**INTRODUCTING THE I-GUIDE DATA ETHICS TOOLKIT (I-DET)**

The I-GUIDE Data Ethics Toolkit (I-DET) comprises four tools, developed to support reproducible and ethical geospatial data science. They are practical and lightweight, providing a structured way to document key actions and ensure alignment with open science expectations, and funder and legal requirements. All four tools work together, and each serves a distinct purpose:

1. **I-GUIDE RESEARCH PRODUCT MANAGEMENT PLAN (RPMP)**

The RPMP helps researchers plan how to reach target audiences. It also helps in planning acquisition, production, and management of datasets and models. RPMP helps researchers keep track of which datasets and models are used to support different publications. It supports reproducibility and compliance with funding agencies’ data sharing and open science policies.

1. **I-GUIDE MANAGING ETHICAL GEOSPATIAL AI AND DATA SCIENCE (MEGAID) TOOL**

The MEG-AID tool is a project management tool designed to foster integration of ethical considerations into geospatial data science projects. By documenting actions at each stage of a project’s lifecycle, the tool promotes reproducibility and ethical practice. This toolkit should be used at the start of a project to anticipate ethical issues, and to plan how to address these issues.

1. **I-GUIDE DATA CARDS**

Adapted from Google’s *Data Playcards* (CC BY-SA 4.0), the I-GUIDE Data Cards have been simplified and tailored to geospatial research. They document the datasets used, and actions taken to address key ethical issues.

1. **I-GUIDE MODEL CARDS**

I-GUIDE Model Cards are an adaptation of the I-GUIDE Data Cards, structured to support model documentation. They capture information about a model’s architecture, purpose, data, validation, ethical issues, and interpretability. Each model used in a project should have a Model Card.

A glossary defining key terms is available, along with concise primers on the following topics:

* **EU AI Act;**
* **EU General Data Protection Regulation;**
* **Fair Inclusion in Geospatial Data Science;**
* **Informed Consent in Geospatial Science.**

**LICENSING AND ATTRIBUTION**

All four tools, as well as the primers and glossary, are licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).  
More information: <https://creativecommons.org/licenses/by-sa/4.0/>

A logo with text on it

AI-generated content may be incorrect.

**DEVELOPING THE TOOLKIT**

This toolkit was developed following surveys, in-depth interviews, and long-term observation, of I-GUIDE platform users. This research provided detailed examples of potential ethical dilemmas that arise during platform use. This research helped translate theoretical frameworks into practical, actionable guidelines tailored to the needs of geospatial data science researchers.

Drawing on this research, two types of frameworks were synthesized and tailored to I-GUIDE. One is AI lifecycle models (Figure 1), e.g., the International Organization for Standards and International Electrotechnical Commission’s Artificial Intelligence Data Lifecycle (ISO/IEC 8183:2023). The other is research data lifecycle models (Figure 2),e.g.,the Digital Curation Center’s lifecycle model (DCC 2007) and the Capability Maturity Model for Scientific Data Management (Crowston & Qin 2011).

A diagram of a process

AI-generated content may be incorrect.*Figure 1: ISO/IEC 8183:2023 Artificial Intelligence Data Lifecycle*

A diagram of a process

AI-generated content may be incorrect.

*Figure 2: Digital Curation Centre Data Lifecycle Model (DCC 2007)*

ISO/IEC 8183:2023 is a framework for managing artificial intelligence (AI) projects, providing a structured approach to defining, building, testing, deploying, and maintaining AI systems. Data lifecycle models assist in data management, from initial generation and collection, via processing, storage, retrieval, analysis, dissemination, and ultimately to archiving or destruction.

# **I-GUIDE DATA SCIENCE LIFECYCLE MODEL**

This section presents the stages of the I-GUIDE Data Science Lifecycle Model (Figure 3).

