**I-GUIDE DATA CARD**

The I-GUIDE Data Card is an easy-to-use tool that will allow you to create documentation for each dataset that you create or use in a project.

Using this tool will help facilitate transparency and reproducibility of your project. It will also help you comply with data management and sharing policies of journals, funding agencies, and universities.

The Data Card applies to:

1. **Secondary Datasets**: Data sourced from external repositories or other researchers;
2. **Primary Datasets**:
   * Data collected through experiments, fieldwork, or user-generated sources;
   * Data obtained via web scraping, API collection, or similar automated means.

**Data Card Attribution**

This Data Card template is an adapted version of materials originally developed by Google’s *Data Cards Playbook*, available at https://pair-code.github.io/datacardsplaybook/.  
Adaptations have been made to support the I-GUIDE platform, with a focus on simplifying structure, tailoring prompts for geospatial data, and aligning with ethical and FAIR data practices in scientific research.

This work is shared under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).  
License details: <https://creativecommons.org/licenses/by-sa/4.0/>

A logo with text on it

AI-generated content may be incorrect.

* + 1. **BASIC INFORMATION**

|  |  |
| --- | --- |
| Data Card ID Number | DC-02 |
| Dataset Name | Alpha Earth Embeddings |
| Dataset Version | v1\_annual |
| Persistent Identifier | ee.ImageCollection("GOOGLE\_SATELLITE\_EMBEDDING\_V1\_ANNUAL") |
| Outputs Supported | DeepMind blog: “AlphaEarth Foundations helps map our planet in unprecedented detail.” (2025) |
| Data Card Author | IGUIDE- Team 1: Jennifer Marlon, Deepika Pingali, Surabhi Upadhyay, Emine Senkardesler, Pratyush Tripathy, Okikiola Michael Alegbeleye |

* + 1. **DATASET OVERVIEW**

|  |  |
| --- | --- |
| Dataset Owners and Publishers | Google DeepMind & Google Earth Engine |
| Source and Acquisition Method | ☑ *Secondary dataset (from external source)* |
| Terms of Use, or Data Sharing Agreement | Creative Commons Attribution Non Commercial 4.0 International (CC BY-NC 4.0) |
| Storage Location | ☑ Repository: (Earth Engine Data Catalog: <https://developers.google.com/earth-engine/datasets/catalog>) |
| Access Control Policies | ☑ Open |

1. **DATASET CHARACTERISTICS**

|  |  |
| --- | --- |
| Data Subjects | ☑ Object (Spectral, temporal, contextual features of Earth’s surface)  ☑ Places (Global) |
| Dataset Size | 2017–2025 (~1.4 trillion pixels) |
| Spatial Data | ☑ Yes  If “Yes”:  Coordinate Reference System (CRS): Native UTM  Spatial Resolution: (10m*)*  Temporal Resolution: *(2017 - 2024)* |
| Data Modality | ☑ Image (Raster with 64 bands)  ☐ Time series |
| Variables | *(Complete for each variable. Table for recording this information is appended at the end of this document)*  Variable name: Bands  Brief description: B0 - B63 (Learned embedding dimensions encoding surface context |

1. **PROVENANCE**

|  |  |
| --- | --- |
| Methods of Collection | ☐ Sensor-based |
| More About Methods | Alpha Earth embeddings are generated using deep neural networks trained on multispectral satellite imagery to produce high-dimensional latent vectors representing land surface features (e.g., vegetation, urban structure, water bodies) at each pixel. These embeddings are context-rich but not human-interpretable at the band level. |
| Tools and Libraries Used | Google Earth Engine |
| Collection Policies (if data collected using web scraping or other digital methods) | Processed using Google Earth Engine |

