(m)App my data! Developing a Map-ability Rating and App to Rapidly Communicate Data Quality and Interoperability Potential of Open Data

Claire Ellul[a*], Paul Reynolds[b], Leonardo Vilardo[a]

[a] Dept. of Civil, Environmental and Geomatic Engineering, University College London, UK - (c.ellul, leonardo.vilardo.21)@ucl.ac.uk
[b] Dept. of Computer Science, University College London, - paul.reynolds.22@ucl.ac.uk

Keywords: Data Integration, Interoperability, GIS, Data Engineering, Map, Open Data, Smart Cities, Environment, Education

Abstract

In an age of burgeoning open data sources, ranging from authoritative platforms to crowd-sourced contributions, the need for data integration is paramount - only with integrated, combined, data can today's complex problems be addressed. However, assessing the quality of data and its potential for integration/interoperability is complex. FAIR (findable, accessible, interoperable and re-useable) approaches go some way to help, but most repositories still only offer text-based search results and require the user to download the data and assess its quality and fitness-for-purpose manually. This paper examines whether exploiting geospatial approaches - specifically, understanding whether a dataset can be mapped and hence integrated with other datasets - could address these issues, in particular for non-expert users. We explore challenges related to open data for education and environmental sustainability and introduce a novel, visual map-ability rating to assist data users in rapidly understanding the data quality and interoperability potential of data they wish to use. This rating system has been developed by researchers from outside the location science domain, to better reflect what is important to the wider community, and derives from a review of 104 open datasets. As well as providing useful insight to data users, the ratings can be used to guide data publishers as to how to improve their data offering.

1. INTRODUCTION

The amount of open data available to the public and researchers has been constantly increasing (e.g. Open Data Watch note an increase in the United Kingdom's overall Open Data Inventory [Open Data Watch, 2022] from 52 to 62.5 [Open Data Watch, 2023]. From authoritative sources to crowdsourced contributions and social media, a diverse array of platforms have become data reservoirs, with many countries having national data repositories (e.g. data.gov.uk or dati.gov.it, regional or municipal repositories, researcher created repositories and more). These increasing quantities and diversity of data can provide evidence to help address a multitude of problems facing society - such as climate change, housing, environmental issues, education and many more. Resolving these problems relies heavily on the integration of multiple sources of data from different silos.

FAIR assesses the findability, accessibility, interoperability and re-usability of data sources and was introduced in 2016 [Wilkinson et al., 2016]. It provides an easy-to-understand assessment process for data sharing [Wilkinson et al., 2016]. It is now a recommended rating system for open data stewardship (UK Open Research Data Taskforce [Open Research Data Taskforce, 2018] and the European Research Executive Agency [Open Science in Horizon Europe, n.d.]). An increasing scale of FAIRness has been identified, ranging from data which cannot be reused to data that is open access and functionally linked [Mons, 2018].

However, FAIR precedes implementation [Wilkinson et al., 2016]. The principles do not suggest any specific technology, standard or solution [Wilkinson et al., 2016] and do not provide detailed, evidence-driven guidelines as to where repository owners should focus their efforts to target the data engineering challenges that are the greatest blockers to data use. *Data engineering* is the term used for the tasks undertaken to prepare data to the point that it is ready for use in analysis. These may include downloading or linking to data, transforming its format, semantic or syntactic mapping to achieve interoperability and more [Reis and Housley, 2022]. FAIR is also agnostic to the nature of the data in the repository. Even in highly FAIR repositories, users (data scientists) often have to download data, transform it and load it into their own systems to assess its quality and potential for interoperability with other datasets they are using for analysis, taking data scientist/user's time away from their core work. These tasks are made more challenging as they are not daily tasks for these users, so an element of re-learning of skills is required every time. Currently, these tasks are also repeated by any data user wishing to use a specific dataset - resulting in duplicated time-wasting.

Accessing and assessing data via a map-based interface rather than as a list could offer a way to rapidly evaluate data quality and fitness-for-purpose, providing the ability to assess completeness, presence of extraneous data, spatial and temporal coverage at a single glance. Once mapped, data can much more rapidly be assessed for fitness-for-purpose (e.g. does it cover the correct area of the globe, at the appropriate granularity). Presenting the results of these tests in a way that can be rapidly reviewed by data scientists will save them time in carrying out preliminary assessments of the quality and fitness-for-purpose of the data.

Additionally, location information is common among different datasets and can be used to combine and integrate them [Geospatial Commission, 2023]. The term generally refers to data that contains geographical coordinates or can somehow be associated with coordinates indirectly (through a link/join with other data). Data integration through the power of location is one of the most meaningful ways to obtain insights and make evidence-based decisions [Geospatial Commission, 2023]: 'Since location is a common attribute among different

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W4-2024-49-2024 | © Author(s) 2024. CC BY 4.0 License.

datasets it can be used to combine and integrate them.'

