ICS 07.040
CCS A77

CH

中 华 人 民 共 和 国 测 绘 行 业 标 准

CH/T XXXX—XXXX

地理人工智能样本数据库建设规范

Specification for training sample database construction of geospatial artificial intelligence

（送审讨论稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中华人民共和国自然资源部　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中华人民共和国自然资源部提出。

本文件由全国地理信息标准化技术委员会卫星应用分技术委员会（SAC/TC230/SC3）归口。

本文件起草单位：暂略

本文件主要起草人：暂略

地理人工智能样本数据库建设规范

1 范围

本文件规定了地理人工智能样本数据库的总体要求、样本数据内容、样本库系统设计、样本库建库、样本库系统集成、样本库测试与验收以及样本库安全与运行维护等内容。

本文件适用于地理人工智能样本数据库（以下简称样本库）建设与共享服务。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 17798 地理空间数据交换格式

GB/T 18316 数字测绘成果质量检查与验收

GB/T 21336.1 地理信息 数据质量 第1部分：总体要求

GB/T 33453—2016 基础地理信息数据库建设规范

GB/T 41867 信息技术 人工智能 术语

CH/T 9007 基础地理信息数据库测试规程

3 术语和定义

GB/T 41867界定的以及下列术语和定义适用于本文件。

3.1

地理人工智能样本数据库 training sample database of geospatial artificial intelligence

用于收集、存储和管理地理人工智能机器学习/深度学习所需样本数据的数据库系统。

注：包括样本入库、浏览、查询、统计、表达、下载、更新等管理、维护与分发功能的软件和支撑环境等。

3.2

样本 training data/sample

服务于地理人工智能机器学习/深度学习模型训练、验证、测试的数据统称。

注：由地理数据和标签数据组成。其既区别于传统地学野外采集的物理样本，也不等同于纯粹的统计抽样。

3.3

标签 label

从监督学习的角度对样本数据赋予的已知或期望的值。

3.4

标注 labeling

通过全人工标注和半自动标注等方法生成样本标签的行为。

3.5

场景分类 scene classification

根据对地观测影像中拍摄内容所表达的场景信息对影像进行分类。

注：场景分类的标签是对一幅影像表达场景的理解，通常表示为某个特定的语义类别。

3.6

目标检测　object detection

从对地观测影像中识别感兴趣的目标，并确定目标对象所属语义类别和位置的过程。

3.7

土地覆盖/土地利用分类　land cover/land use classification

根据对地观测影像中拍摄的不同地物的特征,将每个像素归类为某一种具有自然属性或人工属性的地物类别。

注：与计算机视觉中的“语义分割”任务相对应。其中土地覆盖针对地物在地球表面的自然属性，土地利用针对地物在人类活动中用途。

3.8

变化检测　change detection

利用多时相获取的覆盖同一地表地区的影像，确定和分析该地区在某段时间内地表变化情况。

3.9

多视三维重建　multi-view stereo reconstruction

利用多个视角（或相机）观测物体或场景的影像还原和创建物体或场景的三维几何结构。

注：主要通过多视图影像的密集匹配来实现。

4　缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）
E-R：实体-关系（Entity-Relationship）
LC：土地覆盖（Land Cover）
LU：土地利用（Land Use）
ML：机器学习（Machine Learning）
OGC：国际开放地理信息协会（Open Geospatial Consortium）
UML：统一建模语言（Unified Modeling Language）

5　总体要求

5.1　空间参考

5.1.1　坐标系采用 2000 国家大地坐标系。必要时可采用经批准的其他坐标系，但应与 2000 国家大地坐标系建立联系。

5.1.2　高程基准采用 1985 国家高程基准。采用其他高程基准时，应与 1985 国家高程基准建立联系。

5.2　时间参考

日期应采用公历纪元，时间应采用北京时间。

5.3　数据格式

样本库应支持矢量、栅格、点云等多种数据格式，数据交换格式应符合 GB/T 17798 的要求。

5.4　安全保密

安全保密应按照 GB/T 33453 中的相关规定执行。

6 样本数据内容

6.1 样本数据集

多个样本数据单元的总体集合。根据需要可分为三个子集：训练集、验证集和测试集，三个子集的样本不重复。在没有足够样本的情况下，样本数据集也可分成两个子集：训练集和测试集，或训练集和验证集。

6.2 样本数据单元

样本数据集中的单个样本，是地理AI模型输入的最小数据单元，描述单个训练/验证/测试样本的基本属性，包括原始数据信息和对应的若干个标签信息。

6.3 标签信息

样本数据单元的标注结果，描述通过人工解译等标注行为生成的语义信息，语义信息表达样本原始数据的某种特征。一般分为场景级、目标级、像素级标签。

6.4 任务信息

描述样本数据集可用于完成AI/ML训练的任务。例如场景分类、目标检测、LC/LU分类、变化检测和多视三维重建等。

6.5 样本数据质量

描述样本数据集或样本数据单元的质量信息，包括若干个质量评估指标，以及指标的分析结果，分析结果可以是定性或定量的。记录场景级、目标级、像素级标签的完整性、逻辑一致性、专题质量、位置精度、时间质量以及元质量等。

6.6 溯源信息

描述样本生产过程的信息，包括标注者以及使用何种特定程序或方法标注样本数据。

6.7 样本变更记录集

记录样本数据集两个版本之间样本的变化，包括增加的样本、修改的样本和删除的样本。

6.8 样本数据要求

样本数据应符合以下要求：

a) 各数据内容之间的关系应符合 8.2 的要求；
b) 样本数据质量宜符合 GB/T 21336.1 的要求；
c) 样本数据编码宜按照 OGC 地理人工智能样本标记语言标准。

7 样本库建设流程

样本库建设流程见图1，主要内容如下：

a) 根据样本库用户调查和需求分析结果，进行样本库的总体设计，包括概念设计、逻辑设计、物理设计和安全设计等；
b) 根据设计要求建立集成化软硬件环境，创建样本库库体结构，将各种数据在经过入库检查和数据处理后加载到样本库中，并进行数据集成；
c) 经样本库测试验收后，进行样本库的运行、服务和维护。

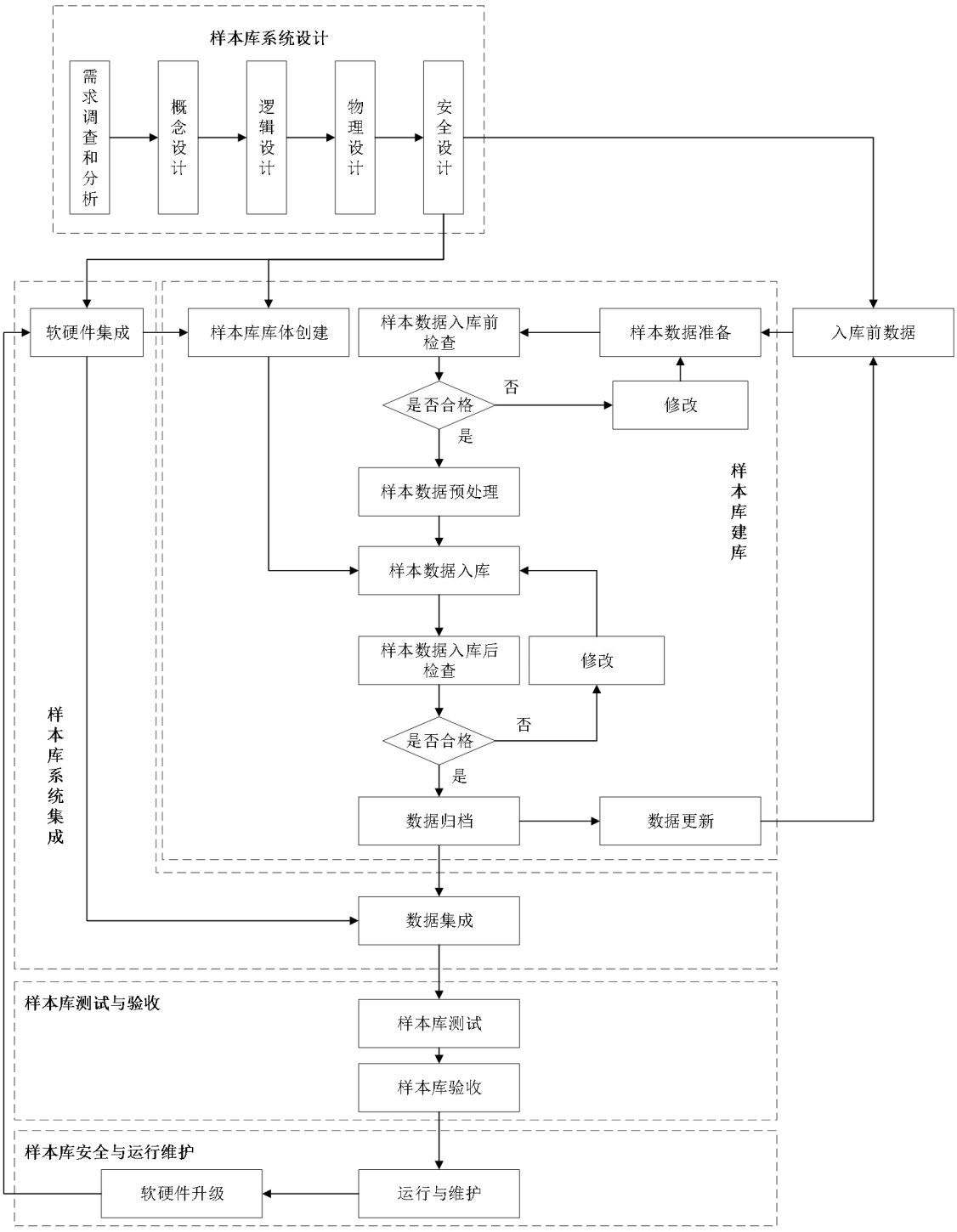


图 1　样本库建设流程图

8　样本库系统设计

8.1　需求调查和分析

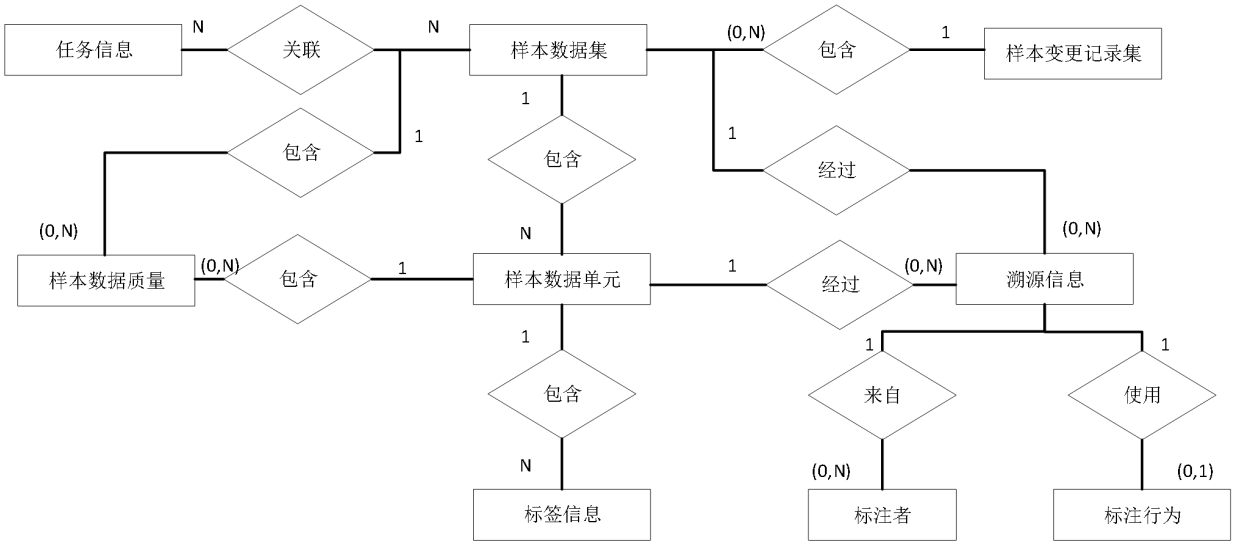
除应符合 GB/T 33453-2016 中 6.2 的规定，还应按照下列内容进行需求调查，并根据需求调查结果，编写需求分析报告：

a） 任务类型；
b） 样本分类体系；
c） 影像传感器类型；
d） 影像分辨率；
e） 数据量预测；
f） 数据覆盖范围；
g） 数据类型；
h） 数据格式；
i） 数据内容；
j） 数据质量。

