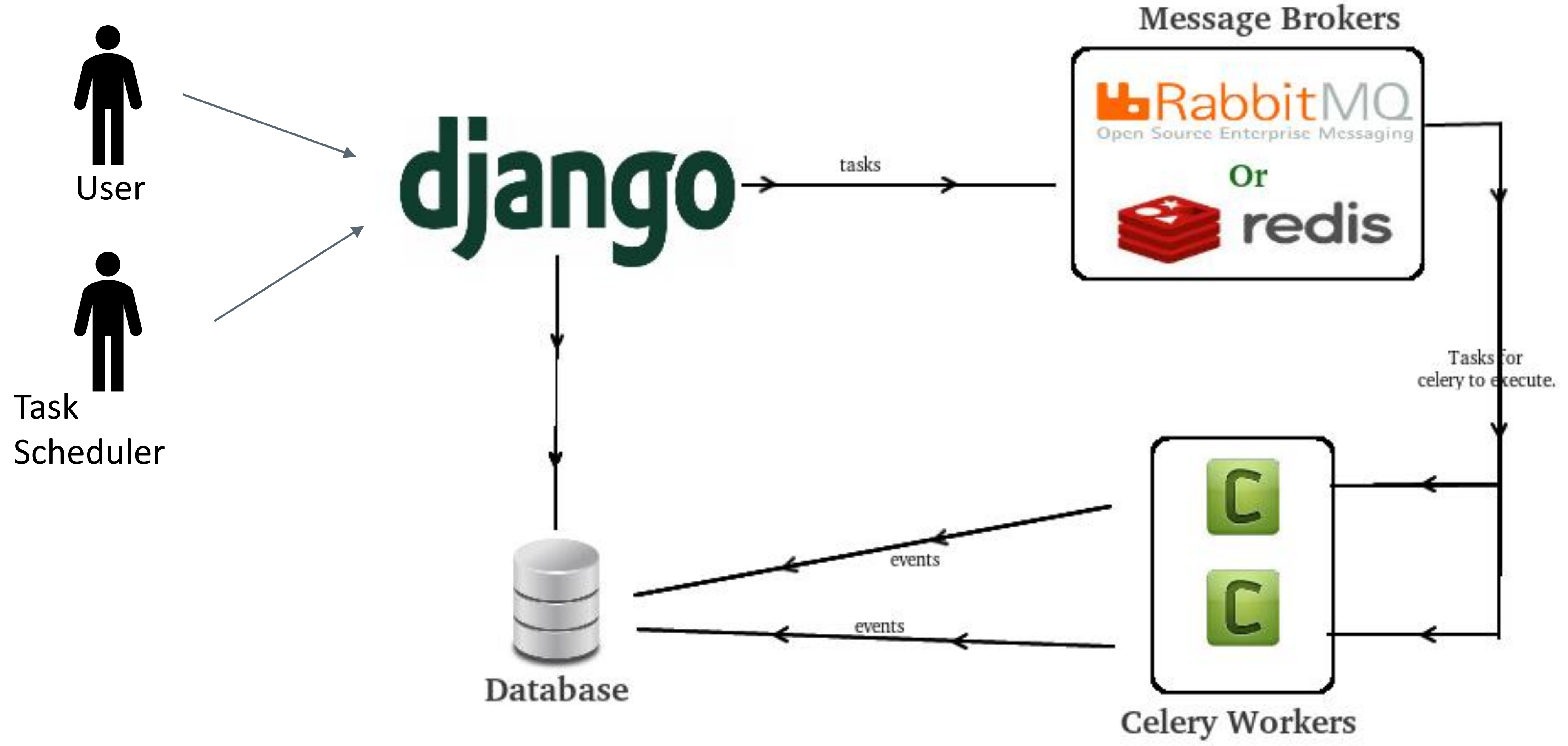


image source: internet articles, unknown author

Urban Big Data Centre

Software Installation

- Throughout the following slides we will demonstrate the following steps
  - Install and configure UBDC AirBnB Data Collection Platform
  - Create of an area of interest (AOI) for data collection
  - Run several harvesting operations
    - Discover listings within a designated AOI
    - Collect details of those listings
    - Collect corresponding calendar information
    - Collect corresponding reviews
    - Collect booking quotes
  - Configure for scale and production
    - Additional workers
    - Scheduling

Urban Big Data Centre

First Steps

```
docker-compose -f .\docker-compose.yml -f .\docker-compose-local.yml up db rabbit worker
```

