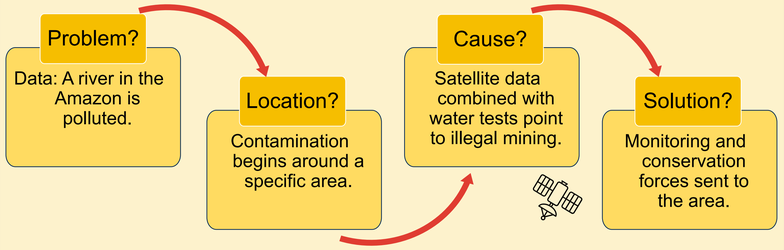
(m)App My Data: The Power of Location as a Data Integrator

Authored by Leonardo Vilardo

Supervised by Dr. Claire Ellul

INTRODUCTION

- It is estimated that **90%** of the global data was created over the **last 2 years**.[1]
- Data becomes more **valuable** when **integrated** with other data, like how Google Maps combines GPS and traffic information to find the best route.
- Useful for environmental and urban studies: Data that includes a location component improves the **spatial understanding** of problems - enabling better **decision-making**.



- Integrating location-data can lead to significant **social** and **environmental** benefits. Knowing this, countries and organisations are creating **Geospatial Strategies** - plans that use spatial data to achieve goals and desired outcomes.
- Yet, integration is **not easy to achieve**. Some data formats cannot be read by computers or lack geographical coordinates, requiring resources and reformatting to be usable.

UK's Geospatial Strategy is investing **billions** of pounds in data-integration technologies, to enhance sectors like transport and the economy.[2]

- Data **fragmentation** is another challenge: Disconnected sources, like governments and private companies, **withhold** data from each other and the public.
- On average, data scientists must **spend** around **60%** of their time manipulating and formatting datasets.[3]
- Bridging gaps and integrating different datasets is crucial to unlock their full **potential**.

AIM

What?

Designed an app and system to **rate** how easily **data** can be put into **maps**. Then, applied it to assess the official **British source** of open data.

Why?

To **promote** productive data sharing, enabling data to be chosen better, create deeper insights and lead to more **successful** and **sustainable strategies** for the future.

- Therefore, this study created a standardised **system** to rate how easily datasets can be mapped and applied it to **evaluate** data.gov.uk – Promoting the official FAIR principles of data sharing.[4]

Findable
Accessible
Interoperable
Reusable.[4]

Related literature

1. Marr, B. (2018). 'How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.
2. British Government. (2023). 'UK Geospatial Strategy 2030', *Geospatial Commission*.
3. Forbes. (2016). 'Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says'. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=753d074d6f63
4. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship'. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18
5. U.S. Geological Survey. (2023). Earthquake Hazards Program. Available at: https://www.usgs.gov/programs/earthquake-hazards

LAIDLAW SCHOLARS

METHODOLOGY

1st STEP

Collect and attempt to map all datasets tagged with “London” and “Environment” from data.gov.uk.

2nd STEP

Record the common factors and challenges of the mapping process for the rating system, continuously improving it along the process – Inductive approach.

3rd STEP

Code the rating system into an App and apply it to the collected data.

- The rating system was coded into a free web app, which enables the map visualisation of the data and its rating.

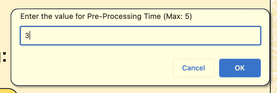
THE APP

- The 'Data's Format' can be judged **objectively**, while other factors are **subjective** and context-dependent. For example, the processing time varies based on a person's expertise.
- Each rating factor has a **per cent weighting** and **grades** from 0 to 5, from worst to optimal conditions.

Data mapping rating system

30% Data format	30% Pre-processing time	20% Spatial density	10% Geographical relevance	10% Choice of Representation
Format should allow instant automated mapping (GeoJSON).	Ideally instant and automated process. No required time for format conversion.	Rich enough to give meaningful insights. No duplicate nor redundant information	Relevant data must be within the studied area	Elements should represent data appropriately. Points for sites, polygons for areas, etc.

