



GRANULAR

SCREENING RURAL DATA SOURCES

D 3.1

MARCH 2023



Funded by the
European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. UK participants in the GRANULAR project are supported by UKRI- Grant numbers 10039965 (James Hutton Institute) and 10041831 (University of Southampton).

D3.1 SCREENING RURAL DATA SOURCES

Project name	GRANULAR: Giving Rural Actors Novel data and re-Useable tools to Lead public Action in Rural areas
Website	www.ruralgranular.eu
Document type	Data/Report
Status	Final
Dissemination level	Public
Authors	McCallum I., Hoffer M., Laso Bayas J.C. (IIASA), Kull M., Vihinen, H. (LUKE), Ysebaert R., Guérois M., Giroud T., Viry M., Lambert N. (RIATE), Voepel H., Steele J. (UoS), Sorichetta A. (UniMi), Tapia C., Cuadrado A. (Nordregio), Miller D., Hopkins J., Farinelli V. (HUT), Ulman M., Simek P., Motyckova V. (CZU), Panoutsopoulos H., Chitos A., Fournarakos A., Zafiraki P. (AUA), Berchoux T. (IAMM)
Work Package Leader	International Institute for Applied Systems Analysis (IIASA)
Project coordinator	Mediterranean Agronomic Institute of Montpellier (IAMM)

Citation: McCallum, I., Hoffer, M., Laso Bayas, J. C., Kull, M., Vihinen, H., Ysebaert, R., Marianne, G., Giraud, T., Viry, M., Lambert, N., Voepel, H., Steele, J., Sorichetta, A., Tapia, C., Cuadrado, A., Miller, D., Hopkins, J., Farinelli, V., Ulman, M., ... Berchoux, T. (2023). Screening rural data sources. GRANULAR. <https://doi.org/10.5281/zenodo.13838524>



This license allows users to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.



Table of contents

1. Executive summary	3
2. Introduction	3
3. Methods	6
3.1. Initial data screening protocol	6
3.2. Evaluation Framework and Template	7
3.3. Study Limitations	8
4. Results and Discussion	9
4.1. Evaluation of data screening	9
4.2. Coverage of a Rural Compass	12
4.3. Additional review of methods and data	13
4.3.1. Accessibility methods and data	14
4.3.2. Human mobility methods and data	14
4.3.3. Earth Observation methods and data	14
4.3.4. Nowcasting and webscrapping methods and data	15
4.3.5. Crowdsourcing data and methods	16
4.4. Data repository	16
5. Conclusions	17
6. References	17

1. Executive summary

This document presents an initial screening of datasets that are relevant to capture rural diversity and to create novel indicators for rural areas. Following a semi-structured format of discovery and evaluation, we have documented 90 different datasets to date which are either already used to characterise rural areas, or could underpin novel indicators. In addition to identifying the datasets themselves and their locations, we provide a suite of associated meta-data. Evaluating the findings of this effort, we demonstrate that the majority of the datasets identified have regional to global coverage, have Local Administrative Unit to gridded (10m - 10km) granularity, are provided annually, are free and open and of moderate relevance in terms of indicator generation for rural areas. With the completion of this deliverable, exploration can begin on the development of the next generation of rural indicators.

2. Introduction

For much of Europe's rural areas, detailed socio-economic and environmental data is scarce, heterogeneous and static (Andersson et al., 2017). Yet, policy-makers and rural actors depend on rigorous evidence to adequately address the challenges that rural communities across Europe face to allow for the prioritization of interventions or enabling innovations. The project GRANULAR aims to address this need by identifying emerging data sources and methods to develop new indicators of sustainable rural development. This deliverable is the first step in this process, documenting a wide array of existing data sources that could potentially underpin these new and novel indicators.

As part of the conceptualisation of rural diversity undertaken in GRANULAR through the elaboration of a Rural Compass, a suite of topics relevant to decision-making in rural areas is being developed (Figure 1). For some of the topics, data already exists from which indicators are generated and openly available (e.g. Eurostat <https://ec.europa.eu/eurostat/web/regions-and-cities/overview>). In addition, the Joint Research Centre of the European Commission recently launched the [Rural Observatory](#), which offers a suite of indicators specific to rural areas. It may, however, be desirable to increase the granularity of some of the indicators (i.e. spatially or temporally) or to add novel indicators where gaps are identified. GRANULAR is also exploring potential indicators from the standpoint of policy needs, both at local and EU levels. Hence, we are considering these broad indicator needs in our assessment of data availability and potential methods.