This paper proposes map-ability rating system to allow data scientists to rapidly evaluate a dataset's potential for location-based integration, and hence interoperability, alongside its general fitness-for-purpose. Our multifactor rating system has been developed using a bottom up approach - driven by the characteristics of the datasets) and based on issues encountered when working with data, and co-developed by two researchers with little location data expertise, in order to ensure that it reflects the needs of a broader community of data users. We embed our ratings into a web App to demonstrate their scope. This results in data users having a rapid visual report for a dataset, saving them time reviewing extensive metadata and downloading the dataset to assess it. The ratings can be used by data producers to identify where to prioritise efforts to encourage reuse,

2. Literature Review

Currently, data users carry out the following steps to retrieve, assess and work with data (adapted from [Mons, 2018]):

- Search for data - e.g. using google, but also by searching various repositories
- Once a potentially relevant dataset is identified, review the metadata to understand more about the dataset and how to license and obtain the data
- Download or otherwise access and transform data it into a format that works with a tool they habitually use
- Open the data and assess its fitness for their task (quality)
- If the data is fit for purpose, further transform the data to render it interoperable with other datasets

Given that data scientists only perform these tasks relatively infrequently, they may not have the expertise to efficiently and effectively perform them. Two time-consuming issues include: understanding data quality and understanding the potential for interoperating the data with other datasets. It may also take more than one iteration of the above workflow until a relevant and fit-for-purpose dataset is identified.

2.1 Understanding Data Quality and Fitness-for-Purpose

The concept of fitness-for-use is central in evaluating the appropriateness of datasets for use by data scientists and relates in particular to the challenges of using a dataset created for one purpose in a different study or context, potentially months or years after it was created. Several studies have delved into evaluating the appropriateness of datasets - e.g. studies by [Jonietz and Zipf, 2016] and [Ahmed et al., 2014] assessed fitness-for-use of Points of Interest (POI) datasets and the performance of map construction algorithms using vehicle tracking data. [Wang et al., 2015] and [Soliman et al., 2022] demonstrated the importance of comparing different methods and assessing accuracy through logistic regression for landslide susceptibility maps and Land-Use Land Cover datasets for flood modelling, respectively. More broadly, a number of key categories of assessment are used when describing data quality - ISO 19157 [ISO, 2013] describes six high level elements: completeness, thematic accuracy, logical consistency, temporal quality, positional accuracy and usability. The task of exploring the data to understand its quality is also challenging - in a recent test on data interpret-ability, only 36.4% of participants were able to quickly clean sample data to achieve a simple visualisation to evaluate a dataset, with participants noting "much physical labor [sic] would be needed to perform the task" and "correcting the data format would be extremely difficult" [Barcellos et al., 2022]. Metadata (describing the quality of the data) - which should help interpret-ability and allow users to assess datasets without the need for download and processing - is both complex to create and use [Ellul et al., 2013]. [Moellering et al., 2005] lists over 70 metadata standards, ranging from domain specific through national to international. [Ellul et al., 2013] shows that decision makers only require "minimal metadata" - e.g. date created, source. They ignore most metadata due to time and complexity or decoupling from the source data.

2.2 Interoperability Challenges

Interoperability principles are widely considered the hardest to address in FAIR [Hong et al., 2020]. Standards are usually considered vital, and extensive efforts have been undertaken to create mappings between various standards that relate to different aspects of physical objects. Examples from the built environment illustrate these concepts [Yu et al., 2022, Siew et al., 2021, Kumar et al., 2019, El-Mekawy and Östman, 2010, Sani et al., 2022, Saquicela et al., 2022].

More recently, "minimal interoperability" mechanisms" [Mulquin, 2023] have emerged, which focus on "good enough" integration and interoperability by identifying what is common across datasets (location/coordinate information in our case). This defers the need for up-front, costly investment in full semantic interoperability until the cost/benefit of an expensive and time-consuming full interoperability mapping are fully understood. These have yet to be explored in the context of geospatial data but geospatial data epitomises minimal interoperability - all that is required is for both datasets to contain coordinate values that relate to their location, so that location commonality can be identified.

3. Method and Results

A four-stage process was followed for this study, with each stage depending on the outcome of the previous one

- Stage 1 - Repository Selection
- Stage 2 - Topic Selection and Manual Dataset Review
- Stage 3 - Developing the Map-Ability Rating (arising empirically through the Stage 2 review process)
- Stage 4 - M(App)-ability App Demonstration

Given the interdependencies of subsequent stages on the previous one, the method summary and results for each stage are presented sequentially.

3.1 Stage 1 - Repository Selection

This initial study focussed on repositories in the United Kingdom (UK), a country which scored 62.5 - i.e. middle range - on Open Data Watch's Open Data Inventory in 2022 . To further narrow down dataset choices, London was selected as the focal area of the study. The city presents a favourable setting to study the integration of data within a dynamic urban landscape, offering significant lessons into the challenges and opportunities for sustainable urban development. In order to test the map-ability concept in multiple contexts, the selected repository should offer a wide range of data types and themes. Repository selection involved the identification and review of three open data repositories from both government and academic sources that meet these criteria.

3.2 Stage 1 - Repository Selection - Results

Three national repositories were considered for this study:

- *data.gov.uk* hosts metadata relating to over 50,000 government centric datasets (non-personal data), with links to data where available. The repository was established in 2010. Datasets are subdivided into 14 categories including 'crime and justice', 'towns and cities', 'transport', 'education' and 'environment'.
- The *UK Data Service* is a comprehensive repository that provides access to a wide range of social, economic, and population data. It offers datasets from various sources, including government surveys, international organisations, and academic research.
- The *ONS Open Data* repository provides access to a wide range of official statistics and data produced by the Office for National Statistics, responsible for the census in the UK. It offers datasets on various topics, including population, economy, and society.

data.gov.uk was selected as both the *UK Data Service* (focussed on social science research) and the *ONS Open Data* repository (focussed on census-derived statistics) repositories are narrower in scope. Additionally, the selection of data.gov.uk as our primary data repository has benefits that include:

- The breadth of data topics covered [Eranki and Reddy, 2012]
- Rich metadata, which is essential for understanding and evaluating the quality and relevance of the datasets [Kuzma and Mościcka, 2020].

3.3 Stage 2 - Topic Selection/Manual Dataset Review

To focus the scope of this preliminary study, two themes were selected for in depth review. Following this, in order to develop a bottom-up rating methodology, a manual attempt to map available datasets was made. In order to establish a comprehensive and uniform rating system for mapping data, datasets were systematically analysed to find common factors influencing the ability to map data and determine appropriate weightings for each of these factors. The process - and in particular challenges encountered - form the basis of the ratings. To ensure that the results reflect situations encountered by, and expertise of, general data scientists, the tasks were carried out by researchers with no specific training in geospatial data engineering.

3.3.1 Search The following steps were carried out to conduct an initial search for datasets on the data.gov.uk:

- 'Environment' theme, filtered by 'London' in the search bar
- 'Education' theme, filtered by 'London' in the search bar

The datasets resulting from the search were systematically listed and metadata (e.g. title, date added to data.gov.uk, date last updated) captured. Links to each dataset were then followed to identify whether the dataset was available (via download or API) for further review.

3.3.2 Dataset Review The review of accessible datasets focussed specifically on whether the dataset can be mapped. Table 1 shows the list.

Map-Ability	Criteria
Directly mappable	Contains co-ordinates
Indirectly mappable	Geo-referencing is possible via linking to another dataset
Not mappable	No geographic information present
n/a	Dataset not accessible

Table 1. Map-Ability of Datasets

When many formats were available for the same dataset, the one with the shortest pre-processing time (judged holistically based on experience gained during this study) for mapping was recorded as the format type. Excel arbitrarily took precedence over CSV.

3.4 Stage 2 - Topic Selection/Manual Dataset Review - Results

3.4.1 Environment Out of 54 London environmental datasets resulting from the search in June 2023, approximately 20 per cent (11) were available for open access, containing downloadable data. In contrast, 7 per cent (4) imposed access restrictions, requiring specific accreditation for access, 4 per cent (2) were marked as "not released". While the vast majority, 69 per cent (37) resulted in error messages upon following links, so no data could be found. Among the error messages: "site can't be reached", "resource not found", "file does not appear to have any style information associated with it", "invalid dataset", "File or directory not found".

Out of the 11 available datasets, 55 per cent (6) were directly mappable, while 45 per cent (5) of these datasets lacked geographical coordinates, necessitating further manipulation and linking to other information sources to enable mapping.

Further investigation also identified that other datasets related to the topic were not displayed in the search results. For instance, datasets such as 'Trees of City of London 2023' and 'Trees of Camden 2023' were not listed, highlighting a general lack of location awareness in the search algorithm (as Camden is part of London). The search also included an unrelated dataset, ambiguously labelled as "Physical Environment", about the use of vacant land in the city of Plymouth.

3.4.2 Education The education category, filtered by London, only yielded 14 datasets. A combined search of education and *London* was thus carried out. A total of 1045 datasets resulted, of which the first 50 were sampled.

Out of 50 London educational datasets, approximately 66% per cent (33) were available for open access. Out of these, 91% (30) were directly accessible from the data.gov.uk website, while 9% (3) could be accessed via an external archive link. 34% (17) of the datasets were not available. Specifically, 76% (13) were marked as "not released", and 24% (4) were inaccessible due to issues such as "page/server not found".

Out of the 33 available datasets, 33% (11) were not mappable at all, and contained no geographic component, 66% (22) could be mapped indirectly, requiring additional processing or georeferencing. None of the datasets could be mapped directly.

As with the environment data, there were datasets in the filtered search that did not align with the "Education" filter e.g. 'Average age of rolling stock', a dataset about trainlines, which does not correlate to the education context.

3.5 Stage 3 - Developing a Map-Ability Rating

In contrast to many ratings and standards, which are developed by data producers, a data driven, user-focussed approach was taken to this task. Throughout the dataset evaluation process in Section 3.4, challenges encountered were noted and used to underpin the development of a rating system that directly reflects how map-able (and hence potentially interoperable) the dataset is, while providing clear information on areas where the producer of the dataset can make improvements. By synthesising insights from Stage 2 results, consistent patterns and themes were identified. These were used to develop an understanding of the relative importance of each factor and stage in the mapping and quality assessment process. In other words, as more datasets were mapped, the proposed factors were refined.