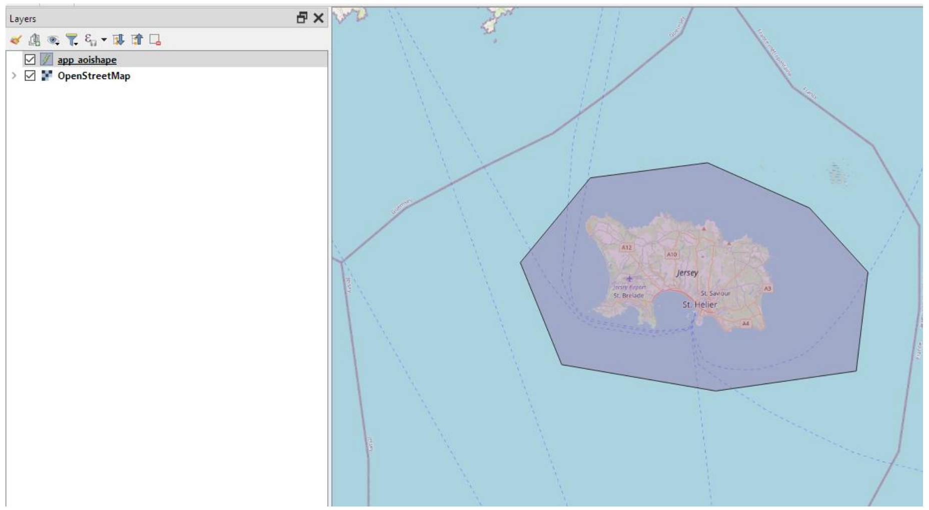
- that should create locally a database, a message broker and one worker
- many default settings can be defined from the docker-compose.yml file
  - default database password/username: postgres/Airbnb
  - default database name: postgres

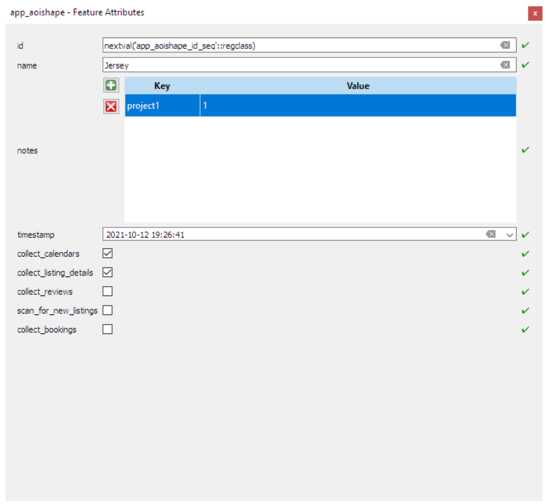
```
docker-compose -f .\docker-compose.yml -f .\docker-compose-local.yml -f .\docker-compose-dev.yml run --rm worker load-mask --only-iso JEY
```

- Load Mask
  - Loads from GADM files the mask for JERSAY
  - Mask defines areas over land valid to scan for listings.

Defining area of interest

Using QGIS, we need to load the AOISHAPE table from the database as a layer in qgis. From there, we can draw our area of interest as polygon.





app_aoishape - Feature Attributes

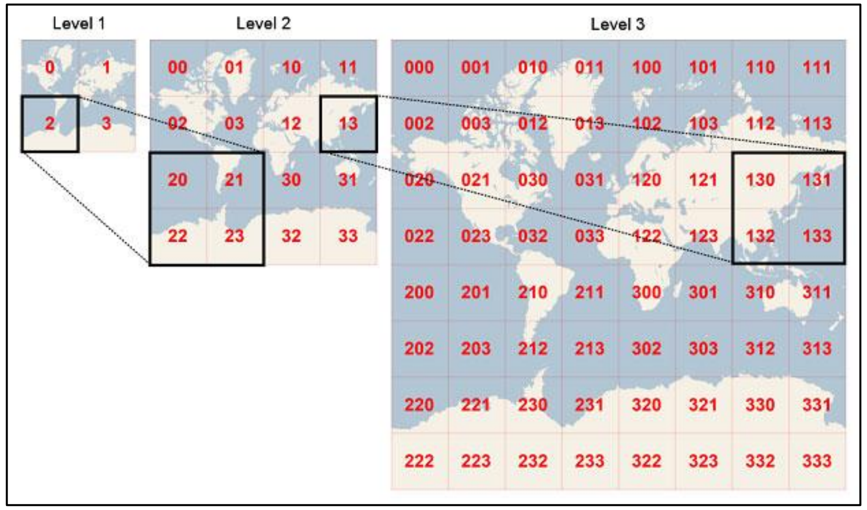
id	nextval('app_aoishape_id_seq'::regclass)
name	Jersey
notes	
timestamp	2021-10-12 19:26:41
collect_calendars	☑
collect_listing_details	☑
collect_reviews	☐
scan_for_new_listings	☐
collect_bookings	☐

