
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

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

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.

A trip-based approach to understanding recreational use of nature in Finland using survey and social media data - Data management plan (after 6 months)

1. Data Summary

Date of the DMP

2026.05.19

DMP version

Version 01

1. Data Summary

This study will re-use the existing data, including the survey data and social media data.

The generated datasets can include trip-level structured datasets, activity categories, social composition classifications, demographic estimates. Formats include CSV, JSON, GeoJSON, and image/video file (JPG/MP4).

The re-used and generated data support the core objective of the project: to identify nature-based recreation trip types using both traditional survey data and digital behavioural traces.

- Survey data provides a structured baseline of recreational behaviour in Finland.
- Social media data provides dynamic, spatially explicit, and behavioural representations of recreation.
- Derived datasets enable integration of both sources to identify trip types and examine their characteristics across populations and platforms

The expected size of the data is as below:

- LVVI3 survey dataset: relatively small (hundreds of respondents, structured tabular data)
- YouTube dataset: ~3,800 videos (including metadata and processed frames)
- Flickr dataset: ~16,000 images (plus metadata; cleaning ongoing)
- Derived datasets and model outputs: moderate-sized structured tables

Overall storage requirement is estimated at several tens of gigabytes, primarily driven by multimedia files.

The origin/provenance of the data is described below:

- Survey data: Natural Resources Institute Finland (Luke), LVVI3 survey (2019–2021)
- YouTube data: publicly available content accessed via YouTube Data API
- Flickr data: publicly available content accessed via Flickr API
- Derived data: generated within the project using computational analysis (LLMs, computer vision models, and statistical modelling)

All social media data originate from user-generated content that has been publicly shared by platform users.

The derived and aggregated datasets may be useful for:

- researchers in nature-based tourism, recreation ecology, and environmental sociology
- land and park management authorities for understanding visitor patterns and demand
- researchers developing computational social science and multimodal AI methods
- policymakers interested in monitoring recreational use of natural areas
- method developers working on human–environment interaction modelling and spatial behaviour analysis

Raw data (social media content and survey microdata) are not intended for external reuse due to licensing, privacy, and platform restrictions; however, aggregated outputs and code may support broader methodological reuse.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Where permitted by licensing and ethical constraints, derived datasets and research outputs will be deposited in trusted repositories that assign persistent identifiers such as Digital Object Identifiers (DOIs). Analysis code archived through GitHub–Zenodo integration will also receive a DOI to support long-term citation and reproducibility. Raw survey data and original social media content will not receive public persistent identifiers within this project due to access restrictions, copyright, and platform terms of service.

Rich metadata and supporting documentation will be created to ensure discoverability, interpretability, and reuse of derived datasets and research outputs. The metadata will include:

- dataset title and description
- creators and affiliations
- temporal and spatial coverage
- data sources and provenance
- variable descriptions and codebooks
- data processing and modelling workflows
- software, model, and version information
- licensing and access conditions

Search keywords will be included to improve discoverability and reuse. Keywords may include terms such as:

- nature-based tourism
- outdoor recreation
- social media data
- human–nature interactions
- multimodal analysis
- YouTube
- Flickr

Metadata deposited in trusted repositories such as Zenodo will be openly accessible and machine-readable, allowing harvesting and indexing through standard repository protocols and scholarly search services. This supports discoverability through search engines, repository aggregators, and research data catalogues.

2.2. Making data accessible

Derived datasets, metadata, and analysis code will be deposited in trusted repositories such as Zenodo, depending on suitability. These repositories assign DOIs and ensure long-term access and resolvability of deposited digital objects. Arrangements with these repositories are standard and supported through institutional research infrastructure.

Not all data will be openly available. Openly shared data will include derived and aggregated datasets, metadata, and code.

Restricted data includes:

- LVV13 survey data (licensing restrictions)
- Raw YouTube and Flickr media (copyright and platform terms)

Open data will be accessible via standard repository download (HTTPS) in formats such as CSV and JSON. Restricted data will remain in CSC secure storage and local computer. Identity of requesters will be verified through institutional affiliation and contact details. No formal data access committee is planned.

Metadata will describe data provenance, variables, processing steps, and access conditions, and will include information needed to locate or request access to the data.

Metadata and DOI records will remain available even if underlying data become restricted or removed.

Documentation (README files, codebooks, and software/version information) will be provided alongside datasets, and analysis code will be shared where possible via GitHub and archived in Zenodo.