To ensure standardisation, and to reinforce the principles of cognitive ease and user-centric design. all identified factors were assigned a scale from 0 to 5, representing the range of usability and practicality. A score of 5 signifies optimal conditions, while 0 indicates inaccessible or non-existent data. In the context of designing user-friendly rating systems, recent empirical studies and theoretical frameworks advocate for the utilisation of a consistent 0-5 rating scale across diverse categories. This body of research underscores the multifaceted advantages of uniform rating scales, linking them to enhanced user comprehension, reduced cognitive load, and greater overall satisfaction ([Lewis and Sauro, 2009]; [Yuan and Recker, 2015]). Specifically, [Praseptiawan et al., 2023] demonstrated that the use of the System Usability Scale (SUS) with a consistent 0-5 rating scale provided valuable insights into users' desires, emotions, perceptions, and habits in a study on mobile application interface redesign. This focus on consistency across dimensions of user experience, including usability and learnability, translates into a more accurate representation of various aspects and the reliability of evaluations ([Lewis and Sauro, 2009]).

Each factor was then weighted according to its importance in terms of creating a map with the dataset. An average calculation was employed to determine the overall rating for each dataset. Weightings were chosen based on the relative importance of each factor in the mapping process, established through the review, mapping, and analysis of many open datasets.

3.6 Stage 3 - Developing a Map-Ability Rating - Results

A total of six separate ratings resulted from the process and issues documented in Stage 2.

The 'choice of representation' factor considers if the choice of the element was appropriate, for instance, information about large areas is more detailed when they are represented by polygons and not points. This is an indication of the granularity, or level of generalisation, of the dataset. For example, if the dataset relates to the entire London Borough of Camden it would achieve a higher score if it is mapped using the Borough boundary polygon or subdivisions of the polygon than if it is mapped using a centroid point, The more granular the dataset, the greater its potential for integration with other datasets, as it can be aggregated to larger spatial units. Similarly, a continuous dataset is perhaps not best represented using a vector approach.

After an initial system of 5 factors was devised, further examination of data highlighted the importance of the time component. 'Time density' was introduced as the sixth and final factor.

The rating system underwent a validation process to verify its accuracy across diverse datasets from data.gov.uk. The validation process comprises the following steps:

- Selection of representative data - including a mix of formats, spatial and temporal characteristics, temporal characteristics and data purpose
- The datasets were then mapped where this was possible
- The weighted rating was generated

3.6.1 The Final Ratings The final weighted, six-factor rating system is as follows:

Factor 1: Data format GeoJSON is widely used in web-based and other GIS systems and also offers an easy representation of one geometry:many attributes (e.g. a time series of data points collected by a sensor). This is particularly helpful for indirect mapping (see Table 1) where many rows of data may be linked to a single geometry. It contrasts with traditional table-based formats used in GIS (e.g. .shp or relational databases) where identical geometry is sometimes stored multiple times for many:one situations. The chosen weighting is 25% which reflects the importance of format in accessing data. Description:

0 Not possible to process into a mappable format
1 Virtually unusable without significant effort
2 Extremely difficult, manual conversion needed
3 Complicated but mostly automated
4 Quick and automated
5 No pre-processing required

Factor 2: Pre-processing time This determines how easily the data can be mapped - specifically whether it can be geo-referenced or not, by specifying the amount of time that is required to undertake geo-referencing. A weighting of 25% is assigned due to it's importance in assessing potential interoperability of the dataset. Description:

0 No geo-referencing possible at all
1 Indirect mapping possible, geo-referencing is complex
2 Indirect mapping possible, geo-referencing is straightforward
3 Direct mapping possible, but format conversion is complex
4 Direct mapping possible, with simple format conversion
5 Direct mapping without conversion, format is GeoJSON

Factor 3: Spatial density This reflects the richness and the depth of the insights that can be derived from visualising the data, and hence being able to asses the overall fitness-for-purpose of the dataset for a specific task. The 'spatial density' rating will depend on the purpose/topic of the dataset. For example, the visualisation of UK UNESCO heritage sites would achieve the best rating for spatial density when it is low and points are sparse, with just a few dozen locations in the entire UK, while a map of protected listed buildings would need to be highly dense to be complete, with hundreds of thousands of data points in the country. Given the importance of this factor for fitness-for-purposes, it is given a weighting of 20%. Description:

0 Non-existent
1 Very low, sparse data points
2 Low
3 Moderate
4 High
5 Optimal, very dense data points

Factor 4: Geographical relevance This indicates whether all the data is within the stated area - for example, if the metadata states that the data relates to England but data also appears in Ireland or only for London when mapped this would result in a lower score as it indicates that quality control may not have been performed on the dataset. Weighting: 10% - given that irrelevant data can diminish the value of the visualisation. Description:

0	0-19% coverage of area of interest
1	20-35% coverage
2	36-50% coverage
3	51-80% coverage
4	81-95% coverage
5	96-100% coverage

Factor 5: Choice of representation How suitable the representation is for the type of data. Certain data types require specific representations for clarity - e.g. points, lines and polygons, perhaps changing at different scales. Weighting: 10% - as representation affects interpretation and clarity. Description:

0	Non-existent or irrelevant representation
1	Unusable representation for the data type
2	Improper but somewhat decipherable
3	Just feasible, could be improved
4	Mostly appropriate for the data type
5	Fully appropriate and clear representation

Factor 6: Time density The frequency with which data is collected. Frequent data collection can provide more up-to-date visualisations. Weighting: 10% - as consistent updates enhance the relevance of the mapping. Description:

0	Non-existent or very sporadic updates
1	Very infrequent updates
2	Infrequent updates
3	Moderate frequency of updates
4	Frequent updates
5	Very frequent, almost real-time updates

3.6.2 Meta-Factors and Multiple Ratings Per Dataset

The **‘data format’ and ‘preprocessing time’** were given the highest relative weightings, as they directly indicate if data can be visualised and the potential time required for to achieve this. However, it should be acknowledged that ’preprocessing time’ is subject to the individual’s proficiency with location data - e.g. the time spent on manipulating data decreases with experience. All other ratings are to a certain extent dependent on the context/use of the data.

To overcome this issue, multiple ratings could be developed for the same dataset, with users contributing ratings dependent on their expertise and the purpose for which the data was used. These meta-factors can then guide other users to a rating that is closest to their expertise and research topic.

3.7 Stage 4 - Map-Ability Rating Visualisation

To demonstrate the rapidity of a visual approach to dataset rating, the rating system was integrated into a web application (App) interface, which produces a chart that combines characteristics of both pie and radar charts. The size of each sector in the chart corresponds to the weighting of each factor. Concentric circles, numbered 0 to 5 to represent ratings, further detail each factor’s rating. The decision to use a hybrid of pie and radar charts was based on a comprehensive assessment of various visualisation techniques. Pie charts effectively represent proportions, helping users easily understand factor weightings. Radar charts, commonly used in data science for multi-variable analysis, allow for a comparative view of individual factors.

The application also calculates a final weighted mean rating for the dataset, factoring in the pre-defined weightings for each rating parameter. This grade reflects the overall quality of the dataset in relation to the rating criteria. Where the data is mappable, the App also allows the user to map the data.

3.8 Stage 4 - Map-Ability Web-Based Demonstration - Results

As noted in 3.4 the percentage of ‘open’ datasets listed on the data.gov.uk that were actually available was relatively low, particularly for the environment topic. To illustrate the application of the rating system, this section presents two examples of the results obtained: an available open dataset from the environment results and a listed but unavailable dataset from the education results.

3.8.1 Environment: London Air Quality Data - Listed and Available

The ”London Air Quality Network Camden - Last Updated 2015” dataset [1] offers data in XML, CSV, RDF, and JSON; however, GeoJSON, our preferred format for spatial visualisation. The dataset contains coordinate information in British National Grid and in WGS84 and is thus directly mappable. QGIS was used to convert CSV to GeoJSON. Table 2 shows the ratings and Figure 1 shows the resulting visualisation.

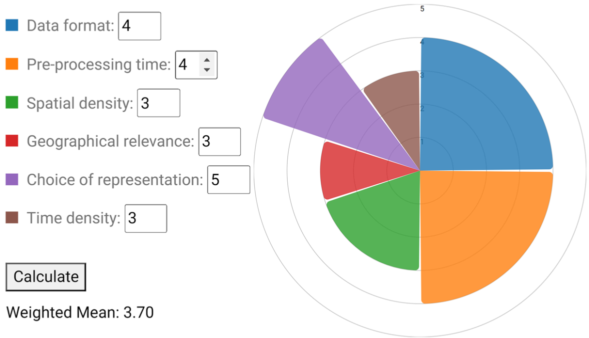
Factor	Rating	Reason
Data format	4	Dataset was CSV format and contained coordinate information so can be directly mapped
Pre-processing time	4	Dataset conversion was relatively simple - CSV to GeoJSON
Spatial density	3	More collection points would greatly improve the dataset’s coverage of Camden
Geographical relevance	3	Covers all of the Camden area
Choice of representation	5	Dataset was presented as points
Time density	3	The two datasets from different years revealed identical information, indicating no new additions despite the declared temporal difference.

