ICS 35.240.99
CCS TP39

团　　体　　标　　准

T/CES XXX-2025

电力非结构化规范存储的元数据管理要求

Electric Non-Structured Standardized Storage Metadata Management Requirements

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　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。

本文件由福建亿榕信息技术有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：福建亿榕信息技术有限公司。

本文件主要起草人：梁懿、苏江文、宋立华、伍臣周、陈又咏、郑略省、李建华、丘志强、邢国用、陈江海、林钊、俞成强、张晓东、吕志超、王燕蓉。

本文件为首次发布。

电力非结构化规范存储的元数据管理要求

1 范围

本文件规定了电力非结构化规范存储的内容结构要求和规范性描述、扩展原则要求、信息质量评估方法。

本文件适用于指导国家电网有限公司所属各级单位企业，外部关联上下游厂商的文本、图像、音频、视频四类常见非结构化数据的接入、存储与治理工作。

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 5271.1-2000 信息技术 词汇 第1部分：基本术语

GB/T 7408-2005 数据元和交换格式 信息交换 日期和时间表示法

GB/T 18391.3-2009 信息技术 元数据注册系统（MDR） 第3部分:注册系统元模型与基本属性

GB/T 25100-2010 信息与文献 都柏林核心元数据元素集

GB/T 3792-2021 信息与文献 资源描述

GB/T 7408.1-2023 日期和时间 信息交换表示法 第1部分:基本原则

3 术语和定义

下列术语和定义适用于本文件。

3.1 非结构化数据 unstructured data

非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据，包括文本、图像、音频和视频等等常见数据格式。

3.2 元数据 metadata

元数据是描述数据的数据，用于提供数据的结构化信息，帮助识别、管理和利用数据资源。按用途可以分成管理元数据、业务元数据、技术元数据三类。管理元数据记录和维护数据资源的管理属性。业务元数据描述数据的业务含义、业务规则。技术元数据描述数据的内容特征，是元数据价值挖掘的主体，如文本的主题，内容分类，摘要等；图像的主题，分类，目标实体或内容描述等；音频的文本表述，主题等；视频的关键帧等。

3.3 元素 element

数字资源元数据的基本语义单位,描述数字资源元数据框架内的基本实体。

3.4 修饰词 modifier

当元素无法满足资源对象的精确描述需要时进一步扩展出的术语。

3.5 元素修饰词 element modifier

对元素的语义进行修饰，提高元素的专指性和精确性。

4 符号和缩略语

下列符号和缩略语适用于本文件。

DC：都柏林核心元数据（Dublin Core）

5 元数据内容结构要求和规范性描述

5.1 元数据内容结构分类

以都柏林 DC 通用元数据规范为基础，设计非结构化元数据的内容结构由管理、业务和技术元数据三类组成，并按元数据信息内容分为一级和二级层次关系的元数据元素，共计 20 个二级元数据元素。如有特别需要，可遵循本规范中的扩展规则进行本地扩展。

表 1 非结构化元数据内容结构表

序号	元数据分类	元数据一级元素	元数据二级元素	元数据注释	备注
1	管理元数据	管理基本信息	主数据	主文件唯一标识 ID	引用都柏林元数据标准
2			文件名称	文件的名称	引用都柏林元数据标准
3			存储链接	文件存储地址	引用都柏林元数据标准
4			入库方式	数据接入方式，包括线下、外网、一级部署、二级部署、自建系统	
5			初次上传时间	文件初次上传时间	引用都柏林元数据标准
6			最后更新时间	文件最后更新时间	
7		认责信息	管理部门	管理部门名称，负责日常维护、权限申请、数据接入、清退等管理	
8	业务元数据	业务基本信息	源端业务系统名称	源端业务系统的名称	引用都柏林元数据标准
9			单位名称	单位名称	
10		系统模块信息	来源系统业务模块	按照业务系统“模块（一级菜单）-功能（二级菜单）-子功能（三级菜单）”的三级功能结构建立数据目录,此项填写业务系统的业务模块。	
11		业务领域信息	文件业务类型	业务文件类型，如合同（含采购合同、服务合同等）、记帐凭证、身份证等	
12	技术元数据	公共技术元数据	文件类型分类	文件类型分类为文本、图像、音频、视频等	引用都柏林元数据标准
13			文件格式	表示文件后缀或格式,方便关联在线浏览软件	引用都柏林元数据标准
14			文件大小	文件存储大小	
15		文本特征元数据	主题词	文档中能够代表其内容特征的、最能说明问题的、起关键作用的词	引用都柏林元数据标准
16			纯文本信息	提取的纯文本信息	
17			文本摘要	文档数据的摘要信息	
18		图像特征元数据	图像文本信息	图像包含的文字信息	

