ICS 35.240
L70

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

团　体　标　准

T/CESA XXXX—XXXX

信息技术 人工智能 面向机器学习的数据标注规程

Information technology- Artificial intelligence- Guideline of data annotation for machine learning

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国电子工业标准化技术协会　发布

目　　次

前　言

本部分按照GB/T　1.1—2009给出的规则起草。

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这些专利的责任。

本标准由中国电子技术标准化研究院提出并归口。

本标准起草单位：

本标准主要起草人：

信息技术　人工智能　面向机器学习的数据标注规程

1　范围

本标准给出了面向机器学习的数据标注流程框架，包括数据标注前期准备、数据标注任务执行以及标注数据结果输出三个阶段。

本标准适用于面向人工智能研究或开发应用等需要实施数据标注的企业、高校、科研院所、政府机构等。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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。

3　术语和定义

下列术语和定义适用于本文件。

3.1

数据标注　data　annotation

对文本、图像、语音、视频等待标注数据进行归类、整理、编辑、纠错、标记和批注等操作，为待标注数据增加标签，生成满足机器学习训练要求的机器可读数据编码。

3.2

标签　label

标识数据的特征、类别和属性等，可用于建立数据及机器学习训练要求所定义的机器可读数据编码间的联系。

3.3

标注任务　annotation　task

按照数据标注规范对指定数据集进行标注的过程。

3.4

数据标注员　data　labeler

负责对文本、图像、语音、视频等待标注数据进行归类、整理、编辑、纠错、标记和批注等操作的工作人员。

3.5

标注工具 annotation tool

数据标注员完成标注任务产生标注结果时所需的工具和软件。

注 1：标注工具可生成标签并提供参考模板。

注 2：不同的数据类型和标注任务需要不同的标注工具。标注工具按自动化程度可分为手动、半自动、自动三种。

3.6

标注平台 annotation platform

开展标注任务的系统化框架。

注：标注平台在包含标注工具全部功能的基础上将所有标注环节工具化，可有效地对标注任务进行全局管理和跟踪。

3.7

标注说明规则 annotation instruction

数据需求方用于明确标注任务和标注数据的书面陈述，包含执行标注任务所需的标注工具、任务描述、标注方法、正确示例、常见错误等内容。

3.8

标注方法 annotation method

定义数据标注员进行数据标注时的环境和流程，应包含标注对象定义、所用标注工具和标注平台、标注格式、标注前的准备工作、标注后的处理工作等。

3.9

众包标注 crowdsourcing annotation

数据需求方公开发布标注任务，数据标注员申领标注任务并在规定时间内完成标注任务发回数据数据需求方，数据需求方收集整理后获得用于机器学习训练的标注数据集的数据标注过程。

3.10

半自动标注 semi-automatic annotation

使用人工结合自动化工具的方式进行数据标注。

4 概述

本标准给出了数据标注的流程框架，它包括标注项目的前期准备工作(包括对于所需数据的定义、标注规则的制定、标注人力的确定)；标注任务的创建、分发，开展、回收和标注结果的质检和质量控制；标注结果输出的建议格式和交付。数据标注流程框架见图1：

a) 定义所需数据；

b) 确定标注说明规则；

c) 确定标注人力的供给方式；

d) 标注工具和平台的选择；
e) 标注任务的创建、分发、开展和回收；
f) 标注结果的质检和质量控制；
g) 标注结果的输出格式建议；
h) 标注数据的交付和验收。

