Data Management Plan for Use Cases of the GeoKur Project

on suitability of global land use data to assess relationships between land use, degradation, pollination and human migration

Authors: Lukas Egli[1], Julia Fischer[1], Juliane Groth[1], Stefano Della Chiesa[2], Christin Henzen[3]
(ordered by institution and alphabetically)
Affiliations: [1]Helmholtz Centre for Environmental Research, Leipzig, [2]Leibniz Institute of Ecological Urban and Regional Development and previously Technische Universität Dresden, [3]Chair of Geoinformatics, Technische Universität Dresden

Publication date: February 2022

Table of Contents

Executive Summary

The BMBF project GeoKur aims to support the curation and quality assurance of Earth System Science (ESS) data sets, focusing on the suitability of geospatial time-series of global land use data by analysing human-environment relations such as land degradation, biodiversity, human migration and ecosystem services. This DMP describes two use cases of the GeoKur project. The use cases use various existing and publicly available datasets of land-use, net-migration, crop yield, etc., to showcase best practices to determine their fitness for use. The data will be used to 1) identify spatial patterns of land degradation processes and in-migration in Sub-Saharan Africa between 2000 and 2015 and to 2) investigate the effects of agricultural management and pollination-related variables on crop-specific yields.
Within the project two partners collaborate on the two use cases: a team of researchers collects, analyses and provides data, scripts and related publications and a team of data stewards and software engineers provides discipline-specific guidance, adapted tools, and develops specific methods and tools based on the researcher needs. Thus, **this data management plan (DMP) strongly focusses metadata, software, and technical aspects for ESS projects, instead of describing common RDM practice and methods**. Moreover, it serves as example to develop ESS discipline-specific guidance and tools for data management.
This DMP follows the Science Europe Template. The two use cases are compliant with the Principles for the Responsible Handling of Research Data at the UFZ. The present principles are based on the Guidelines of the Helmholtz Association on the Management of Research Data, on the Guidance of the European Commission on Data Management according to the FAIR Principles and the Deutsche Forschungsgemeinschaft (DFG) Research Data Guidelines and Guidelines for Safeguarding Good Research Practice.
The project uses publicly available geospatial datasets and provides produced datasets as open data, compliant with the FAIR (Findability, Accessibility, Interoperability, Re-usability) principles. During the project, datasets will be managed data management system (DMS) implemented as open source catalogue CKAN with spatial extensions facilitating direct metadata and data access via an Application Programming Interface (API). For long-term storage, selected results will be stored in the institutional data management system, called UFZ Data Management Platform[1] including raw data after the project ends, resp. published on the Earth & Environmental Science data repository PANGAEA without raw data. The researchers develop data preparation and analysis scripts using the language R. Scripts will be managed on GitHub and published via Zenodo following reproducible research approaches by including links to the well-documented open-source GitHub repository and used datasets to ensure reproducibility of the applied approaches. The data management of the project focuses on discipline-specific provenance and quality tracking for produced and collected datasets and documentation. Therefore, all datasets will be described using a project-specific geospatial extension for the Data Catalog Vocabulary (GeoDCAT) metadata profile with linked provenance ontology (PROV-O) and Data Quality Vocabulary (DQV). Metadata in the GeoDCAT format will therefore be automatically extracted or tracked by GeoKur-specific tools or extended by manual processes, when extraction or tracking is not possible. A specific quality register facilitates the curated management of quality measure descriptions by linking from dataset metadata to the descriptions. Thus, a specific quality assurance workflow is developed, e.g. managing use case-specific quality measures and activities.

[1] The data management is restricted for UFZ intern access, the linked data investigation portal is available here: https://www.ufz.de/drp/de/index.php?drp_data[mvc]=

1) Data description and collection or re-use of existing data

1a) How will new data be collected or produced and/or how will existing data be re-used?

Guidance:
Explain which methodologies or software will be used if new data are collected or produced.
State any constraints on re-use of existing data if there are any.
Explain how data provenance will be documented.
Briefly state the reasons if the re-use of any existing data sources has been considered but discarded.