8.2 概念设计

同类样本数据不同尺度、不同类别、不同传感器之间应建立明确的集成关系。概念模型应包括样本数据集、样本数据单元、标签信息、任务信息、样本数据质量、溯源信息和样本变更记录集。样本库概念模型通过 E-R 图描述，样本库概念模型见图 2。各实体之间应符合如下关系：

a） 任务信息与样本数据集：一个样本数据集对应多个任务信息，一个任务信息可属于多个样本数据集；
b） 样本变更记录集与样本数据集：一个样本变更记录集对应多个样本数据集，一个样本数据集属于一个样本变更记录集；
c） 样本数据集与样本数据单元：一个样本数据集应由多个样本数据单元组成；
d） 样本数据单元与标签信息：一个样本数据单元应对应一个或多个标签信息，标签信息应包含类别属性，类别属性应包括类别名称及对应的编码；
e） 样本数据集与溯源信息：一个样本数据集可由一个或多个标注过程组成，标注者和标注行为构成样本数据集的标注过程溯源信息；
f） 样本数据单元与溯源信息：一个样本数据单元可由多个标注过程组成，标注者和标注行为构成样本数据单元的溯源信息；
g） 样本数据集与样本数据质量：一个数据集可包含多个质量评估信息；
h） 样本数据单元与样本数据质量：一个样本数据单元可包含多个质量评估信息。



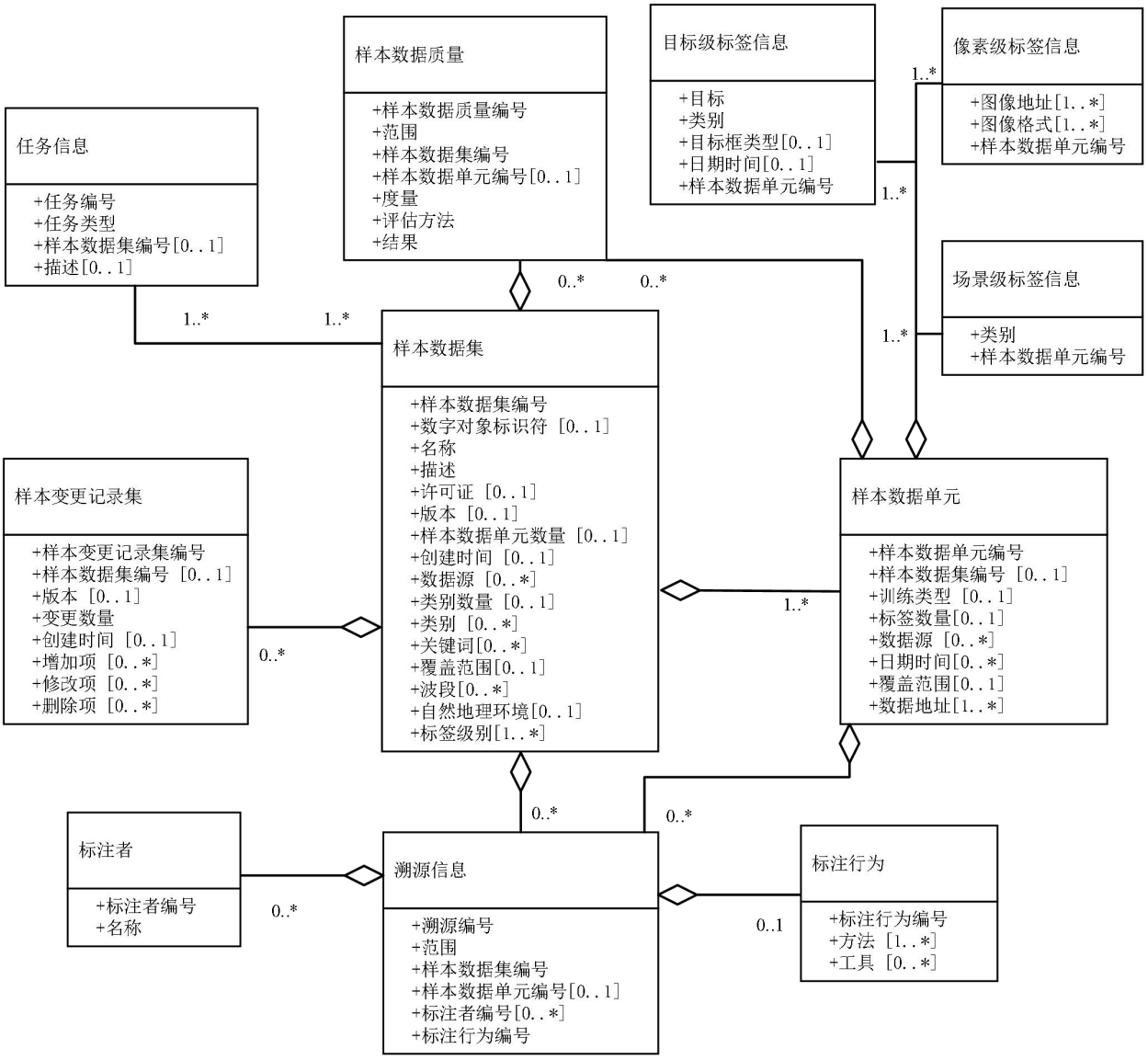
注：1 表示一个实体，N 表示多个实体，（0，1）表示零个或一个实体，（0，N）表示零个或多个实体。

图 2 样本库概念模型

8.3 逻辑设计

8.3.1 逻辑模型设计

按照 8.2 规定的概念模型设计逻辑模型，通过 UML 描述，如图 3 所示。逻辑模型的数据字典应符合附录 A 的要求，其中样本数据集信息和标签信息中的类别可以关联不同类别体系。



注：0..1 表示零个或一个实体，0..*表示零个或多个实体，1..*表示一个或多个实体，[0..1]表示零个或一个属性值，[0..*]表示零个或多个属性值，[1..*]表示一个或多个属性值。

图 3 样本库逻辑模型

8.3.2 样本数据组织

样本数据组织要求如下。

a) 原始数据组织。宜采用文件系统或数据库组织方式。

b) 标签数据组织。宜采用文件系统或数据库组织方式：

1） 场景级的标签数据以类别文本形式记录，采用文件或数据库形式组织方式；

2） 目标级的标签数据以文本或矢量形式组织，采用文件或数据库方式存放；

3） 像素级的标签数据通常以栅格形式组织，采用文件或数据库方式存放；
4） 目标级别的标签数据可通过将目标框或目标边界转换为栅格数据，从而生成像素级别的标签数据；
5） 像素级别的标签数据可通过矢量化生成目标框或目标的边界，从而形成目标级别的标签。

c） 元数据组织。应记录使用样本数据必要和可选的元数据信息，包括数据集标识、数据集适用的任务类型、类别信息、时间范围、空间范围、波段信息等内容。

8.3.3 数据关联

样本库中数据应建立如下关联：

a） 任务与样本数据集的关联：在任务信息（见表 A.6）中，通过“样本数据集编号”字段建立与样本数据集信息（见表 A.1）的关联；
b） 样本数据集与样本数据单元的关联：在样本数据单元信息（见表 A.2）中，通过“样本数据集编号”字段建立与样本数据集信息（见表 A.1）的关联；
c） 样本数据单元与标签的关联：在标签信息（见表 A.3、表 A.4 和表 A.5）中,通过“样本数据单元编号”字段建立与样本数据单元信息（见表 A.2）的关联。

8.4 物理设计

8.4.1 系统软硬件选型

系统软件选型应符合 GB/T 33453-2016 中 6.6.1 规定。系统硬件选型除应按照 GB/T 33453-2016 中 6.6.2 规定执行外，还应对总数据量进行估计，系统存储容量宜为总数据量的 5～10 倍。

8.4.2 库体结构设计

库体结构应包括：

a） 基于逻辑设计提出的模型，按照软硬件配置、数据量估算，分配样本库、软件、工作区的物理空间，确定各种数据的目录结构和存储位置；
b） 样本数据应分别设计数据表结构，样本数据表结构实现示例见附录 B；
c） 确定各数据表的数据项名称、类型、宽度和值域，选定相应的索引关键数据项。

8.4.3 索引库设计

对样本数据单元和目标级别标签建立空间索引。根据表 A.2 中的样本覆盖范围（EXTENT）字段，建立样本数据单元的空间索引。根据表 A.4 中的目标（OBJECT）字段，建立目标级标签的空间索引。

8.5 安全设计

安全设计应符合 GB/T 33453-2016 中 6.7 规定。

9 样本库建库

9.1 样本库库体创建

根据样本库的概念设计、逻辑设计和物理设计，通过数据库管理系统对每类数据进行物理空间的分配和相关参数的设置，创建数据表，建立数据表关联等，物理空间分配时应考虑样本库的扩充性。宜同时进行样本数据准备、样本数据入库前检查、样本数据预处理、样本数据集质量检查。完成后可进行样本数据入库。