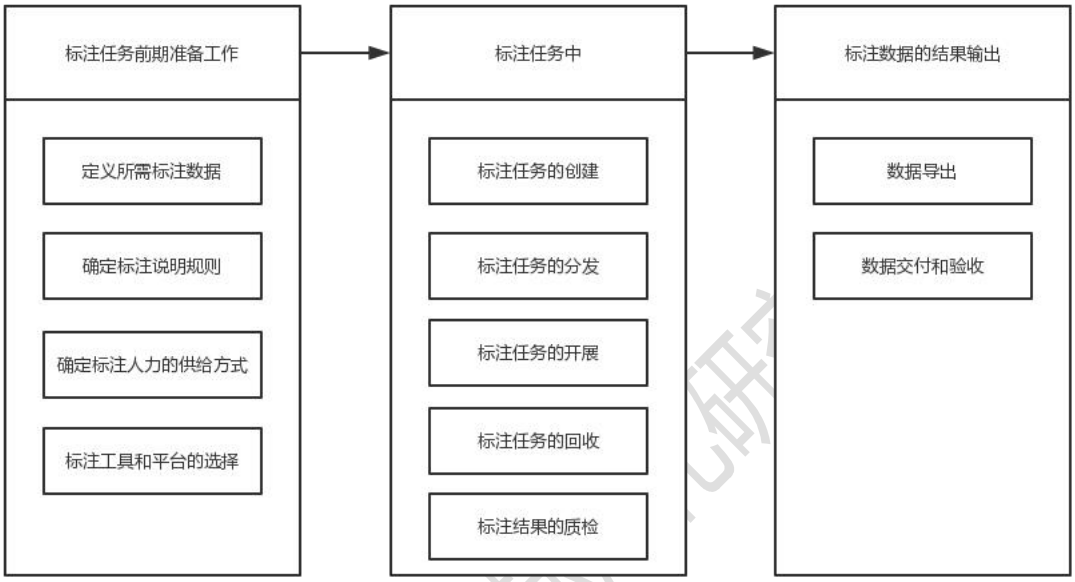

图 1　数据标注流程框架

5　数据标注流程

5.1　定义所需数据和预估数据量

数据标注前应完成以下五项准备工作：

a) 对解决的问题进行分析，明确机器学习和模型训练过程中所需的标注数据类型、量级、用途及应用场景。分析维度包括：业务场景的针对性、标注样本的平衡性、前期经验及改进措施的借鉴等；
b) 对数据进行整理，明确数据与标签文件存放目录结构，在任务分配与回收的时候应按指定的目录进行数据组织；
c) 明确数据与标签文件的命名规则，命名规则应避免数据更新迭代时的重名，便于数据追踪、标注追踪，数据文件名与标签文件名应保持一致；
d) 根据标注任务的人力获取模式、工具的选择、标注任务的类型、算法选择以及整个项目的成本对所需标注数据量进行预估；
e) 与标注人员沟通，明确标注数据的定义并确定最终的需求量。

5.2　确定标注说明规则

5.2.1　标注说明规则的职责分工

数据需求方即业务数据需求方，指需要利用人工智能技术解决实际业务问题的业务团队。数据需求方应负责确保数据标注的规则符合该领域的业务和专业常识，并根据标注规则，检查所标注的数据是否满足数据需求方。

数据使用方指需要使用标注数据训练人工智能模型的研发团队。数据使用方应从机器学习算法角度，确保标注规则可满足机器学习模型的训练要求，并根据该标注规则，检查标注的数据支撑机器学习模型达到数据需求方期望的精度。

数据需求方、数据使用方及数据标注团队应共同参与标注说明规则的制定、调整、迭代、执行的各个环节。数据标注团队应从实际标注角度出发，确保标注规则清晰、明确，以避免数据标注员理解偏差，进而导致标注结果不符合预期。

