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

郑祎杰 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

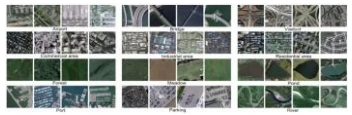
四、AIRD 遥感数据中心

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

99.54 MB



WHU-RS19 数据集
武汉大学
2011

3.24 GB

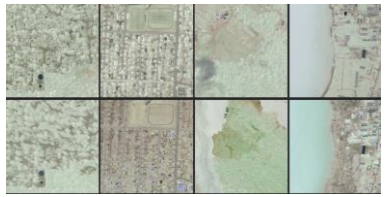
AID Dataset
华中科技大学，武汉大学
2017

6.91 GB

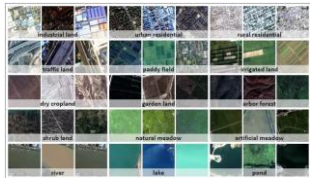
DIOR Dataset
西北工业大学
2019

30 GB

xBD Dataset
卡内基梅隆大学
2020



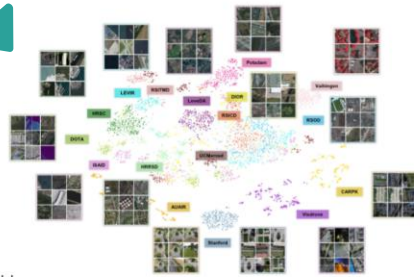
43 GB



Gaofen Image Dataset
武汉大学
2020

111 GB

RemoteCLIP
河海大学
2024



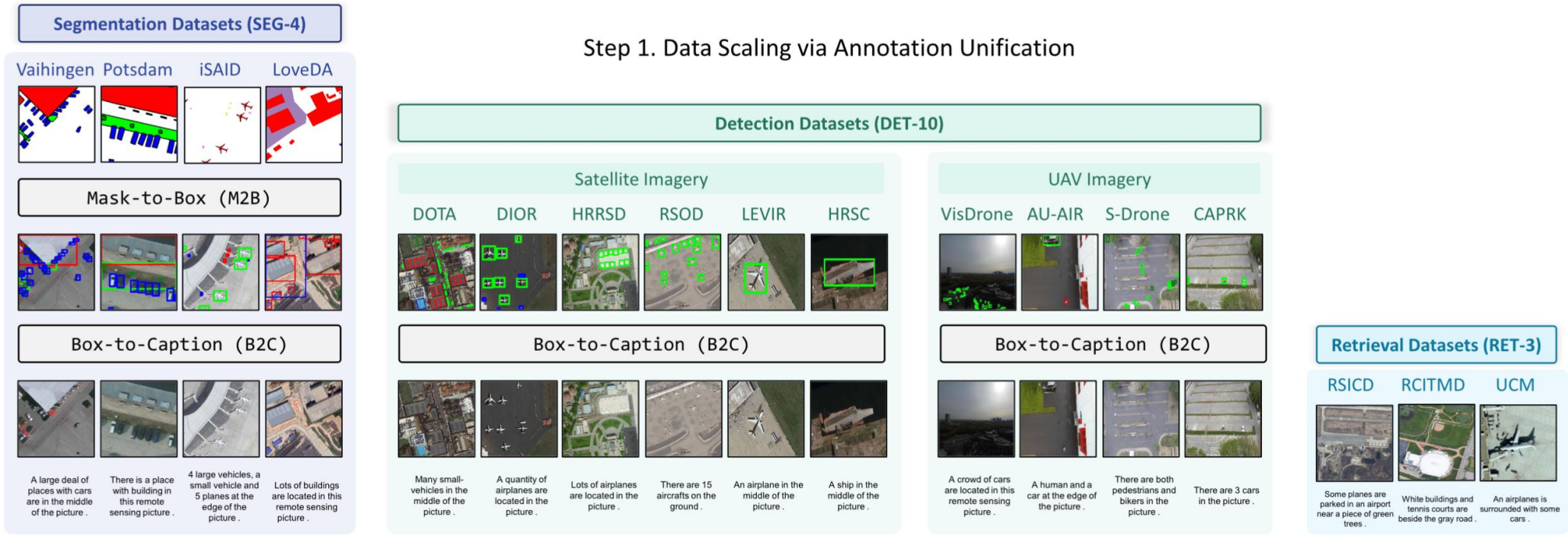
一、AI遥感数据发展现状

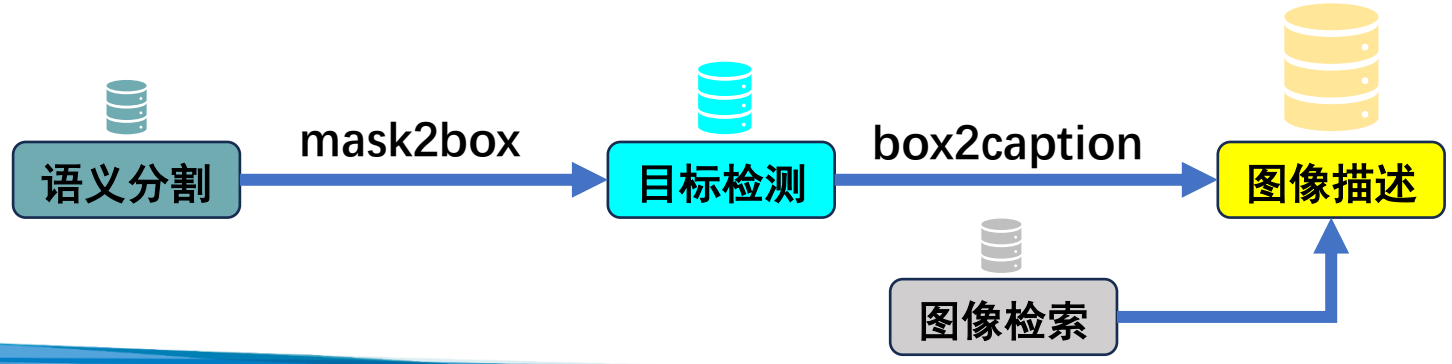
标注类型日益丰富

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