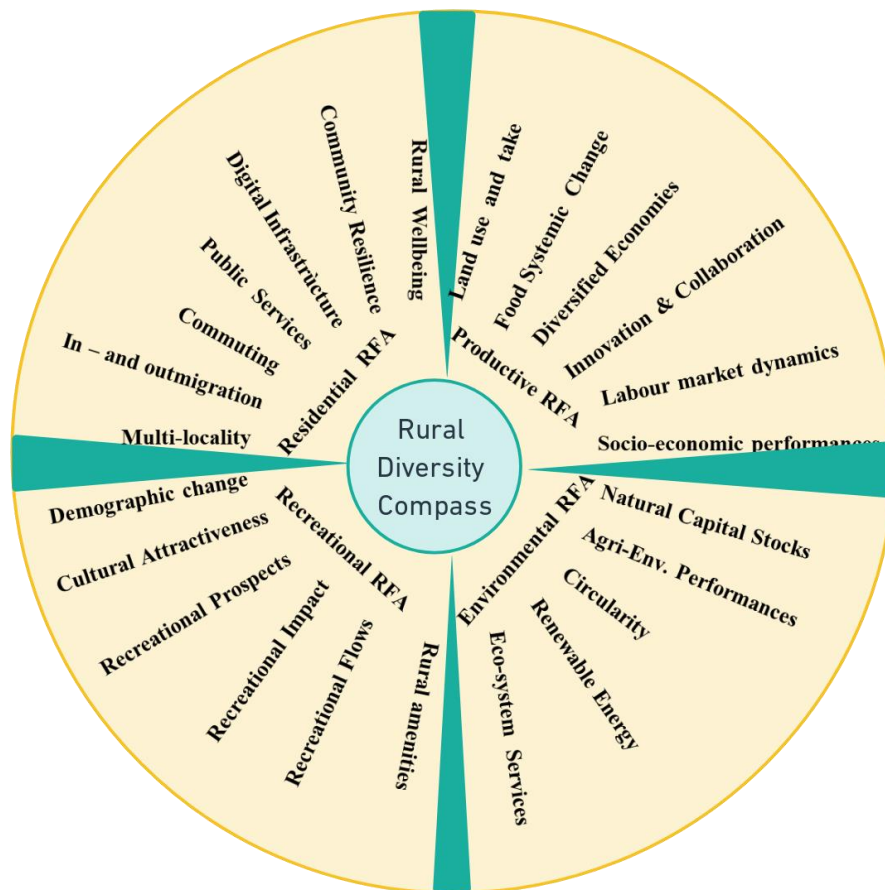


Figure 1. Potential rural development indicators generated via the GRANULAR Rural Compass, classified by functional areas. These are broadly classified as either: residential, productive, recreational or environmental.

In addition to the Rural Compass, 7 Living Labs (LLs), with specific regional policies have elicited the issues that are of most concern to them with regards to rural development, and their associated needs for data and tools (Table 1). The LL priorities served as a starting point for this benchmark to identify existing datasets that match their data needs.

Table 1. Living Labs rural issues, data needs and priorities.

Lab	Rural Issues, Needs, Priorities
Netherlands	Well-being: Methods exist and work well in urban areas, but more difficult to apply in rural areas
Spain	Well being, mobility, tourism, land use, health, infrastructure, education, digitalisation, forest fires, e-governance
France	Numerous topics re. mobility, environment, tourism, marine boating, beach and coastal usage
Poland	Food chain characteristics

Scotland	Natural capital(NC), change over time in value in NC, (e.g. air quality, water quality) Indicators related to people, assets, business related, population
Sweden	Composition of business environment, accessibility of public services, territorial innovation, economic performance
Italy	Climate change exposure, food chains characteristics, tourism

Considering the above-mentioned topics and indicators, WP3 proceeded with a semi-structured format, as described below, to catalogue the variety of potential datasets that can capture the aforementioned needs.

3. Methods

This Deliverable has screened a wide range of existing and emerging data sources and methods that are viable options for consideration within GRANULAR. The project partners were called upon to apply their domain expertise to explore the entire spectrum from conventional to unconventional data. Significant amongst those data are national surveys, censuses, micro-data, Earth observation (e.g., Copernicus, Galileo, EGNOS), online text, IoT, crowdsourcing, and hybrid data (Figure 2). A strong focus was placed on open data complying with the FAIR Principles of data management, however we also screened proprietary data where possible (e.g. mobile data). A strong emphasis was placed upon themes for which there is a lack of data on several aspects at appropriate geographic scales, in particular, on climate and environmental performance and on social challenges, quality of life and well-being.