5.2.2 标注说明规则的定义

标注说明规则应明确项目的背景、意义及数据的应用场景，且包含项目的标注工具、任务描述、标注方法、正确示例、常见错误等内容。

标注说明规则应有可变更性，该变更应由相关方评审同意后，再更新规则文档，且相关方应沿用制定规则时的基本原则及方法。

5.2.3 标注说明规则的内容

标准说明规则包括但不限于：

a) 项目背景：概述标注项目的背景或数据标注需求产生的场景；
b) 版本信息：标注该说明的当前版本编号、发布日期、发布人、发布备注（发布原因或迭代原因）及历史迭代信息（历代版本编号、发布日期、发布人、发布备注等）；
c) 任务描述：概括标注项目的主要任务，包括标注项目的关键信息、数据形式、标注平台、主要标注方法、期望交付时间、正确率要求等；
d) 保密责任：对于数据的密级程度，数据需求方须在规则中列明，明确保密责任，标注方对当前承担的数据标注任务承担保密职责（例如雷达数据标注等任务需要）；
e) 标注方法：给出数据需求方所需数据对象的严谨定义，明确在协定的标注平台上使用何种标注组件、标签及全部操作。标注方法的衡量标准，以标注人员掌握标注方法后，能立刻正确操作一次标注；
f) 正确示例：通过图片、图文、视频等的形式，示范正确的标注方法或成果，数据需求方应明确数据产出，标注方应明确标注认识，标注样例应覆盖特殊样本的标注示例；
g) 注意事项：标注方的错误预警有警示作用，规则制定者在注意事项中，列出标注方应避免的错误、标注方法中应注意的细节及额外处理方式等；
h) 质量要求：数据标注规则应对项目的预期质量有合理的定量预估。审核质检应遵循质量要求。

5.2.4 执行方法及注意事项

数据标注员应学习规则文档，执行培训以保证每个标注人员理解标注说明规则和满足技能要求。

数据需求方宜要求标注方检验标注培训的效果，在标注之前及时发现问题，并把问题及应对措施，整理归档。数据需求方宜要求标注方对含特殊样例的小样本数据集进行预标注，并对标注结果进行审核。标注方满足审核标准后，数据需求方再正式向其分发标注任务。

标注方按照给定规则标注时发现存疑数据，应及时记录。数据需求方应明确此类数据的记录规则、保存路径及后续处理方法等。采用多人标注或定期集中反馈等方法，处理问题数据。

标注说明规则的细则应有可调整性，对调整后的规则细则，应保证参与者及标注方充分理解。发现规则未涵盖的情况或实例时，标注方应及时向数据需求方反馈、沟通和处理。

5.2.5 标注说明中术语体系的规范化

术语体系的规范化至少应满足：

a) 应遵从国家法规和行业规范；
b) 应建立统一的标注术语字典，确保数据标注人员对术语定义的理解一致；
c) 在学习标注说明规则及进行相应的培训后，数据标注人员能够规范地使用标注术语完成任务；
d) 应被标注项目的相关方认可。

5.3 确定标注人力供给方式

应根据标注任务的数据量级、保密性与资质要求、对业务流程的理解程度、成本预算以及交付时间等各类因素评价并确认标注人力供给方式。标注人力模式可包括：内部自营标注、第三方标注、众包标注等。标注人力模式的特点见表1。

表 1 标注人力模式

类型	适合任务（并列表示“和/或”）	特点
内部自营标注人力	要求熟悉业务流程并及时沟通反馈的标注任务	（1）符合业务流程需求； （2）沟通协调效率高
第三方标注人力	（1）对业务流程理解要求低的标注任务； （2）内部自营标注人力不擅长的标注任务； （3）有专业资质要求的标注任务。	（1）项目管理成本低； （2）可作为其他标注人力的补充或作为有资质的审查人员参与质量控制和检查环节。
众包标注人力	（1）时间紧迫且标注数据量大的标注任务； （2）需从大量用户或场景中采集或标注的任务； （3）保密和隐私要求低的标注任务。	（1）成本低，速度快，标注质量参差不齐； （2）难以满足保密性及专业资质要求。

5.4 标注工具和标注平台的选择

标注工具应满足以下条件：

a) 易操作性：标注工具应降低标注人员的操作难度，提供交互方式的自有标注；
b) 输出数据的规范性：标注工具的数据导出格式，应满足或可转换到本标准指定的格式要求；
c) 高效性：标注工具应保证标注任务的完成效率。