This project collects various publicly available datasets and generates new outputs throughout data analysis to highlight best practices determining their fitness for use and addressing information needs. Data analysis including preprocessing steps to overcome the heterogeneity, e.g. in spatial or temporal resolution, coordinate reference systems (CRS), units, data type, will be implemented in R and published as open-source on GitHub.
To facilitate comparing datasets and evaluating their fitness for use, metadata for existing datasets will be enriched by manually evaluating publications, reports and given websites as well as using specific open-source tools, e.g. for quality extraction and for provenance tracking[2]. The purpose of enriching the metadata is to develop a reusable metadata profile focusing on extended quality and provenance information for environmental data by using community standards (see Section 1c).
Existing datasets are discarded for the following reasons: missing or limiting usage license, poor or missing provenance and/or quality information and/or inadequate spatial or temporal extent, spatial, temporal, or thematic resolution, positional or thematic accuracy, or completeness.

1b) What data (for example the kind, formats, and volumes), will be collected or produced?

Guidance:
Give details on the kind of data: for example, numeric, (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media.
Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example pdf, xls, doc, txt, or rdf).
Justify the use of certain formats. For example, decisions may be based on staff expertise within the host organization, a preference for open formats, standards accepted by data repositories, widespread usage.
Give preference to open and standard formats as they facilitate sharing and long-term re-use of data (several repositories provide lists of such 'preferred formats').
Give details on the volumes (they can be expressed in storage space required (bytes), and/or in numbers of objects, files, rows, and columns).

Collected datasets include time-series geographic layer as raster or vector, and tabular data and are provided as GeoPackage, GeoTIFF, TIFF (with TFW), ESRI Shapefile or CSV file (Table 1). The proprietary and semi-open data format ESRI Shapefile is commonly accepted in the ESS community and manageable by several open-source geoinformation systems.
Geographic layers and tabular data contain categorical (land use, land cover, etc.), numerical (Averages, Percentages, etc.) or statistical outputs (regression equation, RMSE, significance, etc.).

[2] https://github.com/GeoinformationSystems/ProvIt, https://github.com/GeoinformationSystems/R2ProvO

Table 1: Collected datasets

Name	DOI or Link
Africa - Admin Level 0	http://geoportal.icpac.net/layers/geonode%3Aafr_g2014_2013_0
Database of Global Administrative Areas (GADM)	https://gadm.org/
MapSPAM Datasets: Yield, Production, Harvested Area, Physical Area	https://doi.org/10.7910/DVN/PRFF8V
Global Synergy Cropland Map, v3	https://doi.org/10.7910/DVN/ZWSFAA
Gridded Population of the World, v 4	https://doi.org/10.7927/H4F47M65
Global Map of Irrigation Areas Version 5, v 5.0	https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/f79213a0-88fd-11da-a88f-000d939bc5d8
World Database on Protected Areas, v 1.6	https://www.protectedplanet.net/en/thematic-areas/wdpa?tab=WDPA
Producer Prices (FAOSTAT)	https://www.fao.org/faostat/en/#data/PP
Harvested Area and Yield for 175 Crops	https://doi.org/10.1029/2007GB002947
Global data set of Monthly Irrigated and Rainfed Crop Areas around the year 2000 (MIRCA2000), version 1.1	https://doi.org/10.1029/2008GB003435
Visitation probability of pollinators	https://doi.org/10.1016/j.ecolind.2013.07.014
Pollinator abundance Europe	https://doi.org/10.5061/dryad.6tj407n
Pollinator abundance Europe (PREDICTS)- predicts-database	https://data.nhm.ac.uk/dataset/the-2016-release-of-the
Bumblebee abundance Europe	https://www.science.org/doi/10.1126/science.aaa7031
JRC Net Migration Grids	https://migration-demography-tools.jrc.ec.europa.eu/data-hub/
Global Forest Change 2000-2012	https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/download.html
Soil salinity dynamics	https://data.mendeley.com/datasets/v9mgbmtnf2/1
NPP MODIS (Gap Filled)	https://search.earthdata.nasa.gov/search?q=C1631984056-LPDAAC_ECS
Global Soil Erosion map	https://esdac.jrc.ec.europa.eu/content/global-soil-erosion

Created datasets will be released using open file formats (CSV, GeoTIFF) with an unknown size (Table 2). The overall file size is estimated as less than 1TB.