- On the app, the user can upload their file to visualise it on the map and enter a grade for each factor, to get a **final rating:**

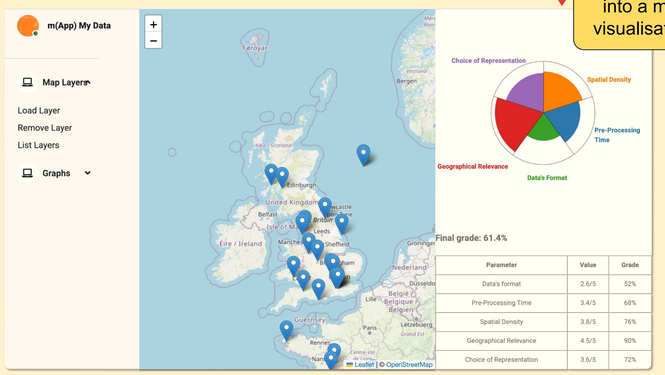


This is what a GeoJSON file looks like (pure text):


```
{"type":"Feature","properties":{"mag":3.2,"place":"5 km N of Hockley, United Kingdom","time":1645484367833,"updated":1668202804973,"tz":null,"url":"https://earthquake.usgs.gov/earthquakes/eventpage/us7000gmts","detail":"https://earthquake.usgs.gov/fdsnws/event/1/query?
```


GeoJSON file about British earthquakes.[5]

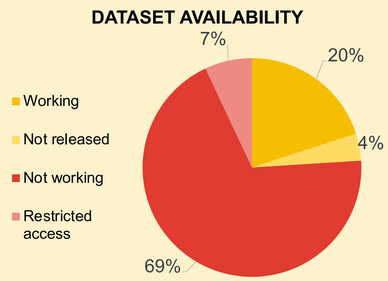
On the app, It gets converted into a map visualisation:

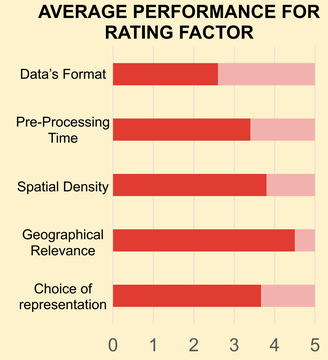


App screenshot of map with the Earthquake dataset - All occurrences from 2018 to 2023.[5]

FINDINGS

- **Search tool** and **filters** showed **illogical** results and **hid** key datasets. For example, 'Trees of City of London 2023' was not included in the search while a dataset from the city of Plymouth was.
- Out of all 54 London environmental datasets: **20%** (11) freely worked; **69%** (37) led to error messages (no data could be found); **7%** (4) was access restricted to common users (some accreditation was necessary); **4%** (2) was “not released”.
- **Duplicates** and **wrong labels**: “London Air Quality 2015” report had the exact same information as the 2016 report. The 2014 report was “not released”.
- **Improper** format: Out of the few working datasets**, 55%** (6) were directly mappable while **45%** (5) did not have coordinates, requiring additional work to be indirectly mapped (manipulated and linked to other information to enable mapping).
- Lack of meaningful **metadata** and **titles**, such as “Physical environment 2016”, a dataset about vacant land.
- The average grades for each factor were: Data's format **52%,** Pre-processing Time **68%**, Choice of Representation **73%,** Spatial density **76%** and Geo Relevance **90%**.
- Total average: **61.4%**.





DISCUSSION

- The choice to **focus** on **London's environmental** datasets, as a prominent city with abundant data resources, offers **insights** to guide future **urban development**, for integration of location-data is essential in decision-making.
- The **suboptimal** results from data.gov.uk reveal **only 11%** (6) of the datasets could be directly mapped. This **contrasts** with the UK Geospatial Strategy of "unlocking the power of location data and technologies".
- This situation of data **inaccessibility** points to an **urgent** need for improvements in the standards of data sharing. Valuable information is being restricted, wasting resources and time from data scientists, and may **hinder** research and lead to **missed opportunities.**
- Therefore, this study aims to **promote** FAIR and productive **data sharing**, enabling researchers, urban planners and policymakers to better choose data, gain deeper insights and develop more **effective** and **sustainable** strategies for the **future**.