标注平台包含标注工具全部功能、团队管理、任务分发、质量审核等环节的模块，且将所有标注环节工具化。规模较大的平台可完成图像、文本、语音或视频等不同任务的标注。对保密数据，标注平台要保证标注数据的安全性。

当数据量相对较小、数据类型相对单一、标注周期较短时，宜选择标注工具进行标注。当标注量较大、数据类型较多、标注难度较大且周期较长时，宜选择标注平台进行标注。

在医学、金融和其它关键领域，标注工具或平台应满足相关法规要求，具备资质/资格证书、许可证等。如：当涉及医学伦理标注时，标注工具或平台的使用应通过相应机构的伦理委员会的论证流程。

5.5 标注任务的创建、分发、开展和回收

5.5.1 标注任务的创建

创建标注任务前，将待标注数据上传。上传的导入方式有两种：本地上传(适用于数据在本地设备上，包括电脑、U盘、移动设备等)；云端上传 (适用于数据在云端，包括公有云和私有云)。当待标注数据量较大时，采用云端上传数据。

标注数据上传成功后，当仅靠标注工具完成标注时，在创建任务的过程中，任务责任人要事先明确标注任务的目的以及标注规范等。当使用标注平台进行标注时，可根据上传的不同类型的数据，划分不同任务模块，再进行相关任务的创建。

创建任务包括：

a) 明确任务基本信息：包含任务目的、任务需求(任务的优先级，对标注人员能力要求的级别等)、任务描述等；
b) 任务配置：根据不同的任务需求，匹配不同的标注工具，添加与标注任务相关的标注标签；
c) 将数据路径上传至平台；
d) 通过版本控制，确保版本编号的一致。

5.5.2 标注任务的分发

根据任务发布者确定的参数及需求，将标注任务分发给标注人员。

标注任务发布者在发布数据时，要明确以下几项与标注任务相关的参数：

a) 参与标注人数；
b) 任务中子任务数量；
c) 数据标注员每人每天工作量；
d) 回收子任务时间点；
e) 任务结束时间点。

标注任务的分发对象包含标注人员和审核人员。标注任务分发给标注人员时，也应将任务分发给审核人。在标注过程中，同时进行标注的审核工作，以便及时发现和解决问题，提高标注效率。

在任务分发前，需确定每一个子任务分发标注的人数，如同一个子任务分发给多人参与，则需对每个子任务的回收结果进行比对，不同标注任务可根据具体情况 (如成本和时间需求)决定同一个子任务是否需多人标注。

分发时，按照任务具体信息和标注需求，分配给相应的数据标注员，实现数据标注任务的优化调度，提高数据标注的效率和质量。

不同标注人力的供给方式也会影响标注任务的分发形式：如使用第三方标注服务公司的服务，则只需把标注任务发送给第三方标注服务公司，它会将标注任务分发到具体标注参与人员。

在标注分发过程中，采用主动学习技术将提升标注任务分发的效率。完成数据标注前，通过标注平台的主动学习，模型可在剩余的待标注数据中，筛选出对模型重要的数据，优先分发给标注人员；其它数据则可延后分发，或不再分发给标注人员。

5.5.3 标注任务的开展

标注任务中数据标注方法大致分为两种：全人工标注；半自动标注。

全人工标注的方式主要依靠人力进行标注，其标注的数据较精准，当标注数据量较大时，会耗费较多人力。

半自动标注的方式采用训练好的模型对目标数据进行检测，并用标注工具完善。半自动标注适用于标注数据量较大，标注任务较简单的标注。半自动标注建立在较成熟模型的基础上，若检测结果的准确度不够，会增加工作量。

在全人工标注中若对标注结果准确率要求较高，在标注前需对标注人员进行相关任务培训。培训内容为标注工具或平台的使用方法及规定、标注的任务目的、标注内容和标准（依据不同标注任务制定不同标注计划）。