Table 2. Weightings for London Air Quality Data

3.8.2 Education Contracts - Listed but Unavailable

Within the *education* search results, the ‘Department for Edu-

[1] Available from https://www.data.gov.uk/dataset/8c69f2f0-ba2a-44b7-9e3e-576d41cd553d/london-air-quality-network-camden, Accessed 3rd January 2024



🟦 Data format: 4
🟧 Pre-processing time: 4
🟩 Spatial density: 3
🟥 Geographical relevance: 3
🟪 Choice of representation: 5
🟫 Time density: 3

Calculate

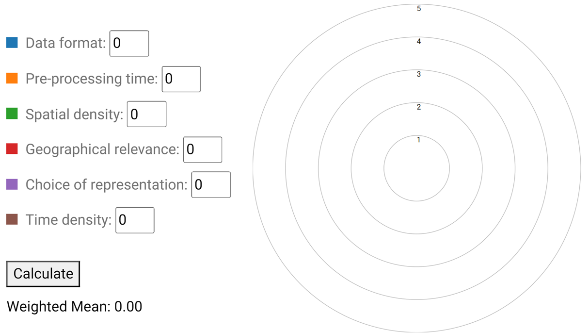
Weighted Mean: 3.70

Figure 1. Visual representation of the integrated rating system in the web application, combining features of pie and radar charts for dataset evaluation for Air Quality in Camden.

cation - Contracts' [2] dataset was listed, with a note that it has not been released by the publisher. Table 3 shows the resulting ratings - in this case all zero - and Figure 2 shows the resulting visualisation.

Factor	Rating	Reason
Data format	0	No georeferencing possible
Pre-processing time	0	Not possible to create a map
Spatial density	0	Non-existent
Geographical relevance	0	0% coverage of the area of interest
Choice of representation	0	Non-existent
Time density	0	Non-existent

Table 3. Weightings for Education Contracts Data



🟦 Data format: 0
🟧 Pre-processing time: 0
🟩 Spatial density: 0
🟥 Geographical relevance: 0
🟪 Choice of representation: 0
🟫 Time density: 0

Calculate

Weighted Mean: 0.00

Figure 2. Visual representation of the integrated rating system in the web application, combining features of pie and radar charts for the Department for Education Contracts Data.

4. Discussion

This research developed a multifactor rating system to provide initial insights into the fitness-for-purposes and potential for interoperability of open datasets, with the aim to reduce the time data scientists/users spend on data engineering by allowing them to - through a visual interface - rapidly determine whether it is worth spending time working with a specific dataset.

Using a bottom-up approach, derived by exploring the data and by two researchers with no specific location data expertise, a six-factor rating was developed – including factors that relate to the time required to map the data and the ease of the process, as well as factors relating to data quality - additional data, spatial and temporal density is proposed. A web-based App has also been developed to showcase the potential of the concept and allow any mappable data to be visualised.

Flexibility is in-built - the rating system is designed to be populated by - and used by - data scientists/users, potentially working in different contexts/on different topics, and with different levels of data engineering expertise.

4.1 Benefits of the Rating

The rating factors were selected as those that could be measured through a visual examination and preliminary exploration of the data – i.e. by accessing and mapping the data, with the factors and their weightings reflecting tasks that are performed by data scientists/users in their initial data review, thus providing significant time-savings to data scientists/users.

The m(App)ability App was developed using a responsive mode Bootstrap template, which means that it can be run on a desktop, laptop or mobile device and can easily be incorporated into a data repository as an alternative way to present search results.

Overall, the visualisation provides insight into the accessibility and usability of open data and the power of visualising the ratings can be seen by comparing Figures 1 and 2. The rating system enables a rapid visual assessment of the dataset, allowing the user to rapidly decide whether they should actually spend additional time reading metadata or attempting to download and evaluate the dataset. This is particularly beneficial in situations where the user is not an expert data engineer.

4.2 Enhancing Current Approaches to Data Evaluation and FAIRness

The rating system is designed to be easy to understand and to be flexible. It is not intended to replace full, standards-based metadata or dataset evaluation, but rather to offer a **precursor** that can be easily deployed by a data provider and by which a data scientist/user can make a rapid evaluation of the data and decide whether to proceed to a full, more time-consuming, evaluation.

Comparing the remaining rating factors to ISO 19157 it can be seen that the chosen factors provide a precursor to full quality and fitness-for-purpose assessment, focussing primarily on measures that relate to completeness in the ISO 19157 approach (e.g. do the locations covered by the data correspond to those expected? Are there missing data elements? Are there extra data elements? What is the temporal quality of the dataset?). The rating does not replace a full quality assessment, but provides information to users as to whether an in-depth, time-consuming, exploration of the data and its quality is warranted.

In terms of FAIRness, the rating system is particularly relevant for *accessibility* and *interoperability*, indicating whether the data can be accessed/downloaded and a minimal interoperability approach (Section 2.2) used to find common location elements. Our rating does not replace a FAIRness assessment but rather takes it further providing an implementable approach that

[2] https://www.data.gov.uk/dataset/d5c0a927-8fb8-45d1-b508-4a1d446e84dd/department-for-education-contracts, Accessed 3rd January 2024

can provide insight into FAIRness and assist a data user in deciding whether time-consuming and potentially expensive work towards full interoperability would be beneficial.