9.2 样本数据准备

按照样本库设计要求，收集所需要的各类数据和资料，并整理、建档和备份，将待入库样本数据存放在专设的存储空间。主要内容如下：

a） 样本标注所用的地理空间数据，包括遥感影像和矢量数据等；
b） 包含原始数据某种特征的标签数据；

c） 与样本元数据相关的资料或文档。

9.3 样本数据入库前检查

入库前的样本数据检查应按照 GB/T 18316 的规定执行，对不合格的数据进行修改，合格后进行样本数据预处理。

9.4 样本数据预处理

样本数据入库前应进行必要的预处理，包括样本类别映射、编码转换、标签数据格式转换、坐标转换、投影转换和数据压缩等，应符合第 8 章样本库设计的各项要求。

9.5 样本数据入库

根据样本数据组织形式采用手动添加或软件程序按照如下要求和方法进行样本数据入库。

a） 样本数据集信息和任务信息按照表 A.1 和 A.6 的规定录入。

b） 遍历数据集样本单元，样本数据单元信息按照表 A.2 的要求录入，其中样本数据单元的标签信息按照任务类型分别录入：

1） 场景分类样本按照表 A.3 的规定录入场景级标签信息；

2） 目标检测样本按照表 A.4 的规定录入目标级标签信息；

3） LC/LU 分类、变化检测和多视三维重建样本按照表 A.5 的规定录入像素级标签信息。

录入形式包括但不限于以下两种：

1） 以属性形式录入；

2） 以文件地址形式录入。

c） 样本数据质量信息按照表 A.7 的规定录入，样本生产单位未提供质量信息时可不录入。

d） 溯源信息按照表 A.8 的规定录入，标注者信息按照表 A.9 的规定录入，标注行为信息按照表 A.10 的规定录入，样本生产单位未提供溯源信息时可不录入。

e） 当数据集更新时，样本变更记录集信息按照表 A.11 的规定录入，数据集没有更新时可不录入。

f） 入库完成后应记录入库日志。

9.6 样本数据入库后检查

样本数据入库后检查要求如下：

a） 样本数据是否存放在规定的数据表中；

b） 入库后样本数据是否完整；

c） 和入库前样本数据是否一致；

d） 样本数据是否重复入库；

e） 入库参数是否正确。

对不合格的数据进行修改，合格后进行入库样本数据归档。

9.7 数据归档

入库样本数据归档应符合 GB/T 33453-2016 中 7.8 的要求。

9.8 数据更新

样本库应具备增加、修改和删除样本数据的功能，还应具备保存更新记录的功能，更新后的数据应按照图1所示建库流程的规定重新入库，保证数据的准确性、时效性和安全性。

10 样本库系统集成

按照GB/T 33453-2016中第8章的规定执行。

11 样本库测试与验收

样本库测试除按照CH/T 9007的规定执行外，还应进行相关联数据的查询并符合8.3.3的要求。样本库验收按照GB/T 33453-2016中第10章规定执行。

12 样本库安全与运行维护

样本库安全与运行维护除应按照GB/T 33453-2016中第11章的规定执行外，还应包括以下内容：

a） 空间位置、类别等涉密的样本数据应采用内外网物理隔离措施；

b） 样本数据集更新应同步更新样本变更记录集信息。

附 录 A
（规范性）
样本数据字典

A.1 样本数据集信息

样本数据集信息相关数据字典见表A.1。

表 A.1 样本数据集信息

序号	名称（中文）	名称[a]（英文）	定义	主/外键	出现次数范围[b]	数据类型[c]	域[d]
1	编号	ID	样本数据集唯一标识符	主键	1	字符串型	未指定域
2	数字对象标识符	DOI	样本数据集的数字对象标识符	—	0/1	字符串型	未指定域
3	名称	NAME	样本数据集的名称	—	1	字符串型	未指定域
4	描述	DESCRIPTION	样本数据集描述信息	—	1	字符串型	自由文本
5	许可证	LICENSE	样本数据集的许可证信息	—	0/1	字符串型	未指定域
6	版本	VERSION	样本数据集的版本信息	—	0/1	字符串型	未指定域
7	样本数据单元量	SAMPLE_COUNT	样本数据集中样本数据单元的数量	—	0/1	整型	数字
8	创建时间	CREATE_TIME	样本数据集创建的时间	—	0/1	类	DateTime日期时间
9	数据源	DATA_SOURCES	原始数据来源	—	0/N	类	JSON字符串或ID或URL
10	类别数量	CLASS_COUNT	样本数据集中样本的类别的数量	—	0/1	整型	数字
11	类别	CLASSES	样本数据集中样本的类别信息	—	0/N	类别数组	未指定域，数组单元可以是类别名称，也可以是类别名称与代码构成的键值对。类别可以通过关联外部类别信息表示，不同数据集可以使用统一类别体系，也可以采用不同分类体系。
12	关键词	KEYWORDS	样本数据集的关键词	—	0/N	字符串型	未指定域
13	覆盖范围	EXTENT	样本数据集的覆盖范围	—	0/1	类	EX_Extent（ISO 19115）
14	波段	BANDS	样本遥感波段信息	—	0/N	类	MD_Band（ISO 19115）
15	自然地理环境	REGION_ENVIRONMENT	样本数据所处的自然地理环境信息	—	0/1	字符串型	未指定域
16	标签级别	LABEL_LEVEL	样本标签级别	—	1/N	字符串型	取值“SCENE”、“OBJECT”和“PIXEL”
17	场景	SCENE	场景级样本标签	—	—	字符串型	—
18	目标	OBJECT	目标级样本标签	—	—	字符串型	—
19	像素	PIXEL	像素级样本标签	—	—	字符串型	—

[a] 英文名称不应出现空格，应是多个字段通过符号“_”连接，每个字段宜全是大写或全是小写。

[b] 在出现次数范围中，最多出现1次用“0/1”表示，只出现一次用“1”表示，重复出现，但不固定次数，且最少出现1次，用“1/N”表示，如果最少可出现0次，用“0/N”表示。

[c] 在数据类型中，定义一组不同的值表示元素，例如整型、实型、字符串型、日期时间型和布尔型。

[d] 在域中，规定可赋的值。“自由文本”是用一种或多种语言表述的无限制文本信息。可用于描述字段的内容或者描述可以是任意字母数字字符集的“未指定域”。

A.2 样本数据单元信息

样本数据单元信息相关数据字典见表A.2。

表 A.2 样本数据单元信息

序号	名称(中文)	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	编号	ID	样本数据单元唯一标识符	主键	1	字符串型	未指定域
2	样本数据集编号	DATASET_ID	样本数据集唯一标识符	外键	0/1	字符串型	A.1样本数据集编号
3	训练类型	TRAINING_TYPE	样本数据单元训练的类型	—	0/1	字符串型	取值“TRAINING”、“TEST”和“VALIDATION”
4	标签数量	LABEL_COUNT	样本数据单元的标签数量	—	0/1	整型	数字
5	数据源	DATA_SOURCES	原始数据来源	—	0/N	类	JSON字符串或ID或URL
6	日期时间	DATE_TIME	原始数据生成的时间	—	0/1	类	DateTime日期时间
7	覆盖范围	EXTENT	样本数据单元的时空范围	—	0/1	类	EX_Extent（ISO 19115）
8	数据地址	DATA_URL	样本数据单元使用的原始数据存放的网络链接地址或本地文件路径列表	—	1/N	字符串型	URL，可以是绝对路径，也可以是相对路径，应满足系统可识别
9	训练	TRAINING	样本数据单元用于训练	—	—	字符串型	—
10	测试	TEST	样本数据单元用于测试	—	—	字符串型	—
11	验证	VALIDATION	样本数据单元用于验证	—	—	字符串型	—

A.3 标签信息

A.3.1 场景级标签信息

场景级标签信息相关数据字典见表A.3。

表 A.3 场景级标签信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	样本类别	CLASS	样本类别信息	—	1	字符串型	未指定域
2	样本数据单元编号	SAMPLE_ID	样本数据单元唯一标识符	外键	1	字符串型	A.2样本数据单元编号

A.3.2 目标级标签信息

目标级标签信息相关数据字典见表A.4。

表 A.4　目标级标签信息

序号	名称(中文)	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	目标	OBJECT	样本中的目标	—	1	字符串型	未指定域
2	类别	CLASS	样本类别信息	—	1	字符串型	未指定域
3	目标框类型	BBOX_TYPE	样本中目标框的类型	—	0/1	字符串型	未指定域
4	日期时间	DATE_TIME	目标的生成时间	—	0/1	类	DateTime 日期时间
5	样本数据单元编号	SAMPLE_ID	样本数据单元唯一标识符	外键	1	字符串型	A.2样本数据单元编号

A.3.3　像素级标签信息

像素级标签信息相关数据字典见表A.5。

表 A.5　像素级标签信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	图像地址	IMAGE_URL	像素分类图存放的网络链接地址或本地文件路径	—	1/N	字符串型	URL，可以是绝对路径，也可以是相对路径，应满足系统可识别。在A.1样本数据集中的“CLASSES”属性以键值对组织“类别名称”及对应“编码”。
2	图像格式	IMAGE_FORMAT	像素分类图的存放文件格式	—	1/N	字符串型	未指定域
3	样本数据单元编号	SAMPLE_ID	样本数据单元唯一标识符	外键	1	字符串型	A.2样本数据集编号

A.4　任务信息

任务信息相关数据字典见表A.6。

表 A.6　任务信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型a	域
1	任务编号	ID	任务唯一标识符	主键	1	字符串型	未指定域
2	任务类型	TASK_TYPE	地理人工智能任务的类型信息	—	1	字符串型	未指定域
3	样本数据集编号	DATASET_ID	样本数据集唯一标识符	外键	0/1	字符串型	A.1样本数据集编号
4	描述	DESCRIPTION	任务描述信息	—	0/1	字符串型	自由文本

A.5 样本数据质量信息

样本数据质量信息相关数据字典见表A.7。

表 A.7 样本数据质量信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	编号	ID	质量唯一标识符	主键	1	字符串型	未指定域
2	范围	SCOPE	质量评估作用域	—	1	类	MD_SCOPE（ISO 19115）
3	样本数据集编号	DATASET_ID	样本数据集唯一标识符	外键	1	字符串型	A.1样本数据集编号
4	样本数据单元编号	SAMPLE_ID	样本数据单元唯一标识符	外键	0/1	字符串型	A.2样本数据单元编号
5	度量	MEASURE	所使用度量的参考	—	1	字符串型	未指定域
6	评估方法	EVALUATION_METHOD	评估方法描述	—	1	字符串型	未指定域
7	结果	RESULT	评估结果	—	1	字符串型	未指定域

A.6 溯源信息

溯源信息相关数据字典见表A.8，标注者信息相关数据字典见表A.9，标注行为信息相关数据字典见表A.10。

表 A.8 溯源信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	溯源编号	ID	溯源唯一标识符	主键	1	字符串型	未指定域
2	范围	SCOPE	溯源信息作用域	—	1	类	MD_SCOPE（ISO 19115）
3	样本数据集编号	DATASET_ID	样本数据集唯一标识符	外键	1	字符串型	A.1样本数据集编号
4	样本数据单元编号	SAMPLE_ID	样本数据单元唯一标识符	外键	0/1	字符串型	A.2样本数据单元编号
5	标注者编号	LABELER_ID	标注者唯一标识符	外键	0/N	字符串型	A.9 标注者编号
6	标注行为编号	PROCEDURE_ID	标注行为唯一标识符	外键	0/1	字符串型	A.10 标注行为编号

表 A.9 标注者信息

序号	名称(中文)	名称(英文)	定义	主/外键	出现次数范围	数据类型	域
1	标注者编号	ID	标注者唯一标识符	主键	1	字符串型	未指定域
2	名称	NAME	标注者名称	—	1	字符串型	未指定域

表 A.10 标注行为信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	编号	ID	标注行为唯一标识符	主键	1	字符串型	未指定域
2	方法	METHODS	标注方法	—	1/N	字符串型	未指定域
3	工具	TOOLS	标注工具	—	0/N	字符串型	未指定域

A.7 样本变更记录集信息

样本变更记录集信息相关数据字典见表A.11。

表 A.11 样本变更记录集信息

序号	名称（中文）	名称（英文）	定义	主/外键	出现次数范围	数据类型	域
1	编号	ID	变更记录集唯一标识符	主键	1	字符串型	未指定域
2	样本数据集编号	DATASET_ID	样本数据集唯一标识符	外键	0/1	字符串型	A.1样本数据集编号
3	样本数据集版本	VERSION	样本数据集的版本信息	—	0/1	字符串型	未指定域
4	变更数量	CHANGE_COUNT	样本数据集变更的样本数量	—	1	整型	数字
5	创建时间	CREATE_TIME	变更记录集创建的时间	—	0/1	类	DateTime日期时间
6	增加项	ADD	增加的样本	—	0/N	字符串型	A.2样本数据单元编号
7	修改项	MODIFY	修改的样本	—	0/N	字符串型	A.2样本数据单元编号
8	删除项	DELETE	删除的样本	—	0/N	字符串型	A.2样本数据单元编号

附 录 B
（资料性）
样本数据表结构实现示例

B.1 样本数据集信息实现示例

样本数据集信息实现示例见表B.1，其中“CLASSES”可以通过键值对表示，也可以通过建立类别体系表，以及样本数据集与类别体系关联表，以此来表示样本数据集的类别列表。本示例通过键值对表示样本数据集的类别列表。

表 B.1 样本数据集信息实现示例

序号	ID	NAME	DESCRIPTION	LICENSE	VERSION	SAMPLE_COUNT	CREATE_TIME	DATA_SOURCES	CLASS_COUNT	CLASSES	EXTNET	LABEL_LEVEL
1	00000	WHU-RS19	WHU-RS19 拥有 19 类从 Google Earth获取的遥感影像,用于场景分类。	CC BY-SA 4.0	1.0	1013	2010-01-01	—	19	["Airport","Beach","Bridge","Commercial","Desert","Farmland","footballField","Forest","Industrial","Meadow","Mountain","Park","Parking","Pond","Port","railwayStation","Residential","River","Viaduct"]	—	SCENE
2	00001	DOTA	用于目标检测的航拍图像数据集。包含16个常见类别，2,806个图像和403,318个实例。	CC BY-SA 4.0	1.5	2806	2019-03-05	—	16	["small-vehicle","large-vehicle","plane","storage-tank","ship","harbor","ground-track-field","soccer-ball-field","tennis-court","swimming-pool","baseball-diamond","roundabout","basketball-court","bridge","helicopter","container-crane"]	—	OBJECT
3	00002	GID	高分影像数据集(GID)是利用高分二号 (GF-2) 卫	CC BY-SA 4.0	1.0	150	2018-01-01	[{"sensor":"GF-2"	5	[{"built-up":"RGB(255,0,0)" },{"farmland":"RGB(0,255	—	PIXEL