Key	Value
project1	1

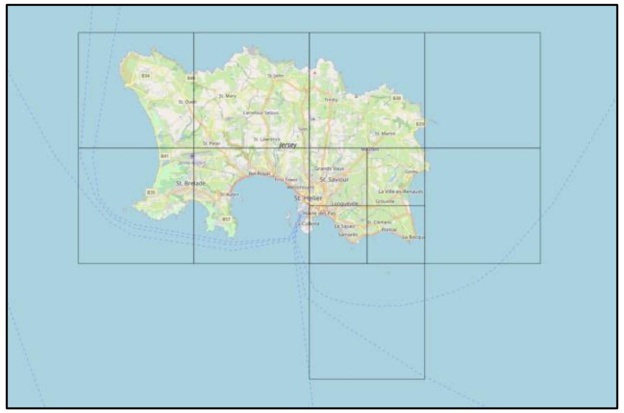
The layer contains some attributes, that defines how it should be used. For example, we could tick the appropriate boxes if we wish to use that AOI to continuously scan for any new Listings that might appear in the future

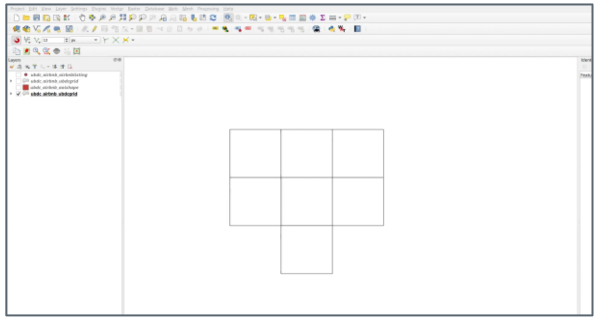
Urban Big Data Centre

Bite sizing



source: https://docs.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system





Urban Big Data Centre

Bite sizing

```
docker-compose -f docker-compose.yml -f docker-compose-
local.yml run --rm worker prep-grid <AOI-ID>

docker-compose -f docker-compose.yml -f docker-compose-
local.yml run --rm worker sense-aoi <AOI-ID>
```