Re-usability is also a key component of FAIR. The rating system provides a structured approach for evaluating datasets, combining objective and subjective factors. This provides inherent flexibility to generate multiple ratings based on the specific context of the dataset's potential use and the researcher's expertise in data processing. This interpretative flexibility recognises that datasets may be rated differently based on the specific use case and the data processing capabilities of the researcher. Multiple rating sets can be associated with the same dataset, each one reflecting a different use case or user expertise. Additionally, depending on user needs, different elements of the ISO 19157 standard could be selected to be the main focus of the visual representation.

4.3 Is Open Data Available and Map-Able?

Our findings show that out of the 54 datasets reviewed under the 'environment' topic, only 11 were directly available. Of these, 55 per cent (6 datasets) could be directly mapped. Similarly, for 'education' of the 50 datasets identified, 33% were available for testing, and 22% of these could be indirectly mapped. The results may have been impacted as titles and labelling of some datasets were ambiguous and vague, the search tool was ineffective, omitting important results while showing out-of-context ones, and some data was duplicated, which could be misleading.

Over 55% of the tested environment data and 66% of the education data could be mapped to a greater granularity than just 'London' (which would be true for all the data).

These results highlight that open data is not reaching its potential in the UK (and perhaps reflect the middle ranking 62.5 Open Data Indicator score [Open Data Watch, 2022]). This is a surprisingly low score, and indicates that significant additional work is required by data producers and curators.

4.4 Future Work

The outcomes presented here are a proof of concept, demonstrating potential benefits of a simplified, visual, approach to rating data. However, flexibility of the rating is a key feature and the weightings of each factor and/or the specific factors chosen will vary depending on users and context. Further comparison - in particular A/B testing - with current approaches to data quality review and in particular to full standards-based metadata are required, with the outcome being a context-specific sliding scale of rating detail from full metadata down to the restricted set presented here.

A number of elements of the rating could be generated automatically - e.g. it should be possible to determine the format of the data, and to parse the data and metadata to determine whether it can be mapped (directly or indirectly) by querying the API offered by repositories such as data.gov.uk.

Similarly, this conversion/pre-processing could be automated. While it was relatively easy to convert the air quality data to the required format this did require some understanding of the use of GIS software for conversion and also of coordinate reference systems. To make the conversion simpler, this process could be automated via a search for columns with appropriate titles (Easting, Northing, latitude, longitude) or values within the known coordinate ranges and systems for a particular country. This automation could be further extended to search for place names in the dataset (or even in the metadata) greatly increasing the percentage of data that can be mapped. This could be facilitated by the use of a standard set of UK geographies (e.g. census output areas, town boundaries, parishes, councils, wards, counties and so forth).

The selected repository - data.gov.uk - may not include as extensive coverage of social science data as the UK Data Service or the Office of National Statistics, both of which are likely to be extensively used by data scientist. In particular, it can be expected that the ONS data would be far more map-able, given that it relates directly to statistical units of geography. Further research, using other repositories both in the UK and elsewhere, is required to establish exactly how map-able open data is. This would also go some way to understanding whether the apocryphal '80% of data is geospatial' figure is valid.

As noted above, the mix of objective/subjective ratings was a deliberate choice for this project. We fully acknowledge that ratings are not a one-size-fits-all, however and the ratings could be extended in future – e.g. by subdividing or weight the rating factors by whether the mapping task can be achieved by 'expert' 'average' and 'novice users or by the specific application to which the dataset will be applied. This flexibility can be embedded into the App - further research is required to determine whether this additional level of complexity is counterproductive in terms of the simplicity of our current approach.

5. Conclusion

The work presented in this paper provides a starting point to fill the gap between the concepts presented by FAIR and providing specific implementation guidelines for data repository owners and developers. In particular, it takes into account the fact that data scientists/users are time-poor and also may not have strong data engineering skills, having to re-learn tasks such as data transformation every time they start a new project. An integral development in this study was the creation of a rating system that reflects both a summary of the quality of the data and the potential for minimal interoperability, generated by exploring the data itself and also by researchers whose expertise in location data is low. This rating system and the associated App provides researchers with a structured and comprehensive evaluation mechanism, enhancing the integration of location-based data. By offering a visual representation combining features of pie and radar charts, the rating system aids researchers in rapidly assessing dataset quality and relevance, thereby streamlining the process of data integration.

However, in attempting to demonstrate the power of location to underpin a 'minimal interoperability' approach, this research also highlighted limited data accessibility and demonstrates the urgent need for improvements in data-sharing. Valuable information remains restricted and data continues to be fragmented, leading to inefficiencies in resource allocation, prolonged data manipulation efforts by data scientists, and potential obstruction of research progress. It also represents missed opportunities to leverage data for informed decision-making. As the volume of data continues to surge, it is imperative that concerted efforts are made to enhance data-sharing, making critical information readily available to those who seek to build smarter and more sustainable urban environments.

6. Acknowledgements

The authors would like to thank the University College London's Laidlaw Foundation Scheme, which provided the funding for this research `https://www.ucl.ac.uk/laidlaw-scholarships/`

References

Ahmed, M., Karagiorgou, S., Pfoser, D., Wenk, C., 2014. A comparison and evaluation of map construction algorithms using vehicle tracking data. *GeoInformatica*. https://doi.org/10.1007/s10707-014-0222-6.

