GRANULAR

PRACTICE ABSTRACT

Web scraping tourism platforms for rural data

What it delivers, what it costs and when it breaks

Official tourism statistics are published at coarse scales, arrive late and say little about what actually attracts visitors. **GRANULAR tested whether public data from an online tourism platform could fill that gap, using TripAdvisor as the example.**

A Python crawler collected every listed establishment in a study region, with address, coordinates, category and rating, then gathered the visitor reviews attached to each one, including the reviewer's country, month of travel and type of trip.

The results were georeferenced, aggregated to administrative regions and classified into cultural, recreational and nature-based tourism.

The French dataset alone held around 50,000 establishments and almost 2 million reviews. Checked against official figures for rural NUTS2 regions between 2020 and 2022, review activity tracked overnight stays closely, with a correlation of 0.85 and one extra review corresponding to roughly 2,310 overnight stays. The method worked less well where attractions are few or the platform is little used, as in the Finnish regions.

The honest conclusion is that this data is powerful but fragile. A scraper is built for one website and breaks when that website changes, and during the project TripAdvisor added commercial anti-bot protection that stopped the production workflow.

⬢ **Recommendations**: budget for maintaining the collection pipeline, not only for analysing the data; always validate a platform indicator against official statistics before relying on it; expect thin coverage where attractions are few; respect platform terms of service, crawl slowly and cache what you download; and start from the open tutorial and source code the project has published rather than building from scratch.

Authors: M. Hofer, I. McCallum (IIASA)

Funded by the European Union

UK Research and Innovation

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).

Methodological context and rationale

Tourism information is often difficult to obtain for rural areas because official statistics are published at large geographic scales and provide limited detail. Online platforms such as TripAdvisor contain valuable information on attractions and visitor reviews, but these data are not available as downloadable datasets.

To address this, GRANULAR developed **a web-scraping workflow that collects publicly available tourism information from TripAdvisor and converts it into GIS-ready datasets for regional analysis**. This is one of the novel open data sources GRANULAR set out to test, alongside mobility, Earth observation and citizen science data. What it adds is **descriptive detail that official statistics do not carry**: not how many visitors came, but what they came for, when, and how one rural destination differs from another.

Description of the method

A **Python script** automatically visits TripAdvisor pages within a study area and collects information on tourism establishments such as attractions, museums, beaches and tours. For each establishment, the scraper extracts information such as **name, location, category and visitor ratings**. Visitor reviews can also be collected, including their publication date and available travel information.

The collected data are cleaned and exported into **GIS-compatible formats for mapping and further analysis**.

A simplified version of the workflow is available as an open tutorial together with the source code. Advanced features used in the project, such as VPN rotation, domain switching, multilingual crawling and address geocoding, were omitted to keep the tutorial easy to understand, replicable and with fewer dependencies.

Capacity-building activities at local level aim to bridge the gap between data-driven local initiatives — which seek to collect, use or process detailed territorial information — and existing, accessible datasets or innovative methodologies that local users may not yet know or be able to use.

Testing and application in Multi-Actor Labs

The method was applied in **France, Finland and Spain**, where it successfully collected large datasets of tourism establishments and visitor reviews.

The resulting datasets supported **tourism mapping and regional analysis** and were presented during Living Lab activities. Feedback confirmed that the approach provides **useful local tourism information that is often missing from official statistics**. Applying the same workflow in France, Finland and Spain, with no change beyond the definition of the study area, was itself the replicability test. The code transferred directly between the three; what varied was the usefulness of the output, which depends on how intensively the platform is used locally and was weakest in the Finnish regions, where attractions and reviews are fewer.

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).

Lessons learned and practical recommendations

- Web scraping can provide **detailed tourism information** that is otherwise difficult to obtain and can **complement official tourism statistics**. However, the method requires **regular maintenance** because every website has its own structure, and even small changes can stop a scraper from working.
- During the GRANULAR project, TripAdvisor introduced stronger anti-bot protection (DataDome), making automated data collection much more difficult. This illustrates an important **challenge of relying on commercial websites**: access to data can change without warning, and long-term data collection cannot be guaranteed.
- Anyone applying this method should expect to spend **time maintaining the scraping workflow, adapting it to website changes, and checking data quality**. For long-term monitoring, it is worth considering multiple data sources so that analyses are less dependent on a single platform.

Scalability and replicability



The approach can be adapted to other regions and tourism websites, although each website requires its own implementation. **Basic Python knowledge** is needed to modify and run the scraper. The **published tutorial** provides a practical starting point for researchers and practitioners who want to build similar workflows.

By making **visitor patterns visible** to the people managing a destination, the method supports the **digital transition of rural tourism management**, and indirectly its ecological dimension, since seasonality and visitor concentration are the starting point for keeping tourism within local environmental limits.

Further information:

- GRANULAR project: www.ruralgranular.eu
- Tutorial and source code: https://github.com/FlightlessAlbatross/tripadvisor-scraper-tutorial

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).

GRANULAR

CIHEAM MONTPELLIER · tetis · CIHEAM CHANIA · Nordregio · WAGENINGEN UNIVERSITY & RESEARCH · International Institute for Applied Systems Analysis IIASA www.iiasa.ac.at · PAGE

RiATE · cnrs · Université Paris Cité · ECORYS · Universidad de Vigo · Luke Natural Resources Institute Finland · The James Hutton Institute

IN SUPREMÆ DIGNITATIS 1343 · Agricultural University of Athens · CZU Czech University of Life Sciences Prague · aeidl · Università degli Studi di Milano · Politechnika Koszalińska

THÜNEN · PREPARE Partnership for Rural Europe · region västerbotten · P10 · Pays Pyrénées Méditerranée · University of Southampton · Distretto Rurale e Biologico della Val di Cecina

www.ruralgranular.eu