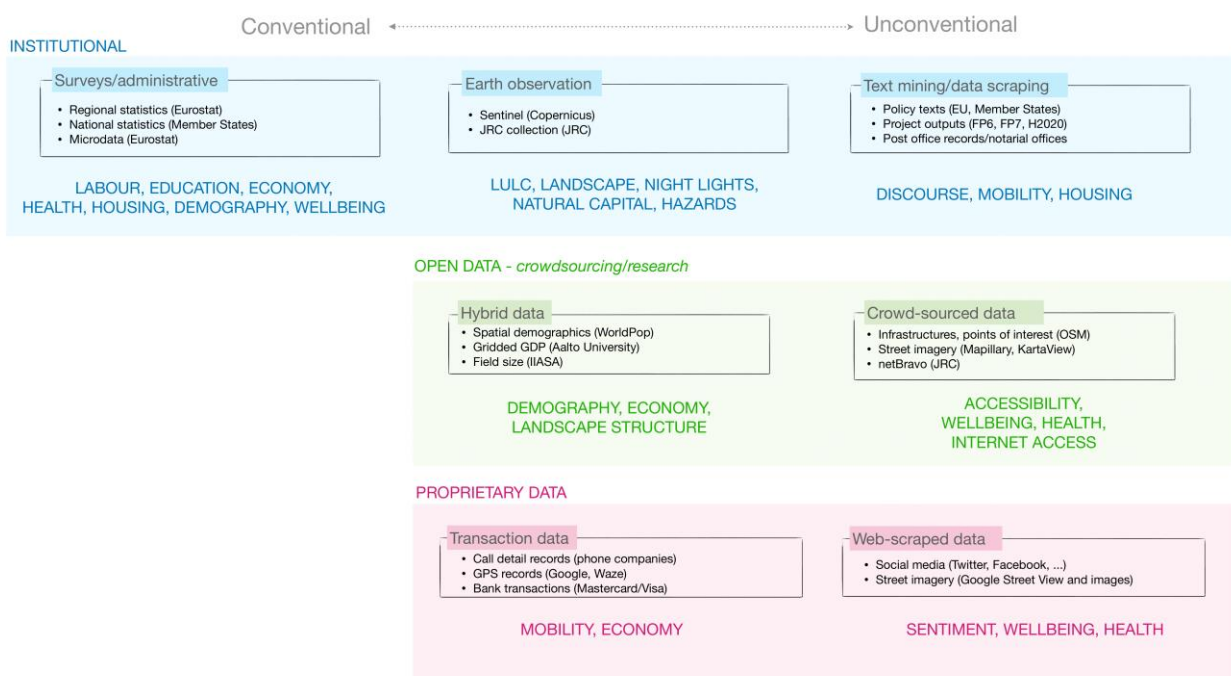


Figure 2. Broad data categories that were screened. Data types considered included institutional data, open data and proprietary data and cover the full spectrum from conventional to unconventional in nature.

3.1. Initial data screening protocol

To develop our initial screening protocol, we started from the prototype Rural Diversity Compass described above. Using the broad concept of functional areas, namely residential, productive, recreational or environmental, we were able to define the semi-structured screening method and protocol. For the evaluation framework and semi-systematic review we refer to Allen *et al.*, (2021); An and Alarcón (2020); and Hargreaves and Watmough (2021).

We initially relied upon GRANULAR partner domain expertise to identify relevant datasets to address the rural diversity compass. Complementing this approach, we applied a semi-structured screening protocol using the Google Dataset Search <https://datasetsearch.research.google.com/>. An initial search was performed using the term “rural data europe”. We set the timeline for last update to “past 3 years”, and selected all download formats, usage rights, topics and open data. We then performed a series of searches for the broad functional areas and selected relevant datasets.

3.2. Evaluation Framework and Template

In order to catalogue the relevant datasets and apply an evaluation framework, we employed a modified approach of Allen *et al.*, 2021 (Table 2). Based on this and partner expertise and feedback, we established a template for the collection and storage of meta-data.

Table 2. Framework of attributes and criteria for evaluating datasets. Modified framework based on Allen *et al.*, 2021.

CHARACTERISTICS	Description/categories
Indicator class	Main thematic addressed
Data Class (s) ¹	<ol style="list-style-type: none"> 1. Administrative (e.g. census, national public surveys) 2. Commercial or transactional data arising from the transaction between two entities (e.g. credit cards, online transactions, scanner data) 3. Satellite imagery and Earth Observation (e.g. Landsat, Sentinel, night lights etc). 4. Other sensor networks (e.g. road sensors, climate sensors, air pollution sensors, smart meters) 5. Tracking device (e.g. tracking data from mobiles, GPS) 6. Behavioural data (e.g. online searches, online page views) 7. Opinion data (e.g. social media)
Title	Title of the dataset
Description	Brief description of what the dataset captures
Authors and ID	People/organisations behind the dataset and identifiers (DOI/API)
...	Additional data characteristics
CRITERIA	Description/Scoring
1. Scope/Coverage	Relates to the geographic coverage of each study from local to global: 4 – global (includes good coverage of most countries globally) 3 – regional (multiple countries covering a geographic region) 2 – national (limited countries – 1 or a few countries) 1 – local/sub-national area within 1 country 0 – don’t know or N/A
2. Granularity	Relates to the granularity of the study or level of spatial disaggregation – ranging from national, sub-national (admin levels 1 to 3). Subnational admin levels correspond to NUTS used in EU: 5 – gridded data (specify) 4 – admin level 3 (eq. LAU) 3 – admin level 2 (eq. NUTS3) 2 – admin level 1 (eq. NUTS2) 1 – national 0 – don’t know or N/A

¹ Based on UNITED NATIONS STATISTICAL COMMISSION 2014. Big data and modernization of statistical systems; . *Report of the Secretary-General. E/CN. 3.2014/11 of the forty-fifth session of UNSC 4-7 March 2014*. New York: United Nations.