Barcellos, R., Bernardini, F., Viterbo, J., 2022. Towards defining data interpretability in open data portals: Challenges and research opportunities. *Information systems*, 106, 101961.

El-Mekawy, M., Östman, A., 2010. Semantic mapping: an ontology engineering method for integrating building models in ifc and citygml. *3rd ISDE DIGITAL EARTH SUMMIT, 12-14 June,*.

Ellul, C., Foord, J., Mooney, J., 2013. Making metadata usable in a multi-national research setting. *Applied ergonomics*, 44(6), 909–918.

Eranki, K., Reddy, A., 2012. Geo-spatial library: a geospatial educational tool for knowledge management and capacity building. *2012 IEEE International Conference on Engineering Education: Innovative Practices and Future Trends (AICERA)*. https://doi.org/10.1109/aicera.2012.6306753.

Geospatial Commission, 2023. UK Geospatial Strategy 2030 – Unlocking the Power of Location.

Hong, N. C., Cozzino, S., Genova, F., Hoffmann-Sommer, M., Hooft, R., Lembinen, L., Martilla, J., Teperek, M., Holl, A., 2020. Six recommendations for implementation of FAIR practice.

ISO, 2013. Geographic information - Data quality. Standard, International Organization for Standardization, Geneva, CH.

Jonietz, D., Zipf, A., 2016. Defining fitness-for-use for crowdsourced Points of Interest (POI). *International Journal of Geo-Information*. https://doi.org/10.3390/ijgi5090149.

Kumar, K., Labetski, A., Ohori, K. A., Ledoux, H., Stoter, J., 2019. Harmonising the OGC standards for the built environment: A CityGML extension for Landinfra. *ISPRS International Journal of Geo-Information*, 8(6), 246.

Kuzma, M., Mościcka, A., 2020. Evaluation of metadata describing topographic maps in a national library. *Heritage Science*. https://doi.org/10.1186/s40494-020-00455-3.

Lewis, J., Sauro, J., 2009. The Factor Structure of the System Usability Scale. *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International.* `https://doi.org/10.1007/978-3-642-02806-9_12`.

Moellering, H., Aalders, H., Crane, A., 2005. *World spatial metadata standards: scientific and technical descriptions, and full descriptions with crosstable*. Elsevier.

Mons, B., 2018. *Data stewardship for open science: Implementing FAIR principles*. CRC Press. .

Mulquin, M., 2023. Interoperability and the minimal interoperability mechanisms. *Personal Data-Smart Cities: How cities can Utilise their Citizen's Personal Data to Help them Become Climate Neutral*, River Publishers, 135–149.

Open Data Watch, 2022. Indexes of Data Quality and Openness.

Open Data Watch, 2023. Open data bienniel reports - 2016, 2022. Technical report, Open Data Watch.

Open Research Data Taskforce, 2018. Realising the potential - Final report of the Open Research Data Task Force.

Open Science in Horizon Europe, n.d. Open Science in Horizon Europe.

Praseptiawan, M., Untoro, M., Fahrianto, F., Prabandari, P., Wisnubroto, M., 2023. Redesigning UI/UX of A Mobile Application Using Task Centered System Design Approach. *Applied Information System and Management (AISM)*. https://doi.org/10.15408/aism.v6i1.24665.

Reis, J., Housley, M., 2022. *Fundamentals of Data Engineering*. " O'Reilly Media, Inc.".

Sani, M. J., Musliman, I. A., Abdul Rahman, A., 2022. IFC to CityGML Conversion Algorithm Based on Geometry and Semantic Mapping. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 46, 287–293.

Saquicela, V., Vilches-Blázquez, L. M., Freire, R., Corcho, O., 2022. Annotating OGC web feature services automatically for generating geospatial knowledge graphs. *Transactions in GIS*, 26(1), 505–541.

Siew, C., Abdul Halim, N., Karim, H., Zain, M., Looi, K., 2021. CityGML Application Domain Extension for 3D Strata Representations in The Smartkadaster System: Towards Beyond Cadastre Purpose In Malaysia. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 8, 99–103.

Soliman, M., Morsy, M., Radwan, H., 2022. Assessment of implementing land use/land cover LULC 2020-ESRI global maps in 2D flood modeling application. *Water*. https://doi.org/10.3390/w14233963.

Wang, L., Guo, M., Sawada, K. e. a., 2015. A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. *Geosciences Journal*. https://doi.org/10.1007/s12303-015-0026-1.

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G.and Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*.

Yu, S. C., Ahn, J. W. et al., 2022. Design of 3D Data Model of Underground Utilities in Korea Using CityGML Application Domain Extension. *Sensors & Materials*, 34.

Yuan, M., Recker, M., 2015. Not All Rubrics Are Equal: A Review of Rubrics for Evaluating the Quality of Open Educational Resources. *The International Review of Research in Open and Distributed Learning*. https://doi.org/10.19173/irrodl.v16i5.2389.