Table 2: Potentially produced datasets

Name	Short description
Average migration per hotspot	Geospatial dataset
Pixel-based degradation process per hotspot	Geospatial dataset

Hotspot joined data (incl. migration, land degradation processes)	Tabular dataset: Average positive net-migration and area (km2 and % of total hotspot area) affected by specific degradation process per hotspot
Extracted land-use and pollination variables	Tabular dataset: extraction of the relevant variables of the collected datasets for each pixel (one row represents one pixel)
Multiple regression output	Tabular data: effect of pollination and irrigation onf crop-specific yields

2) Documentation and data quality

2a) What metadata and documentation (for example the methodology of data collection and way of organizing data) will accompany the data?

Guidance:

Indicate which metadata will be provided to help others identify and discover the data.
Indicate which metadata standards (for example DDI, TEI, EML, MARC, CMDI) will be used.
Use community metadata standards where these are in place.
Indicate how the data will be organized during the project mentioning, for example, conventions, version control, and folder structures. Consistent, well-ordered research data will be easier to find, understand, and re-use.
Consider what other documentation is needed to enable re-use. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, and so on.
Consider how this information will be captured and where it will be recorded (for example in a database with links to each item, a 'readme' text file, file headers, code books, or lab notebooks).

Focusing a mostly automated and tool-supported RDM in ESS projects, the GeoKur project provides a specific understanding of data documentation and data quality and general documentation and data quality aspects are not covered here.
Data documentation will be provided via structured metadata, scientific publications and documented source code. Additional readme or guidance documents will not be provided. For collected datasets existing metadata will be enriched by manual information gathering and automated extraction (see Section 1a). For produced datasets, several approaches and tools to track metadata directly during the data creation phase will be used and implemented, in particular for quality or provenance tracking. Several administrative, descriptive, structural and technical information will be collected by using a project-specific GeoKur metadata profile based on GeoDCAT, PROV-O and the Data Quality Vocabulary (DQV). Aim of the GeoKur profile is to provide extended quality and provenance information, reduce the number of required fields for general information, e.g. abstract or provider's contact information, and focus on reusing existing information by linking them, e.g. linking to a data provider profile via orcID instead of including all characteristics of the provider in the GeoKur profile. Thus, the profile also facilitates collecting and referencing existing documentations, e.g. websites, reports or other documentations for the description of input data and referencing related scientific publications for both inputs and created datasets. Furthermore, datasets are described using controlled vocabularies or ontologies for thematic categories, e.g. Agrovoc, GEMET or ESA.
A project-specific data management system (DMS) - a modified open-source CKAN - is used to manage collected and created datasets and their metadata as well as separated (linked) metadata entries for processes and workflows, which summarize dataset and process descriptions for a provenance graph.

Within CKAN, datasets can be structured hierarchically by linking parent and child metadata, organized as provenance graph by linking input and output datasets for a certain process, or versioned by linking metadata to previous version's metadata.
Process metadata refer to a GitHub script repository, if available. The repository contains the source code, a short description, installation notes, and license and contact information. Final versions of the scripts will be published on Zenodo. Detailed information about research aspects and the developed analysis will be provided as scientific publication referencing data and code repositories.
The DMS includes a central quality register for managing quality measure descriptions. It enables data providers to add (specific) quality measures and reuse the descriptions in several dataset metadata.
Metadata managed in the CKAN can be accessed via Web-based CKAN user interface or via CKAN APIs. By using the ckanr library the researchers can directly gather and publish metadata from their analysis scripts in R to the CKAN (see Section 2b).

2b) What data quality control measures will be used?

Guidance:
Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardized data capture, data entry validation, peer review of data, or representation with controlled vocabularies

A project-specific quality assurance process implementing standard roles, like data provider, data curator and data publisher, will be applied to all datasets and will extend typical internal reviews. For produced datasets, the QA process includes activities and measures for each phase of the data lifecycle. Data quality and data maturity are frequently monitored and evaluated by using a developed checklist managed in the software Research Data Management Organiser (RDMO). FAIRness and Openness aspects are predominantly taken into account during data collection and publication.
Moreover, collected datasets are described with different metadata sets in the original repository or website. Metadata extraction tools, like MetadataFromGeodata, and manual processes, scanning publications, reports and website, will therefore be applied to enrich existing meta information and re-publish them in the GeoKur DMS, enabling comparison by using the GeoKur metadata profile.
Created datasets will be described with metadata from different sources. Consistency in terminology and definitions will be achieved by using controlled vocabularies, e.g. for spatial reference systems.
Provenance information will be tracked when running the developed data processing or analyzing scripts using specific provenance tracking tools or by using the R package ckanr. Quality information will be captured during data processing or analysing by using the modified CKAN API or will be directly extracted from the dataset files using the MetadataFromGeodata tool.
For adding quality measure descriptions to the central quality register a two-step process is defined, entering the descriptions first, and reviewing and enable publishing by a data steward afterwards.