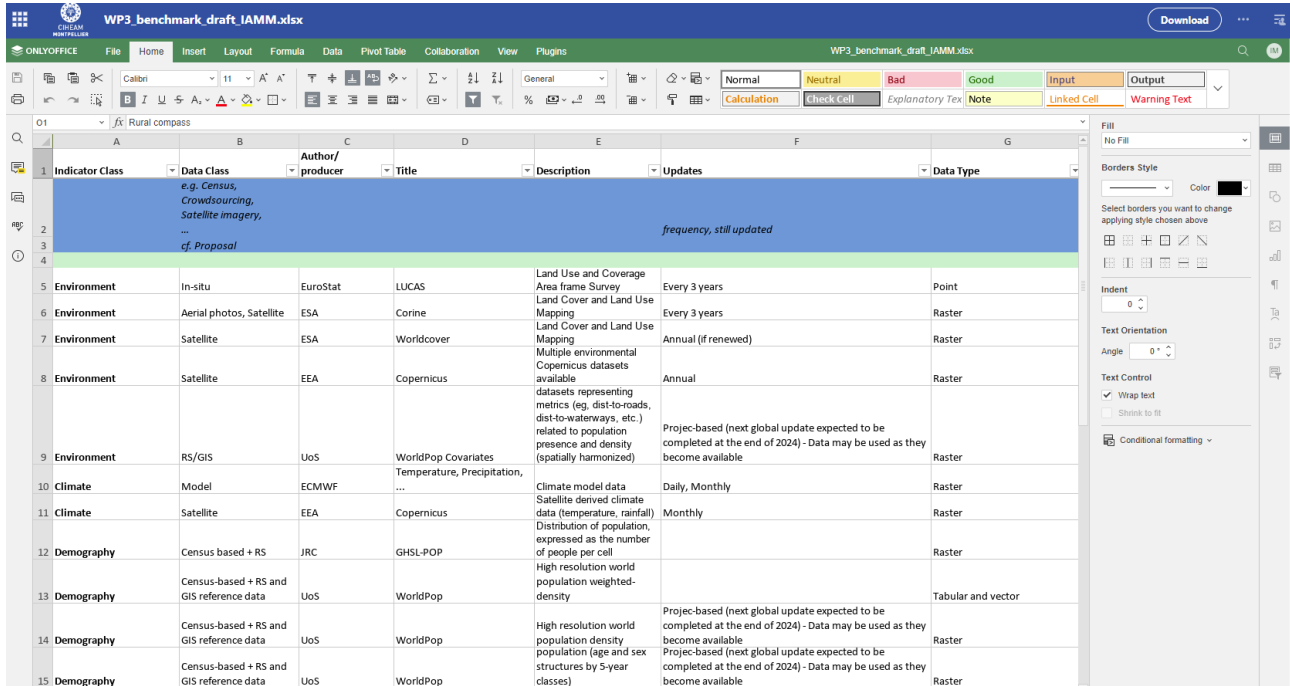
3. Frequency/Timeliness	<p>Relates to the frequency of production of the dataset: can support national-level report (Yes/No) and latest data point.</p> <p>Additional frequency evaluation:</p> <p>5 – weekly/daily 4 - monthly 3 - quarterly 2 – annual 1- > 1 year 0 – don't know or N/A</p>
4. Costs/Access/Replicability	<p>Relates to the free and open access and availability of the derived datasets, as well as raw datasets (for the analysis) and the model or code:</p> <p>Yes – derived dataset made open access, preferably with input datasets and code needed to reproduce results</p> <p>No – not open source; derived datasets and code are not made available</p>
5. Relevance	<p>Relates to the degree of relevance of the study in terms of supporting the rural compass:</p> <p>4 - Very high relevance – can be used to monitor >2 compass components and fill a current priority gap</p> <p>3 - High relevance – can be used to monitor 1 or 2 compass components</p> <p>2 - Moderate relevance – can be used as complementary data source for 1 or more compass components (i.e. partial dataset for deriving an indicator, or additional disaggregation of an official indicator).</p> <p>1 - Lower relevance – does not correspond to a compass indicator, but can be used as a proxy or complementary new dataset for monitoring a target or goal</p> <p>0 – don't know or N/A</p>

3.3. Study Limitations

At this stage we were able to perform a semi-structured screening of the relevant rural datasets for Europe applicable to the objectives of GRANULAR. These efforts are ongoing, with account taken as new datasets are discovered and new rural indicators are identified and prioritised. Future efforts will include the publication of this exercise in the form of a structured review of European rural data.

4. Results and Discussion

Based on the methodology described above, a table was designed to capture as wide an array as possible of potential datasets, along with their associated meta-data (Figure 3). This effort will continue to evolve over the course of the project, ultimately populating the GRANULAR data repository and digital platform.



Indicator Class	Data Class	Author/producer	Title	Description	Updates	Data Type
	e.g. Census, Crowdsourcing, Satellite imagery, ... <i>cf. Proposal</i>				<i>frequency, still updated</i>	
Environment	In-situ	EuroStat	LUCAS	Land Use and Coverage Area frame Survey	Every 3 years	Point
Environment	Aerial photos, Satellite	ESA	Corine	Land Cover and Land Use Mapping	Every 3 years	Raster
Environment	Satellite	ESA	Worldcover	Land Cover and Land Use Mapping	Annual (if renewed)	Raster
Environment	Satellite	EEA	Copernicus	Multiple environmental Copernicus datasets available	Annual	Raster
Environment	RS/GIS	UoS	WorldPop Covariates	datasets representing metrics (eg. dist-to-roads, dist-to-waterways, etc.) related to population presence and density (spatially harmonized)	Projec-based (next global update expected to be completed at the end of 2024) - Data may be used as they become available	Raster
Climate	Model	ECMWF	...	Temperature, Precipitation, ...	Daily, Monthly	Raster
Climate	Satellite	EEA	Copernicus	Satellite derived climate data (temperature, rainfall)	Monthly	Raster
Demography	Census based + RS	JRC	GHSL-POP	Distribution of population, expressed as the number of people per cell		Raster
Demography	Census-based + RS and GIS reference data	UoS	WorldPop	High resolution world population weighted-density		Tabular and vector
Demography	Census-based + RS and GIS reference data	UoS	WorldPop	High resolution world population density	Projec-based (next global update expected to be completed at the end of 2024) - Data may be used as they become available	Raster
Demography	Census-based + RS and GIS reference data	UoS	WorldPop	High resolution world population (age and sex structures by 5-year classes)	Projec-based (next global update expected to be completed at the end of 2024) - Data may be used as they become available	Raster

Figure 3. Screenshot of the GRANULAR rural data table (March, 2023).

4.1. Evaluation of data screening

As of March 2023, we have collected more than 90 datasets (and associated metadata) based upon the protocol outlined above. These represent a wide range of indicator classes spanning from energy to health, the environment and well-being, with demography, infrastructure and environmental indicator classes comprising half of the total datasets (Figure 4a). For each dataset, an extensive meta-data collection is provided spanning data extent, spatial resolution, temporal extent, etc. For the majority of datasets, URLs and DOIs are provided to locate the data.