19		音频特征元数据	音频转文本信息	音频转成文本信息	
20		视频特征元数据	视频关键帧	视频的关键帧	

5.2 管理元数据规范描述

a)主数据
名称: metaData
标签:主数据
定义:赋予文件唯一标识 ID 编码。
注释:文件唯一标识 ID 编码。
术语类型:元素
元素修饰词:无
值域:数据唯一标识，组织编码结构如下图所示：

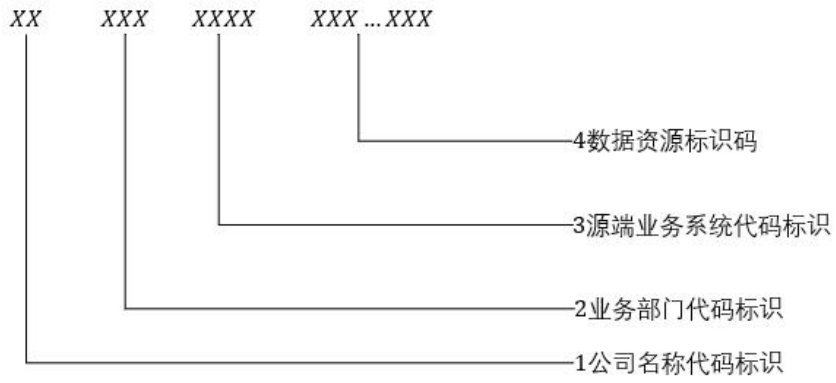

图 1 组织编码结构示意图

注 1：公司名称代码标识以两位编码表示，如总公司为 01、北京分公司为 02、天津分公司为 03，类推。

注 2：业务部门代码标识以三位编码表示，如公司的营销部门为 001、研发部门为 002、财务部门 003，类推。

注 3:源端业务系统代码标识以四位编码表示,如业务部门中的营销系统为 0001、研发系统为 0002、财务系统为 0003，类推。

注 4：数据资源标识码是非固定码，按具体需求设置位数，可按照自行制定上报数据接收方，保证主数据是唯一性即可。

b)文件名称
名称:fileName
标签:文件名称
定义:文件的名称。
注释:文件的名称。
术语类型:元素
元素修饰词:无
值域:自由文本，限制 25 个字符以内。

c)存储链接
名称: storeLink
标签:存储链接

定义:文件物理存储地址。
注释:文件物理存储地址，用于索引、获取文件。
术语类型:元素
元素修饰词:无
值域:文件链接地址

d)入库方式
名称: storeMethod
标签:入库方式
定义:数据来源、接入方式。
注释: 数据接入方式。
术语类型:元素
元素修饰词:无
值域: 线下、外网、一级部署、二级部署、自建系统

e)初次上传时间
名称: firstUploadTime
标签:初次上传时间
定义:数据首次上传的时间。
注释: 数据首次上传至数据中台的时间，时间格式采用 yyyy-mm-dd hh:mm:ss。
术语类型:元素
元素修饰词:无
值域:时间格式采用 yyyy-mm-dd hh:mm:ss

f)最后更新时间
名称: lastUpdateTime
标签:最后更新时间
定义:数据最后上传更新的时间。
注释:数据最后上传数据中台的更新时间，时间格式采用 yyyy-mm-dd hh:mm:ss。
术语类型:元素
元素修饰词:无
值域:时间格式采用 yyyy-mm-dd hh:mm:ss

g)管理部门
名称: mgtDepartment
标签:管理部门
定义:数据的管理部门名称。
注释:管理部门名称，负责日常维护、权限申请、数据接入、清退等管理。
术语类型:元素
元素修饰词:无
值域:业务部门编码

5.3 业务元数据规范描述

a)源端业务系统名称
名称: systemName
标签:源端业务系统名称
定义:数据产生的源端业务系统名称。
注释:数据产生的源端业务系统名称，如经法系统。
术语类型:元素
元素修饰词:无

