Sharing Geospatial Data at Scale

Jorge Gomes <jorge@lip.pt>

IBERGRID 2026, Santiago de Compostela







Initial motivation

- Mainland Portugal didn’t had an homogeneous, up-to-date, and accurate digital terrain model.
  - Rigorous knowledge of the territory's altimetry is fundamental for many sectors of activity.
- Direção-Geral do Território (DGT) contracted a LiDAR survey of the Portuguese mainland.
  - LiDAR (Light Detection and Ranging) technology allows for knowledge of the terrain's relief in a reduced timeframe and with high altimetric positional accuracy.
  - LiDAR technology is significantly more efficient than other methodologies that can be used to obtain digital models of the Earth's surface.

- Agriculture
- Environment
- Infrastructure
- Archaeology
- Public Works
- Energy
- Land Use Planning
- Forests
- Nature Conservation
- Civil Protection
- Telecommunications
- Urban Planning
- Municipal Management
- Insurance
- ...

How is it done ...



Piper PA-31T2 Cheyenne II XL



PA31T Cheyenne



Partenavia P-68



Galaxy T2000
Teledyne Optech



RIEGL VQ-780 II-S
RIEGL



PhaseOne iXU-RS1000



PhaseOne iXM-RS 150F – RGB
PhaseOne iXM-RS 100 – NIR

Precise geospatial positioning:

- GNSS
- Inercial system(IMU/INS)

LiDAR sensors:

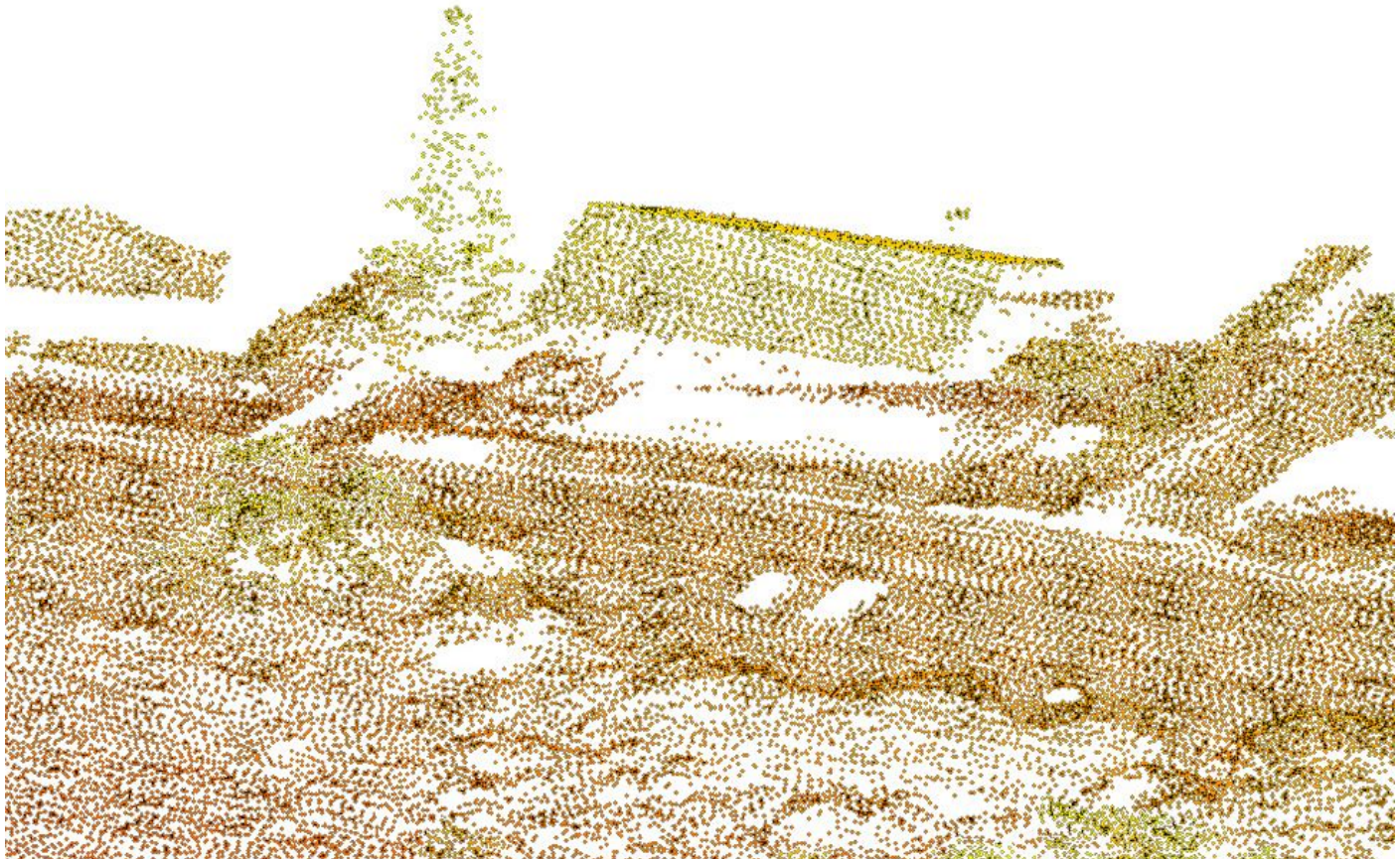
- Linear mode sensors
- Wavelength in the infrared
- Pulse frequency >= 400 kHz
- Record >= 4 returns p/pulse

Capacity to capture images with:

- Spatial resolution >= 0.25 m
- Radiometric resolution >= 8 bits
- Spectral composition: Red, Green, Blue and Near Infrared

PRR Plano de Recuperação e Resiliência · República Portuguesa · Financiado pela União Europeia NextGenerationEU · SMOS Sistema de Monitorização da Ocupação do Solo · d.gTerritório · LIP · CNCA

What do you get as output



Output is a LiDAR point cloud:

- detailed 3D digital representation of a physical space or object
- consisting of thousands or millions of individual georeferenced data points
- Each point represents a precise location where a laser beam emitted by a LiDAR sensor bounced off a surface and returned to the scanner.
- Average density 10 points/m^2

LAS data format

3D point cloud in LAS format
(las is just the file extension)

LAS Specification v.1.4 - R15

ASPRS - American Society for Photogrammetry & Remote Sensing

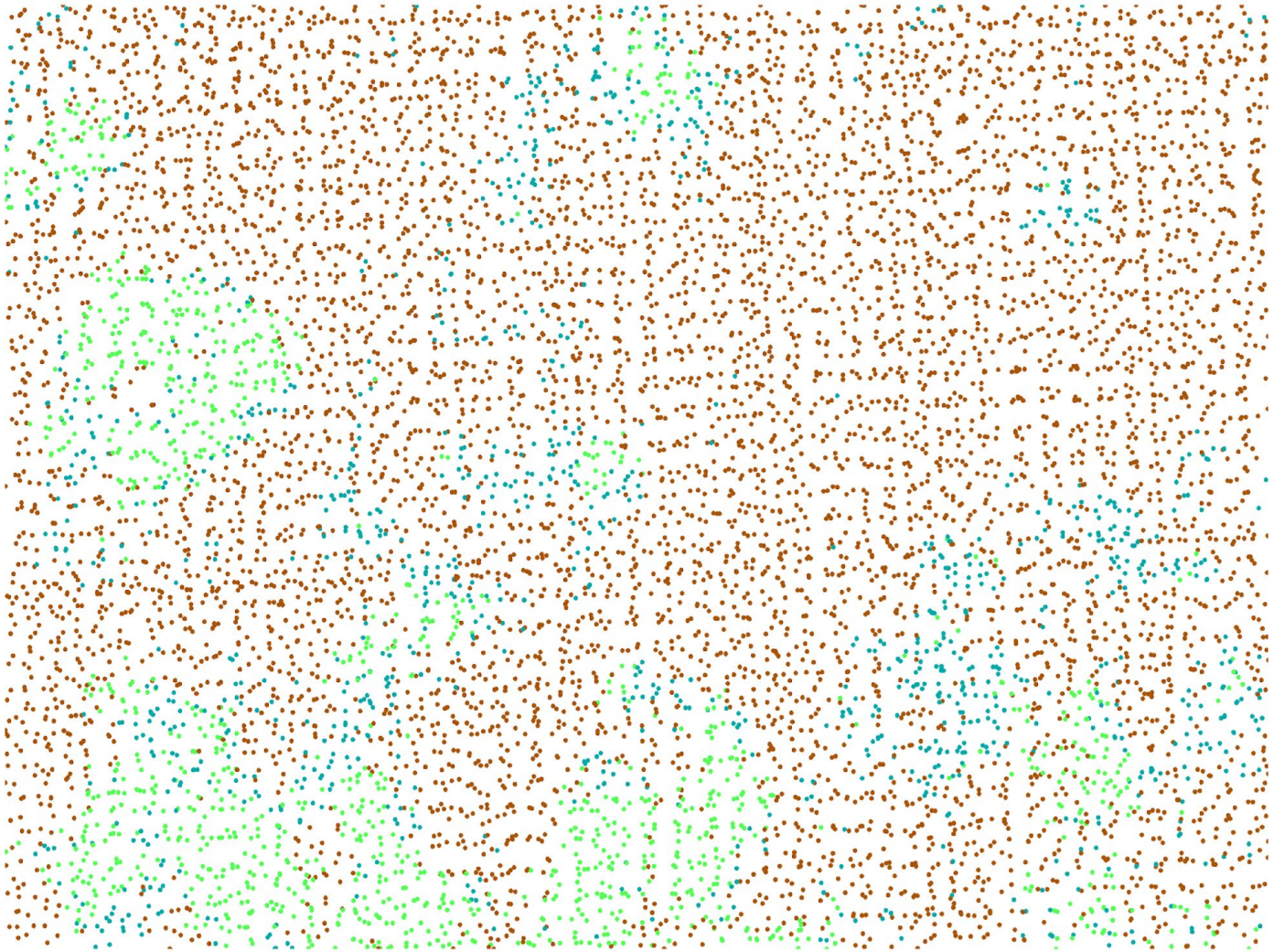
Point Data Record Format 8

We use the compressed format LAZ
LAZ is the official lossless compressed version of LAS

LAS and LAZ are now OGC standards

Item	Format
X	long
Y	long
Z	long
Intensity	unsigned short
Return Number	4 bits (bits 0-3)
Number of Returns (Given Pulse)	4 bits (bits 4-7)
Classification Flags	4 bits (bits 0-3)
Scanner Channel	2 bits (bits 4-5)
Scan Direction Flag	1 bit (bit 6)
Edge of Flight Line	1 bit (bit 7)
Classification	unsigned char
User Data	unsigned char
Scan Angle	short
Point Source ID	unsigned short
GPS Time	double
Red	unsigned short
Green	unsigned short
Blue	unsigned short
NIR	unsigned short

Classification



Classification of the points

1 – Other

2 – Terrain

3 – Low Vegetation (0 m – 0.5 m)

4 – Medium Vegetation (0.5 m – 2 m)

5 – Tall Vegetation (above 2 m)

6 – Buildings

7 – Noise

9 – Water

26 – Bridges

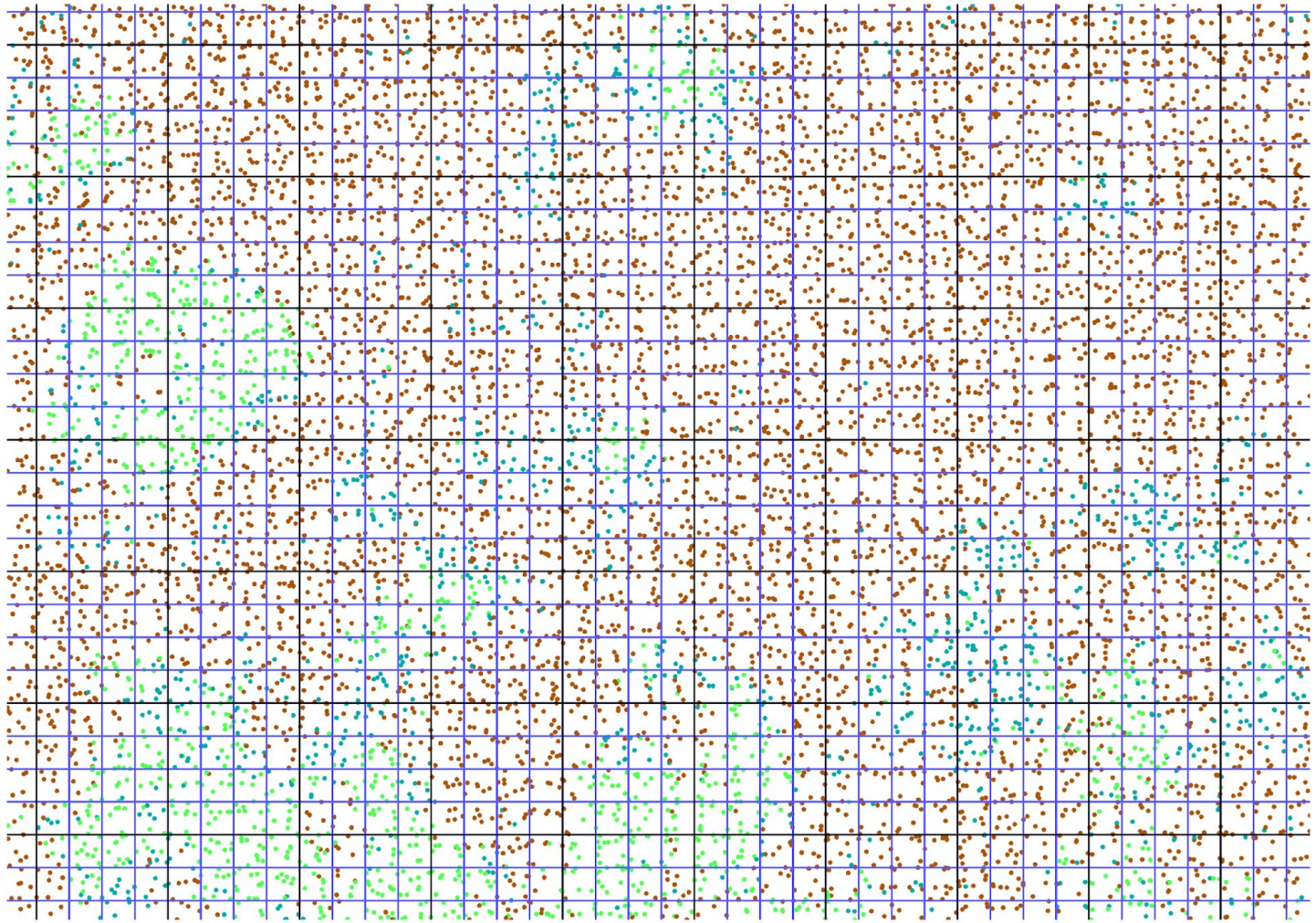
Classification accuracy

"Terrain" class: 1%

Other classes: 10%

Digital Elevation Models and Resolution

Resolution: 2 m **Resolution: 50 cm**



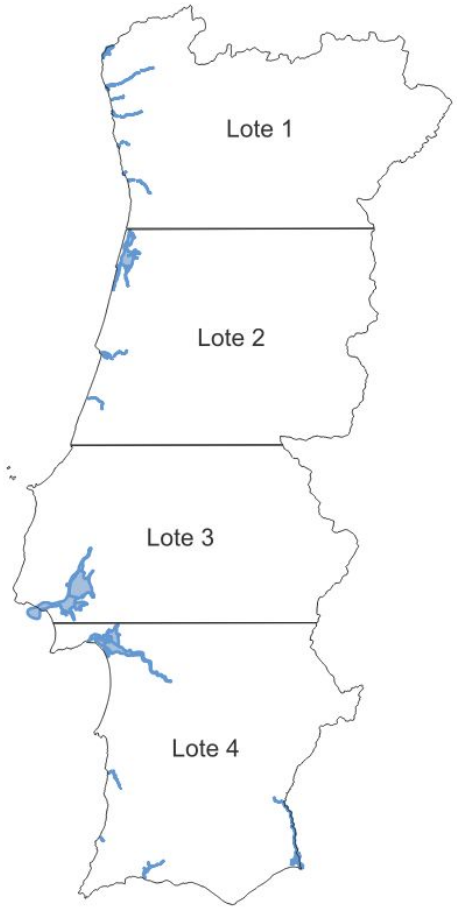
Additional products:

- **Digital Terrain Model (DTM)**
  - 2 – Terrain
- **Digital Surface Model (DSM)**
  - 1 – Other
  - 2 – Terrain
  - 3 – Low Vegetation
  - 4 – Medium Vegetation
  - 5 – Tall Vegetation
  - 6 – Buildings
  - 26 – Bridges

PRR Plano de Recuperação e Resiliência · República Portuguesa · Financiado pela União Europeia NextGenerationEU · SMOS Sistema de Monitorização da Ocupação do Solo · dgTerritório · LIP · CNCA

LiDAR data volume

4x batches



Data volume

Altimetry - final data products

LAZ compressed files: aprox. 15,5 TB

DTM and DSM : aprox. 2,3 TB

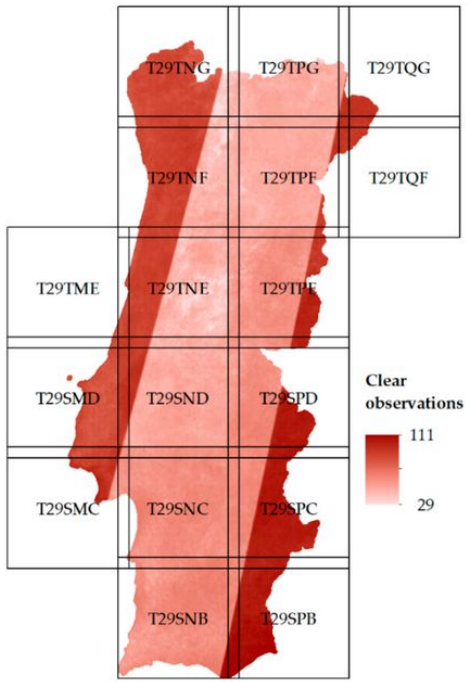
Total number of files: 455.974

Intermediate Products

Aprox. 100 TB

But at the same time LIP was collaborating with DGT on processing of Sentinel 2 data to remove clouds

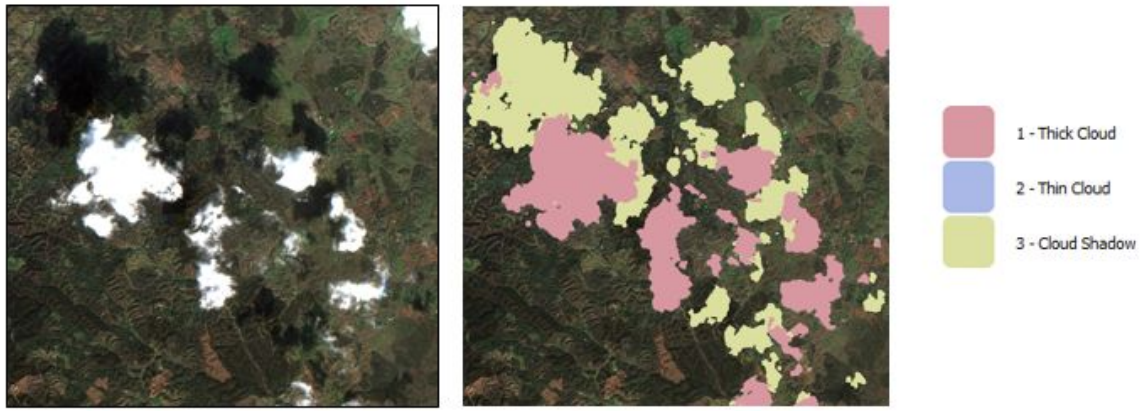
Sentinel 2 data L1C e L2A ⇒ 70TB



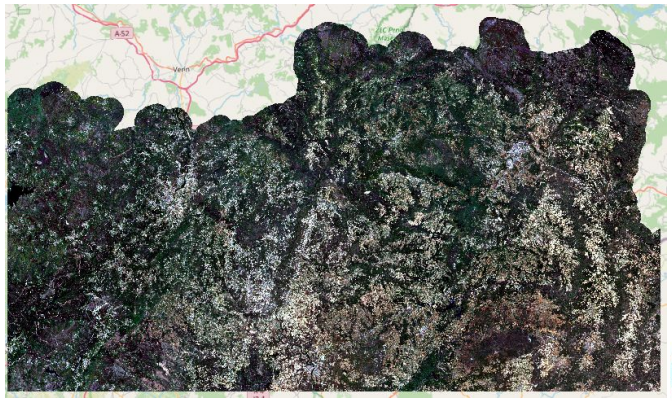
2015	0.92TiB
2016	2.8TiB
2017	5.4TiB
2018	6.1TiB
2019	6.2TiB
2020	6.1TiB
2021	6.2TiB
2022	12 TiB
2023	9.7TiB
2024	6.2TiB
2025	8.4TB

and beyond …

Produce tiles without clouds using omnicloudmask



Monthly composite tiles without clouds



Continuous pipeline processing all new Sentinel 2 data for Portugal.

But at the same time LIP was collaborating with DGT on processing of Sentinel 2 data on vegetation indexes

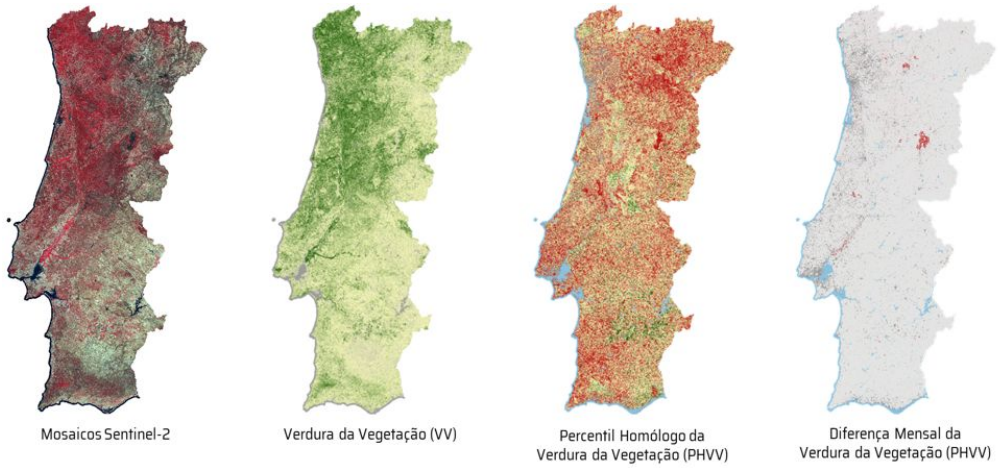
Intra-Annual Vegetation Status Maps (IAEVM) characterize the vegetative vigor of vegetation using the Normalized Difference Vegetation Index (NDVI). Three types of products are available:

- Vegetation Greenness (VV) - Characterizes the current vigor of vegetation using NDVI calculated from the Sentinel-2 Monthly Composite.
- Vegetation Greenness Homologous Percentile (VGHP) – Compares the current vigor of vegetation with the homologous historical values of Monthly Composites.
- Vegetation Greenness Monthly Difference (VGDM) – difference between the vegetation vigor of the current month and the previous month. This product uses post-processed daily Sentinel-2 images instead of Monthly Composites.

Composite tiles without clouds

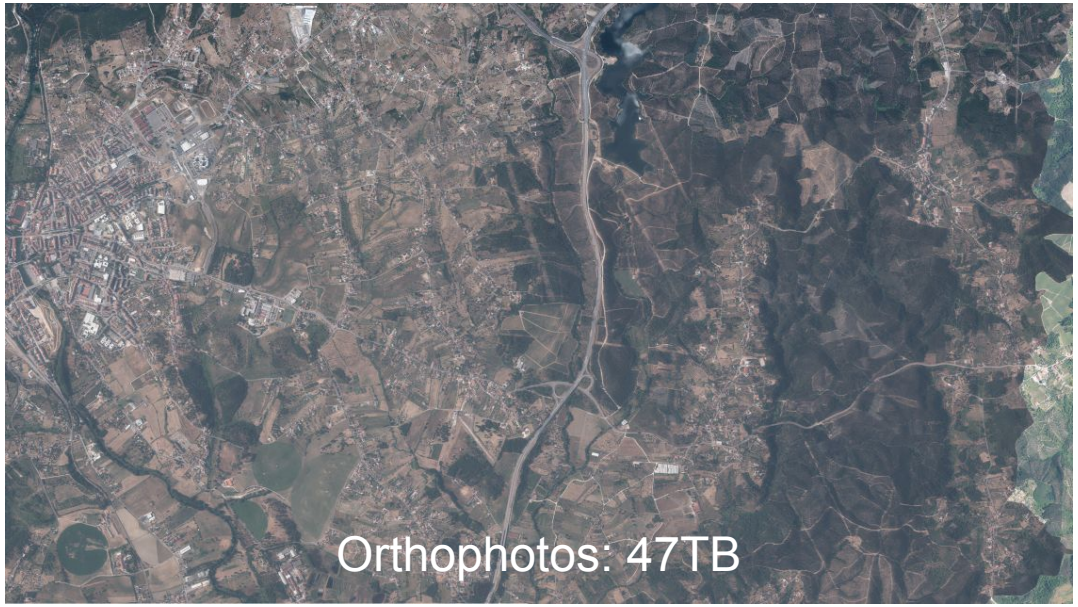


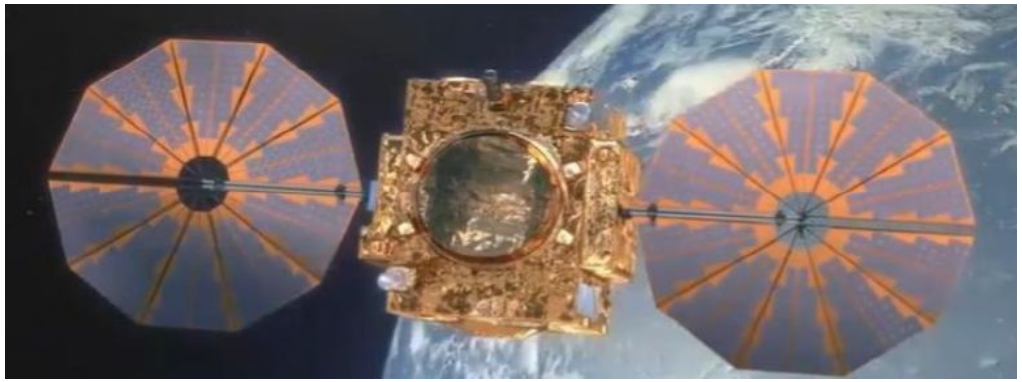
Intra-Annual Vegetation Status Maps



Also orthoimages ...

- Aerial or satellite photograph that has been geometrically corrected so that it has a uniform, consistent scale and can be measured like a map.
  - Ortho photos: 1995, 2004-2006, 2007, 2010, 2012, 2015, 2018, 2021, 2025, …
  - Ortho satellite: 2023 (with sharing restrictions ⇒ access control)





Ortho Sat 2023: 4.884 files, 30cm/pixel, 18TB

Challenges and requirements

- **Goal: safely store and share the data:**
  - **Growing needs:** current datasets will grow and new ones will need to be made available
  - **Different types of data:** point clouds, raster, vectorial
  - **Enable data access:** download to user premises
  - **Enable processing:** allow users to process data efficiently in the datacenter
  - **Scalability and robustness:** potentially many large simultaneous downloads
  - **Homogeneous interface:** standards based, common interface for all products and data types
  - **Easy access:** friendly web frontend
  - **Programmatic access:** APIs
  - **Support access restrictions:** authentication, authorization, integration with autenticação.gov.pt
  - **Provide metrics:** downloads per region, district, municipality and time interval, etc

Joint collaboration for data access and distribution

Joint collaboration between DGT, LIP and CNCA

LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS

A leading institution for experimental particle physics and associated technologies in Portugal.
Particle Physics, Instrumentation and Methods, Computing

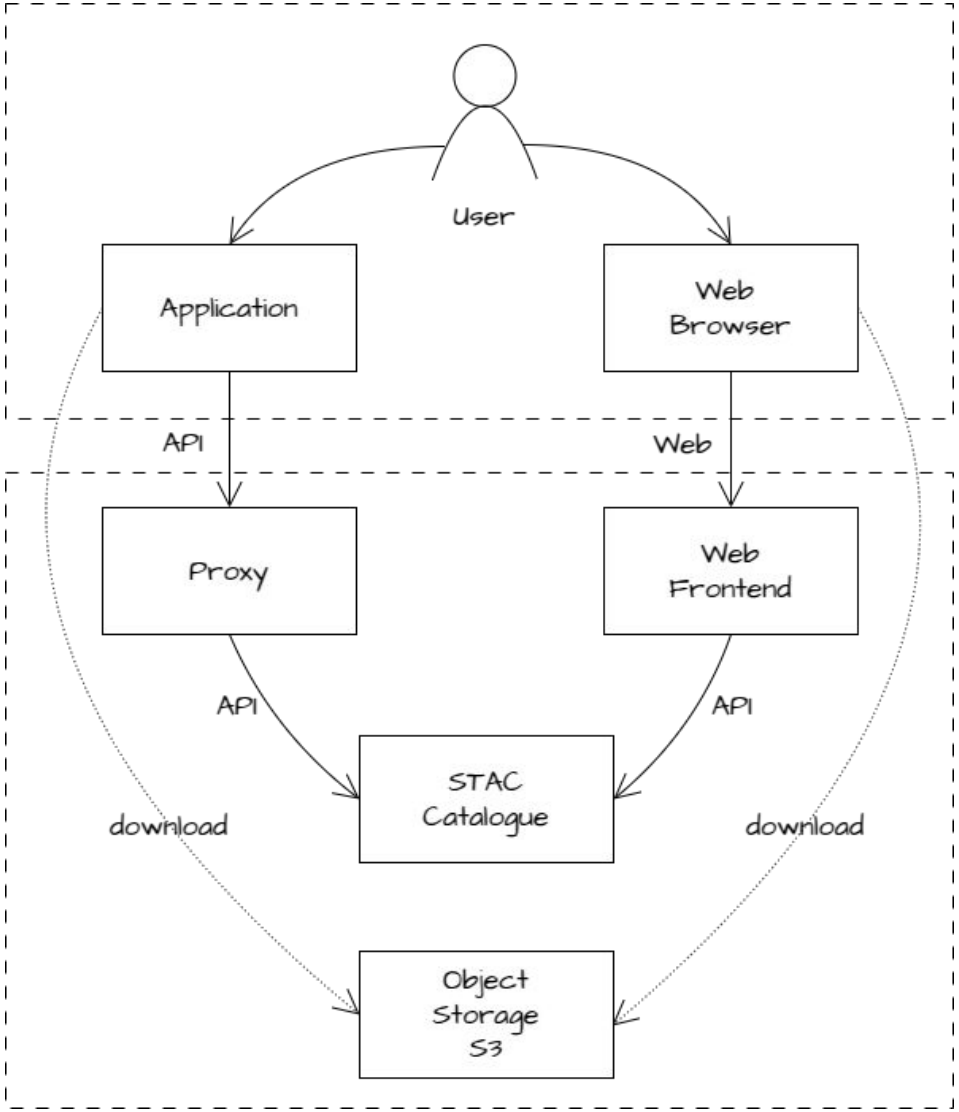
Development of the CDD Platform
Data Handling, Processing and Visualization services

https://www.lip.pt/

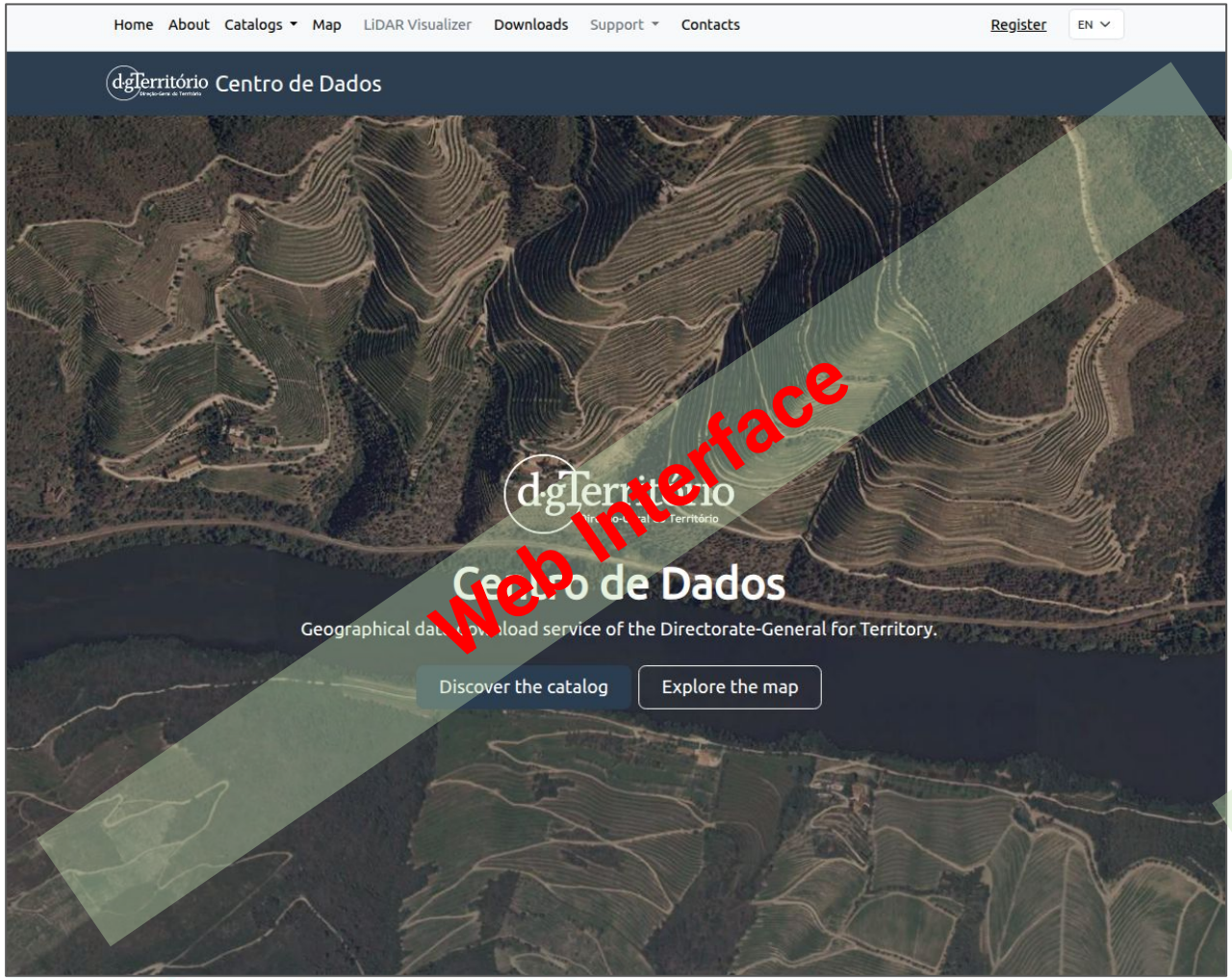


https://www.acnca.pt

Initial concept



The CDD platform (CDD = Centro de Dados)
https://cdd.dgterritorio.gov.pt



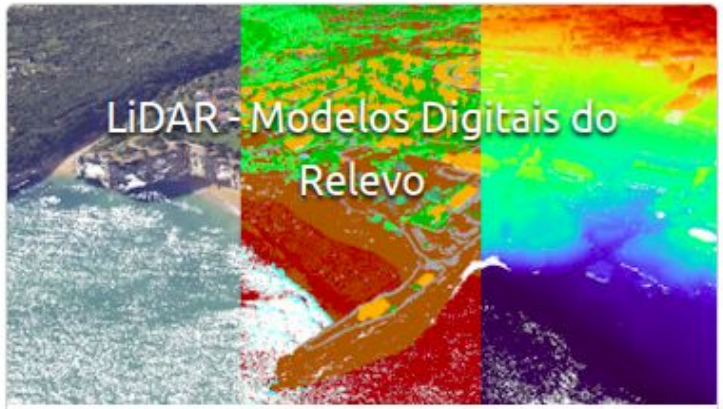


PRR · REPÚBLICA PORTUGUESA · Financiado pela União Europeia NextGenerationEU · SMOS Sistema de Monitorização da Ocupação do Solo

LIP · d.gTerritório · CNCA

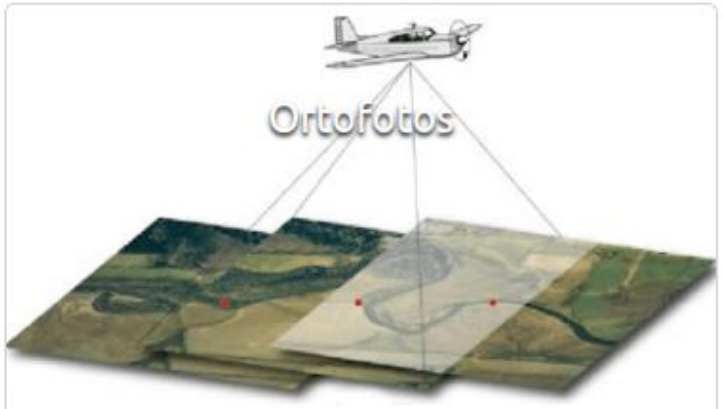
Data catalogs in the CDD

Geographical datasets from the Direção-Geral do Território available for download.



Modelos digitais do relevo obtidos com LiDAR aerotransportado para o território de Portugal continental. Coleções de Modelos Digitais do Terreno (MDT), Modelos Digitais de Superfície (MDS) e Nuvens de pontos. Zona Noroeste ainda indisponível.

View collections



Coberturas de Ortoimagens do território de Portugal Continental produzidas a partir de fotografias áreas recolhidas com câmaras aerofotogramétricas digitais e posterior mosaico das fotografias ortorretificadas.

View collections



Modelos digitais do relevo da Região Autónoma dos Açores, obtidos com LiDAR aerotransportado, incluindo Coleções de Modelos Digitais do Terreno (MDT), Modelos Digitais de Superfície (MDS) e Nuvens de pontos.

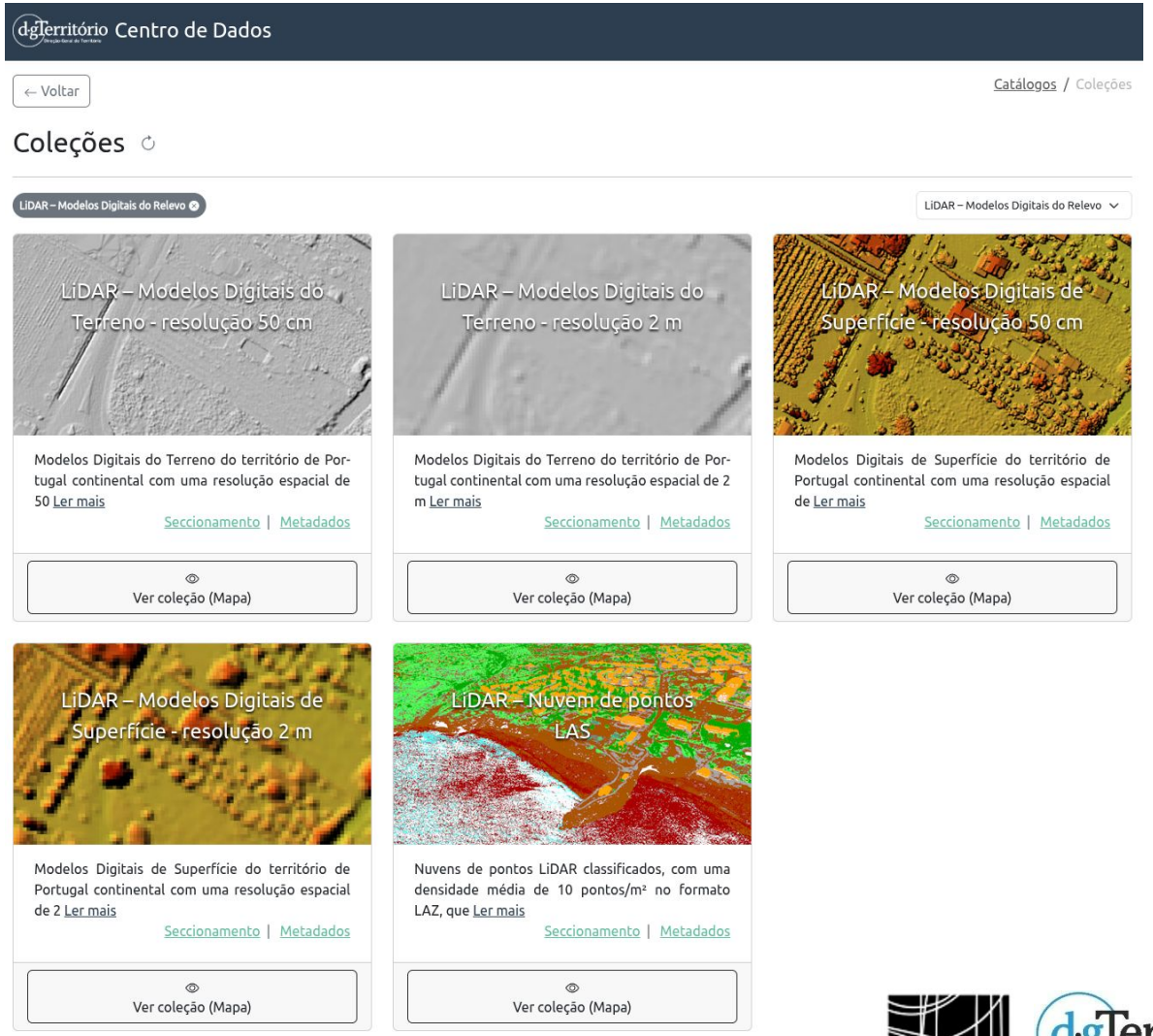
View collections

Example:
The data collections in the CDD LiDAR catalog

Digital elevation models for mainland Portugal.

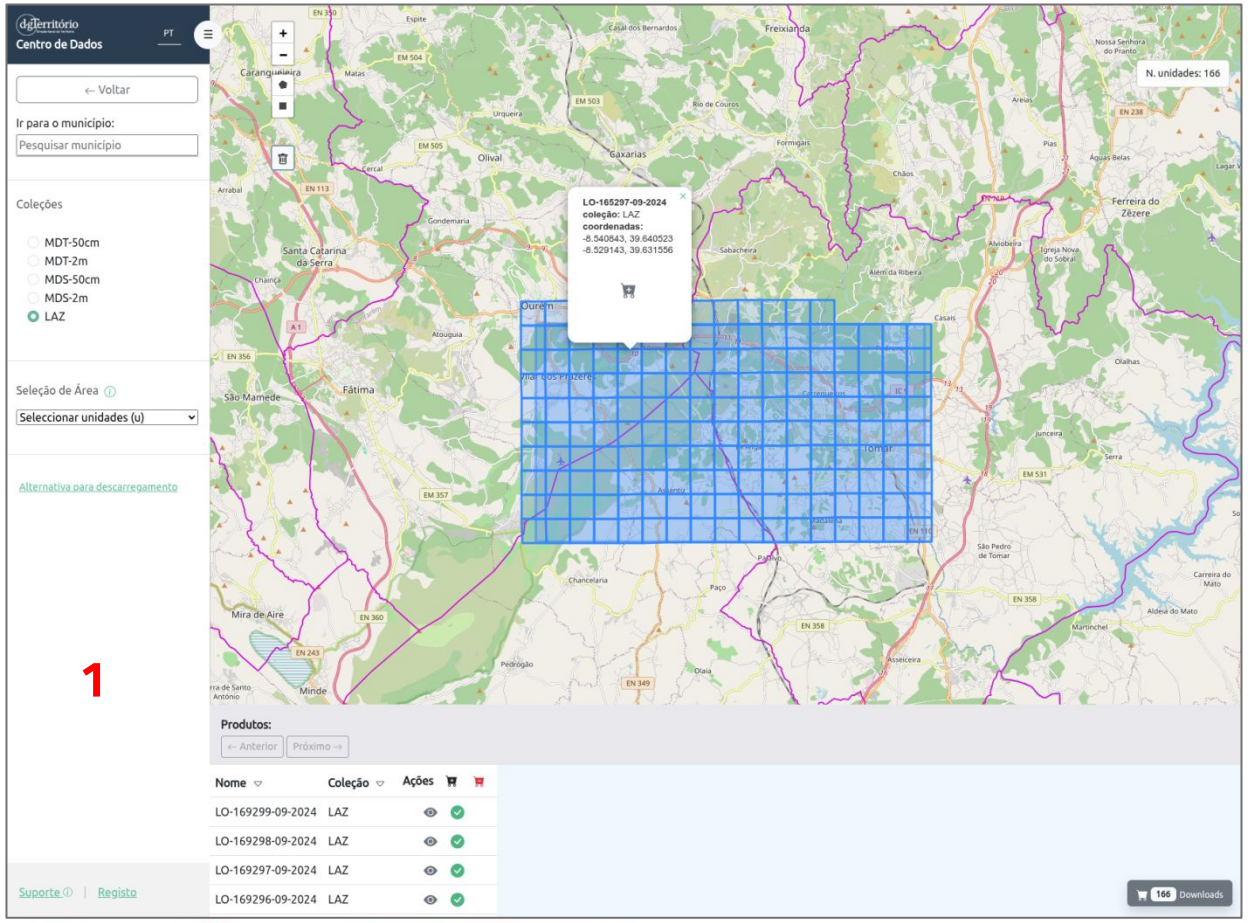
Available collections:

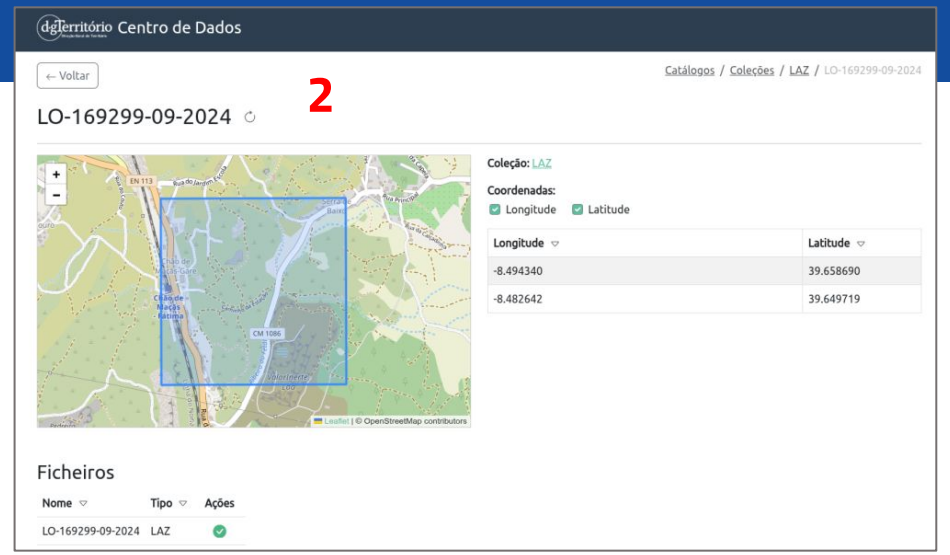
1. **DTM 50 cm**
2. **DTM 2m**
3. **DSM 50 cm**
4. **DSM 2 m**
5. **LAZ**

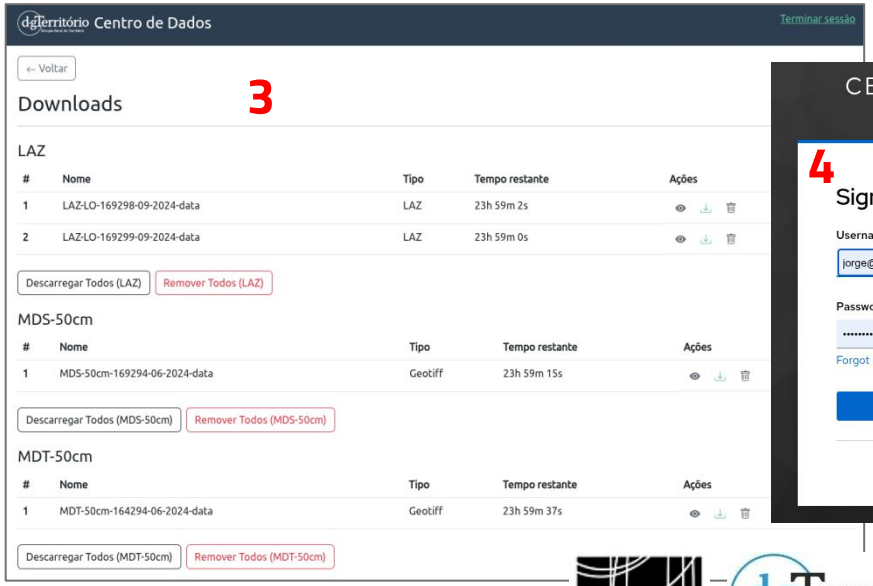


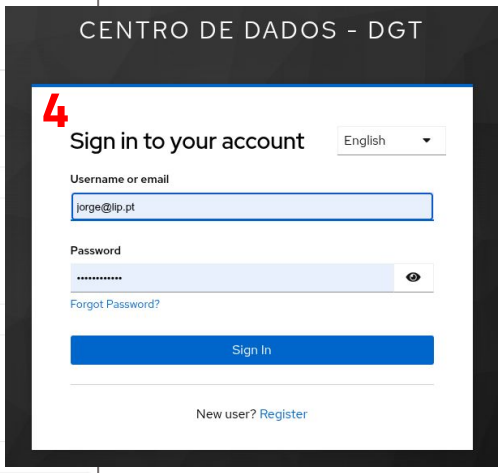
Downloading

https://cdd.dgterritorio.gov.pt/





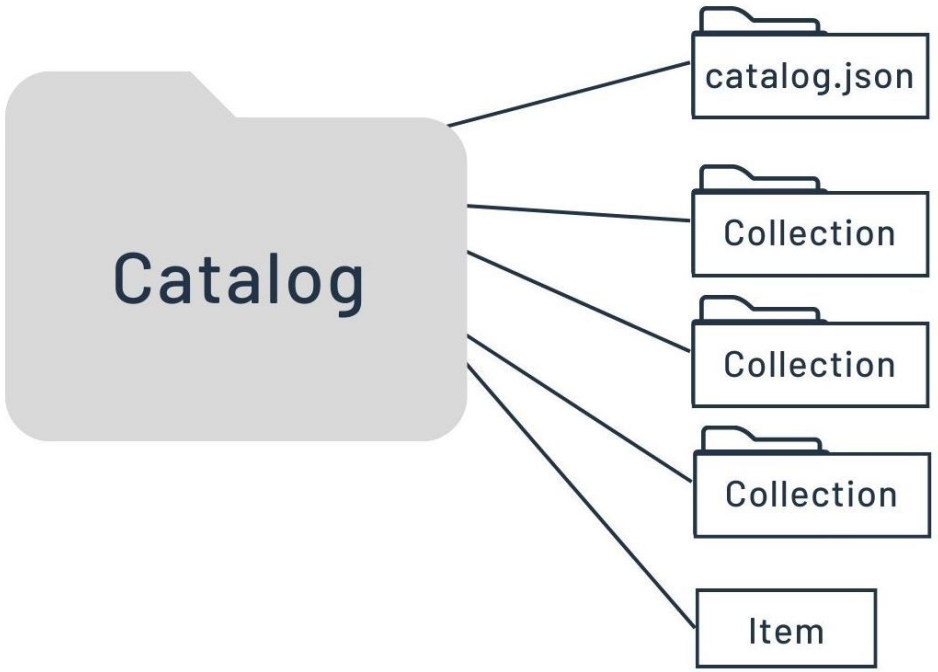




PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU
SMOS Sistema de Monitorização da Ocupação do Solo
LIP
dgTerritório
CNCA

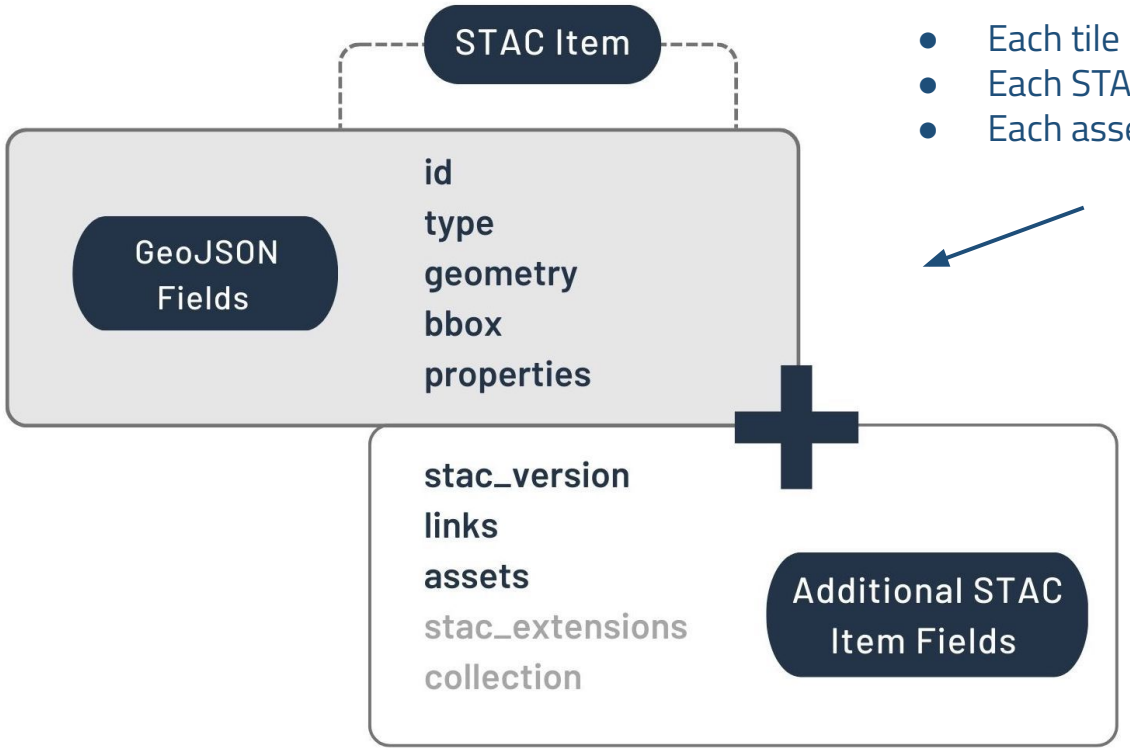
Spatio Temporal Asset Catalogue (STAC)

- STAC is a network of JSON files
  - that reference other JSON files,
  - each JSON file adheres to a specific core specification
    - depending on which STAC component it is describing.
  - The core JSON format can also be customized
    - to fit differing needs
- There are three component specifications
  - Catalog
  - Collection
  - Item
- Each component can be used alone
  - but they work best in concert with one another.



- A Catalog is usually the starting point for navigating a STAC
- catalog.json file contains links to some combination of:
  - other STAC Catalogs,
  - Collections,
  - and/or Items.
- There are no restrictions on the way STAC Catalogs are organized.
  - But in practice it may depend on the implementation

STAC item

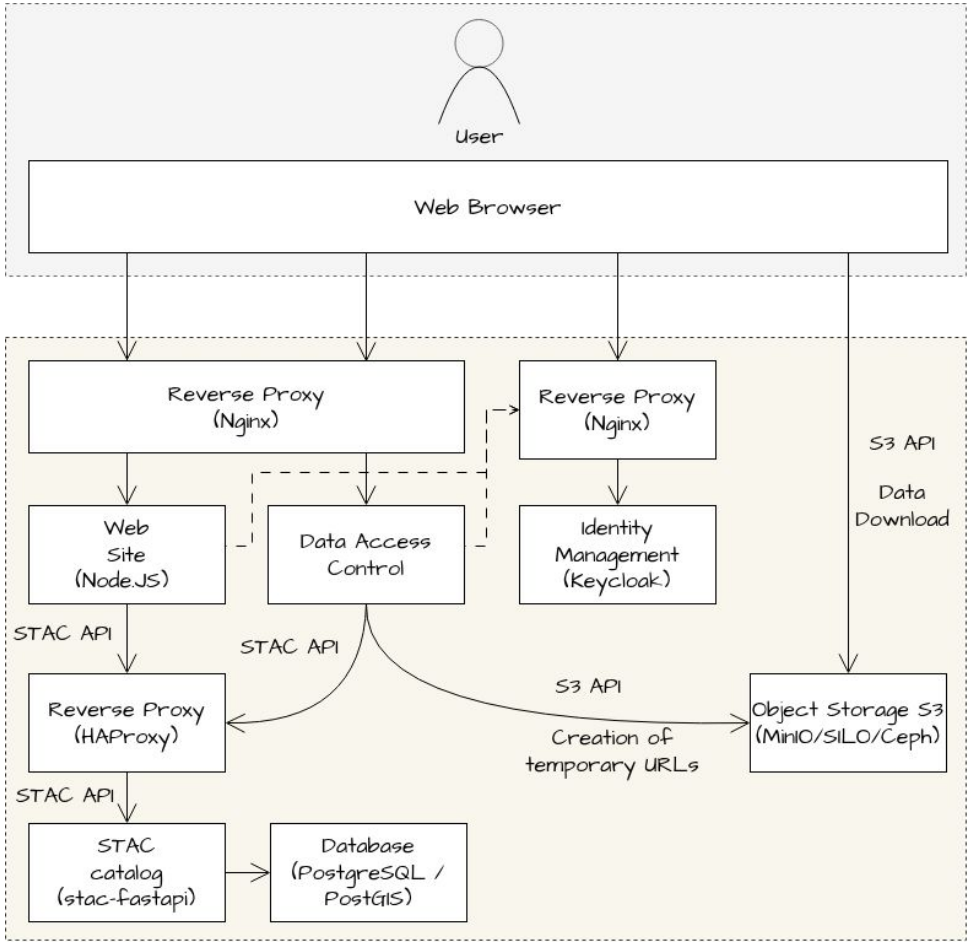


- Each tile is a STAC item
- Each STAC item has an asset
- Each asset links to data in S3

- A STAC Item is the foundational building block
- Is a GeoJSON
  - supplemented with additional metadata
  - enables to traverse through catalogs.
- The STAC Item JSON
  - uses standard GeoJSON fields
  - plus few additional information fields
- Can be easily read by any modern GIS or geospatial library

Architecture

https://cdd.dgterritorio.gov.pt/



- **STAC Catalog:**
  - Based on eoAPI
- **Data Storage:**
  - S3 object storage: Ceph, MinIO => SILO
- **Web Interface:**
  - JavaScript + Node.JS + Express and EJS
  - Leaflet + Openstreetmaps
  - Keycloak
  - HAProxy
  - Nginx
- **Data ingestion:**
  - Python scripts
  - PDAL, GDAL
- **Code Management and Deployment:**
  - GitLAB + Containers

Jupyter Notebooks

JupyterHub platform hosted at CNCA customised by LIP:

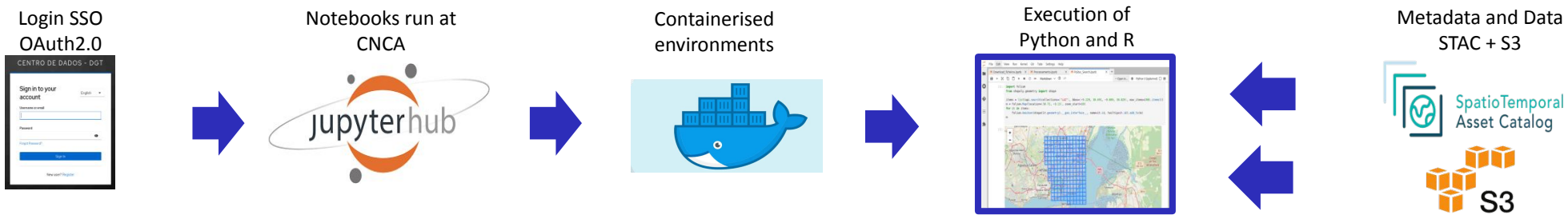
- for local data processing close to the data
- uses CNCA computing resources

Provides fast direct access to the CDD:

- STAC catalog metadata
- data stored at the CNCA object storage

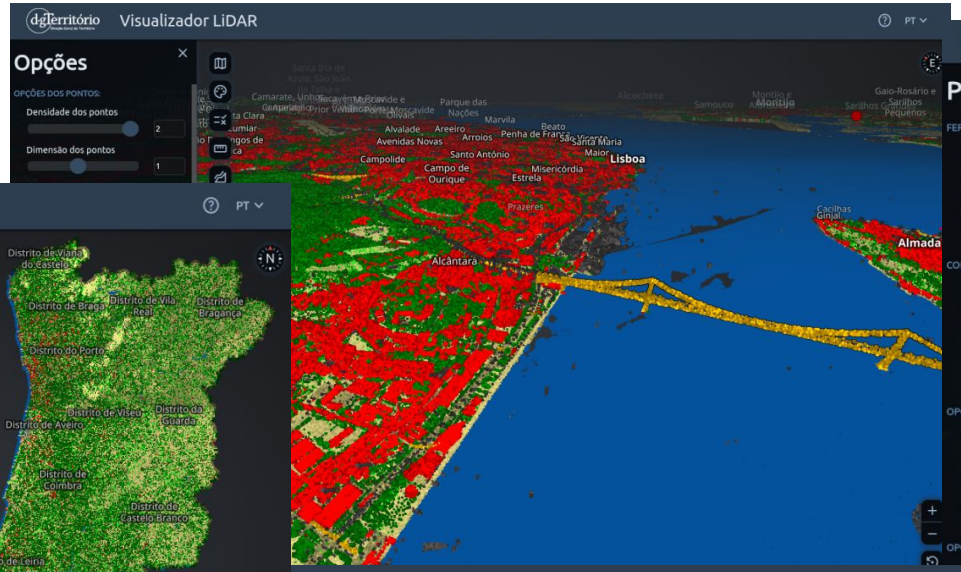
Allows users:

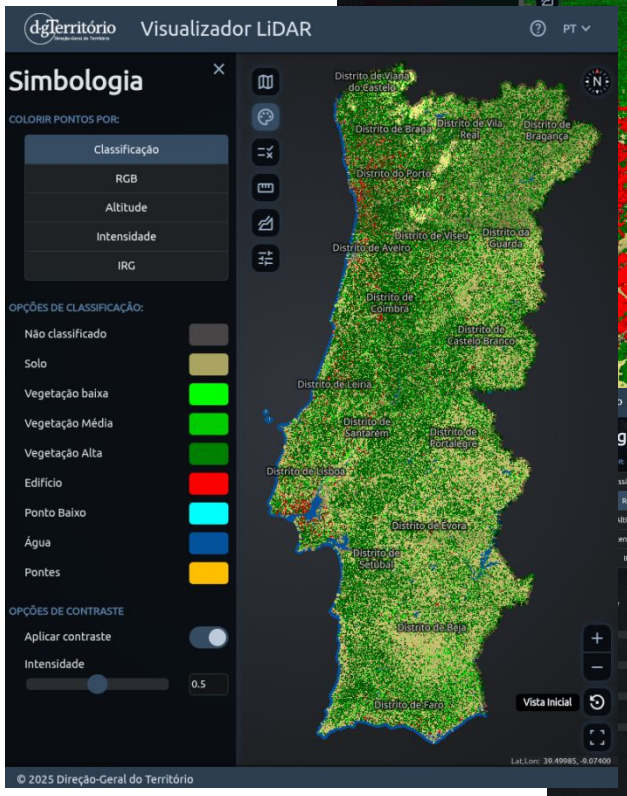
- use a pre-configured software environment
- creation and execution of user codes in Python and R
- faster exploration of the CDD data:
  - high network bandwidth
  - access to HPC and cloud computing resources

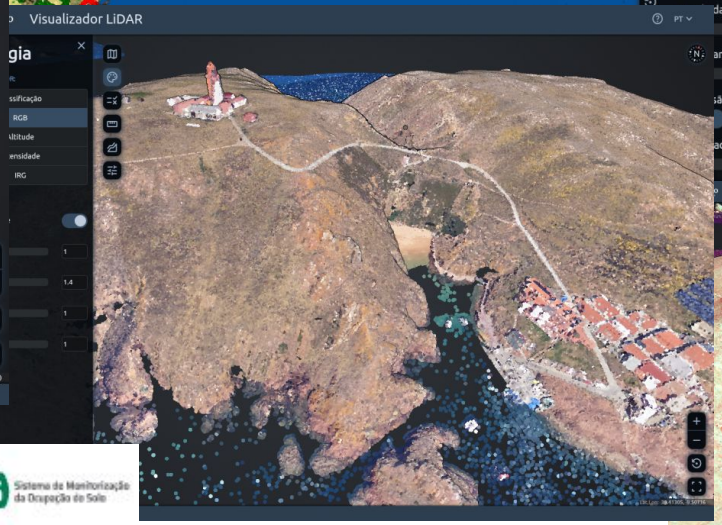


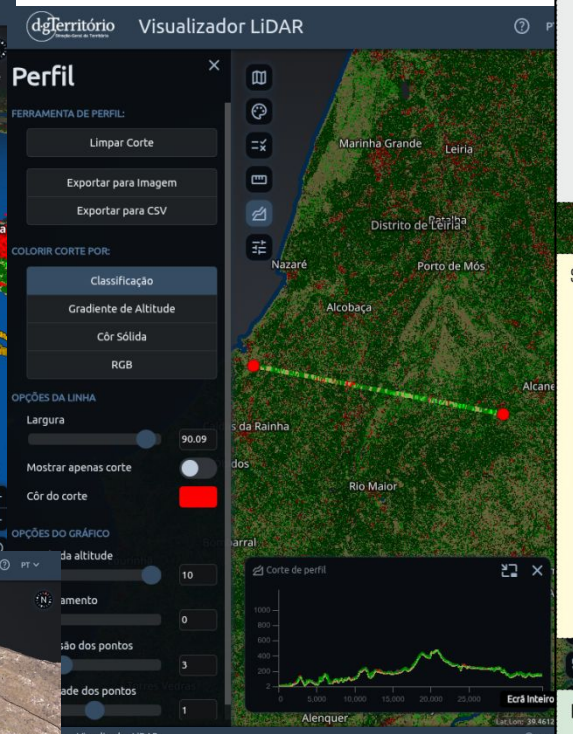
LiDAR visualisation capabilities

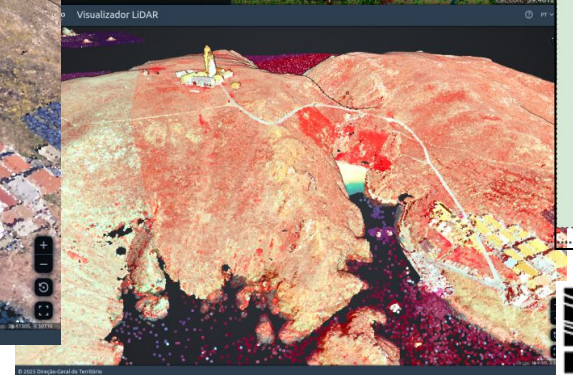
O visualizador permite explorar e navegar interativamente através de um browser web.