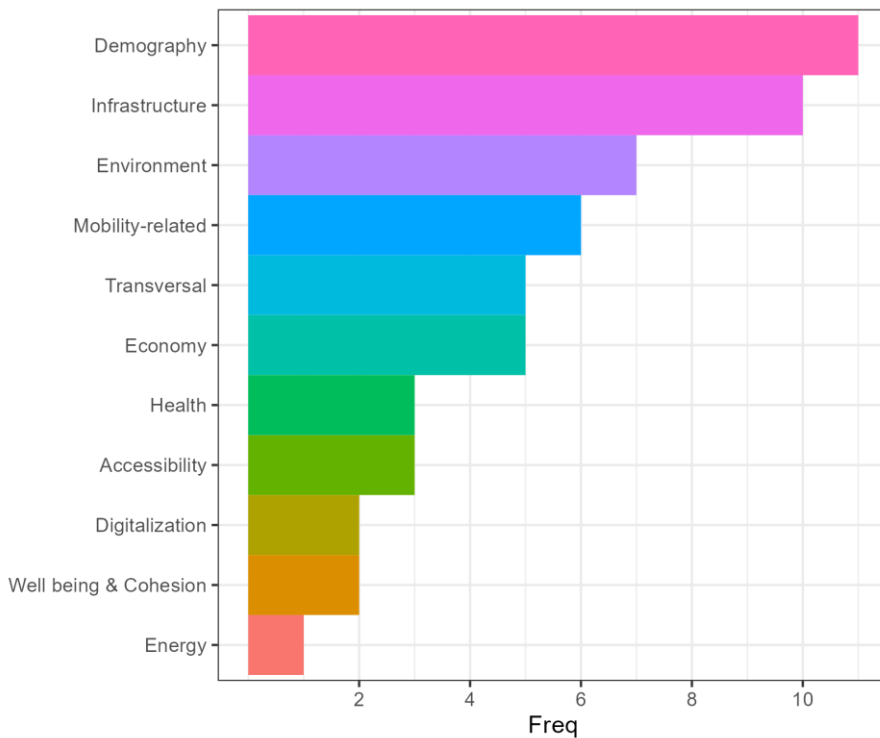


Figure 4a. Frequency of indicator classes captured within the WP3 data screening.

Figure 4b represents the variety of data classes present in the data screening, with more than half of the classes comprised of statistical, census and remote sensing derived datasets.

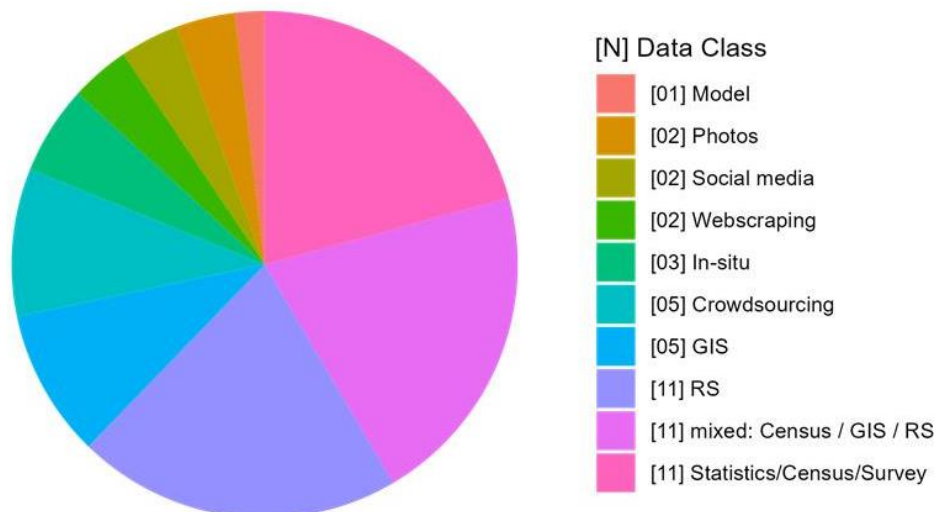


Figure 4b. Frequency of data classes captured within the WP3 data screening.

Figure 4c. depicts the data producer classes present in the current screening, with almost half of the datasets provided by various EU institutions (e.g. Eurostat). This indicates the important role that the European Commission plays in data acquisition and provision.

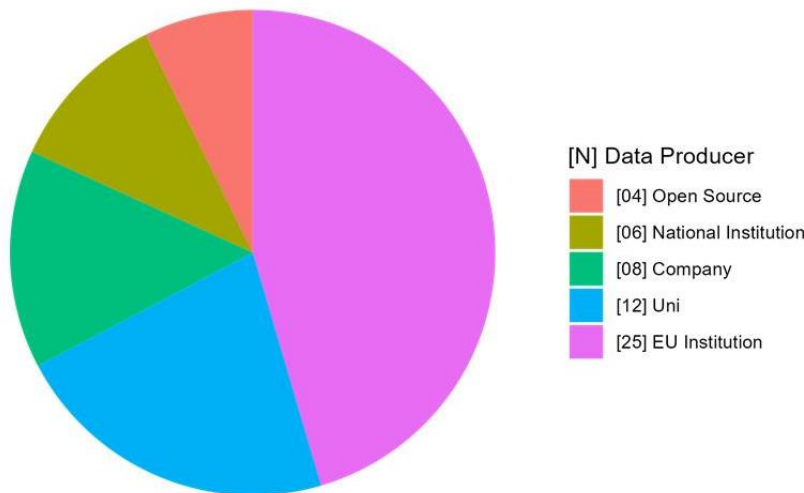


Figure 4c. Frequency of data producer classes captured within the WP3 data screening.