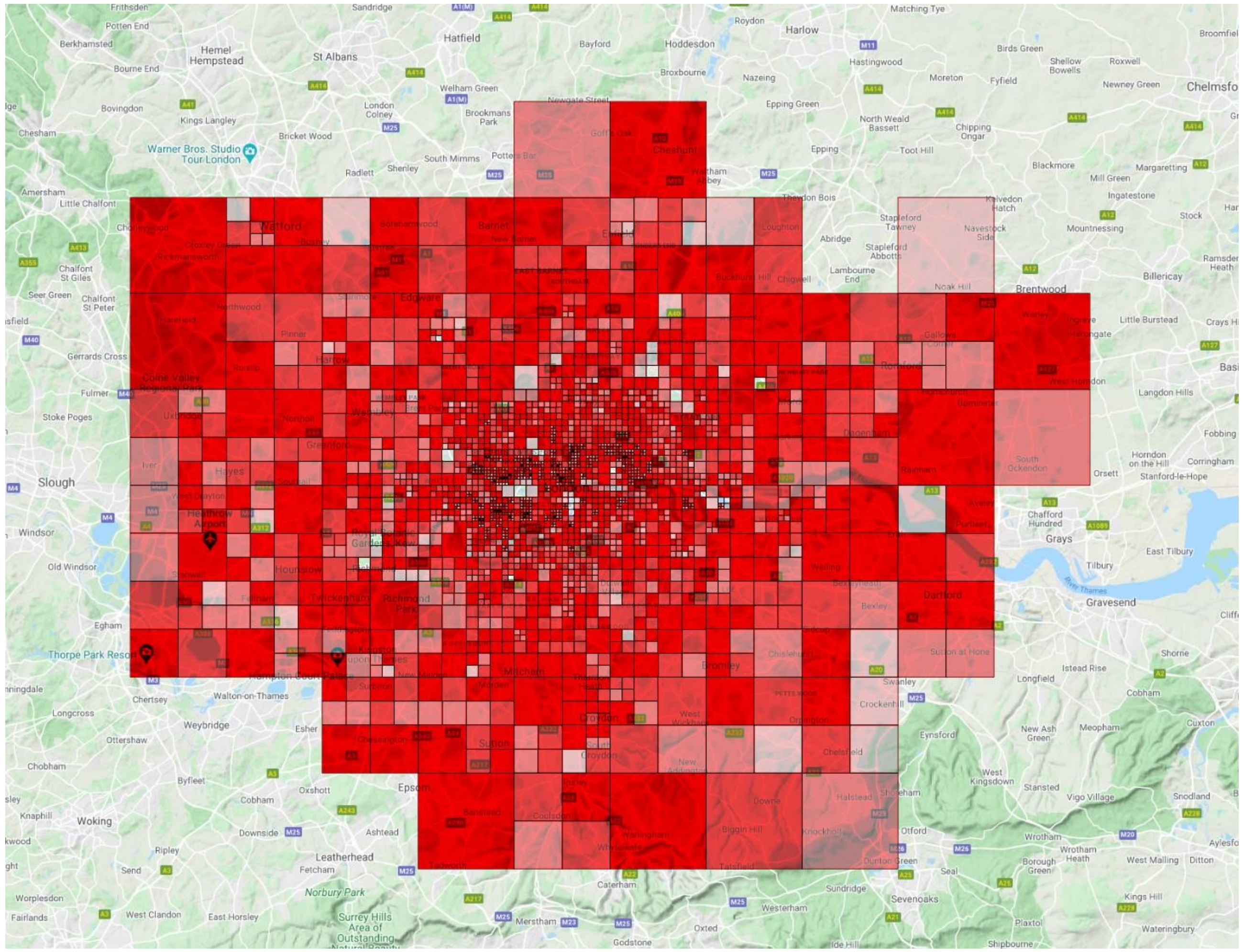
Bite sizing

London
Each Grid has 50 listings of less

They don't overlap.

The grids are regular, which means you could use them to cover the whole world if you wished without overlaps.

Quadtrees rule.

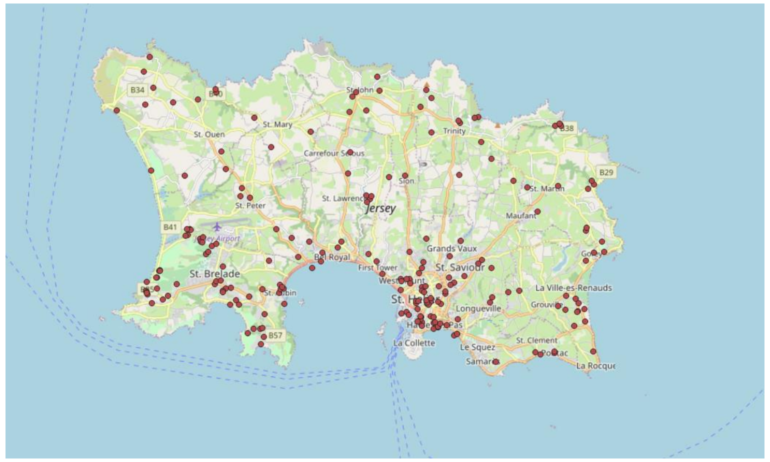


Urban Big Data Centre

Identifying Listings Inside our AOI

When the following command is issued, points representing AirBnB listing locations will start to populate the database table. These points can be visualised from QGIS.

```
docker-compose -f docker-
compose.yml -f docker-compose-
local.yml run --rm worker find-
listings <AOI-ID>
```



Or or all activated aois

docker-compose -f docker-compose.yml -f docker-compose-local.yml run --rm worker send-task discover-listings

Listing Details

```
docker-compose -f docker-compose.yml -f docker-compose-
local.yml run --rm worker fetch-listing-detail
<LISTING-ID>
```

```
docker-compose -f docker-compose.yml -f docker-compose-
local.yml run --rm worker send_task get-listing-details
```

Calendars

```
docker-compose -f docker-compose.yml -f docker-
compose-local.yml run --rm worker fetch-calendar
<LISTING-ID>
```

```
docker-compose -f docker-compose.yml -f docker-
compose-local.yml run --rm worker send_task get-
calendars
```

Users and Reviews

```
docker-compose -f docker-compose.yml -f docker-
compose-local.yml worker run --rm fetch-reviews
<LISTING-ID>
```

```
docker-compose -f docker-compose.yml -f
docker-compose-local.yml run --rm worker send_task
get-reviews
```

Production – Scheduling Data Collection

- The system can be configured to fire tasks on a pre-defined basis
- Schedule specified in **`celery.py`**

```
docker-compose up beat
```

Using Multiple Workers

- With just one worker deployed we'll soon reach the limits of what can be collected
- Our setup facilitates the deployment of additional workers to parallelise the processing
- Limited only by available computational resources

```
docker-compose scale worker=2
```

Questions?

To stay up to date with our research and other activities you can sign up to our newsletter on our website and follow us on social media:

www.ubdc.ac.uk

/urbanbigdata

@UrbanBigData

Linked in Urban Big Data Centre

Urban Big Data Centre

JOINTLY FUNDED BY

UKRI Economic and Social Research Council

University of Glasgow