序号	ID	NAME	DESCRIPTION	LICENSE	VERSION	SAMPLE_COUNT	CREATE_TIME	DATA_SOURCES	CLASS_COUNT	CLASSES	EXTNET	LABEL_LEVEL
			星影像构建的大尺度土地覆盖数据集,具有覆盖范围大、分布广、空间分辨率高等优势					,"resolution":"2m","dataType":"optical"}]		,0)" },{"forest":"RGB(0,255,255)" },{"meadow":"RGB(255,255,0)" },{"water":"RGB(0,0,255)"}]		
4	00003	whu-building	该数据集由2012年4月获得的航拍影像组成,其中包含20.5 km2范围内12796栋建筑(2016年数据集同一区域内16077栋建筑)。	CC BY-SA 4.0	1.0	1	2019-01-01	[QuickBird,Worldview series,IKONOS,ZY-3]	1	[{"building":1}]	[15601 60,517633 8.4,15666 61.4,5179 409.2]	PIXEL
5	00004	DOTA	数据集收集了谷歌地球,GF-2卫星和航空影像。数据集中有18个常见类别,11,268个图像和1,793,658个实例。	CC BY-SA 4.0	2.0	11268	2021-02-25	[{"sensor":"GF-2","resolution":"2m","dataType":"optical"}]	18	["small-vehicle","large-vehicle","plane","storage-tank","ship","harbor","ground-track-field","soccer-ball-field","tennis-court","swimming-pool","baseball-diamond","roundabout","basketball-court","bridge","helicopter","container-crane","airport","helipad"]	—	OBJECT
6	00005	TG1HRSSC	基于天宫一号的多谱段、高空间分辨率、多时相高光谱遥感场景分类数据集。	—	1.0	204	2020-04-01	[{"sensor":"Tiangong-1","resolution":"5m,10m,	9	["city","farmland","forest","culture-pond","desert","lake","river","port","airport]	—	SCENE

序号	ID	NAME	DESCRIPTION	LICENSE	VERSION	SAMPLE_COUNT	CREATE_TIME	DATA_SOURCES	CLASS_COUNT	CLASSES	EXTNET	LABEL_LEVEL
								20m", "dataType":"hyperspectral"}]				
7	00006	WHU_MVS	多视角影像构成的用于多视三维重建的遥感影像数据集。	—	1.0	5680	2021-09-03	—	—	—	—	SCENE

B.2 样本数据单元信息实现示例

样本数据单元信息实现示例见表B.2。

表 B.2 样本数据单元信息实现示例

序号	ID	DATASET_ID	TRAINING_TYPE	LABEL_COUNT	DATE_TIME	EXTENT	DATA_URL
1	0000000	00001	TRAINING	50	—	—	dota-v1.5/image/P0000.png
2	0000001	00001	TEST	111	—	—	dota-v1.5/image/P0001.png
3	0000002	00002	TRAINING	1	—	—	GID/image/GID_1.tif
4	0000003	00003	TRAINING	1	["2012-01-01","2016-01-01"]	[15601 60,517 6338.4 ,15666 61.4,5 179409 .2]	[WHU-building/T1/WHU-building_00001.tif, WHU-building/T2/WHU-building_00001.tif]
5	0000004	00000	TRAINING	1	—	—	WHU-RS19/Airport/Airport_version1_1.jpg
6	0000005	00000	TRAINING	1	—	—	WHU-RS19/Park/Park_version1_1.jpg
7	0000006	00004	TRAINING	6	—	—	dota-v2.0/image/P0000.png
8	0000007	00005	TRAINING	1	2012-09-27	—	TG1HRSSC/Airport/Airport_version1_1.tif

B.3 标签信息实现示例

B.3.1 场景级标签信息实现示例

场景级标签信息实现示例见表B.3，场景级标签宜通过文本形式表示，建立标签与样本数据单元的对应关系。本示例以表形式表示标签信息，通过“SAMPLE_ID”建立标签与样本数据单元的关系。

表 B.3　场景级标签信息实现示例

序号	SAMPLE_ID	CLASS
1	0000004	Airport
2	0000005	Park

B.3.2　目标级标签信息实现示例

目标级标签信息实现示例见表B.4，目标级标签宜通过文本文件或矢量文件形式表示，建立标签与样本数据单元的对应关系。本示例以表形式表示标签信息，通过“SAMPLE_ID”建立标签与样本数据单元的关系。

表 B.4　目标级标签信息实现示例

序号	OBJECT	SAMPLE_ID	CLASS	BBOX_TYPE	DATE_TIME
1	{"type": "Feature", "geometry": {"type": "Polygon", "coordinates": [[906.0, 3250.0], [926.0, 3250.0], [926.0, 3263.0], [906.0, 3263.0], [906.0, 3250.0]]}}	0000000	small-vehicle	Horizontal BBox	2018-03-01
2	{"type": "Feature", "geometry": {"type": "Polygon", "coordinates": [[631.0, 4322.0], [813.0, 4322.0], [813.0, 4485.0], [631.0, 4485.0], [631.0, 4322.0]]}}	0000000	plane	Horizontal BBox	2018-03-01
3	{"type": "Feature", "geometry": {"type": "Polygon", "coordinates": [[447.0, 188.0], [456.0, 188.0], [456.0, 200.0], [447.0, 200.0], [447.0, 188.0]]}}	0000001	large-vehicle	Horizontal BBox	2018-03-01

B.3.3　像素级标签信息实现示例

像素级标签信息实现示例见表B.5。像素级标签宜通过栅格文件形式表示，建立标签与样本数据单元的对应关系。本示例以表形式表示标签信息，通过“SAMPLE_ID”建立标签与样本数据单元的关系，其中“IMAGE_URL”采用相对路径。本示例默认系统已配置根目录，可结合相对路径寻址。

表 B.5　像素级标签信息实现示例

序号	SAMPLE_ID	IMAGE_URL	IMAGE_FORMAT
1	0000002	GID/label/GID_1.png	PNG
2	0000003	WHU-building/Label/WHU-building_00001.png	PNG

B.4　任务信息实现示例

任务信息实现示例见表B.6。

表 B.6　任务信息实现示例

序号	ID	TASK_TYPE	DATASET_ID	DESCRIPTION
1	0000	场景分类	00000	高分辨率卫星图像分类
2	0001	目标检测	00001	卫星影像二维目标检测
3	0002	LC/LU分类	00002	卫星影像LC/LU分类

序号	ID	TASK_TYPE	DATASET_ID	DESCRIPTION
4	0003	变化检测	00003	航空影像建筑物变化检测
5	0004	目标检测	00003	航空影像建筑物提取
6	0005	多视三维重建	00006	高分辨率航空影像三维重建

B.5 样本数据质量信息实现示例

样本数据质量信息实现示例见表B.7。

表 B.7 样本数据质量信息实现示例

序号	ID	SCOPE	DATASET_ID	SAMPLE_ID	MEASURE	EVALUATION_METHOD	RESULT
1	0000000002	dataset	00000	—	场景类别平衡度	全检验法，统计属于各场景类别样本数据的数量，计算变异系数	0.935
2	0000000003	dataset	00001	—	对象位置平均偏移度	抽样检验法，计算500个抽样样本数据单元中对象位置与参考数据相比的中心点平均偏移度	0.023
3	0000000004	sample	00001	0000006	检测目标缺失个数百分比	全检验法，计算样本数据单元与参考数据对比，计算缺失目标数量占总目标数量的百分比	0.02%

B.6 溯源信息实现示例

溯源信息实现示例见表B.8，标注者信息实现示例见表B.9，标注行为信息实现示例见表B.10。

表 B.8 溯源信息实现示例

序号	ID	SCOPE	DATASET_ID	SAMPLE_ID	LABELER_ID	PROCEDURE_ID
1	0000000001	dataset	00002	—	[00000,00001]	000
2	0000000002	dataset	00003	—	[00000,00001,00002]	001
3	0000000003	sample	00004	0000006	[00000]	000

表 B.9 标注者信息实现示例

序号	ID	NAME
1	00000	小明
2	00001	小红
3	00002	小志

表 B.10 标注行为信息实现示例

序号	ID	MRTHODS	TOOLS
1	000	手动标注	Labelme
2	001	半自动标注	Anno-Mage

B.7 样本更新记录集信息实现示例

样本更新记录集信息实现示例见表B.11。

表 B.11　样本更新记录集信息实现示例

序号	ID	DATASET_ID	VERSION	CHANGE_COUNT	CREATE_TIME	ADD	MODIFY	DELETE
1	00000	00001	1.5	9	2019-03-05	—	[0000000, 0000001, ...]	—
2	00001	00004	2.0	8462	2021-02-25	[0000006, ...]	—	—

参 考 文 献

[1] GB/T 19710.1-2023 地理信息 元数据 第1部分：基础
[2] GB/T 41864—2022 信息技术 计算机视觉 术语
[3] IETF RFC 1738 Uniform Resource Locators (URL)

《地理人工智能样本数据库建设规范》
编制说明

行业标准项目名称：地理人工智能样本数据库建设规范

行业标准项目编号：202233009

送审行业标准名称：

（此栏送审时填写）

报批行业标准名称：

（此栏报批时填写）

承担单位：武汉大学

当前阶段：☐征求意见　☑送审稿审查　☐报批稿报批

编制时间：二〇二四年五月

地理人工智能样本数据库建设规范
编制说明

一、 概况

1.1 任务来源

2022 年 9 月 6 日自然资源部下达《自然资源部办公厅关于印发 2022 年度自然资源标准制修订工作计划的通知》（自然资办发〔2022〕39 号），本标准是自然资源部发布的 2022 年自然资源卫星应用行业标准计划项目之一，项目编号：202233009，标准计划名称《地理人工智能样本数据库建设规范》。本标准由全国地理信息标准化技术委员会卫星应用分技术委员会归口，由武汉大学牵头起草。计划周期：24 个月。

1.2 目的意义

作为新一轮科技和产业变革的核心驱动力，人工智能包括深度学习等技术已经成为国际上高新技术竞争的制高点，在我国更是上升为国家战略。地理信息和遥感是与人工智能紧密关联的领域，随着对地观测体系的不断完善和地理国情普查等工程的实施，积累形成了多类型、多粒度、多时相、多模态的遥感影像和海量的基础地理信息。应用人工智能技术实现对地观测数据的智能解译意义重大，地理人工智能技术可广泛应用于自然资源调查、生态环境监测、国防安全等诸多领域。地理人工智技术的成功需要建立在大规模、高质量的样本数据基础之上，鉴于样本的重要性，地理信息和遥感领域的学者和研究机构陆续标注和发布了许多开放的样本数据集。但目前公开的地理人工智能样本数据集分类体系各异、标注方法不同、建库方案多样，难以有效的共享和集成利用已有样本数据，迫切需要建立地理人工智能样本数据库的建设规范。

首先，建立《地理人工智能样本数据库建设规范》可以填补我国地理信息和遥感领域人工智能样本标准规范的空白。国际开放地理信息协会（Open Geospatial Consortium，OGC）启动了地理人工智能样本标准工作组（Training Data Markup Language for AI Standard Working Group，TrainingDML-AI SWG），由武汉大学牵头组织，对地理人工智能样本数据进行规范描述。在国内，2020年8月4日，国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部五部门联合印发了《国家新一代人工智能标准体系建设指南》，旨在推动人工智能产业技术研发和标准制定。我国前期立项研制了《GB/T 30319-2013 基础地理信息数据库标准》、《GB/T 33453-2016 基础地理信息数据库建设规范》等，但目前我国尚未形成针对地理人工智能样本数据的标准。

其次，该标准可以规范化指导地理人工智能样本的组织和管理，促进样本的共享利用水平。当前，现有样本库没有时空相关特性信息，不具备地理位置属性和时间属性，样本时空分布不均，削弱了模型的稳健性。现有样本库大多面向单一任务，如面向场景、目标、像素构建的样本数据集，且缺乏对样本溯源信息（包括标注人员、标注方法和标注流程等）的规范化描述，缺少对应的样本数据质量信息，使得难以对样本库进行溯源与质量评估。

通过统一的、标准化的描述，有助于不同领域研究人员和业务人员对地理人工智能样本库建设形成统一的认知基础，进而促进地理人工智能样本的高效整合与综合应用。基于这一思路，本项目拟联合国内优势的高校和企事业单位，立项制定该标准，作为我国地理人工智能样本库构建与样本数据集成应用的依据，也将成为测绘、自然资源等单位时空数据管理的基础，应用于智慧城市与自然资源环境监测的建设。

1.3 主要起草人及工作分工

编制任务下达后，武汉大学为牵头单位，武汉珞遥信息技术有限公司、广东省国土资源测绘院、国家基础地理信息中心、自然资源部国土卫星遥感应用中心、武汉理工大学、广东省国土资源技术中心、北京吉威数源信息技术有限公司、广东南方数码科技股份有限公司、重庆长安汽车股份有限公司、江苏易图地理信息科技有限公司、湖北大学共同成立了编制组。编制组成员包括总体技术负责人和长期从事卫星应用地理信息和遥感专业领域的专业技术人员和专家分工合作开展标准各章节的编写，编制组主要人员组成及分工见表 1。

表 1　编制组人员分工

序号	姓名	单位	任务分工	备注
1	乐鹏	武汉大学	本标准主编，负责组织标准编制、主要内容及意见讨论、修改及统稿定稿等工作	
2	龚健雅	武汉大学	负责通读全稿并给出修改意见	
3	刘小丁	广东省国土资源测绘院	负责标准 5 章数据内容部分，通读全稿并给出修改意见	
4	武昊	国家基础地理信息中心	负责术语部分	
5	王光辉	自然资源部国土卫星遥感应用中心	负责通读全稿并给出修改意见	
6	姜良存	武汉理工大学	负责标准内容意见讨论并给出修改意见	
7	曹志鹏	武汉大学	负责标准文稿主体内容与技术审改	
8	吴浩儒	武汉大学	负责技术要求测试验证，通读全稿并给出修改意见	
9	雷丽珍	广东省国土资源技术中心	负责标准 5 章数据内容部分	
10	黎珂	北京吉威数源信息技术有限公司	负责技术要求测试验证，通读全稿并给出修改意见	
11	张晨晓	武汉大学	负责技术要求测试验证，通读全稿并给出修改意见	
12	刘瑞祥	武汉大学	负责技术要求测试验证，通读全稿并给出修改意见	
13	王凯旋	武汉大学	负责技术要求测试验证，通读全稿并给出修改意见	
14	吴丽龙	武汉珞遥信息技术有限公司	负责技术要求测试验证，通读全稿并给出修改意见	

15	梁哲恒	广东南方数码科技股份有限公司	负责通读全稿并给出修改意见	
16	郭海京	广东省国土资源测绘院	负责标准 6.3 章内容	
17	张俊	国家基础地理信息中心	负责标准 7.1 章内容	
18	高绵新	广东省国土资源测绘院	负责技术要求测试验证，通读全稿并给出修改意见	
19	马晓黎	广东省国土资源测绘院	负责技术要求测试验证，通读全稿并给出修改意见	
20	李冯	武汉理工大学	负责技术要求测试验证，通读全稿并给出修改意见	
21	金诗程	广东省国土资源测绘院	负责技术要求测试验证，通读全稿并给出修改意见	
22	高时雨	广东省国土资源测绘院	负责技术要求测试验证，通读全稿并给出修改意见	
23	颜凯	重庆长安汽车股份有限公司	负责技术要求测试验证，通读全稿并给出修改意见	
24	孙涛	江苏易图地理信息科技有限公司	负责技术要求测试验证，通读全稿并给出修改意见	
25	张明达	湖北大学	负责技术要求测试验证，通读全稿并给出修改意见	
26	胡磊	湖北大学	负责技术要求测试验证，通读全稿并给出修改意见	

1.4 主要工作过程

1.4.1 征求意见稿阶段

2022 年 3 月，主编单位成立起草组，起草组开展了大量的调研工作，包括国内外有关现有标准，以及地理人工智能样本建设的实际实施情况，编制组开始起草标准草案。

2022 年 3 月-2022 年 8 月，主编单位就草案进行了修改，邀请参编单位和起草人召开了会议，对标准草案进行编研交流，就草案的修改和完善进行了任务分工。

2022 年 9 月-2022 年 10 月，主编单位对建议书进行修改，与参编单位和起草人进行线上讨论，于 2022 年 10 月完成标准实施方案。

2022 年 10 月- 2023 年 11 月，以标准草案为基础，编制组又以电话、社交

软件、电子邮件和视频会议的形式与地理信息和遥感领域生产作业单位、大学、科研院所的多位技术专家和生产专家进行多次交流探讨，并根据专家意见对标准草案进行修改完善，于2023年11月完成了标准征求意见稿和编制说明。

1.4.2 送审讨论稿阶段

2023年11月- 2024年5月，按照全国地理信息标准化技术委员会卫星应用分技术委员会标准化工作管理规定要求，征求意见稿发至卫星应用分技委全体委员、相关测绘单位和相关单位的专家，并在自然资源标准化信息服务平台开始广泛征求有关单位及专家的意见。收到的回函单位或专家数22个，回函并有建议或意见的单位或专家数12个。共收到113条意见，其中采纳意见87条，部分采纳意见12条，未采纳意见14条。编制组按照专家的意见对标准征求意见稿进行了详细的修改，形成送审讨论稿。

二、 标准编制原则和确定标准主要内容的依据

2.1 标准编制原则

（1）先进性

《地理人工智能样本数据库建设规范》与国际标准接轨，填补我国地理信息和遥感领域人工智能样本标准规范的空白。武汉大学牵头制定了国际开放地理信息协会OGC首个地理人工智能样本数据标准Training Data Markup Language，具有全球通用性和国际影响力。该标准通过全面而统一的描述体系，确保了不同领域在地理人工智能样本库建设方面形成了一致的认识和共识。其详尽的规范和明确的指导原则为样本库的构建提供充分的可操作性，极大地促进地理人工智能样本的集成应用。综上所述，标准符合标准编制先进性原则。

（2）适用性

《地理人工智能样本数据库建设规范》考虑不同用户需求，包括地理信息、遥感领域不同的行业，以及专业人员如科研人员、从业工作者和非专业用户等需求，确保标准适用性。如标准中逻辑模型和物理模型能够根据不同情况和需求进行调整和适应，以满足实际需求，并能适应地理人工智能不断发展的变化和需求。

（3）可操作性

《地理人工智能样本数据库建设规范》所制定的标准与当前的技术、生产或服务水平相适应。本标准进行了充分调研和实践，从地理人工智能样本数据的生产、转换入库和应用，均考虑了样本库建设的可操作性，为地理人工智能样本数据库的建设和广泛应用奠定了基础。

2.2 国内外调研情况

在国内，2020 年 8 月 4 日，国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部五部门联合印发了《国家新一代人工智能标准体系建设指南》，旨在推动人工智能产业技术研发和标准制定。我国前期立项研制了《GB/T 30319-2013 基础地理信息数据库标准》、《GB/T 33453-2016 基础地理信息数据库建设规范》等。以上工作和相关标准规范对于本标准规范的立项制定具有参考意义。

在国际上，国际标准化组织（ISO）制定了 ISO 19157:2013[7]确立了描述地理数据质量的原则，可以为地理人工智能样本质量描述提供基本原则和指导。国际开放地理信息协会 OGC 制定了《OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard》，填补了地理人工智能样本数据标准规范的空白。

2.3 主要技术内容的说明

《地理人工智能样本数据库建设规范》主要技术内容包括：

（1）样本库总体要求。

（2）样本数据内容和要求：样本数据主要包括样本数据集、样本数据单元、标签信息、任务信息、样本编码、溯源信息、样本数据质量和样本变更记录集。同时定义了 9 个基本术语，包括地理人工智能样本数据库、样本、标签、标注、场景分类、目标检测、土地覆盖/土地利用分类、变化检测、多视三维重建。

（3）样本库建设流程。

（4）样本库系统设计：包括 1）概念设计、2）逻辑设计、3）物理设计。

（5）样本库建库：主要流程包括样本库库体构建，样本数据准备，入库前检查，样本数据的预处理，样本数据入库和入库后检查。

主要技术内容详细介绍以及其主要依据说明如下。

1. 样本库总体要求

空间参考：坐标系采用 2000 国家大地坐标系。必要时可采用经批准的其他坐标系，但应与 2000 国家大地坐标系建立联系。高程基准采用 1985 国家高程基准。采用其他高程基准时，应与 1985 国家高程基准建立联系。

时间参考：日期应采用公历纪元，时间应采用北京时间。

数据格式：样本库应支持矢量、栅格、点云等多种数据格式，数据交换格式应符合 GB/T 17798-2007 的要求。

安全保密应符合 GB/T 33453-2016 中 4.5 的要求。

样本库总体要求主要依据：

1）其中建设原则主要依据是遵循国家标准《GB/T33453-2016 基础地理信息

数据库建设规范》中 4.1 建设原则。

2) 其中数据格式主要依据是科学数据管理和指导原则 FAIR（Findable, Accessible, Interoperable, Reusable）原则，即可发现、可访问、可互操作和可重用[1]。此外，武汉大学主持国家自然科学基金委重大研究计划集成项目“大规模遥感影像样本库构建及开源遥感深度网络框架模型研究”构建了 LuoJiaSET 大规模遥感样本库（以下简称“LuoJiaSET”）。通过编写代码程序，使用 Python 编程语言构建软件，实现样本库数据格式和入库数据格式转换，如图 1 所示。

图 1 数据格式转换软件界面

2. 样本数据内容和要求：样本数据主要包括样本数据集、样本数据单元、标签信息、任务信息、溯源信息、样本数据质量和样本变更记录集。

（1） 样本数据集：多个样本数据单元的总体集合。根据用途分为训练集、验证集和测试集，三个子集的样本不重复。在没有足够样本的情况下,样本数据集可分成两个子集：训练集和测试集，或训练集和验证集。

（2） 样本数据单元：样本数据集中的单个样本，是深度学习模型输入的最小

数据单元，描述单个训练/验证/测试样本的基本属性，包括原始数据信息和对应的若干个标签信息。

（3） 标签信息：样本数据单元的单个标注结果，描述通过人工解译等标注行为生成的语义信息，语义信息表达样本原始数据某种特征。一般分为场景级、目标级、像素级标签。

（4） 任务信息：描述针对整个样本数据集的任务描述。例如场景分类、目标检测、LC/LU 分类、变化检测和多视三维重建等。

（5） 样本数据质量：描述样本数据集或样本数据单元的质量信息，包括若干个质量评估指标，以及指标的分析结果，分析结果可以是定性或定量的。记录场景级、目标级、像素级标签的完整性、逻辑一致性、专题质量、位置精度、时间质量以及元质量等。

（6） 溯源信息：描述生产样本数据集中样本数据的一次标注行为的信息，记录标注者参与并使用了特定程序或方法来标注样本数据，用于帮助用户了解样本数据集的来源，提高样本数据集的可信度。

（7） 样本变更记录集：记录样本数据集两个版本之间样本的变化，包括增加的样本、修改的样本和删除的样本。

样本数据要求：样本数据应满足以下要求。

a) 各数据内容之间的关系应符合本文件 8.2 的要求；

b) 样本数据质量宜符合 ISO 19157[7]的要求；

c) 样本数据编码宜按照 OGC 地理人工智能样本标记语言标准。

样本数据内容与要求的主要依据:

1) 样本数据集主要依据是对当前已收集的开源遥感人工智能解译样本库收集、整理和转换入库。LuoJiaSET 目前共收集并转换 83 个开源样本数

据集[2]，约 520 万以上样本量，如图 2 所示。其他参与单位根据业务需求制作的样本数据集共有 30 个，约 236 万以上样本量，如表 2 所示，样本数据集包括场景分类、目标检测、地表要素分类和变化检测等多种任务类型，影像成像包括普通图像、视频关键帧、无人机影像和卫星影像，影像分辨率包括米级和亚米级分辨率，数据类型包括自然影像、多光谱影像、高光谱影像和 SAR 影像等，类别即包括自然地物、人工地物，也包括精细的农作物种类，还包括施工车辆、挖掘机、推土车和人等移动目标类别。综上所述，最终确定样本数据集是样本数据内容之一。