1. **SENSITIVE DATA**

|  |  |
| --- | --- |
| Human Subject Identifiability | ☑ None  ☐ Personally identifiable information  ☐ Pseudonymous data  ☐ Anonymous data |
| Other Sensitivity Factors | ☐ Commercially sensitive data  ☐ Health data  ☐ Data about children  ☐ Data about marginalized group: *(Specify)*  ☐ Location sensitive data  ☐ Military or security related data  ☐ Restricted government data  ☐ Surveillance data |
| Measures Taken to Handle Sensitive Data | *(Describe de-identification, anonymization, encryption, and/or access restrictions)* |
| Demographic Variables Represented in Dataset | ☐ Age  ☐ Culture  ☐ Disability status  ☐ Ethnicity  ☐ Gender  ☐ Language  ☐ Nationality  ☐ Race  ☐ Socio-economic status  ☐ Other: *(Specify)* |
| Correlation with Demographic Variables | Embeddings may indirectly correlate with socioeconomic status (e.g., via urbanization patterns) but are not explicitly demographic. |
| Dataset Representativeness | Global coverage with consistent spatial resolution. |
| Information About Ethical Oversight | ☑ Not subject to Institutional Review Board (IRB) approval: This dataset is derived from publicly available data with no human subjects involved.  ☐ Subject to Institutional Review Board (IRB) approval  Name of IRB:  Link to IRB website: *(URL)*  Primary IRB contact: *(Name, Email)*  Approval dates: *(From, To)*  Reference number: |
| Informed Consent Processes | (*Describe how consent was obtained, key elements covered, and extra measures taken to facilitate consent (e.g., with special populations). If waived, note the reason)* |

1. **TRANSFORMATIONS**

|  |  |
| --- | --- |
| Transformations Applied | ☐ Anomaly detection  ☐ Cleaning mismatched values  ☐ Cleaning missing values  ☐ Converting data types  ☐ Data aggregation  ☐ Dimensionality reduction  ☐ Joining input sources  ☐ Redaction or anonymization  ☐ Other: *(Specify)* |
| Description of Transformations | *(Complete one version for each transformation)*  Transformation applied:  Field(s) transformed: *(List affected variables)*  Reason for transformation:  Who carried out transformation:  Methods applied:  Platform, tool, or libraries used (including link): |

1. **SUITABLE AND UNSUITABLE USES OF DATASET**

|  |  |
| --- | --- |
| Suitable Uses | * Earth observation research * AI model inputs for spatial prediction tasks |
| Unsuitable Uses | * **Direct interpretation of embedding dimensions as real-world physical variables** |

1. **ANNOTATION TASKS (only complete if dataset includes labeled data)**

|  |  |
| --- | --- |
| Types of Annotation Performed | ☐ Annotation target in data: *(Specify what was being labeled)*  ☐ Crowdsourced  ☐ Human (expert)  ☐ Human (non-expert)  ☐ Machine-generated *(Describe how system generated labels)*  ☐ Other: *(Specify)* |
| Description of Annotations | *(Complete one version for type of annotation)*  Number of unique annotations: *(Total distinct labels/categories)*  Total annotations:  Platforms, tools, or libraries (include link):  Task description*)*  Methods used:  Inter-rater adjudication policy: |

1. **APPLICATIONS AND BENCHMARKS (only complete if dataset used for AI purposes)**

|  |  |
| --- | --- |
| Relevant AI Model(s) | *(List models trained, tested, or validated on this dataset, or for which this dataset has served as input data)* |
| Use in AI | ☐ Training  ☐ Testing  ☐ Validation  ☐ Fine-tuning |
| Key AI Tasks | *(Describe purposes of the AI application, e.g., forecasting)* |
| Evaluation Results | *(Provide accuracy, precision-recall, F1-score, or other performance benchmarks)* |

**APPENDIX: VARIABLES (see “Variables” field in Section 2, “Dataset Characteristics”; add extra rows if necessary)**

|  |  |
| --- | --- |
| **VARIABLE NAME** | **BRIEF DESCRIPTION** |
| **B0 – B63** | Learned embedding dimensions encoding surface context (Unitless) |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |