



国家对地观测科学数据中心
National Earth Observation Data Center

AI Ready 遥感数据综述和思考

郑伟杰 陆君言 李国庆
国家对地观测科学数据中心

2024-9-20

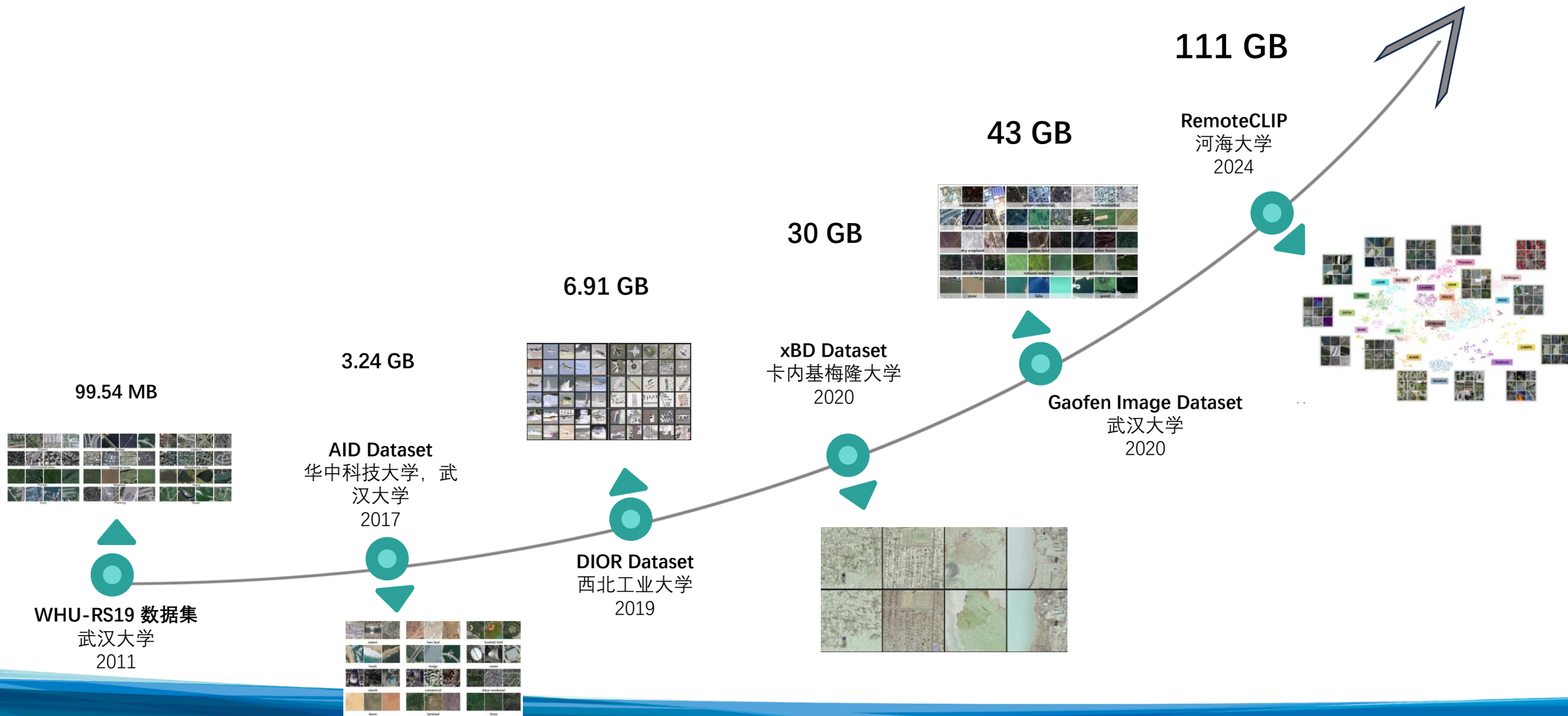


报告提纲

- 一、AI 遥感数据发展现状**
- 二、AIRD 概念发展**
- 三、遥感 AIRD 标准建立**
- 四、AIRD 遥感数据中心**
- 五、遥感 AIRD 未来展望**

一、AI遥感数据发展现状

数据集规模持续扩大



标注类型日益丰富

标注粒度

任务类型

图像级

图像描述

[caption] This is an aerial view of **landslides** area...

视觉问答

[vqa] -Is there any bridges in the image? -No.

场景分类

...

区域级

视觉定位

[grounding] **Landslides**<0.1, 0.2, 0.8, 0.9>,
Building<0.2, 0.32, 0.22, 0.34><0.5, 0.2, 0.51, 0.21>,
Cloud<0.5, 0.67, 0.6, 0.8>

目标检测

...

像素级

语义分割

[segmentation] **Landslides**<0.1, 0.3, 0.11, 0.32, 0.14, 0.35, ...>

变化检测

...

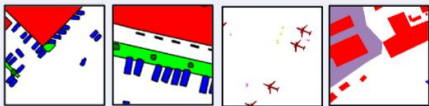


一、AI遥感数据发展现状


数据集互操作需求增加

Segmentation Datasets (SEG-4)


Vaihingen Potsdam iSAID LoveDA



Mask-to-Box (M2B)



Box-to-Caption (B2C)




A large deal of places with cars are in the middle of the picture .
There is a place with building in this remote sensing picture .
4 large vehicles, a small vehicle and 5 planes at the edge of the picture .
Lots of buildings are located in this remote sensing picture .

Step 1. Data Scaling via Annotation Unification


Detection Datasets (DET-10)

Satellite Imagery

DOTA DIOR HRRSD RSOD LEVIR HRSC



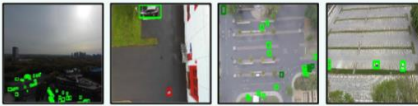
Box-to-Caption (B2C)




Many small-vehicles in the middle of the picture .
A quantity of airplanes are located in the picture .
Lots of airplanes are located in the edge of the picture .
There are 15 aircrafts on the ground .
An airplane in the middle of the picture .
A ship in the middle of the picture .

UAV Imagery

VisDrone AU-AIR S-Drone CAPRK




Box-to-Caption (B2C)



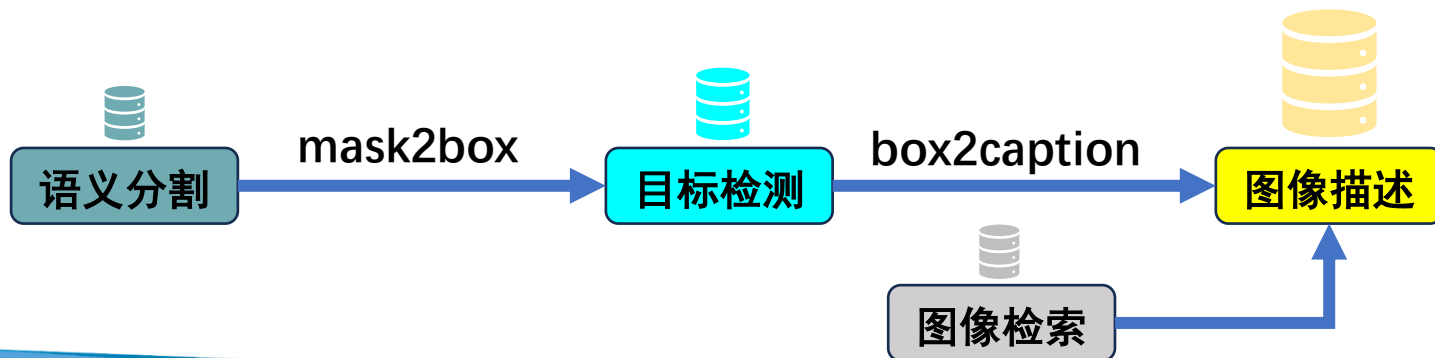
A crowd of cars are located in this remote sensing picture .
A human and a car at the edge of the picture .
There are both pedestrians and bikers in the picture .
There are 3 cars in the picture .

Retrieval Datasets (RET-3)

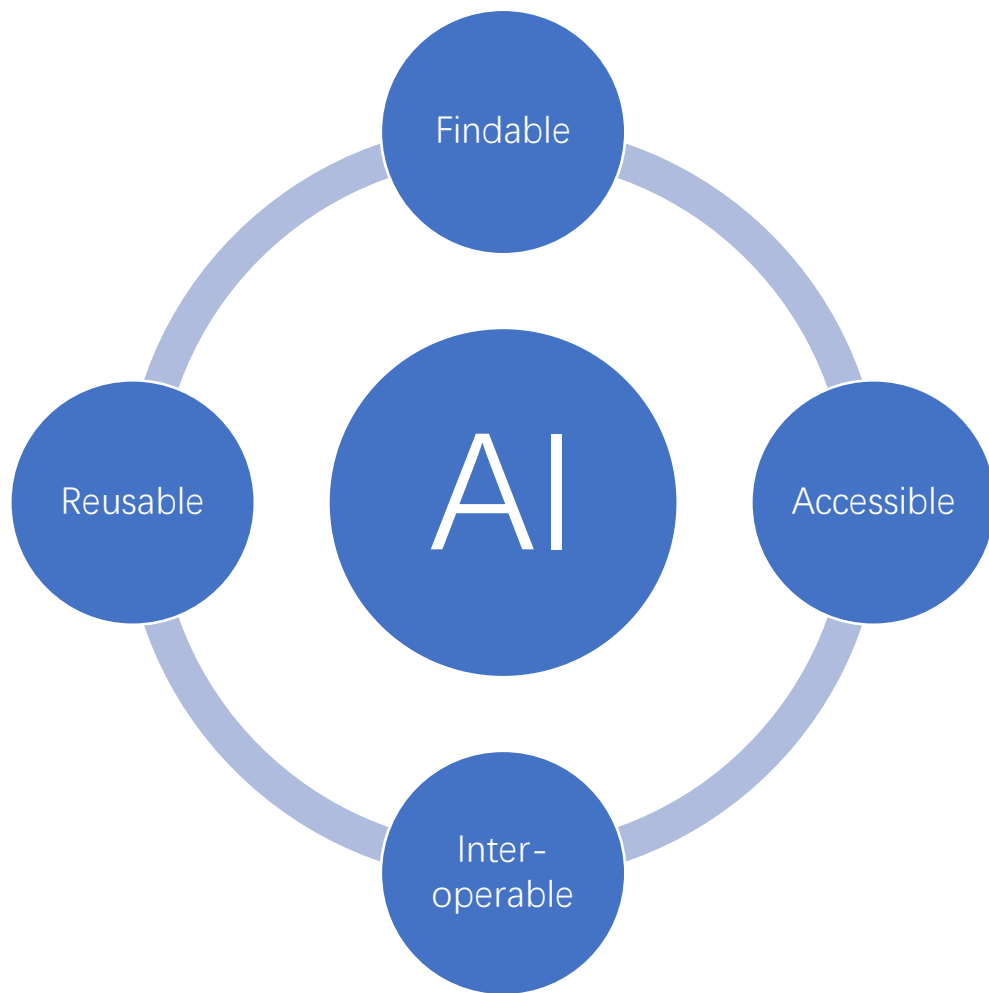
RSICD RCITMD UCM



Some planes are parked in an airport near a piece of green trees .
White buildings and tennis courts are beside the gray road .
An airplanes is surrounded with some cars .



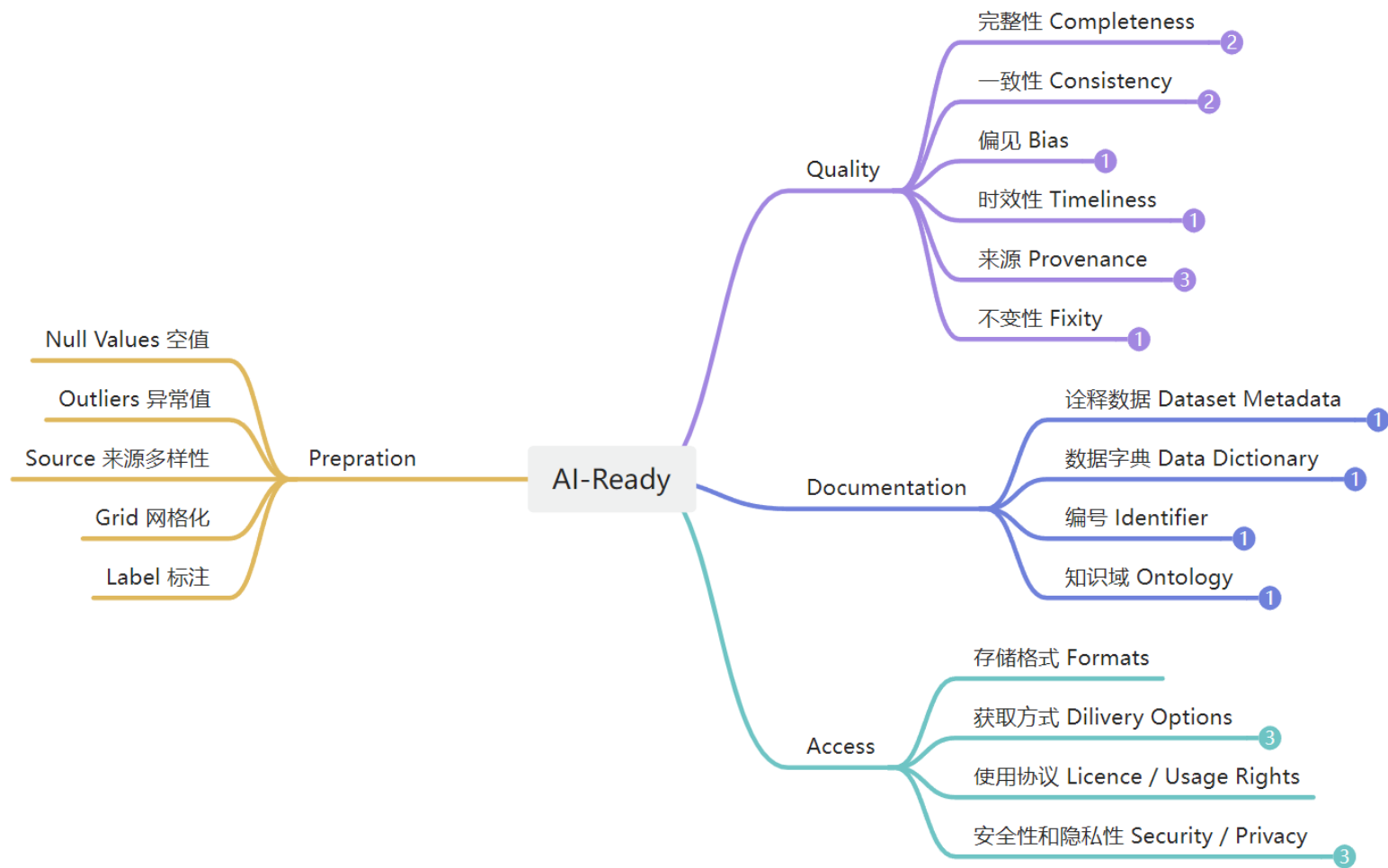
遥感领域 AI 数据存在的困境



- 对地观测领域训练数据集现状
 - 高质量数据缺乏且用户难以获取;
 - 缺少数据管理标准导致异质性数据集;
 - 数据集的发现性和互操作性低;
 - 缺少数据集生命周期的教程;

Open Geospatial Consortium, OGC Testbed-18, 2022

AI Readiness 概念发展



二、AIRD 概念发展

数据质量是 AI Ready 地球科学数据的核心



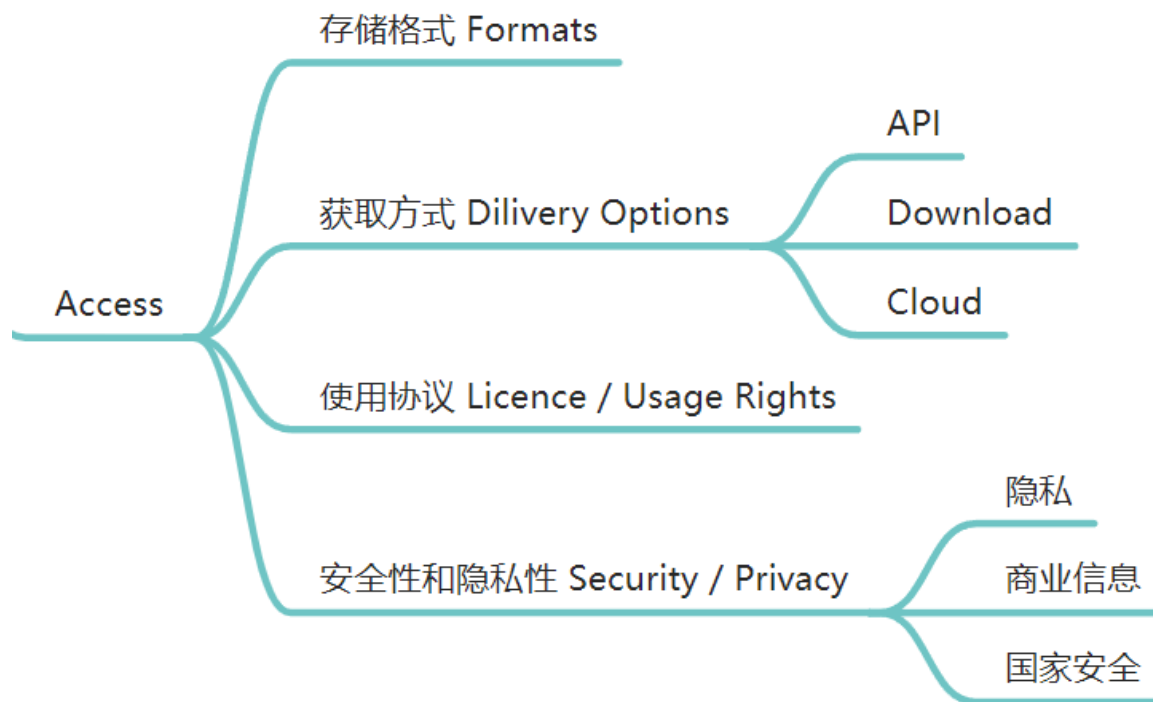
二、AIRD 概念发展

数据文档关乎用途和互操作性



二、AIRD 概念发展

数据获取是实现 FAIR 的关键



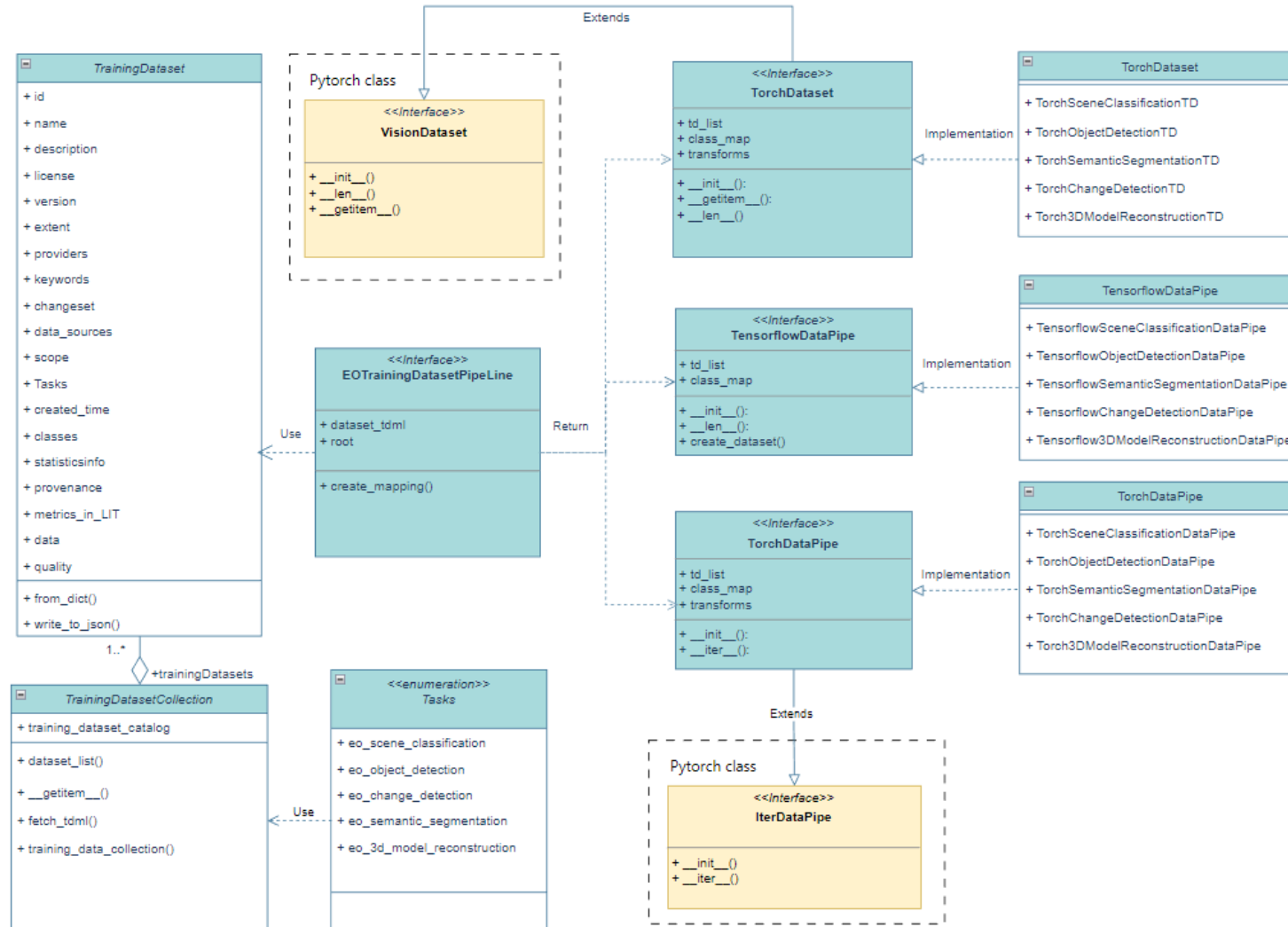
OGC 标准: TrainingDML-AI 标记语言



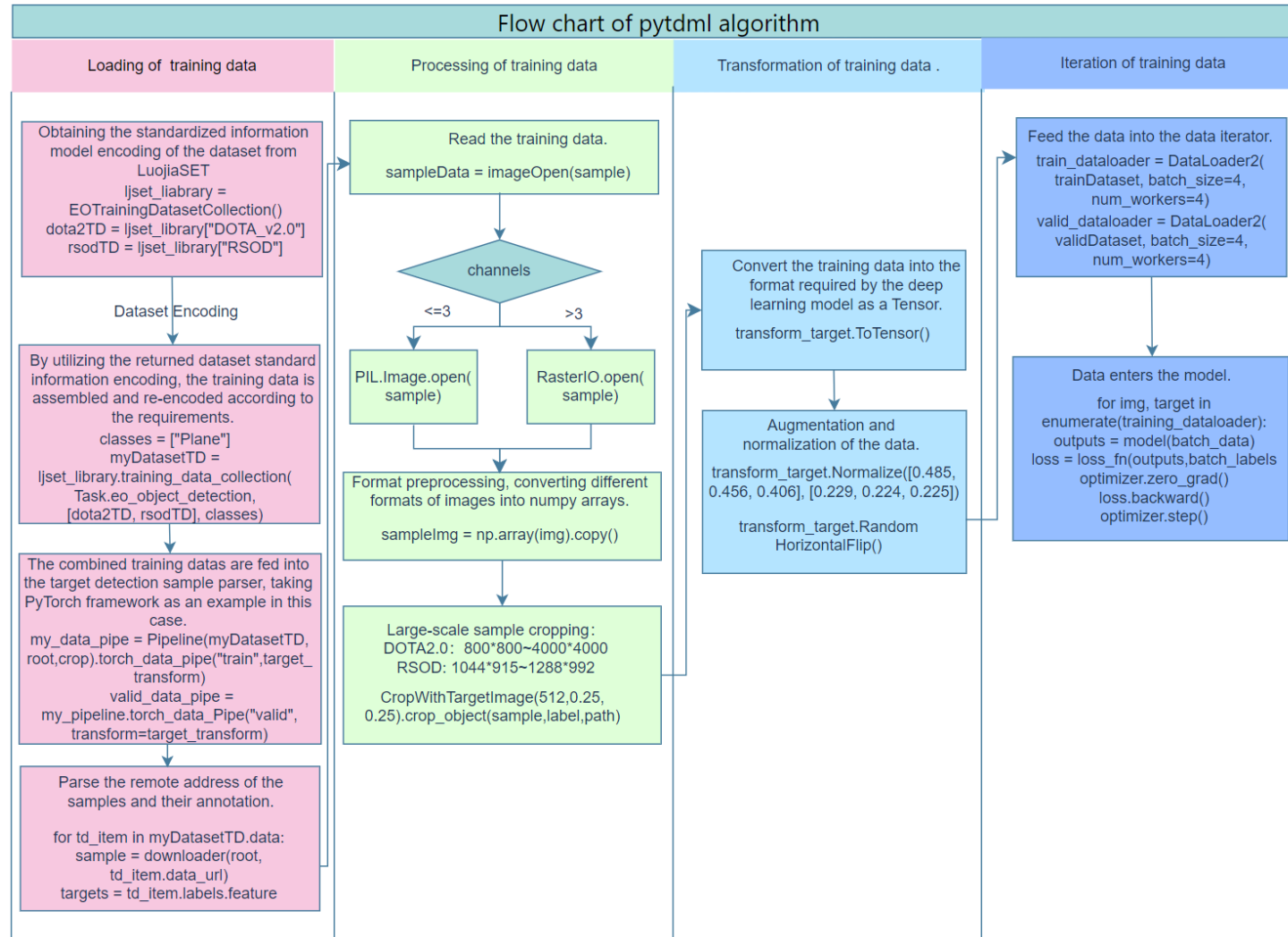
三、遥感AIRD 标准建立



标准化工具：pyTDML



标准化工具: pyTDML



遥感 AIRD 的内在思考

- AIRD和FAIR的差异性在于AI

FAIR 原则

- Findability
- Accessibility
- Interoperability
- Reusability

+ AI

- 遥感 AIRD 的科学性来自于遥感数据的特征
 - 地理位置
 - 时相
 - 空间分辨率
 - 传感器
 -



Figure 4 – Core concepts.

The full overview of concrete classes and attributes are presented in Figure 5. Concepts related to the EO AI/ML applications are defined as