In addition, we evaluated the data collection based on the following criteria, as described above, summarised in Table 3 of: Scope/coverage; Granularity; Frequency/Timeliness; Costs/Access/Replicability; and Relevance.

Table 3. Results of evaluation criteria.

Criteria	Description	Result (Average)
Scope/Coverage	Relates to the geographic coverage of each study from local to global:	3.4 (between regional and global)
Granularity	Relates to the granularity of the study or level of spatial disaggregation	4.5 (between LAU and Gridded)
Frequency/Timeliness	Frequency with which the raw data is produced	2 - Annual
Costs/Access/Replicability	Relates to the free and open access and availability of the derived datasets	Yes (Majority), some partially open data and proprietary data exists
Relevance	Relates to the degree of relevance of the study in terms of supporting the rural compass	2.3 (Moderate relevance)

On average, the majority of our datasets fall between regional and global in scope. In terms of granularity, they are typically of an LAU resolution or a gridded resolution. On average, our screened data has an update frequency of 1 year, meaning annual products are most often available. However, datasets with lower and higher update frequency are common. In terms of access, the majority of datasets are free and open, with several partially open (or open with restrictions). Finally the screened datasets fall under moderate relevance on average in terms of supporting the rural compass. These are subjective scores based on the available metadata and will be refined as the data is tested later in the project.

4.2. Coverage of a Rural Compass

The aim of this Deliverable is to document as wide an array of existing data sources as possible that could potentially underpin new and novel indicators determined via the Rural Diversity Compass. In particular, we measured the coverage of the broad functional areas identified in the Compass by the datasets captured in this screening, namely residential, productive, recreational, environmental or other (Figure 5). Here we see that of the four functional areas, we have identified datasets for three of them, along with several datasets falling under the “other” class. The majority of our datasets refer to the residential class (over half) followed by environmental and productive classes. Currently we have not identified any data classes falling under the recreational classes specifically, although several datasets could be classed under multiple categories (including recreational, e.g. Strava mobility data).

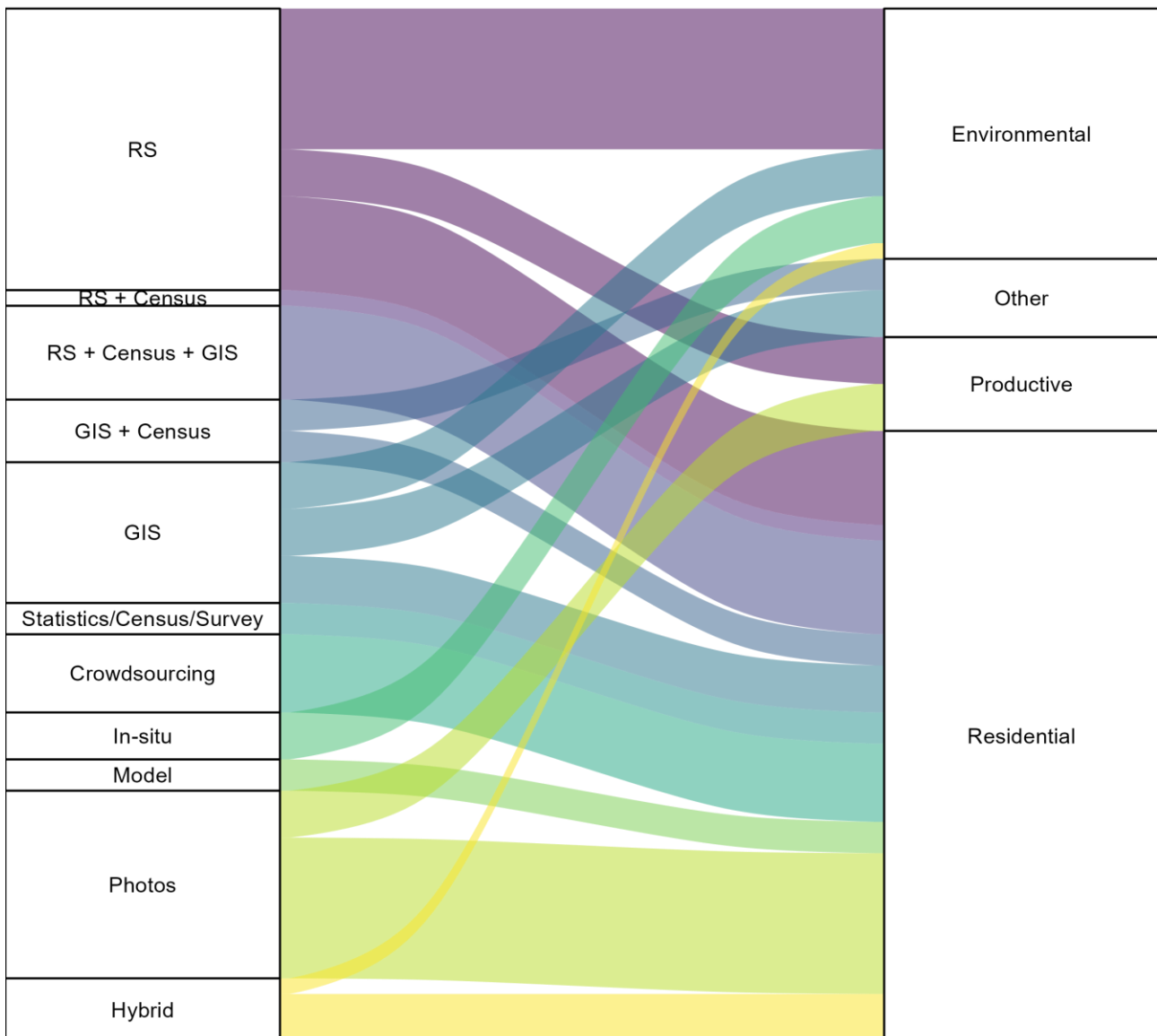


Figure 5. A comparison of the data classes available from the search (left) and the broad rural compass indicators (right) to which they link. The thickness of the lines relates to the number of datasets, relative to the total.