2.3. Making data interoperable

The study follows FAIR principles and established community standards to ensure interoperability and reuse.

Data will be stored in open formats such as CSV/Parquet (tabular data), JSON (structured outputs), and GeoJSON with WGS84 (EPSG:4326) for spatial data. Documentation will include README files, codebooks, and data dictionaries to support reuse.

A project-specific activity taxonomy is used but fully documented. It will not constitute a formal ontology, but definitions and mappings to survey categories will be provided to ensure interpretability. No additional ontologies are developed.

Data will include internal identifiers linking trip-level records to source materials (LVV13 survey and social media) and to processing steps (model versions and workflows), ensuring traceability across datasets while respecting platform and licensing constraints.

2.4. Increase data re-use

All datasets will be accompanied by README files, codebooks, and data dictionaries describing variables, units, and data structure. Data processing, modelling, and analysis workflows will be fully documented and version-controlled to ensure reproducibility.

Where possible, derived datasets and code will be openly shared under standard licences. Raw survey and social media data cannot be openly released due to licensing, copyright, and platform restrictions. Open outputs will be shared via trusted repositories with DOI assignment.

Reusable outputs will be provided in open formats (CSV, JSON, GeoJSON) with full documentation and persistent identifiers. Restricted data will remain securely stored in CSC or locally and not publicly released.

Data provenance will be documented through clear source descriptions (LVV13, YouTube, Flickr), dataset versioning, and tracking of all preprocessing and modelling steps.

Quality control includes automated cleaning, manual validation of a subset of data, cross-checks between model outputs and human labels. Data are securely stored in CSC Allas with restricted access. The project prioritises reproducibility, ethical use of public data, and sharing of code and derived outputs where permitted.

3. Other research outputs

3. Other research outputs

This project generates non-data outputs including Python scripts, analysis workflows, and model configurations.

Where possible, code and workflows will be openly shared via GitHub and archived in Zenodo with DOIs, ensuring findability, versioning, and long-term access. Outputs will be documented to enable reuse, including descriptions of preprocessing steps, model settings, and analytical procedures.

4. Allocation of resources

4. Allocation of resources

Costs are mainly indirect and include CSC storage and computing services, DOI-based archiving (Zenodo), and researcher time for documentation, metadata creation, and curation. No major additional infrastructure costs are expected.

Costs are covered by institutional CSC services and Horizon Europe project funding.

Data management is the responsibility of the doctoral researcher under supervision of the Principal Investigator, who ensures compliance with ethical and funder requirements.

Restricted data will be securely stored in CSC Allas during and after the project for verification purposes. Open outputs (derived data, code, metadata) will be archived in Zenodo with DOIs and preserved.

5. Data security

5. Data security

All data will be stored and processed in secure, access-controlled environments. Active datasets are stored in CSC Allas or locally, with restricted access limited to the research team. Data transfer is done via secure institutional channels, and no sensitive data are transferred via unsecured platforms. Regular backups and version control ensure data recovery and integrity.

Derived and shareable outputs (e.g., aggregated datasets, code, and metadata) will be stored in trusted repositories such as Zenodo, ensuring long-term preservation, curation, and DOI-based access. Restricted raw data will remain securely archived in CSC systems but will not be publicly released.

6. Ethics

6. Ethics

Data sharing is limited by (1) YouTube and Flickr terms of service and copyright, (2) LVVI3 survey licensing restrictions, and (3) privacy/re-identification risks from combining geolocation and visual data. Automated facial analysis further requires that only aggregated, non-identifiable results are shared.

The study does not collect primary data or interact with individuals, so informed consent is not applicable. Data use is based on public availability (social media) and existing licensing agreements (survey data), not participant consent.

7. Other issues

7. Other issues

The project will follow the data management procedures and guidelines of:

- CSC – IT Center for Science (Finland): use of CSC Allas secure storage and computing environment for data processing, storage, backup, and controlled access.
- University of Helsinki research data management guidelines: including requirements for data protection, ethical review compliance, and secure handling of research data and personal data where applicable.
- Luke (Natural Resources Institute Finland) LVVI3 data use agreement: governing access, processing, and non-disclosure of survey microdata.
- EU Horizon Europe / FAIR principles: ensuring open science compliance, including FAIR data management, metadata standards, and open access to research outputs where legally possible.