在标注人员标注前期，需建立标注者与标注数据使用者之间的反馈机制，确保两者间信息同步。这可有效解决标注者在标注过程中出现的信息不对称（如标注数据使用者对标注者最新的标注要求）等问题。

标注时，可根据标注规则对少量样本先行试标注，将试标注结果反馈给数据需求方，确认标注结果正确无误后，再批量开展数据标注任务。

5.5.4 标注任务的回收

在项目协定的任务将要完成时，项目负责人需回收标注作业，且需保证已分配的任务能被完整交付。自营的标注团队可直接向标注人员或标注小组负责人收取；第三方标注服务公司需提前联系项目负责人，保证外部团队能按时交付；众包平台的回收任务只需保证任务完成的时间设置合理、参与者能及时提交任务即可。回收环节中需注意个别情况和变化的出现，如果标注人员未能按时交付，则需由候补成员继续完成剩余任务，以保证标注任务进度。

5.6 标注结果的质检和质量控制

5.6.1 质量检查

质量检查能够确保数据标注结果有价值，符合数据需求方的特定应用目的。根据项目特性，质量检查方法可以归纳为以下几种，标注项目负责人需要根据场景需求及项目特点进行选择：

a) 逐条检查：即对整个标注项目所包含的所有标注子任务逐一核查并确认。适用于项目量级不大、人力资源充沛、时间节点不紧张、对标注数据结果的准确率要求极高的标注项目。这种方法覆盖的质检范围最全，同时也适用于任何形式的数据标注场景。该方法可确保标注数据输出的最高质量，尤其对于数据格式主观成分较多、应用场景较复杂的任务更有效；

b) 按比例抽查：即从全部标注数据中科学地抽取样本，对样本中的数据逐条检查，以此评判全部标注数据的质量。样本量的选择需符合统计学基本原理，足以代表全部标注数据。抽查审核时，项目负责人可指派较有经验的审核员完成，从而确保交付质量；

c) 抽样检验又可分为以下三种：

   1) 简单抽样：以等概率抽取 n 件待检测样本的方法，必须注意：不能有意识抽取好的或差的，也不能仅抽取表面摆放的或容易抽取的；

   2) 系统抽样：每隔一定时间或一定编号进行检测，而每一次又是从一定时间间隔内生产出的产品或一段编号产品中任意抽取一个或几个样本的方法；

   3) 分层抽样：当不同类型产品有不同的加工环境（如操作者、不同算法）时，对其质量进行评估时的抽样方法。

d) 机器验证：通过机器学习，包括使用已训练模型进行检查或使用迁移学习、在线学习等方法对人工标注的数据做质量检查，实现全自动或辅助人工质检方式。机器学习方法输出的准确率不能完全代表数据集的准确率，但能在一定程度上反映数据集的质量。

在质量检查过程中，需要设定质检间隔，防止由于一次性不合格数据积压过多而导致延误交付。还需要根据算法要求设定质检合格率，增加标注人员容错率。

5.6.2 质量控制

与质检面向结果不同，质量控制面向过程，确保标注过程可控，并产生预期的结果。在标注过程中，需要对数据质量及其行为进行规范和检测，及时预警反馈，查明低质量数据原因，以此控制标注数据的质量。质量控制的方法根据项目特性可归纳为以下四种：

a) 多人验证：即在任务进行期间，安排超过一名人员做同一个子任务，通过标注工具的功能自动或人工辅助选择出最优、最正确的标注结果；
b) 埋题验证：即在任务进行期间，除了常规标注子任务外，在任务中混进若干已知结果的测试题，以此验证一线操作标注人员的标注水平。这种方法适用于标注作业进行中，有助于项目负责人监控标注人员的水平，及时发现潜在问题。虽然这种方法不能完全代表标注数据成果的质量，但在一定程度上说明标注人员的认真程度及标注能力；
c) 标注人员状态验证：通过一定方法对标注人员的操作规范性、实时注意力状态、标注准确率等方面进行检查与监测，及时发现操作违规问题，保证数据质量；
d) 机器验证：在任务进行期间使用机器学习方法，得到数据准确率，一旦发现离群点或明显的降低趋势，及时对标注人员预警和警告。