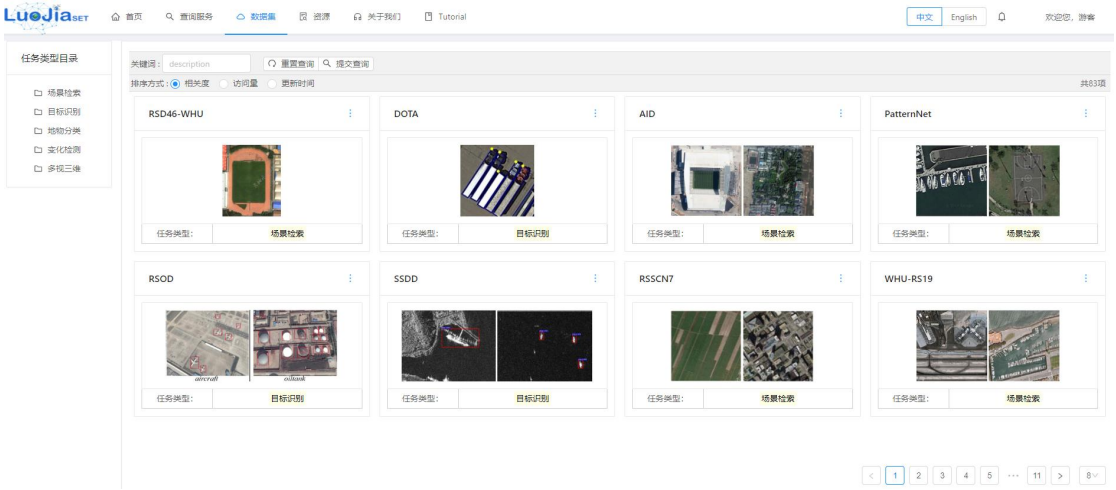


图 2 LuoJiaSET 中的样本数据库

表 2 样本数据集统计信息表

任务类型	样本数据集名称	样本量（万）	概述	制作单位
场景分类	照片场景分类样本	0.27	分别为边坡护理、不改变原用地性质的光伏用地、河堤整治、农村道路、农田水利设施、农业简易棚、农业结构调整、设施农业用地、实地未变化（包括复耕/复绿/建设用地、未利用地、推填土三个子类）、水面简易棚、土地整理。	广东省国土资源测绘院
	高尔夫识别样本	0.5	高尔夫球场	广东省国土资源技术中心
	卫片场景分类数据集	1.7	卫片场景分类数据集为使用常用卫片数据制作的场景分类数据集，类别包括停车场、高速入口、湿地公园、采矿用地等。	北京吉威
目标	视频目标检测样本	7.0	视频目标检测的预警类型分为 8 种，包括房屋建设、	广东省国土

任务类型	样本数据集名称	样本量（万）	概述	制作单位
检测			施工人员、活动板房、堆沙、堆砖头、施工车辆、挖掘机和推土车。	资源测绘院
	卫片目标检测数据集	13	卫片目标检测样本数据集为使用覆盖全国的高分系列、北京系列、高景、吉林、资源系列等多种数据源的卫星数据制作的目标检测样本集，类别涵盖尾矿库、机场、操场、大坝、桥梁、舰船、港口、油罐等十余种目标物。	北京吉威
	航片目标检测数据集	6.4	航片目标检测数据集为使用无人机数据制作的目标检测样本集，类别涵盖飞机、油罐、舰船、电线塔杆、风力发电等多种目标物。	北京吉威
	照片目标检测数据集	1.2	照片目标检测数据集为使用铁塔摄像头视频截图制作的目标检测数据集，类别涵盖工程车、坑塘、大棚、推堆土、烟火、苗木等目标物。	北京吉威
	视频目标检测数据集	865 段视频	视频目标检测数据集为使用铁塔摄像头、车载摄像头、街道摄像头等多种视频数据制作的目标检测数据集，类别涵盖小型汽车、工程机械车辆、烟火等多种动态目标。	北京吉威
	遥感建筑物检测	2.2	数据集为建筑物目标检测数据集，影像为空间分辨率为 1 米的 512*512 的切片。	南方数码
地表要素分类	全要素语义分割样本	7	分辨率为 1 米多光谱影像样本，涉及耕地、园地、林地、草地、建筑物、构筑物、道路、堆掘地表、水面、裸露地表等 10 类。	广东省国土资源测绘院
	橡胶园	3.43	4 波段、16bit、分辨率为 1 米多光谱影像样本。	广东省国土资源测绘院
	城市绿地	1.91	0.2 米航片样本，涉及树林、花圃、草地等绿地类型。	广东省国土资源测绘院
	荔枝（龙眼）	0.027	珠海 1 号 OHS 卫星高光谱地表要素分类样本，分辨率 10 米。	广东省国土资源测绘院
	耕地	0.72	分辨率为 1 米多光谱影像样本。	广东省国土资源测绘院
	全要素语义分割样本	29	涉及耕地、园地、林地、草地、建筑物、构筑物、道路、堆掘地表、水面、裸露地表、光伏板等 11 类。	广东省国土资源技术中心
	15 类地表要素分类数据集	32	15 类地表要素分类数据集为使用常用卫片数据制作的全地物类别分类数据集。	北京吉威
	专题要素提取	58	专题要素提取样本数据集为使用常用卫片数据制作的专题要素提取样本，类别涵盖耕地、林地、园地、建筑、水体、坑塘、采矿用地、光伏用地、高尔夫球场等类别。	北京吉威
	遥感耕地提取	6.40	数据集为耕地语义分割数据集，影像为空间分辨率为 1 米的 512*512 的切片。	南方数码
	遥感林地提取	5.30	数据集为林地语义分割数据集，影像为空间分辨率为 1 米的 256*256 的切片。	南方数码

任务类型	样本数据集名称	样本量（万）	概述	制作单位
	遥感水田提取	1.67	数据集为水田语义分割数据集，影像为空间分辨率为1米的256*256的切片。	南方数码
	遥感道路提取	1.05	"数据集为道路语义分割数据集，影像为空间分辨率为1米的256*256的切片。	南方数码
	遥感水体提取	2.20	数据集为水体语义分割数据集，影像为空间分辨率为1米的256*256的切片。	南方数码
	遥感建筑物语义分割	10.30	数据集为建筑物语义分割数据集，影像为空间分辨率为0.8米的256*256的切片。	南方数码
变化检测	全要素变化检测样本	5.00	分辨率为1米多光谱影像样本，涉及耕地、园地、成林、幼林、草地、建筑物、构筑物、道路、堆掘地表、水面、裸露地表等11类之间的变化512×512切片。	广东省国土资源测绘院
	SAR变化检测样本	0.19	Sentinel-1，植被-推土、植被-建筑、水域-推土、水域-建筑等变化类型。	广东省国土资源测绘院
	全要素变化检测样本	4	涉及耕地、园地、林地、草地、建筑物、构筑物、道路、推填土、水面等9类之间的变化。	广东省国土资源技术中心
	全要素变化检测	30	全要素变化检测数据集为使用常用的卫片数据，以及部分航片数据制作的自然资源调查监测所关注的所有业务变化类别数据集。	北京吉威
	专题要素变化检测	18	专题要素变化监测数据集为使用常用卫片数据制作的专题变化样本，类别包含建筑变化、林地变化、线性地物变化等类别。	北京吉威
	遥感建筑物变化检测	2	数据集为建筑物变化检测数据集，影像为空间分辨率为1米的512*512的切片。	南方数码
多视三维重建	建筑三维重建数据集	0.8	GF7三维重建数据集为使用GF7数据制作的核线影像+DSM数据制作的样本数据，目前类别主要为建筑三维重建。	北京吉威

2）样本数据单元主要依据是LuoJiaSET通过收集转换、人工标注和半自动化标注共包含520万以上样本数据单元，其中目标检测、地表要素分类和变化检测样本数据以512×512统计，可访问LuoJiaSET平台对样本数据单元进行访问[2][3]。其他参与单位标注共包含236万以上样本数据单元，如表2所示。综上所述，最终确定样本数据单元是样本数据内容之一。

3）标签信息主要依据是LuoJiaSET收集并转换的500万以上样本数据单元

包含标签信息。综上所述，最终确定标签信息是样本数据内容之一。

4) 任务信息主要依据是 LuoJiaSET 中样本数据集任务类型包含的场景分类、目标检测、地表要素分类、变化检测和多视三维重建，并对样本进行描述。综上所述，最终确定任务信息是样本数据内容之一。

5) 样本编码主要依据是 TrainingDML-AI SWG[4，5，8]，通过 Java、Python 编程语言实现样本编码，采用 JSON 编码，LuoJiaSET 通过分布式对象存储对原始地理数据/标签数据分层文件目录加元数据文件组织编码模式。综上所述，最终确定样本编码是样本数据内容之一。

3. 样本库建设流程：在确定样本库建设总体目标的基础上,根据样本库用户调查和需求分析,结合数据分析结果,进行样本库的总体设计和详细设计，包括概念设计、逻辑设计、物理设计和安全设计等;根据设计要求建立集成化软硬件环境，创建样本库库体结构,将各种数据在经过入库检查和数据处理后加载到样本库中,并进行数据集成；经系统测试后,开始样本库的运行、服务和维护。具体建设流程如图 3 所示。

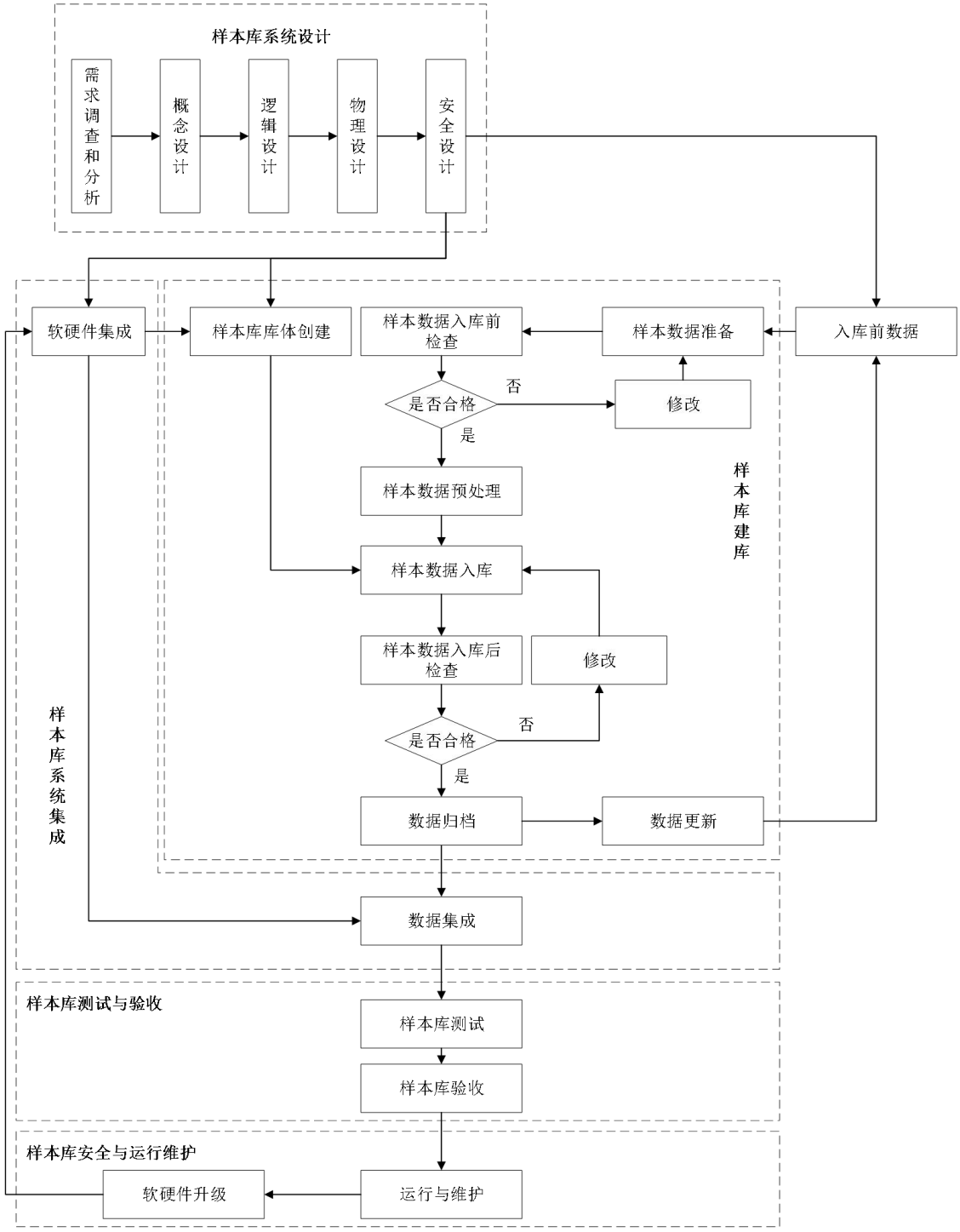
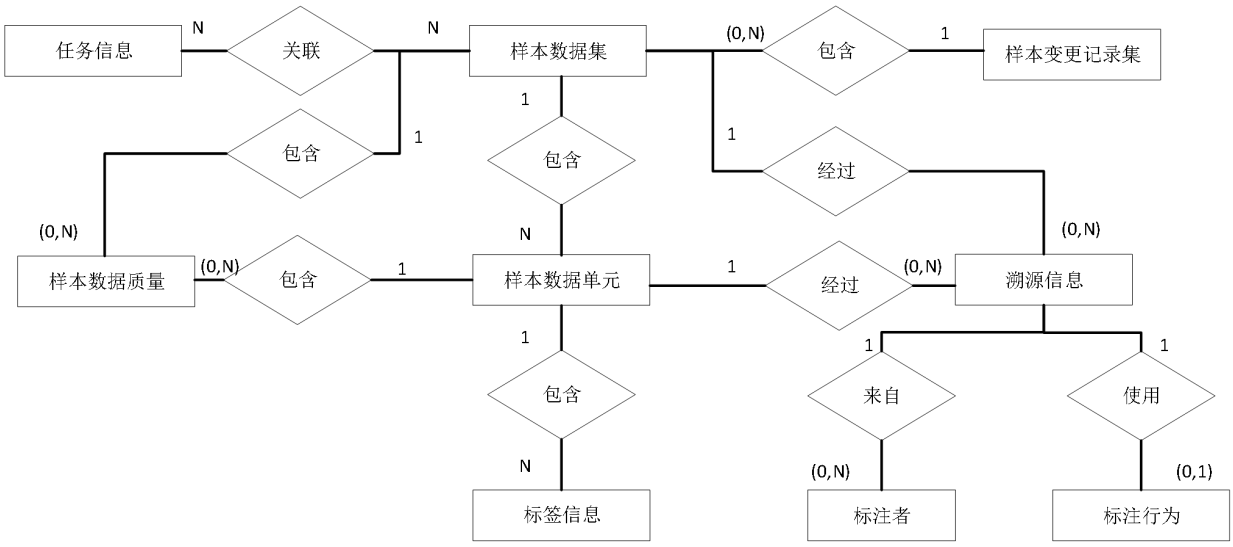