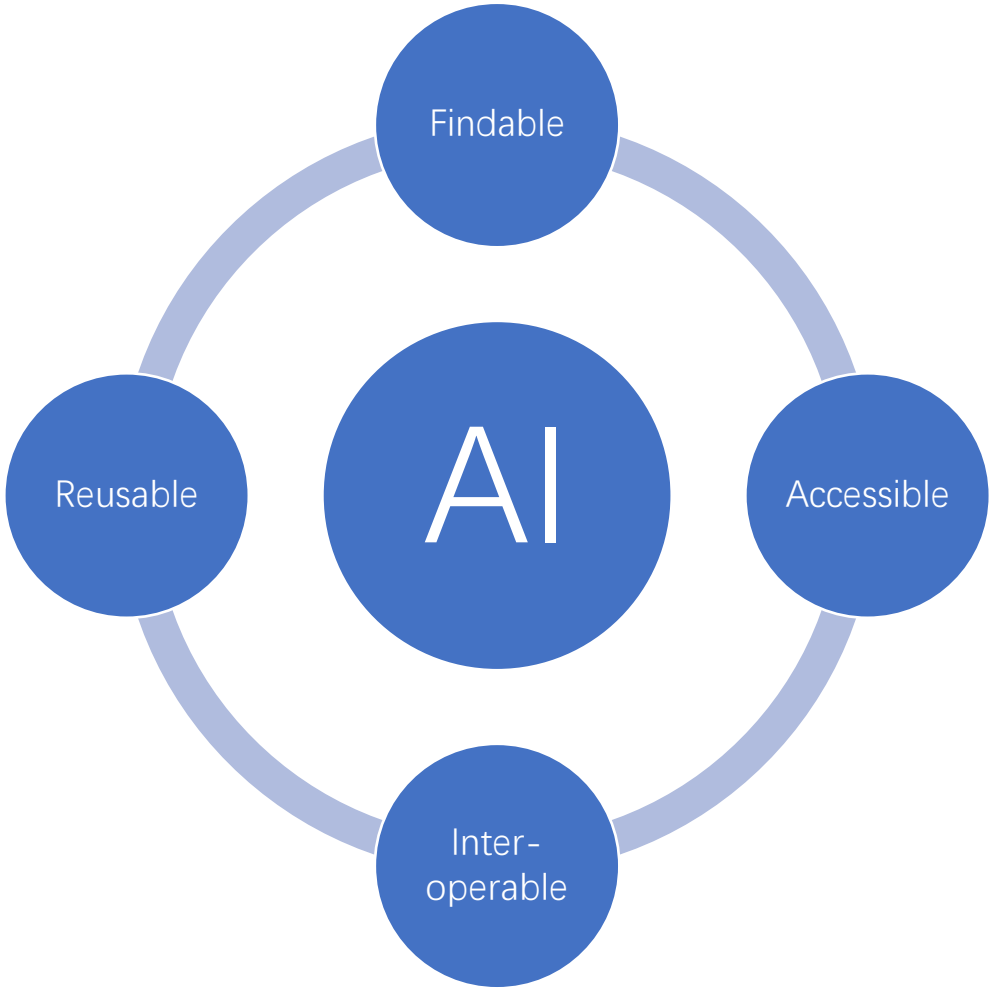




F. Liu et al., RemoteCLIP, TGRS, 2024

一、AI遥感数据发展现状

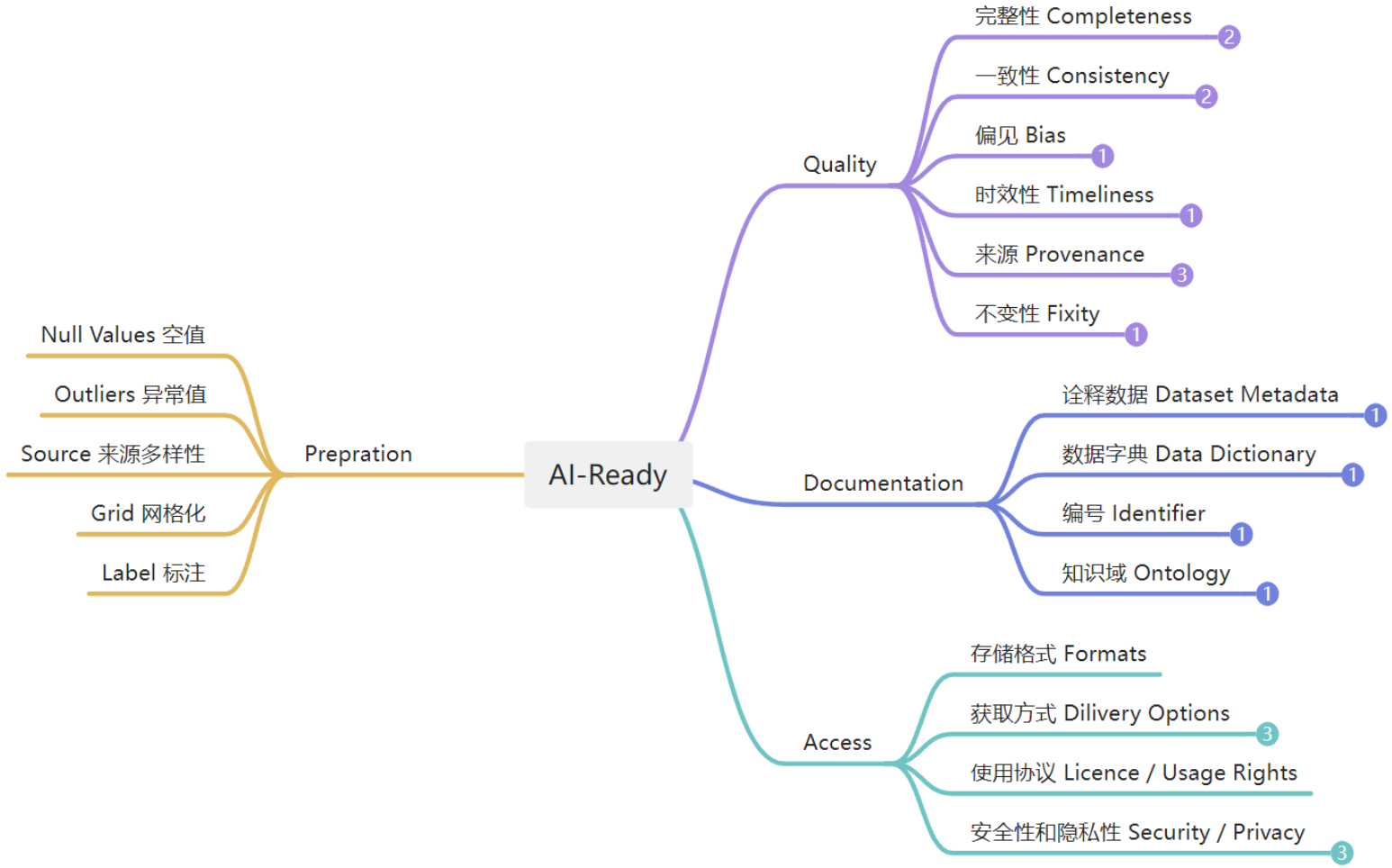
遥感领域 AI 数据存在的困境



- 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



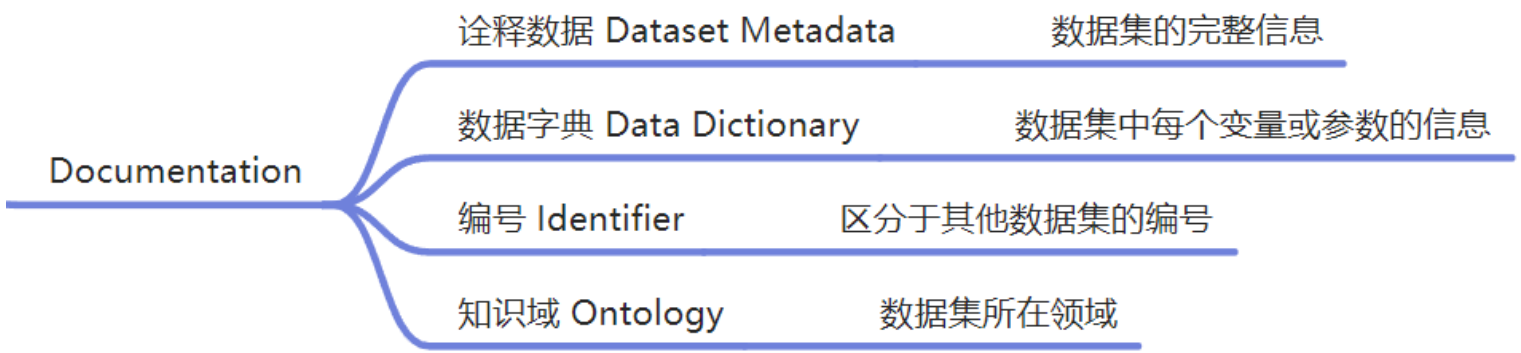
Mills, A. (2022) Are Your Data Ready? Take Stock with ESIP's New AI-Ready Checklist. Earth Science Information Partners

二、AIRD 概念发展

数据质量是 AI Ready 地球科学数据的核心

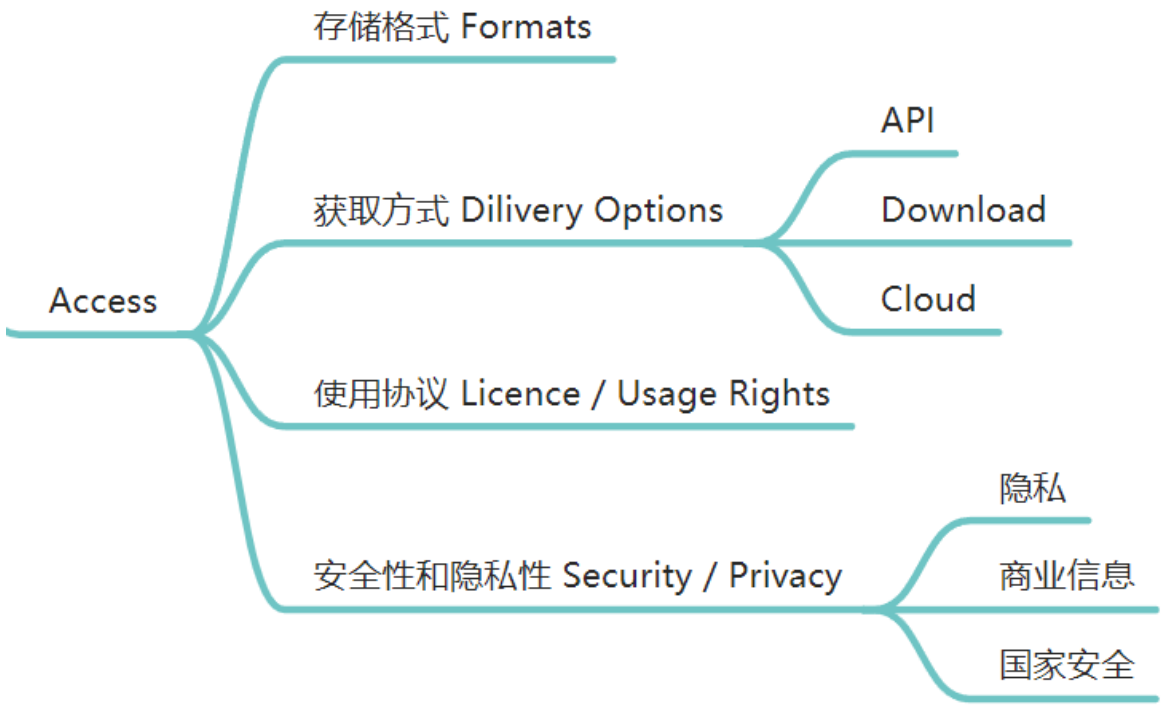
二、AIRD 概念发展

数据文档关乎用途和互操作性



二、AIRD 概念发展

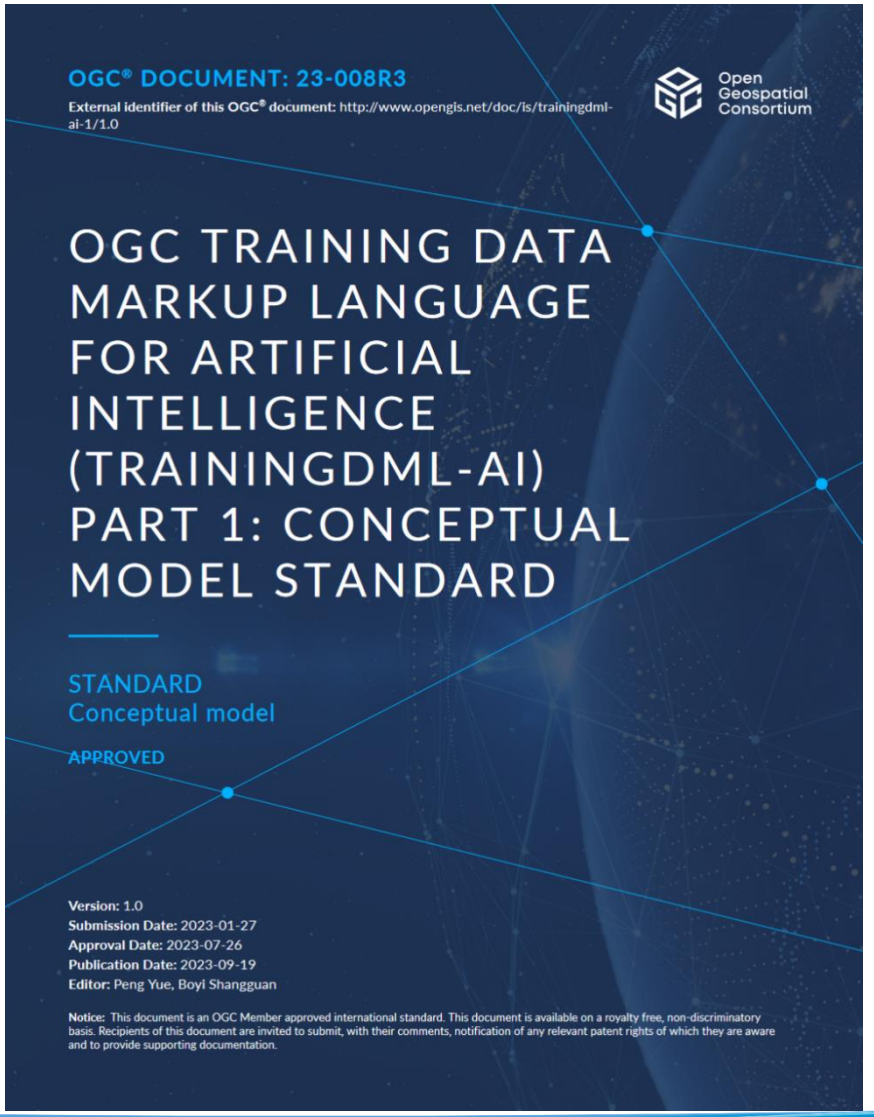
数据获取是实现 FAIR 的关键



三、遥感 AIRD 标准建立

OGC 标准：TrainingDML-AI 标记语言

- 2022.8
  - V 0.1 初稿
- 2022-2023
  - 多轮修改
- 2023.9
  - Part1 概念模型 正式发布
- 2024.4
  - Part 2 JSON 编码标准
  - Part 3 XML 编码标准
- 2024.7
  - PyTDML 发布



三、遥感AIRD 标准建立

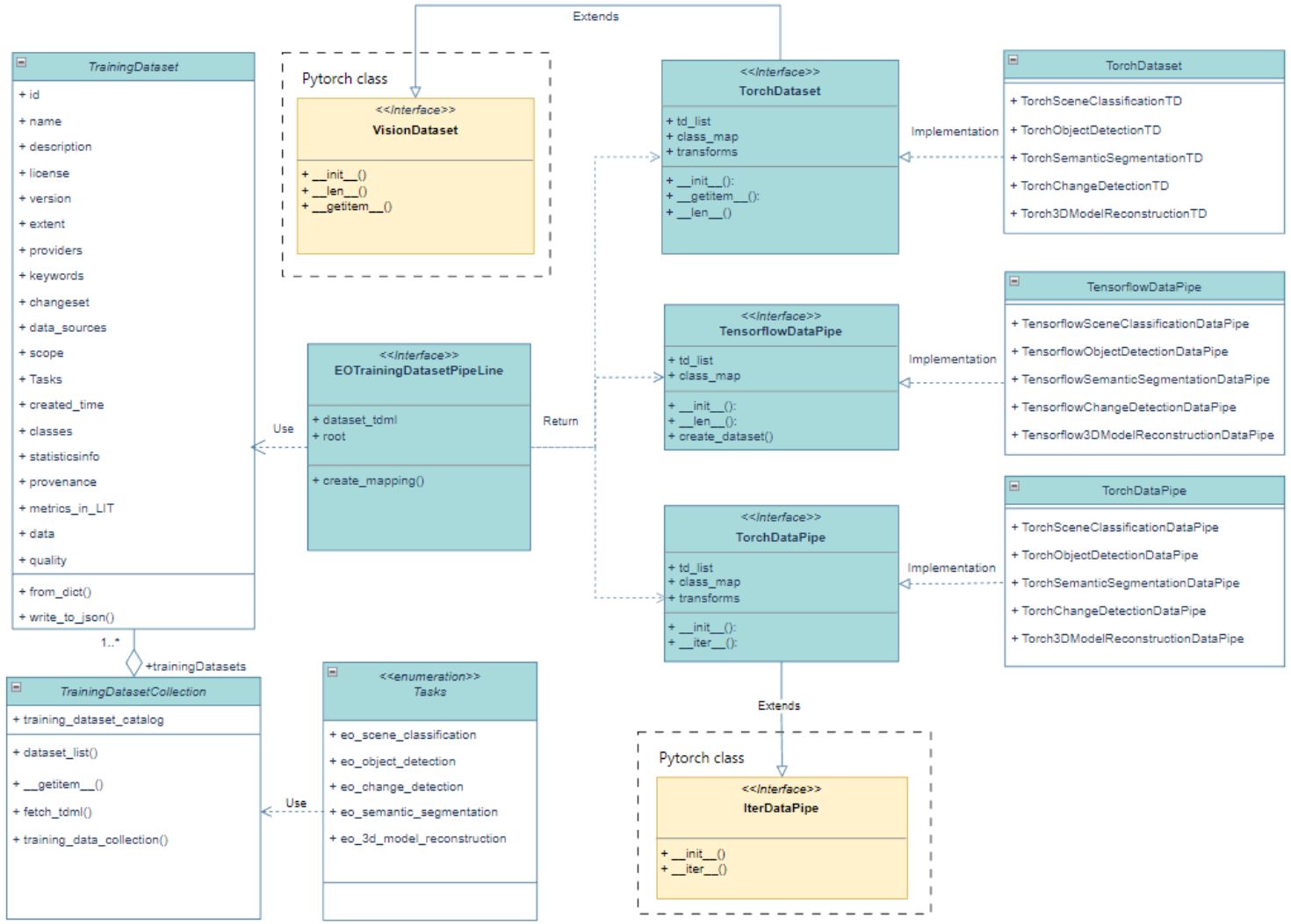
OGC 标准：TrainingDML-AI 标记语言


```
{
  "type":"EOTrainingDataset",
  "id":"whu_rs19",
  "name":"WHU-RS19",
  "description":"Wuhan University-Remote Sensing 19 Categories (WHU-RS19) has 19 classes of remote sensing images scenes obtained from Google Earth",
  "license":"CC BY-SA 4.0",
  "version":"1.0",
  "amountOfTrainingData":1013,
  "createdTime":"2010-01-01",
  "providers":[ ],
  "classes":[ ],
  "numberOfClasses":19,
  "bands":[ ],
  "imageSize":"600x600",
  "tasks":[
    {
      "type":"EOTask",
      "id":"whu_rs19-task",
      "description":"Structural high-resolution satellite image indexing",
      "taskType":"Scene Classification"
    }
  ],
  "labeling":[
    { }
  ],
  "data":[
    {
      "type":"EOTrainingData",
      "id":"airport_01",
      "dataSources":[
        "googleEarth"
      ],
      "dataURL":"image/Airport/airport_01.jpg",
      "labels":[
        {
          "type":"SceneLabel",
          "class":"Airport"
        }
      ]
    },
    { },
    { },
```

三、遥感AIRD 标准建立

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

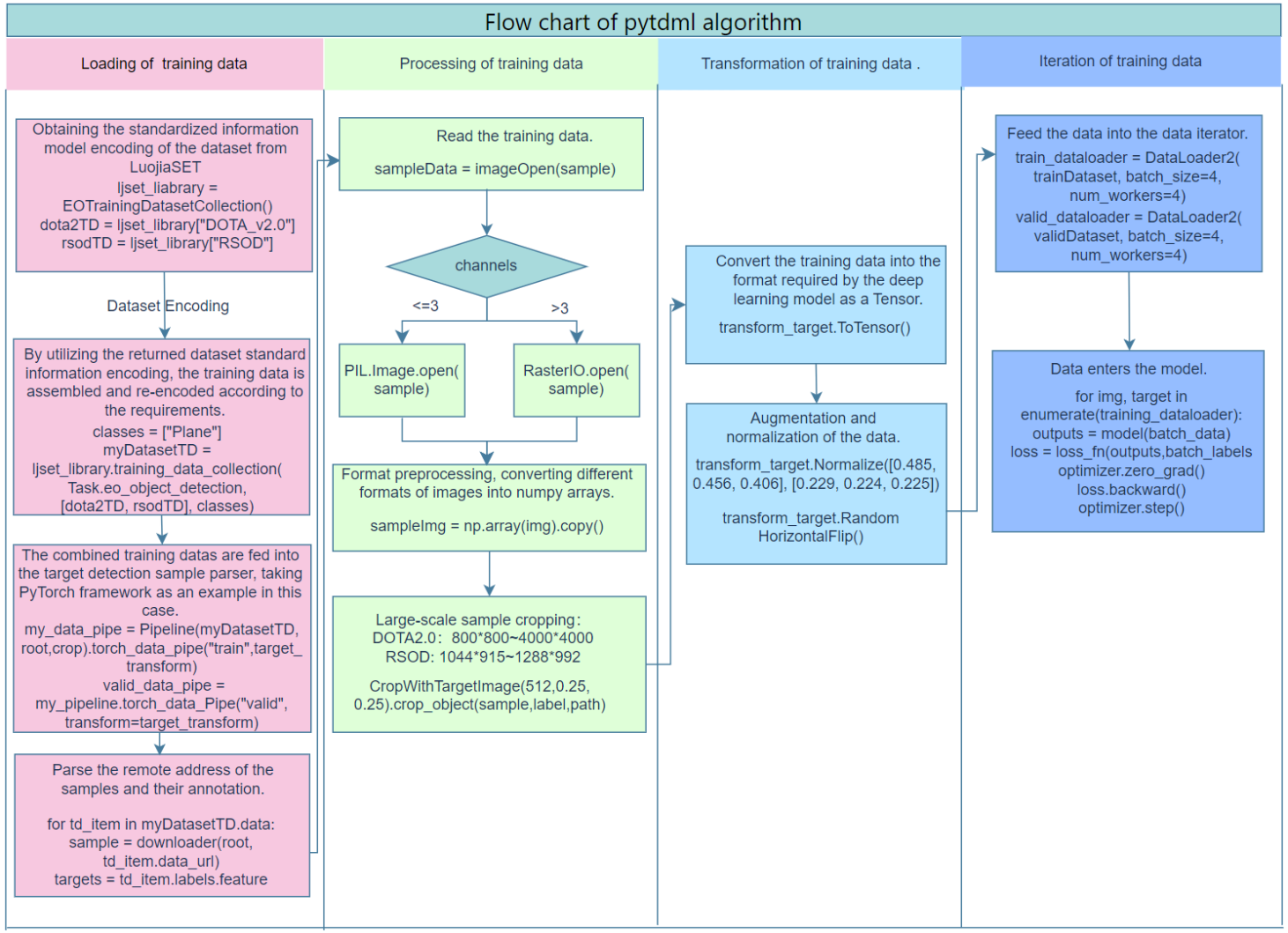
标准化工具：pyTDML



三、遥感AIRD 标准建立

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

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

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.

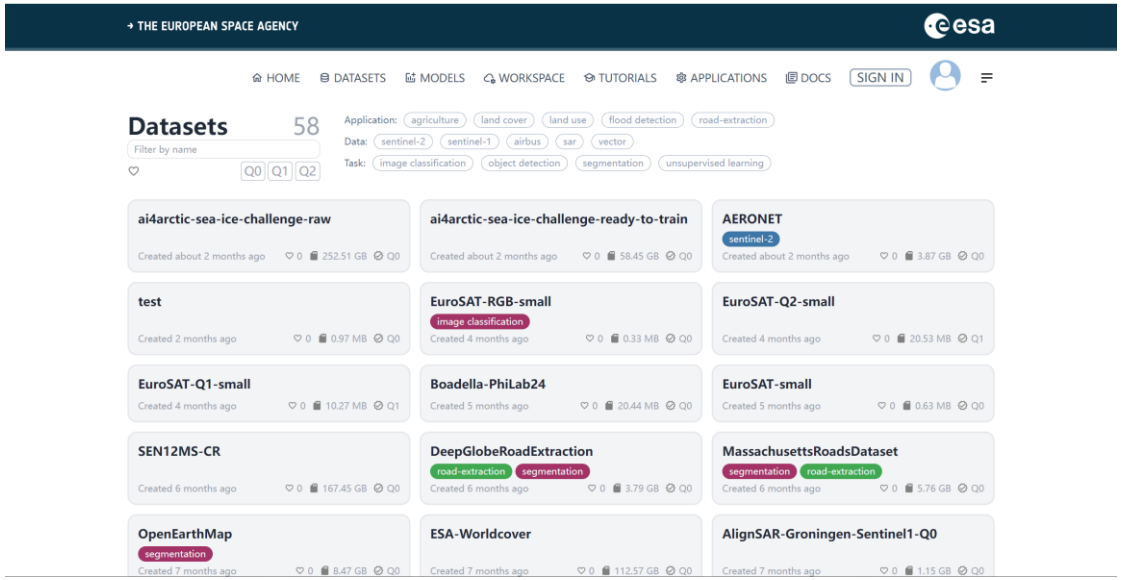
- 遥感 AIRD 的科学性来自于遥感数据的特征
  - 地理位置
  - 时相
  - 空间分辨率
  - 传感器
  - ……

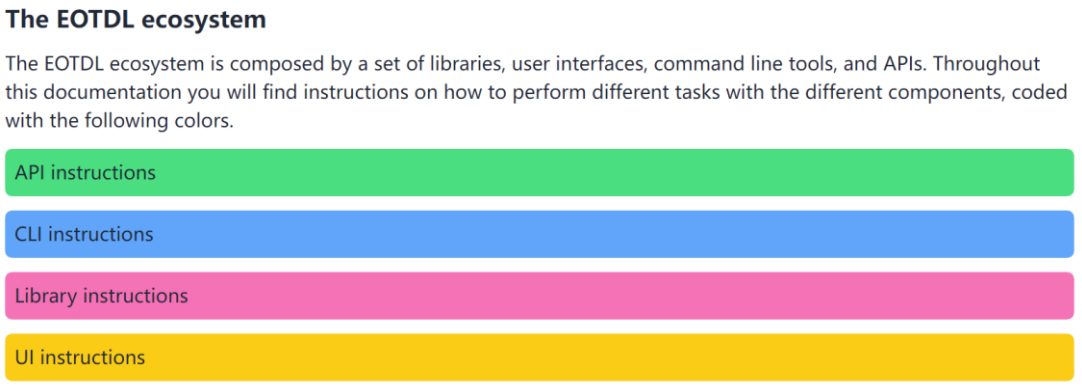
四、AIRD 遥感数据中心

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

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





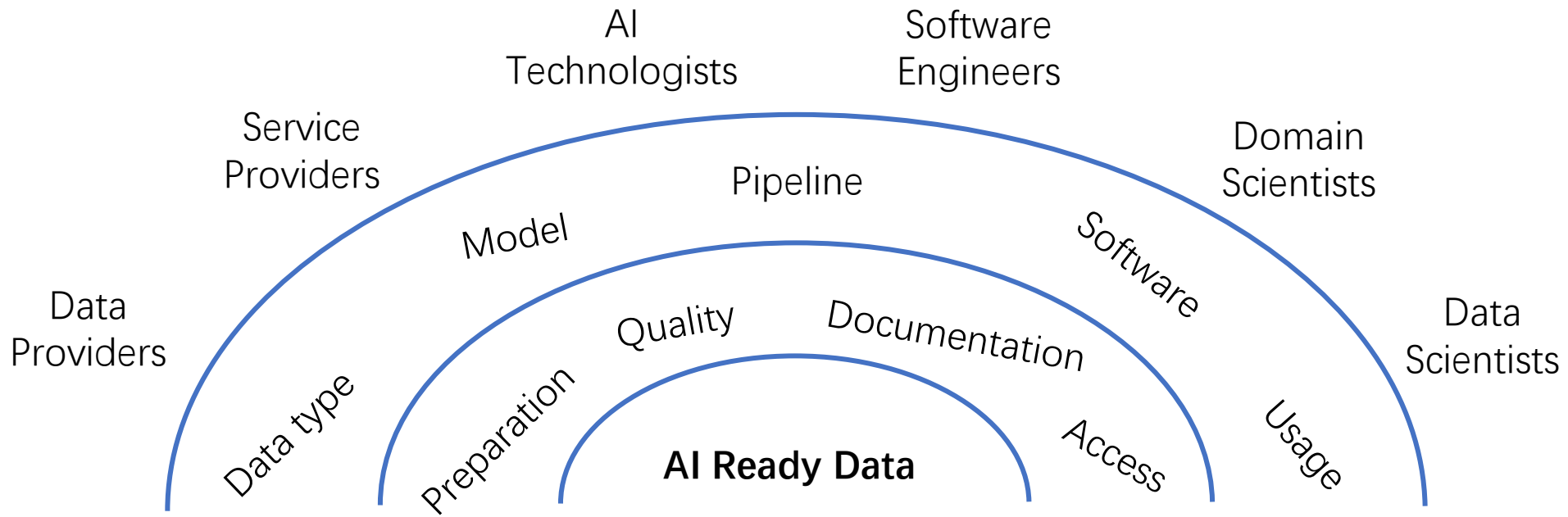
五、遥感 AIRD 未来展望

AIRD 未来研究方向

五、遥感 AIRD 未来展望

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

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