4.3. Additional review of methods and data

The following sub-sections describe specific areas of review for data to address new and novel rural indicators, namely the areas of human mobility, accessibility and earth observation.

4.3.1. Accessibility methods and data

A review of the accessibility literature and data availability was completed, which is available at the following [URL](#). It provides an overview of possible solutions and limitations for creating accessibility indicators at the European context.

The first part of the document presents, at European scale, the policy context and the main initiatives developed so far for proposing harmonized indicators on accessibility. The second part highlights the main issues to be considered when calculating accessibility indicators (origin-destination pairs, routing engines, accessibility indicators computation). The third section makes an overview of existing databases and possibilities that could be considered in a European context for the selection of origins / destinations pairs. The fourth part highlights existing solutions for routing engines according to several transportation modes (road, cycle, transport-transit). The final section discusses possibilities offered in term of indicator creation when the travel time matrix is calculated with a case-study on hospitals in France.

4.3.2. Human mobility methods and data

A literature review for human mobility methods and data is near completion consisting of 150+ papers and reports, 19 of which have associated open source datasets available. A summary of article counts in the review across three broad categories is given here:

- Data type: Facebook (5), Twitter (12), Google LH (10), other Google (8), mobile phone (38), call records (19), WhatsApp (1), social media (13), GPS (27), Wi-Fi (4)
- Location type: Europe/EU (39), urban (36), rural (9)
- Analytical type: Big data (20), COVID-19 (28), AI (3), machine learning (5), functional zones (1)

When complete, the literature review will provide a roadmap for establishing a human mobility database that includes raw and processed data, plus modelled mobility outputs to fill gaps where there is missing or sparse data in the living laboratories. The next steps include assessing available datasets (plus costs if applicable) and establishing select methods for producing modelled mobility outputs.

4.3.3. Earth Observation methods and data

Over the past fifty years, the range and accessibility of datasets suitable for geospatial analysis of rural socio-economic conditions have significantly increased. These datasets include various categories such as optical and radar Earth Observation (EO) satellites and sensors, EO data portals, geospatial population and infrastructure data layers, cloud-based interfaces, and freely available socio-economic survey data.

We identified 11 optical (Sentinel-2, Landsat, ASTER, MODIS, VIIRS, Worldview, GeoEye, Pléiades, Skysat and Flock) and 7 Synthetic Aperture Radar (SAR) (SRTM, ALOS-2, SAOCOM, Sentinel-1, TanDEM-X, TerraSAR-X) EO satellites and sensors that are currently operational and of relevance to monitor rural indicators. Data associated with EO cover a wide range of sectors for rural policy planning, mostly associated with the environmental (climate, soil characteristics, terrain, biodiversity, vegetation, forest, hazards), productive (land use, irrigation, agricultural systems, energy) and residential (settlements, night light, urbanisation, heat islands) functions of rural areas.

In GRANULAR, the focus will be on developing methods that combine sensors (SAR and optical) in a common framework to derive novel indicators. The methods will leverage models for cross-modal image/text analysis with the objective of extending them with EO data to get a spatial representation of rural indicators.

4.3.4. Nowcasting and webscraping methods and data

A review of the literature on nowcasting indicators is ongoing. The starting point for this review includes 297 papers in Scopus found using the following search string: “nowcast* AND indicator*”. The majority (86%) of these documents are academic articles and the rest are conference papers and book reviews. Most of the documents are classified in the subject areas of economics or business management (~35%), mathematics or computer science (~18%), social or decision sciences (~17%), earth science (11%), and environmental science and engineering (6%). Remarkably, the search “nowcast* AND indicator* AND rural*” only yielded two hits on our Scopus search. This suggests that the degree of adoption of nowcasting techniques for rural policy design, monitoring and evaluation is still limited. The unspecificity of nowcasted indicators also highlights the relevance of the ongoing scoping process in the task.

The activity is now progressing in the classification of the indicators found in the literature according to a set of criteria, including, inter alia: (1) theme within the rural policy framework; (2) scale and granularity; (3) data source; (4) methodology, including methods and techniques. This classification will be instrumental in deciding on feasible and relevant nowcasting indicator(s) to be produced and tested in the project. For this aim we shall adopt a staged-based decision process based the following decision criteria: (1) conceptual alignment with the Rural Diversity Compass; (2) policy relevance, according to the EU Rural Vision and other policy documents; (3) applicability at the local level, considering the priorities defined by the Living Labs; (4) coverage by standard statistics and indicators, considering quality and timeliness; (5) technical feasibility and expected accuracy of the prediction model.