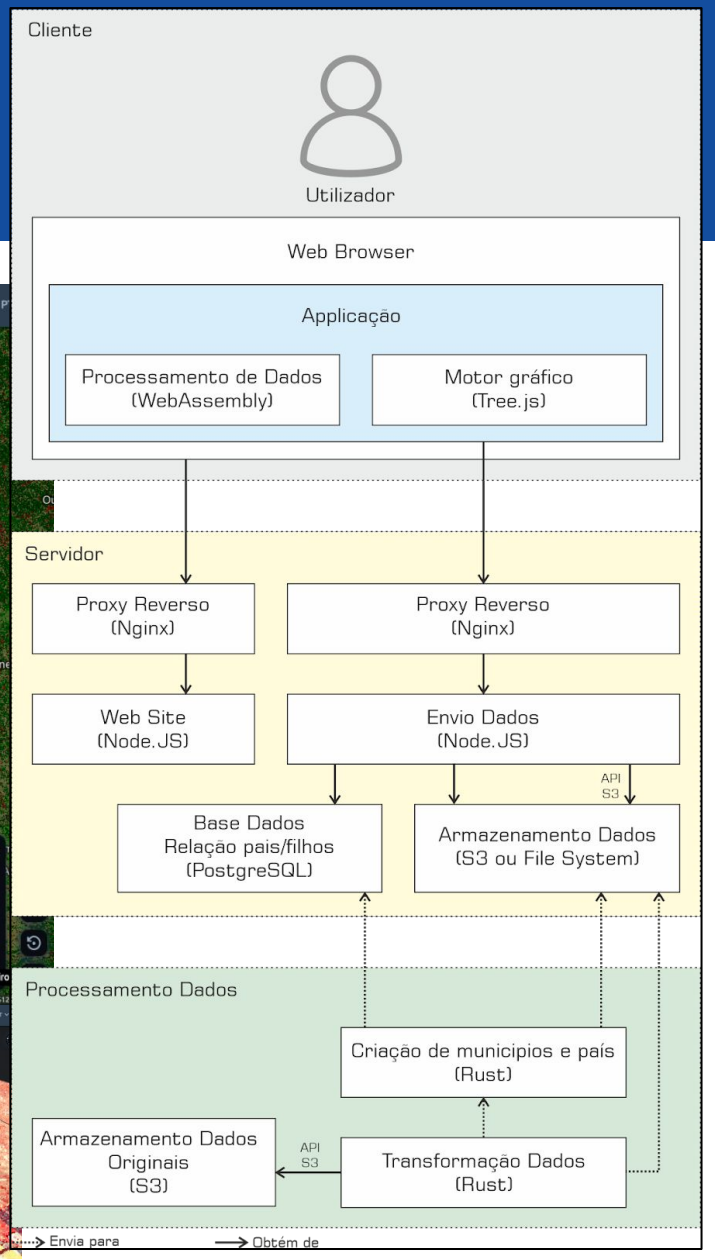






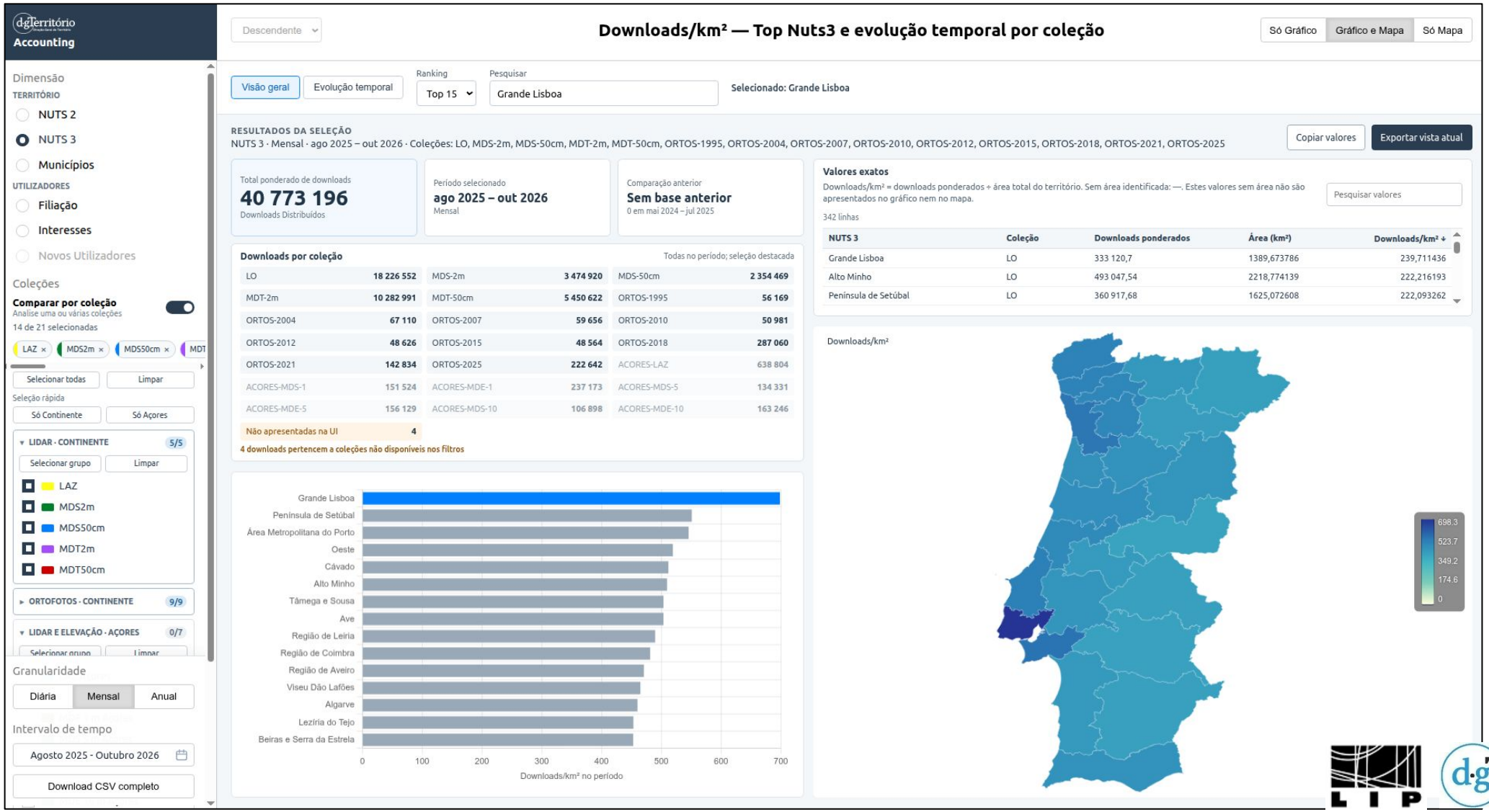




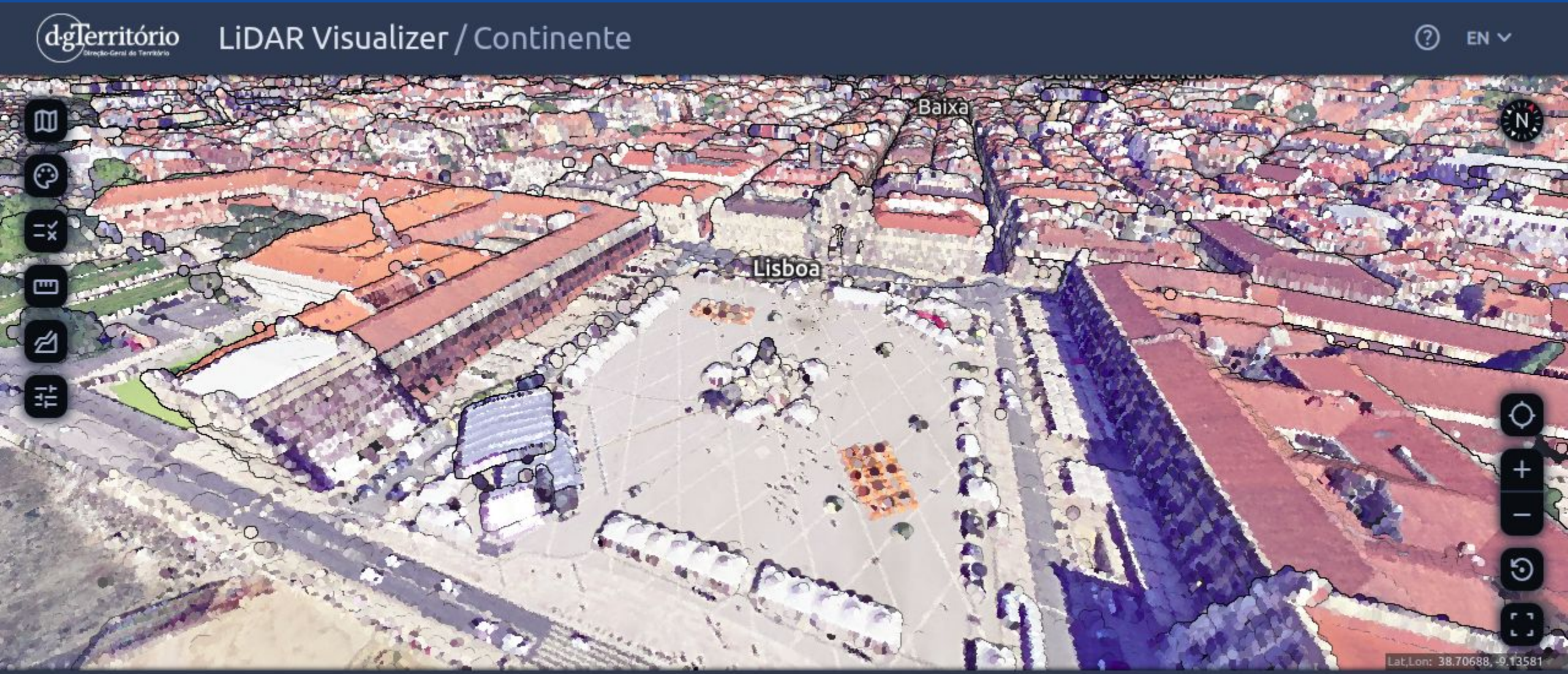


PRR Plano de Recuperação e Resiliência · REPÚBLICA PORTUGUESA · Financiado pela União Europeia NextGenerationEU · SMOS Sistema de Monitorização da Ocupação do Solo · LIP · dgTerritório · CNCA

Accounting



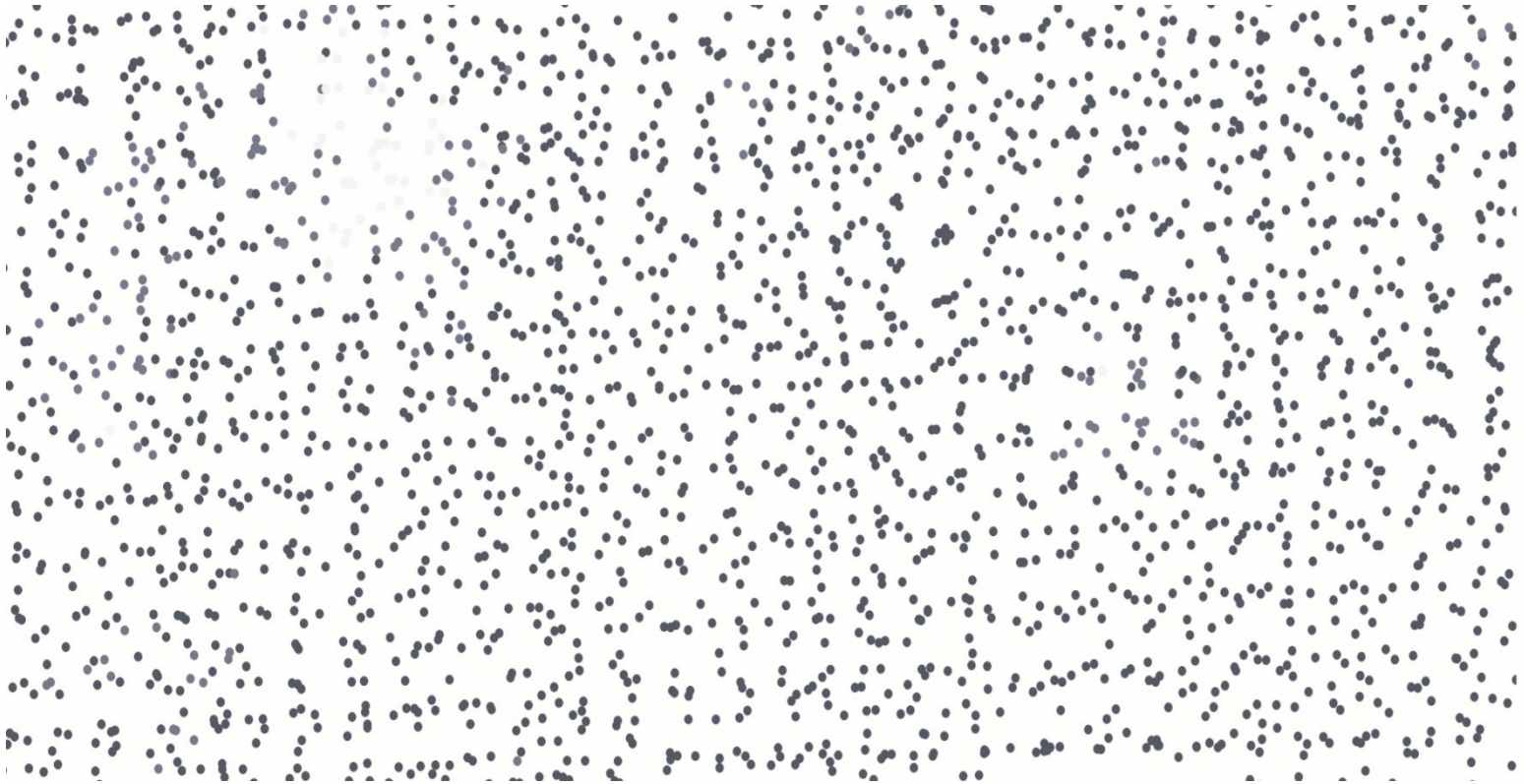
- Parsing and correlation of log entries
- Creation of accounting records in the database
- Views
- Automation for new collections
- Web interface for analysis



Thanks

PRR
REPÚBLICA PORTUGUESA
Financiado pela União Europeia
SMOS Sistema de Monitorização da Ocupação do Solo
LIP
d.gTerritório
CNCA

Density and accuracy



Altimetric accuracy

10 cm

Planimetric accuracy

30 cm

Average density of LiDAR points

10 points/m^2