3) Storage and backup during research process

3a) How will data and metadata be stored and backed up during the research?

Guidance:

Describe where the data will be stored and backed up during research activities and how often the backup will be performed. It is recommended to store data in least at two separate locations.
Give preference to the use of robust, managed storage with automatic backup, such as provided by IT support services of the home institution. Storing data on laptops, stand-alone hard drives, or external storage devices such as USB sticks is not recommended.

Several datasets with different maturity levels will be created or retrieved from existing sources during the project. Hence, datasets with low maturity (interim analysis, interim results, try and error tests, etc.) will be stored on an institutional server with weekly backup on an independent server. Datasets with a higher maturity level in terms of metadata, quality and provenance information will be stored in the GeoKur DMS, which is managed on a virtual machine, hosted by the center for information services and high performance computing with institutional backup and mirroring strategies. Both partner institutions UFZ and TUD provide professional data management infrastructure components and backup strategies.
Final data results will be published in long-term archives with a proper backup strategy (see Section 5a). In addition, researchers will upload their scripts regularly to a GitHub repository.

3b) How will data security and protection of sensitive data be taken care of during the research?

Guidance:
Explain how the data will be recovered in the event of an incident.
Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships.

The researchers will not use or generate sensitive data in this project. The institutional IT services will provide the latest backup/snapshot. In case of problems with the backup, collected datasets can be re-published, using the original source files and created datasets can be re-created by running the script again. The researchers will not have access to the virtual machines, hosting the DMS. Additional cloud storage for data sharing is not necessary.
The GeoKur DMS will be publicly available from the project beginning with an opportunity to manage datasets and metadata for private access only. Each researcher has a specific DMS account structured along the institution to facilitate private data access and data access restricted to institutional colleagues. TUD as DMS hosting institution is responsible to manage all accounts. During the development and analysis phases of the project, access to the data results and scripts will be restricted to the project partners and published afterwards (see Section 5a).

4) Legal and Ethical Requirements, Codes of Conduct

4a) If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Guidance:
Ensure that when dealing with personal data, data protection laws (for example GDPR) are complied with:
Gain informed consent for preservation and/or sharing of personal data.

Consider anonymization of personal data for preservation and/or sharing (truly anonymous data are no longer considered personal data).
Consider pseudonymization of personal data (the main difference with anonymization is that pseudonymization is reversible).
Consider encryption which is seen as a special case of pseudonymization (the encryption key must be stored separately from the data, for instance by a trusted third party).
Explain whether there is a managed access procedure in place for authorized users of personal data.

Personal data are not used and hence, data protection laws do not apply for the project.

4b) How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Guidance:
Explain who will be the owner of the data, meaning who will have the rights to control access:
Explain what access conditions will apply to the data? Will the data be openly accessible, or will here be access restrictions? In the latter case, which? Consider the use of data access and re-use licenses.
Make sure to cover these matters of rights to control access to data for multi-partner projects and multiple data owners, in the consortium agreement.
Indicate whether intellectual property rights (for example Database Directive, sui generis rights) are affected. If so, explain which and how will they be dealt with.
Indicate whether there are any restrictions on the re-use of third-party data.

The datasets and scripts will be produced and owned by researchers from the Helmholtz Centre for Environmental Research. Data will be published open access after submitting related scientific publications. Whenever possible, data will be licensed under the Creative Commons CC BY 4.0 International license. Different data licenses will be applied, if used existing datasets are published under a license with constraints for reusing and publishing. The relevant scripts will be licensed under the GPL 3.0.
Intellectual property rights are not affected. There are no restrictions on the re-use of third-party data.

4c) What ethical issues and codes of conduct are there, and how will they be taken into account?

Guidance:
Consider whether ethical issues can affect how data are stored and transferred, who can see or use them, and how long they are kept. Demonstrate awareness of these aspects and respective planning.
Follow the national and international codes of conducts and institutional ethical guidances, and check if ethical review (for example by an ethics committee) is required for data collection in the research project.

Ethical issues and codes of conduct are not relevant for this project.