4.3.5. Crowdsourcing data and methods

Efforts are underway to review the literature on the impact of crowdsourcing in the rural space. Searching Scopus with the keywords “crowdsourcing AND rural” detects 141 items, of which 67 are articles. Adding the keyword “indicator” drops the number of items to 6, all of which are conference papers. Similar checks of Nature Scientific Data with “crowdsourcing AND rural” detect 7 datasets. Adding the keyword “indicator” reduces the results to 1 dataset. Resulting articles and datasets include themes of biodiversity, rural communities, well-being, air quality and sustainable development.

Next steps will involve analysing the identified studies and datasets, looking for replicable and scaleable datasets and methods that could potentially address the identified indicator needs. In particular we will look at the requirements of the Living Labs, which in part due to their needs for increased granularity, specific local needs and the potential ability of crowdsourcing tools to collect information, lend themselves well to crowdsourcing techniques.

4.4. Data repository

GRANULAR is currently developing a meta-data repository for rural datasets (Figure 6). The GRANULAR repository and digital platform will be linked to <https://www.ruralgranular.eu/> when ready (expected release spring 2023).

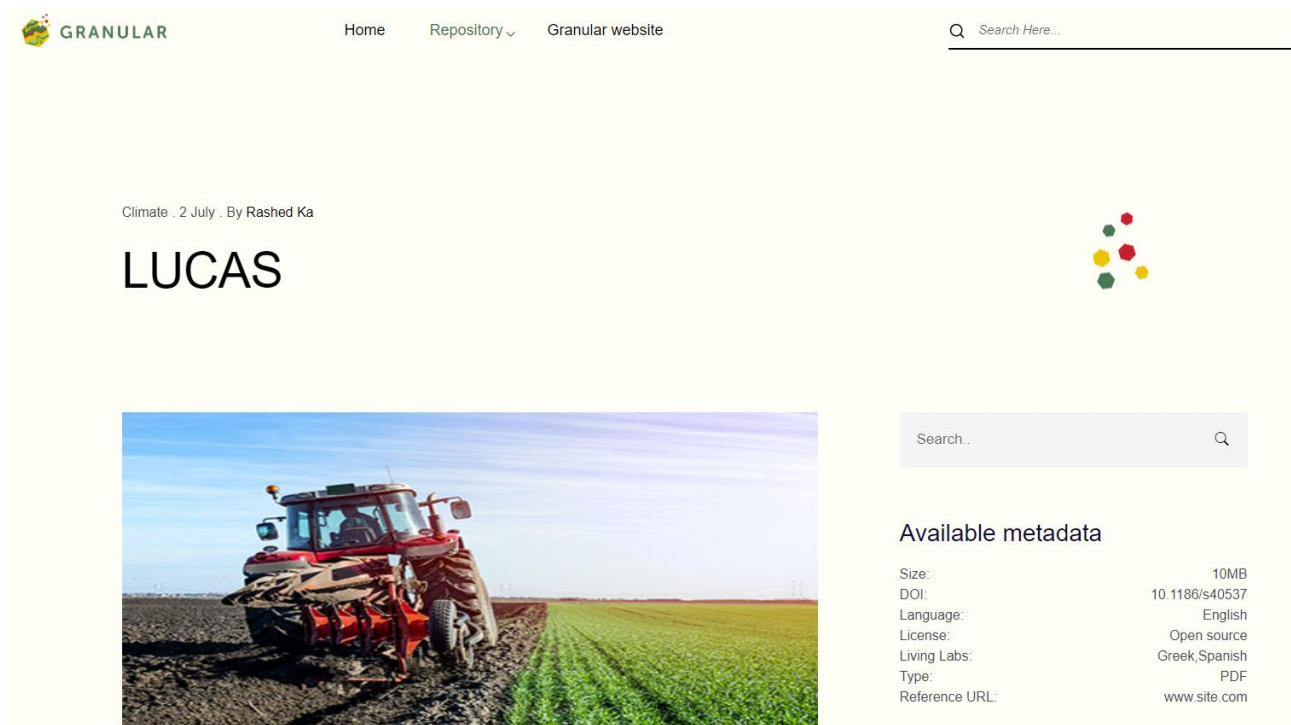


Figure 6. Screenshot of the GRANULAR repository as of March 2023 (under construction).

5. Conclusions

GRANULAR has completed an initial screening of data availability for the generation of new and novel datasets to support indicators of rural sustainability for Europe. Conducting a semi-structured survey and evaluation of more than 90 existing suitable datasets, these datasets, along with accompanying meta-data have been recorded in an online table. The screened datasets address the majority of rural compass indicators, with the majority of datasets representing demography, infrastructure and environment. Most datasets contain a free and open licence or allow partial access, with just a few requiring purchase. This effort is ongoing and a systematic review is planned for publication.

6. References

Allen, C., Smith, M., Rabiee, M. *et al.* (2021). A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals. *Sustain Sci* **16**, 1701–1716. <https://doi.org/10.1007/s11625-021-00982-3>

An, W.; Alarcón, S. (2020). How Can Rural Tourism Be Sustainable? A Systematic Review. *Sustainability* 2020, *12*, 7758. <https://doi.org/10.3390/su12187758>

Andersson, A.; Höjgård S.; Rabinowicz E. (2017). Evaluation of results and adaptation of EU Rural Development Programmes. *Land Use Policy*. 67(298-314). <https://doi.org/10.1016/j.landusepol.2017.05.002>

Hargreaves, P.K. & Watmough, G.R. (2021). Satellite Earth observation to support sustainable rural development. *International Journal of Applied Earth Observation and Geoinformation* 103 e102466. [10.1016/j.jag.2021.102466](https://doi.org/10.1016/j.jag.2021.102466).