Plan Overview

A Data Management Plan created using DMPTuuli

Title: A trip-based approach to understanding recreational use of nature in Finland using survey and social media data

Creator: Leyi Xu

Principal Investigator: Tuuli Toivonen

Data Manager: Leyi Xu

Contributor: Spencer Wood, Anna Hausmann, Liisa Tyrväinen, Marjo Neuvonen

Affiliation: University of Helsinki

Funder: European Commission

Template: Horizon Europe

ORCID iD: 0000-0002-6625-4922

Project abstract:

People’s recreational use of nature is evolving over time. To support their diverse needs, this study combines traditional survey data with social media data to investigate nature recreation through a trip-based approach. Rather than assuming that individuals belong to fixed visitor categories, this study recognizes that nature recreation is situational and the same individual may participate in different types of trips over time. Specifically, this study (1) identifies distinct types of nature recreation trips based on activities and social composition, (2) investigates the characteristics of these trip types, and (3) compares how different trip types are represented across survey and social media data to assess their implications for recreation monitoring.

ID: 26345

Start date: 01-01-2026

End date: 31-12-2026

Last modified: 19-05-2026

Grant number / URL: 101120250

A trip-based approach to understanding recreational use of nature in Finland using survey and social media data - Initial DMP (a part of proposal text)

Research data management and management of other research outputs

Types of data/research outputs (e.g. experimental, observational, images, text, numerical) and their estimated size; if applicable, combination with, and provenance of, existing data.

Research outputs can be

1. Trip types with textual description, which is derived from automated content analysis of social media data and interpretation of survey data
2. Statistical description of each trip type in numerics, including activity patterns, social composition, and demographic characteristics
3. Spatial and temporal distributions of trip types based on the geographical coordinates and timestamps of social media data

Findability of data/research outputs: Types of persistent and unique identifiers (e.g. digital object identifiers) and trusted repositories that will be used.

Research outputs will be made findable through the use of persistent identifiers and trusted research repositories, where permitted by data access and licensing conditions.

Aggregated and derived datasets (e.g., trip-level variables, activity categories, latent class analysis outputs, and summary statistics) will be deposited in a recognised open research repository such as Zenodo.The repository will assign Digital Object Identifiers (DOIs) to deposited datasets, ensuring long-term findability and citability.

Code used for data processing, modelling, and analysis will be made publicly available via a version-controlled repository such as GitHub, with a DOI generated through Zenodo integration to ensure permanent archiving and citation.

The raw survey data (LVVI3) and original social media media files (YouTube and Flickr content) cannot be openly deposited due to data use agreements, copyright restrictions, and platform terms of service. These datasets remain accessible only within the secure CSC Allas environment and are not assigned public persistent identifiers.

Where direct sharing is restricted, metadata descriptions and documentation of datasets (data dictionaries, variable descriptions, and processing workflows) will be published with DOIs to ensure transparency and reproducibility.

Accessibility of data/research outputs: IPR considerations and timeline for open access (if open access not provided, explain why); provisions for access to restricted data for verification purposes.

The survey data is used under a data use agreement with the data provider and is not publicly shared due to licensing and privacy restrictions. Access to this dataset is restricted to the research team and governed by the terms set by the original data controller.

Social media data (YouTube and Flickr) are collected via official APIs and consist of publicly available content; however, redistribution of raw media files is restricted by platform terms of service and copyright considerations. Therefore, raw videos, images, and associated metadata will not be openly shared.

All derived datasets (e.g., trip-level variables, aggregated statistics, and model outputs) will be archived in a de-identified form where possible and may be shared upon reasonable request for research verification, subject to ethical and legal constraints. Access for verification purposes can be granted under controlled conditions (e.g., data-sharing agreement and non-commercial use restriction), ensuring that no identifiable content or raw media is disclosed.

Research results will be made openly available through peer-reviewed publications and aggregated data tables, supporting transparency while complying with intellectual property and privacy requirements.

Interoperability of data/research outputs: Standards, formats and vocabularies for data and metadata.

All research data and derived outputs will be stored and shared using open, non-proprietary formats to ensure interoperability and long-term usability.

Structured datasets (e.g., trip-level data, activity classifications, social composition variables, and model outputs) will be stored in CSV and/or Parquet formats, with accompanying metadata in JSON where appropriate. These formats are widely used in data science workflows and support reuse across platforms and programming environments.

Geospatial information (e.g., coordinates of trips and recreation areas) will follow standard formats such as GeoJSON or CSV with WGS84 (EPSG:4326) coordinate reference system, ensuring compatibility with GIS software and spatial analysis tools.

Metadata will follow established research data documentation practices, including clear variable naming, data dictionaries, and structured README files.

Code and analysis workflows will be documented in reproducible scripts, with version control to ensure transparency and compatibility across computing environments.

Reusability of data/research outputs: Licenses for data sharing and re-use (e.g. Creative Commons, Open Data Commons); availability of tools/software/models for data generation and validation/interpretation /re-use.

Reusability of research outputs is supported through clear licensing of derived datasets, open sharing of analysis code, and documentation of computational workflows.

Derived and non-restricted datasets (e.g., aggregated trip-level variables, activity categories, social composition summaries, and model outputs) will be shared under a Creative Commons Attribution (CC BY 4.0) licence where possible, allowing reuse with appropriate citation. Where repository requirements differ, equivalent open licences (e.g., CC BY-NC 4.0 for non-commercial restrictions) may be applied in accordance with data source agreements.

Original data sources are subject to restrictions: the LVVI3 survey data are governed by the data provider's licensing terms, and raw social media content from YouTube and Flickr cannot be redistributed due to platform terms of service and copyright limitations. These datasets will therefore not be shared publicly.

To support reuse and reproducibility, all code used for data processing, modelling, and analysis will be made openly available via a public repository (e.g., GitHub), including:

- data preprocessing scripts
- activity classification and aggregation procedures
- latent class analysis models
- validation routines

Curation and storage/preservation costs: person/team responsible for data management and quality assurance.

Data management, curation, and quality assurance are the responsibility of the doctoral researcher Leyi Xu, under the supervision of Prof. Tuuli Toivonen, who serves as the principal investigator.

The research team is responsible for:

- collecting and organising survey and social media datasets
- ensuring data quality through cleaning, validation, and documentation of preprocessing steps
- maintaining metadata for all derived variables (e.g., trip-level datasets, activity categories, and model outputs)
- implementing version control for code and processed datasets to ensure traceability and reproducibility
- ensuring compliance with ethical, legal, and institutional data management requirements throughout the project lifecycle

Data storage and preservation are handled using CSC (IT Center for Science) Allas secure storage services, which are already integrated into institutional research infrastructure. Long-term preservation applies primarily to:

- anonymised or aggregated derived datasets (where sharing is permitted)
- analysis code and documentation

Raw data from YouTube, Flickr, and the LVVI3 survey remain under restricted access conditions and are not publicly archived due to licensing and platform restrictions, but are securely stored for the duration of the project as required for verification and research integrity.