图 3 样本库建设流程图

4. 样本库系统设计：样本库系统设计应包括概念设计、逻辑设计、物理设计。

（1） 概念设计：



注：1 表示一个实体，N 表示多个实体，（0，1）表示零个或一个实体，（0，N）表示零个或多个实体。

图 4 样本库概念模型

样本库概念模型应包括样本数据集、样本数据单元、标签信息、任务信息、样本数据质量、溯源信息和样本变更记录集。样本库概念模型如图 4 所示，各实体之间的关系如下：

a) 任务信息与样本数据集：一个样本数据集对应多个任务信息，一个任务信息可属于多个样本数据集；

b) 样本变更记录集与样本数据集：一个样本变更记录集对应多个样本数据集，一个样本数据集属于一个样本变更记录集；

c) 样本数据集与样本数据单元：一个样本数据集应由多个样本数据单元组成；

d) 样本数据单元与标签信息：一个样本数据单元应对应一个或多个标签信息，标签信息应包含类别属性，类别属性应包括类别名称及对应的编码；

e) 样本数据集与溯源信息：一个样本数据集可由一个或多个标注过程组成，标注者和标注行为构成了样本数据集的标注过程溯源信息；

f) 样本数据单元与溯源信息：一个样本数据单元可由多个标注过

程组成，标注者和标注行为构成了样本数据单元的标注过程溯源信息；

g) 样本数据集与样本数据质量：一个数据集可有多个质量评估信息；

h) 样本数据单元与样本数据质量：一个样本数据单元可有多个质量评估信息，样本数据质量由若干个质量评估指标构成。

（2） 逻辑设计：包括逻辑模型设计，样本数据的组织，数据的关联。

逻辑模型设计：根据样本库概念模型设计样本库逻辑模型，如图 5 所示。

样本数据的组织：原始数据、标签数据和元数据的组织要求如下。

a）原始数据组织。宜采用文件系统或数据库组织方式。

b）标签数据组织。宜采用文件系统或数据库组织方式：

1）场景级的标签数据以类别文本形式记录，采用文件或数据库形式组织方式；

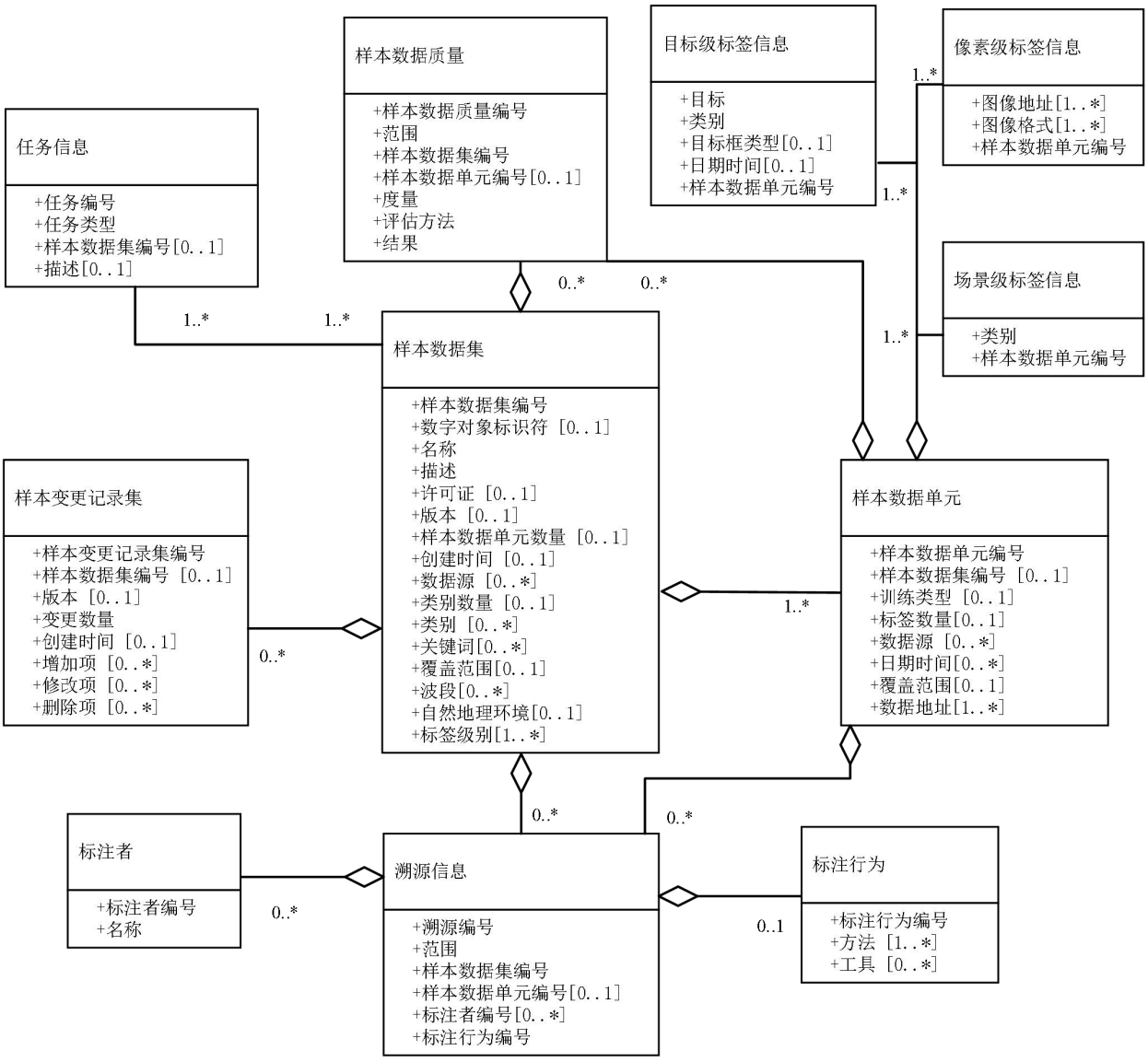
2）目标级的标签数据以文本或矢量形式组织，采用文件或数据库方式存放；

3）像素级的标签数据通常以栅格形式组织，采用文件或数据库方式存放；

4）目标级别的标签数据可通过将目标框或目标边界转换为栅格数据，从而生成像素级别的标签数据；

5）像素级别的标签数据可通过矢量化生成目标框或目标的边界，从而形成目标级别的标签。

c）元数据组织。应记录使用样本数据必要和可选的元数据信息，包括数据集标识、数据集适用的任务类型、类别信息、时间范围、空间范围、波段信息等内容。



注：0..1 表示零个或一个实体，0..*表示零个或多个实体，1..*表示一个或多个实体，[0..1]表示零个或一个属性值，[0..*]表示零个或多个属性值，[1..*]表示一个或多个属性值。

图 5 样本库逻辑模型

样本库中数据应建立如下关联：

d) 任务与样本数据集的关联：在任务信息（见表 A.6）中，通过“样本数据集编号”字段建立与样本数据集信息（见表 A.1）的关联；

e) 样本数据集与样本数据单元的关联：在样本数据单元信息（见表 A.2）中，通过“样本数据集编号”字段建立与样本数据集信息（见表 A.1）的关联；

f) 样本数据单元与标签的关联：在标签信息（见表 A.3、表 A.4 和表 A.5）

中,通过“样本数据单元编号”字段建立与样本数据单元信息（见表 A.2）的关联。

（3） 物理设计：包括软硬件选型，库体结构设计，索引库设计。

样本库建设技术要求主要依据：

6) 概念设计主要依据是 TrainingDML-AI SWG 对概念模型进行设计[4]，LuoJiaSET 概念模型按照图 4 设计。

7) 逻辑设计主要依据是 LuoJiaSET 根据样本库逻辑念模型设计样本库逻辑模型,如图 6 所示。样本数据的组织主要依据是 LuoJiaSET 实现原始地理数据、标签文件和元数据的组织形式。原始地理数据、标签文件采用分布式对象存储，如图 7 所示。数据的关联主要依据是 LuoJiaSET 通过关系数据库和分布式对象存储建立原始地理数据、标签文件与元数据的关联，如图 7 所示。