5.6.3 质量检查与控制中合格标准的确认

在标注结果的质量检查和控制环节，需在抽查前建立并确认合格标准，并在相关环节贯彻实施。合格标准应具备可量化特性；在医学和其他关键行业，数据标注质量的合格标准还需遵从国家法规和行业规范的约束，如数据标注结果需由有资质的第三方邀请有资质和从业经验的专家进行验证。从而确保标注结果的质量，并使得标注结果的质量检查和控制流程有据可依。

5.7 标注结果输出交付规范

5.7.1 图像类型的数据

图像类标注任务的数据结果为带有标签的数据，包含标签的具体内容，及此图像标签对应的图像空间位置（可选）。不同的标注任务和要求会产出不同的结果，但不影响定义数据格式及组成部分。

输出格式推荐使用易解析、易存储的数据格式，格式包括但不限于json或xml。标注文件应该包含标注详细的标签信息。每个独立的标签需包含以下的信息:

a) 标签 id：每个标签的独立编号；
b) 文件路径：待标注图像的名称或路径；
c) 置信度：各标签的置信度；
d) 每个标签中可能包含多个对象，对于每个对象需要定义：
    1) 对象类型：比如 bounding_box 或者 key point；
    2) 对象详情：为对象的空间信息、内容信息，或与其它对象的关系信息。每个对象的详情因其类型而异。

5.7.2 文本类型的数据

文本类标注任务的数据结果包含文本标签的位置和标签的具体内容。不同标注任务和要求会产出不同的结果，但不影响定义数据格式及组成部分。

标注文件的输出格式推荐使用易解析、易存储的数据格式，包括json、xml、txt等。标注文件应该包含详细的标签信息。每个独立的label需包含以下的信息:

a) 标签 id：每个标签的独立编号；
b) 文件路径：待标注文本的文件链接；
c) 原始文本：待标注文本的全部内容（文本标注任务仅需提供文件路径或原始文本中的一个）；

d) 置信度：为标签的置信度；
e) 每个标签中可能包含多个对象，对于每个对象需要定义：
   1) 对象类型：比如 text_classification 或者 text_tag；
   2) 对象详情：对象的具体文本位置和内容信息，或与其它对象的关系信息。每个对象的详情因其类型而异。

5.7.3 语音类型的数据

语音类标注任务的数据结果包含语音标签的时间位置和标签的具体内容（例如转写内容、说话人信息、噪声等）。不同标注任务和要求会产出不同的结果，但不影响定义数据格式及组成部分。

标注文件的输出格式为json文件或其他通用输出格式，其中文件应详细的标签信息。每个独立的标签需包含以下的信息:

a) 标签 id：每个标签的独立编号；
b) 文件路径：待标注音频名称或路径；
c) 置信度：标签的置信度；
d) 如果是单句录音，则每个标签中包含一个对象；如果是多句录音，则每个标签中包含多个对象。每个标注对象应包括：
   1) 对象类型,比如 speech_to_text；
   2) 对象详情，包括对象具体时间位置和内容信息，或与其他对象的关系信息；每个对象的详情因其类型而异，说话者的信息以及噪音标签等都可以放在对象详情中。