5) Data Sharing and LT Preservation

5a) How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Guidance:

Explain how the data will be discoverable and shared (for example by deposit in a trustworthy data repository, indexed in a catalogue, use of a secure data service, direct handling of data requests, or use of another mechanism).
Outline the plan for data preservation and give information on how long the data will be retained.
Explain when the data will be made available. Indicate the expected timely release. Explain whether exclusive use of the data will be claimed and if so, why and for how long. Indicate whether data sharing will be postponed or restricted for example to publish, protect intellectual property, or seek patents.
Indicate who will be able to use the data. If it is necessary to restrict access to certain communities or to apply a data sharing agreement, explain how and why. Explain what action will be taken to overcome or to minimize restrictions.

During the project, datasets will be managed and shared via GeoKur DMS, e.g. as use case for tool developments (see Section 3). Selected outputs will be published in relevant long-term archives, like PANGAEA.
There will be no embargo period. The data and scripts will be freely available for the public audience under the given license.

5b) How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Guidance:
Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes.
Indicate how it will be decided what data to keep. Describe the data to be preserved long-term.
Explain the foreseeable research uses (and/ or users) for the data.
Indicate where the data will be deposited. If no established repository is proposed, demonstrate in the DMP that the data can be curated effectively beyond the lifetime of the grant. It is recommended to demonstrate that the repositories policies and procedures (including any metadata standards, and costs involved) have been checked.

In the project, no dataset has to be retained or destroyed. UFZ as data producer decides about data to keep with respect to publications, to fitness for use in related or follow-up projects. After the project ends, relevant datasets, e.g. raw formats that are not published in the PANGAEA archive, will be managed in the institutional DMS, the UFZ Data Investigation Portal, providing discipline-specific search, filter and visualization functionality. Whenever possible, relevant results will be published via specific web services hosted at the UFZ, e.g. Open Geospatial Consortium Web Map Service, to enable interoperability and machine-actionable usage using discipline-specific interfaces that are typically not provided by the archives.
R scripts for data preparation and analysis will be made available on GitHub and persistently stored and published on Zenodo.
The resulting datasets can be used for similar/further spatial analyses. In particular, datasets with enriched metadata on provenance and data quality can be better evaluated regarding fitness for use and thus, being relevant for the ESS community and applications beyond this project. Developed concepts on fitness for use evaluation can be applied for other projects. Moreover, the developed workflows can be transferred to other world regions for similar research aims.

5c) What methods or software tools are needed to access and use data?

Guidance:
Indicate whether potential users need specific tools to access and (re-)use the data. Consider the sustainability of software needed for accessing the data.

Indicate whether data will be shared via a repository requests handled directly, or whether another mechanism will be used?

The collected and created datasets will use open or semi-open formats, which can be accessed for local editing via editors or geoinformation systems or with well-known API / interfaces (see Section 2). Scripts will be implemented in R with the open-source software RStudio and can be access with R studio or other R compatible editors.

GeoKur DMS and relevant archives provide direct download access and specific APIs for programmatic use of the datasets.

5d) How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Guidance:
Explain how the data might be re-used in other contexts. Persistent identifiers (PIDs) should be applied so that data can be reliably and efficiently located and referred to. PIDs also help to track citations and re-use.
Indicate whether a PID for the data will be pursued. Typically, a trustworthy, long-term repository will provide a persistent identifier.

Persistent identifiers will be provided for datasets by long-term repositories, e.g. PANGAEA, Zenodo.

6) Data Management Responsibilities and Resources

6a) Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Guidance:
Who (for example role position and institution) will be responsible for data management (i.e. the data steward)?

The UFZ researchers as dataset producers will be responsible for data management during the project. They will be guided and supported by the UFZ Research Data Management team, in particular for quality assurance and curation aspects, by the UFZ IT services, for technical support and institutional storage and backup, and by the GeoKur TUD project team, in particular for metadata and quality assurance aspects.

6b) What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Guidance:
What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Reusable)?

The development of RDM best practices and their application is part of the deliverables of the GeoKur project. Therefore, there are no additional costs for the data management and for making the data FAIR. Furthermore, no additional costs are expected for storage and backup during the project execution and will be eventually covered internally by the institution. The GeoKur infrastructure will

no longer be supported after the project's end and related software developer and administrators cannot manage the DMS and other infrastructure components. However, relevant data will be managed in institutional DMS or long-term archives and supported by the responsible parties (see Section 3).