**A diagram of a model

AI-generated content may be incorrect.**

*Figure 3: ISO/IEC 8183:2023 Artificial Intelligence Data Lifecycle*

1. **DEFINING THE RESEARCH PROBLEM**

**Example Ethical Considerations**:

* **Bias/Discrimination**: Are diverse community perspectives included in defining the problem?
* **Transparency**: Is the problem definition process documented and shared with stakeholders?

1. **DATA ACQUISITION AND CREATION**

**2a. Acquiring Secondary Datasets**

**Objective**: Identify and acquire datasets from sources such as open research repositories databases or government data portals.

**Example Ethical Considerations**:

* **Policy Compliance**: Are the datasets being used in accordance with their licensing terms?
* **Fairness**: Are all regions or communities equally represented in the training data?

**2b. Collecting Primary Non-Human Subjects’ Data**

**Objective**: Collect datasets about physical phenomena, whether natural or human-made

**Example Ethical Consideration**:

* **Potential misuse of data:** Could datasets be used to exploit natural resources or compromise national security?

**2c. Collecting Primary Human Subjects’ Data**

**Objective**: Collect data about people using methods such as interviews or surveys

**Example Ethical Considerations**:

* **Consent**: Do the data subjects fully understand the risks and benefits of participation, and do they have the opportunity to withdraw at any stage?
* **Fairness:** Are various demographic groups well-represented?

**2d. Collecting Primary Datasets from Web Scraping or other Digital Activities**

**Objective**: Collect digital datasets using methods such as scraping from websites and social media

**Example Ethical Consideration**:

* **Ethical oversight:** Does the study require oversight from an Institutional Review Board (IRB)?

1. **PROCESSING AND CLEANING DATA**

**Objective**: Process and clean input and training datasets to ensure data quality and suitability for analysis and model training.

**Example Ethical Considerations**:

* **Fairness**: Does the cleaning process maintain the diversity and integrity of the original data?
* **Privacy**: Is sensitive information anonymized before data is made available for analysis?

1. **STORING DATA**

**Objective**: Store processed input and training data securely, ensuring compliance with data privacy regulations and restricted access.

**Example Ethical Consideration**:

* **Policy compliance**: Is data storage compliant with regulations such as GDPR?

1. **MODEL ACQUISITION AND CREATION**

**5a. Acquiring Existing Model**

**Objective**: Acquire existing models that align with the research problem.

**Example Ethical Considerations**:

* **Fairness**: Does the pre-existing model generalize well to the project’s intended uses?
* **Transparency**: Are the model’s assumptions, training data, and limitations clearly documented?

**5b. Building and Training a Custom Model**

**Objective**: Develop and train custom models using training data, ensuring models are accurate, fair, and transparent.

**Example Ethical Considerations**:

* **Fairness**: Are training data biases addressed during model training?
* **Interpretability**: Are the model’s decision-making processes easy to understand for end users of the model’s output?

1. **INTEGRATING MULTIPLE MODELS INTO A SINGLE MODEL**

**Objective**: Combine multiple existing or newly developed models into a unified system to enhance analytical capacity or predictive power

**Example Ethical Considerations**:

* **Transparency**: Is the contribution of each integrated model clearly documented and communicated to users?
* **Fairness:** Could combining models lead to compounding existing biases?

1. **GENERATING AND INTERPRETING MODEL OUTPUTS**

**Objective**: Use the models and input data to generate insights that answer research questions and inform decision-making.

**Example Ethical Consideration**:

* **Explainability**: Are insights, and their limitations, documented in an understandable way?

1. **DISSEMINATING, ARCHIVING, OR DESTROYING RESEARCH OUTPUTS**

**Objective**: Share datasets, models, and research findings with communities and the broader scientific community; ensure their long-term archiving or securely destroy sensitive datasets.

**Example Ethical Considerations**:

* **Access Equity**: Are findings shared in a way that is accessible to all, including non-experts?
* **Transparency**: Are the methods and limitations of the research clearly communicated?