5.7.4 视频类型的数据

视频类标注任务的数据结果可包含视频标签的时间位置、空间位置和标签信息等内容。不同标注任务和要求会产出不同的结果，但不影响定义数据格式及组成部分。

标注文件的输出格式推荐使用易解析、易存储的数据格式，包括json、xml等。标注文件应该包含详细的标签信息。每个独立的标签需包含以下的信息:

a) 标签：id 每个标签的独立编号；
b) 文件路径：待标注视频文件名称或路径；
c) 置信度：为标签的置信度；
d) 每个标签中可能包含多个对象，对于每个对象需包含：
   1) 对象类型：例如 scene_classification；
   2) 对象详情：具体描述对象的时间、空间信息和内容信息，或与其他 object 的关系信息；每个对象的详情因其类型而异。对于视频中起始和结束帧的位置描述也应该放到对象详情中，比如 Object_frame_index_start 以及 Object_frame_index_end。

5.7.5 其它

医学影像数据具有其特殊性，因此在此单独定义输出标准。

对于DICOM类型的数据，按照ISO 12052的要求，参照前述图像、文本、语音和视频数据的输出格式，存储在DICOM数据格式的相应标签和数据集合中。

5.8 数据交付和验收

5.8.1 数据交付

数据交付时，标注团队需对最终提交的数据量进行说明。交付的内容包括：

a) 标注结果 （必选）；
b) 交付和说明文档 （可选）；
c) 关于标注数据的 Metadata （非必选），包括描述原始数据的元信息，比如图像的采集地点、光线、拍摄角度或音频的采集时间、声道数量等；
d) 原始数据 （非必选，有时数据使用方可直接访问原始数据，则无需单独交付原始数据）。

交付的文件存储结构可参考以下：

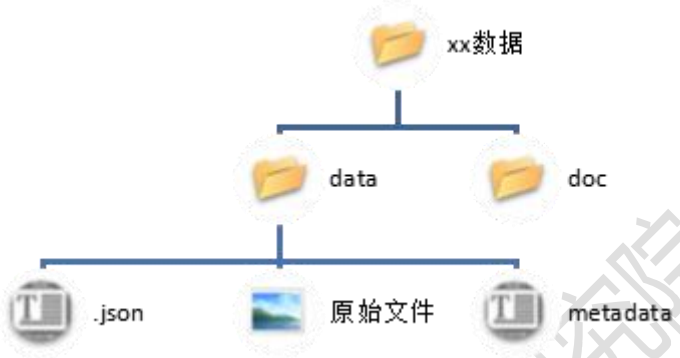

图 2 参考交付文件

说明：
Data——数据文件夹
doc——说明文档文件夹（可选）
.json——(或.xsml 等)标注结果文件,可以每一个 label 单存一个标注结果文件,或者是所有 label 的结果在一个标注文件中
原始文件——为单条标注结果对应的原始文件，如图片、音频、文本、视频
metadata——原始文件元信息 （非必备）

5.8.2 数据验收

数据标注团队在交付数据后，数据需求方应在数据验收期内完成对数据标注结果的验收工作，验收方式包括抽样验收和逐一验收两种。若验收数据质量未达到预期值，数据需求方可要求数据服务提供商对数据进行修正。

中国电子工业标准化技术协会（CESA）是全国电子信息产业标准化组织和标准化工作者自愿组成的社会团体。广泛联系全国电子信息产业标准化机构和标准化工作者，协助政府部门搞好电子信息产业标准化工作，开拓信息技术领域的标准化工作是中国电子工业标准化技术协会的主要工作内容之一。中国境内从事科研开发、制造、营销和服务的企事业单位、高等院校、社会组织和个人均可随时向中国电子工业标准化技术协会团体标准工作部提出团体标准项目建议。

中国电子工业标准化技术协会标准按照《电子工业标准化技术协会协会团体标准管理办法》进行制定和管理。

在本标准实施过程中，如发现需要修改或补充之处，请将意见和有关资料寄至中国电子工业标准化技术协会，以便修订时参考。

中国电子工业标准化技术协会地址：北京市海淀区万寿路27号
电话：010 - 64102952 电子邮箱： standards@cesa.cn
网址：www.cesa.cn