classes extended from abstract classes. Each core concept with related classes will be described in the rest subsections.

Earth Observation Training Data Lab (EO-TDL)

- 维护机构：ESA
- 数据集数量：58
- 数据获取方式：API, CLI, UI等
- 数据文档：自定义标准
- 数据集质量分级：
 - Q0 无元数据
 - Q1 含有STAC元数据
 - Q2 经过自动化的QA程序
 - Q3 经过人工整理

The screenshot displays the 'Datasets' page of the Earth Observation Training Data Lab (EO-TDL). The interface is dark-themed and includes a navigation bar with 'HOME', 'DATASETS', 'MODELS', 'WORKSPACE', 'TUTORIALS', 'APPLICATIONS', and 'DOCS'. A search bar and 'SIGN IN' button are also present. The main content area shows a grid of dataset cards, each with a title, creation date, and size. Filters for 'Application' (agriculture, land cover, land use, flood detection, road-extraction), 'Data' (sentinel-2, sentinel-1, airbus, sar, vector), and 'Task' (image classification, object detection, segmentation, unsupervised learning) are visible. The dataset cards include: ai4arctic-sea-ice-challenge-raw (252.51 GB), ai4arctic-sea-ice-challenge-ready-to-train (58.45 GB), AERONET (3.87 GB), test (0.97 MB), EuroSAT-RGB-small (0.33 MB), EuroSAT-Q2-small (20.53 MB), EuroSAT-Q1-small (10.27 MB), Boadella-PhiLab24 (20.44 MB), EuroSAT-small (0.63 MB), SEN12MS-CR (167.45 GB), DeepGlobeRoadExtraction (3.79 GB), MassachusettsRoadsDataset (5.76 GB), OpenEarthMap (8.47 GB), ESA-Worldcover (112.57 GB), and AlignSAR-Groningen-Sentinel1-Q0 (1.15 GB).

The EOTDL ecosystem

The EOTDL ecosystem is composed by a set of libraries, user interfaces, command line tools, and APIs. Throughout this documentation you will find instructions on how to perform different tasks with the different components, coded with the following colors.

API instructions

CLI instructions

Library instructions

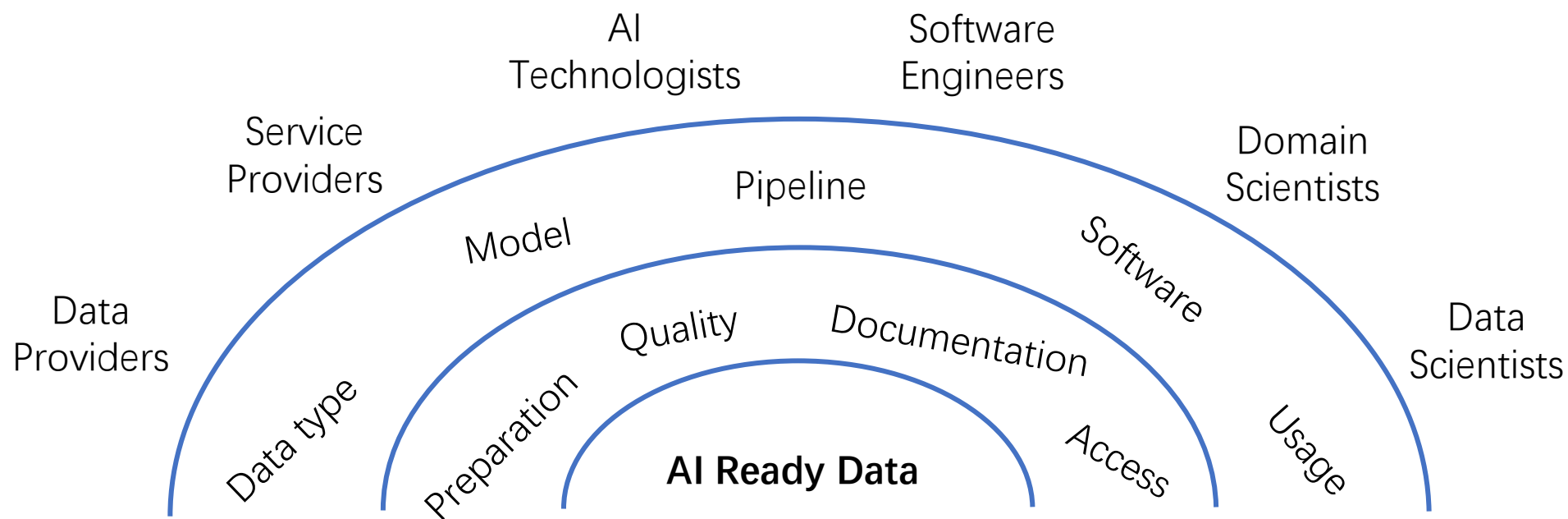
UI instructions

AIRD 未来研究方向



五、遥感 AIRD 未来展望

AIRD 未来发展方向



References

- [1] F. Liu et al., "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3390838.
- [2] Open Geospatial Consortium, "OGC Testbed-18: Call for Participation (CFP) Version 3.0," Oct. 31, 2022. [Online]. Available: https://portal.ogc.org/files/?artifact_id=100034#ML
- [3] ESIP Data Readiness Cluster (2023). Checklist to Examine AI-readiness for Open Environmental Datasets. Version 1.0. Earth Science Information Partners. <https://github.com/ESIPFed/data-readiness> [date accessed].
- [4] W. University, P. Ltd, W. Enterprises, and G. M. University, OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard.
- [5] P. Yue, R. Liu, and B. Shangguan, OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 2: JSON Encoding Standard.
- [6] Earthpulse, "EOTDL - Earth Observation Training Data Lab." Accessed: Sep. 14, 2024. [Online]. Available: <https://eotdl.com>

2024年空天数据技术与数据治理研讨会

Aerospace Symposium on Data Technology and Data Curation, DTDC 2024

11月28日 - 12月2日 · 广州



扫码注册参会!

谢谢!

国家对地观测科学数据中心

National Earth Observation Data Center