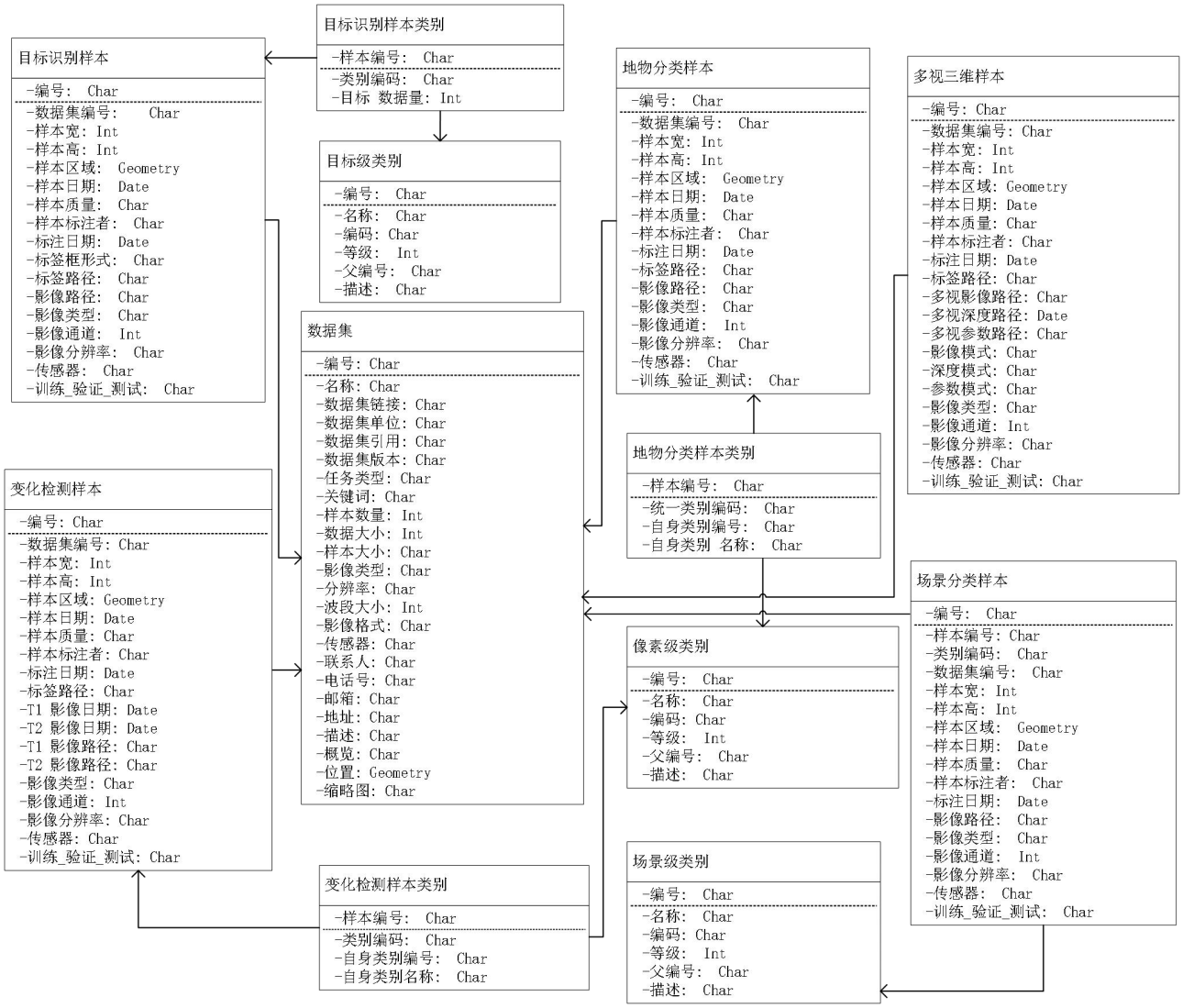


图 6 LuoJiaSET 逻辑模型

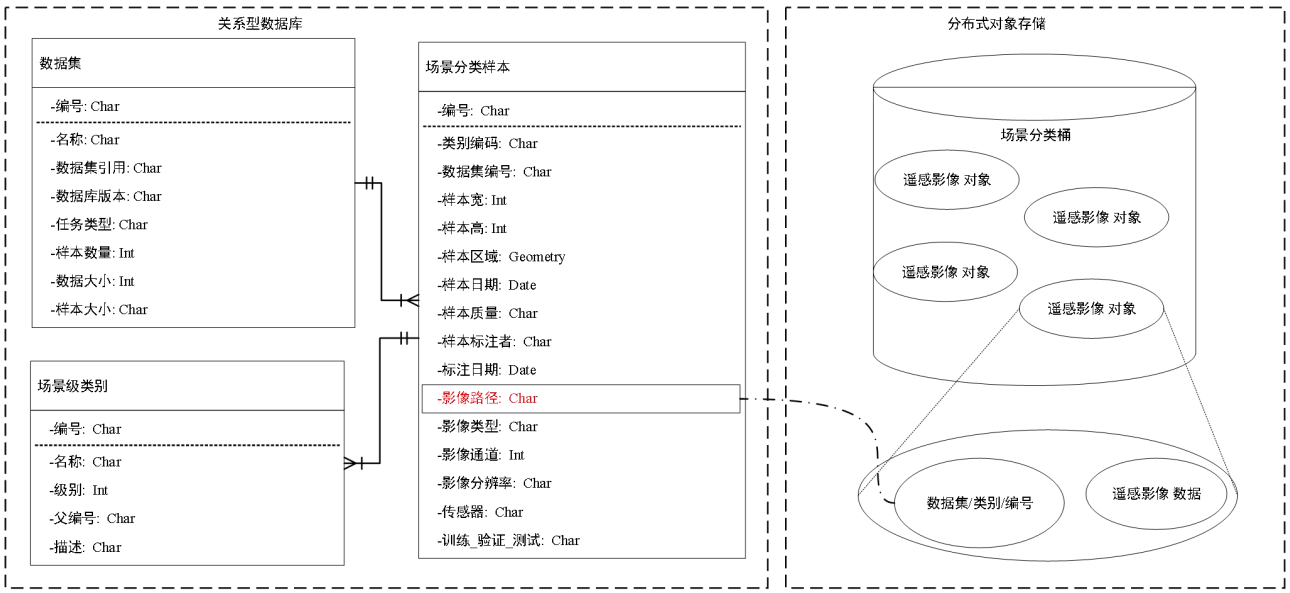


图 7 LuoJiaSET 数据组织（以场景分类样本数据组织为例）

8）物理设计：LuoJiaSET 软件选择包括 PostgreSQL，MySQL，MinIO，Tomcat

等；硬件选型包括使用三台服务器：每个服务器 24 逻辑 CPU，Intel(R) Xeon(R) CPU E5-2692 型号 CPU，2.20GHz 主频；网络带宽 1000Mb/s，使用 PostgreSQL 按照图 7 结构实现库体结构设计，以场景分类样本表为例，如表 3 所示。

表 3 场景分类样本表

序号	属性项	描述	类型	长度	取值说明	约束
1.	ID	样本编号	Text	20	数据唯一 ID，系统维护	Y
2.	CLASS_ID	类别编号	Integer	8	外键	Y
3.	SAMPLE_SIZE	样本大小	Text	30		Y
4.	SAMPLE_AREA	样本区域	Geometry	20		Y
5.	SAMPLE_DATE	样本时间	Date			Y
6.	ANOTATION_DATE	样本标注时间	Date			
7.	IMAGE_PATH	样本路径	Text	300		Y
8.	DATASET_ID	数据来源	Text	20	外键	Y
9.	IMAGE_ TYPE	影像类型	Integer	8		Y
10.	IMAGE_ RESUOLUTION	影像分辨率	Text	20		Y
11.	INSTRUMENT	传感器	Text	20		N
12.	TRN_VAL_TEST	样本用途：训练、验证、测试	Text	1	TRAIN-0; VAL-1; TEST-2;	

5. 样本库建库：主要流程包括样本库库体构建，样本数据准备，样本数据入库前检查，样本数据预处理，样本数据入库和入库后检查。

（1） 样本库库体构建：根据数据库的逻辑设计和物理设计，通过数据库管理系统对每类数据进行物理空间分配和相关参数的设置，创建数据表、建立数据表关联等，物理空间分配时应考虑数据库的扩充性。

（2） 样本数据准备：按照样本库设计的要求，收集所需要的各类数据和资料，并整理、建档和备份，将待入库样本数据存放在专设的存储空间上。主要内容如下：

a) 样本标注所用的地理空间数据，包括遥感影像和矢量数据等；

b）记录包含原始数据某种特征的标签文件；

c）记录样本元数据相关的资料或文档。

（3） 样本数据入库前检查：入库前的样本数据检查应按照 GB/T 18316 的规定执行，对不合格的数据进行修改，合格后进行样本数据预处理。

（4） 样本数据预处理：入库样本数据应按照第 8 章样本库设计要求进行一致性转换，主要包括样本类别映射、编码转换、标签文件格式转换、坐标转换、投影转换和数据压缩等。

样本数据入库：根据样本数据组织形式采用手动添加或软件程序按照如下要求和方法进行样本数据入库。

a） 样本数据集信息和任务信息按照表 A.1 和 A.6 的规定录入。

b） 遍历数据集样本单元，样本数据单元信息按照表 A.2 的要求录入，其中样本数据单元的标签信息按照任务类型分别录入：

1）场景分类样本按照表 A.3 的规定录入场景级标签信息；

2）目标检测样本按照表 A.4 的规定录入目标级标签信息；

3）LC/LU 分类、变化检测和多视三维重建样本按照表 A.5 的规定录入像素级标签信息。

录入形式包括但不限于以下两种：

1）以属性形式录入；

2）以文件地址形式录入。

c） 样本数据质量信息按照表 A.7 的规定录入，样本生产单位未提供质量信息时可不录入。

d） 溯源信息按照表 A.8 的规定录入，标注者信息按照表 A.9 的规定录入，标注行为信息按照表 A.10 的规定录入，样本生产单位未提供溯源信息时

可不录入。

e） 当数据集更新时，样本变更记录集信息按照表 A.11 的规定录入，数据集没有更新时可不录入。

f） 入库完成后应记录入库日志。

（5） 样本入库检查内容包括：样本数据是否存放在规定的数据表中、入库后样本数据是否完整、和入库样本数据是否一致、样本数据是否重复入库，对不合格的数据进行修改，合格后进行入库样本数据归档。

样本库建库主要依据是《GB/T 33453-2016 基础地理信息数据库建设规范》中 7.1 建库流程，并且 LuoJiaSET 平台结合该标准与样本数据的特点，实现建库流程，并将 520 万张样本入库。

三、 验证试验的情况和结果

1. 主要实验分析

通过前期实验和武汉大学主持的国家自然科学基金委重大研究计划集成项目“大规模遥感影像样本库构建及开源遥感深度网络框架模型研究”，构建了 LuoJiaSET 大规模遥感样本库（以下简称“LuoJiaSET”），形成了论文《遥感影像智能解译样本库现状与研究》与软著《遥感智能解译样本库平台[简称 LuoJiaSET]V1.0》。同时标准编制组制定了《OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard》[8]并顺利发布。本文件的实验结论主要来源于此。

2. 技术经济论证

相应的技术指标参考了“大规模遥感影像样本库构建及开源遥感深度网络框架模型研究”项目以及“OGC TrainingDML-AI”标准化项目。在此基础上，

经实际的生产实践和检验，明确了地理人工智能样本数据库建设规范和指标要求，从样本数据内容、样本库建设技术要求和样本库建库等方面对相关内容、技术流程与技术要求进行了规定，详见标准文本。

四、 采用国际标准和国外先进标准的程度，以及与国际、国外同类标准水平的对比情况

国外对于相关地理人工智能样本数据库的规范化非常重视，在此背景下，OGC TrainingDML-AI 标准工作组发布国际标准[8]，为地理人工智能样本数据库的建设和应用提供了明确的指导。

国内尚未形成针对地理人工智能样本数据库的相关技术标准与规范。

本标准是根据我国国情自主研制的标准，同时本标准的技术内容与相关国际国内标准相互协调，具有领先水平。

五、 与现行法规、标准的关系

本标准与现行的法律、法规无冲突和违背，与现行的国家标准不存在冲突。

六、 重大分歧意见的处理经过和依据

未涉及重大分歧意见。

七、 废止现行有关标准的建议

该标准未替代或废止现行相关标准。

八、 实施标准的要求和措施建议

建议作为推荐性行业标准实施。

九、 其他应予说明的事项

无。

十、 参考文献

[1] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship”, Sci Data, vol. 3, no. 1, p. 160018, Mar. 2016, DOI: 10.1038/sdata.2016.18.

[2] 龚健雅等,“遥感影像智能解译样本库现状与研究”, 测绘学报, vol. 50, no. 8, pp. 1013‐1022, 2021.

[3] “LuojiaSET 遥感影像样本服务平台”. http://geos.whu.edu.cn/luojiaSet, http://58.48.42.237/luojiaSet.

[4] P. Yue et al., “Towards a training data model for artificial intelligence in earth observation”, International Journal of Geographical Information Science, vol. 36, no. 11, pp. 2113‐2137, Nov. 2022, DOI: 10.1080/13658816.2022.2087223.

[5] OGC 地理人工智能样本标记语言标准工作组，OGC Training Data Markup Language for AI Standard Working Group (TrainingDML-AI SWG). https://www.ogc.org/projects/groups/trainingdmlswg.

[6] ISO/IEC 21778:2017 Information technology — The JSON data interchange syntax.

[7] ISO 19157 Geographic information — Data quality.

[8] OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard. https://docs.ogc.org/is/23-008r3/23-008r3.html