Research Data Management Plan

2025-04-01

Table of contents

Project: ETL for Maximum Daily Temperature in the Atlántico region, Colombia
Description: This project involves extracting and processing daily maximum temperature data from the IDEAM website for the Atlántico department in Colombia. The data is collected using web scraping with Selenium and organized through a structured ETL pipeline.

Data Management and FAIR Principles

This data management plan is designed in accordance with the FAIR principles:

- **Findable**:
  The data is organized using a clear folder structure (bronze, silver layers) and meaningful filenames. Metadata such as variable name, department, station code, and date range are included both in filenames and in documentation.
- **Accessible**:
  Although the data source IDEAM DHIME does not provide persistent identifiers like DOIs, the access is provided through documented web scraping procedures. The scripts supply transparent, reproducible access to the source data.
- **Interoperable**:
  The final dataset is stored in standardized `.csv` format using UTF-8 encoding and column names (`Fecha`, `Valor`). Data types and formats are consistent across records.

- **Reusable**:
  Reusability is supported through complete documentation, version-controlled code (via GitLab), and modular ETL scripts that allow adaptation for other variables, regions, or time ranges. Future extensions will include metadata schemas and automated update monitoring. Although the original data source does not provide DOIs, we document data lineage and provenance through reproducible scripts and version-controlled pipelines.

Data Sources and Collection

The primary data source is the IDEAM website: IDEAM DHIME. Daily temperature data is extracted through automated web scraping using Selenium. The data is saved in different stages:

- **Bronze layer**: Raw `.zip` file downloaded from the IDEAM portal.
- **Silver layer**: Cleaned `.csv` file containing two columns: `Fecha` (Date) and `Valor` (Temperature).

The extraction process is configured to target:

- Variable: Temperatura
- Parameter: Temperatura máxima diaria
- Department: Atlántico
- Station Code: 29035080

Data Sources and DOI Availability

The data used in this project originates from a hydrometeorological web platform.

Note: The database used in this study does **not provide Digital Object Identifiers (DOIs)**. Instead, detailed metadata and links to the official download interface are included in the repository documentation.

Data Update Policy

The ETL process is designed for **regular updates**. The automated pipeline fetches data from **01/01/2000** up to **15 days before the current date**, guaranteeing both historical coverage and timely updates.

This is handled by the Python automation script in the `orchestrator.py`, which ensures reliability by retrying downloads on failure and using dynamic date handling.

Future extensions may include integration with CI/CD tools to schedule and monitor updates automatically.