值域:源端业务系统编码

b)单位名称
名称: unitName
标签:单位名称
定义:数据产生的单位名称。
注释:数据产生的单位名称，如总部、国网北京市电力公司。
术语类型:元素
元素修饰词:无
值域:总部与省测公司名称编码

c)来源系统业务模块
名称:sourceSystemModule
标签:来源系统业务模块
定义:来源系统业务模块的层级所属。
注释:按照业务系统“模块（一级菜单）-功能（二级菜单）-子功能（三级菜单）...”的多级系统功能结构填写。
术语类型:元素
元素修饰词:无
值域:限制文本

d)文件业务类型
名称:fileBusinessType
标签:文件业务类型
定义:业务文件的数据类型。
注释:业务文件数据类型，如合同（含采购合同、服务合同等）、记帐凭证、身份证等。
术语类型:元素
元素修饰词:无
值域:自由文本

5.4 技术元数据规范描述

a)文件类型
名称: fileTpye
标签:文件类型
定义:数据的文件类型分类。
注释:数据文件类型分类为文本、图像、音频、视频等。
术语类型:元素
元素修饰词:无
值域:文本、图像、音频、视频，单选。

b)文件格式
名称: fileFormat
标签:文件格式
定义:文件的格式。
注释:表示文件后缀或格式，方便关联在线浏览软件，通常以后缀名表示。
术语类型:元素
元素修饰词:无
值域:文件的后缀名

c)文件大小

名称: fileSize
标签:文件大小
定义:资源在物理上的大小程度。
注释:表述数据资源，著录内容的实际物理尺寸。对于数字化图像资源,可著录其存储容量。
术语类型:元素
元素修饰词:无
值域: 单精度浮点数+字节单位（B, KB, MB, GB, TB PB）

d)主题词
名称: subjectWord
标签:主题词
定义:资源内容的主题描述。
注释:是在标引和检索中用以表达文献主题的人工语言，具有概念化和规范化的特征。
术语类型:元素
元素修饰词:无
值域:自由文本，多个主题词之间用全角标点符号顿号“、”分隔

e)纯文本信息
名称: pureTextInfo
标签:纯文本信息
定义:纯文本信息由可打印字符组成。
注释:纯文本信息由可打印字符组成，人可以直接阅读和理解其形式。
术语类型:元素
元素修饰词:无
值域:自由文本

f)文本摘要
名称:textAbstract
标签:摘要
定义:资源内容的概要。
注释:用自由行文的形式简要描述图像资源的内容。
术语类型:元素
元素修饰词:无
值域:自由文本，不超过 200 个文本字符

g)图像文本信息
名称: imageTextInfor
标签:图像文字信息
定义:图像本身的文本内容。
注释:图像本身蕴含的文本内容，如身份证的人名、地址、出生日期等。
术语类型:元素
元素修饰词:无
值域:自由文本

h)音频转文本信息
名称:speechText
标签:音频转文本信息
定义:音频转换为文本信息。
注释:将音频转换为文本内容，文字内容方便进行整理和加工。
术语类型:元素

元素修饰词:无
值域:自由文本

j)视频关键帧
名称: videoKeyFrame
标签:视频关键帧
定义:视频中的静止画面。
注释:截取视频中具有代表视频内容的帧，通常指关键帧。
术语类型:元素
元素修饰词:无
值域:图片，格式为 PNG、JPEG 或 JPG。

5.5 元数据扩展原则要求

a)现有元数据内容结构中，如果没有恰当的元素可供复用,允许自行扩展元素。
b)自行扩展的元素不能和已有的元素有任何语义上的重复。
c)扩展的修饰词必须遵循向上兼容原则，即修饰词在语义上不能超出被修饰词（元素）的语义。
d)新增加的元素和修饰词优先采用其他元数据标准中的元素和修饰词。
e)新增元素如果复用来自其他元数据标准的元素或修饰词,必须说明来源,使用时严格遵循其语义。
f)建立扩展流程报备机制，由企业数据存储规范管理部门统一管理扩展元素的登记、审核和发布，避免扩展滥用或重复定义。

6 元数据访问与安全管理要求

6.1 访问控制与权限管理要求

最小权限原则：基于员工角色设置基础权限，确保仅授权人员访问必要元数据，防止越权操作。
动态权限调整：实时响应职位变动或业务需求变化，自动更新访问权限，减少人为管理漏洞。

6.2 数据安全保护要求

敏感元数据脱敏：对描述个人或业务敏感字段的元数据实施脱敏处理，避免分析过程中泄露关联信息。
加密存储与传输：采用高强度加密算法保护元数据存储及传输过程，防范非法截取。
元数据血缘追踪：记录数据流转路径，快速定位异常访问或篡改行为，支持安全事件溯源。
元数据审计日志：明确日志记录内容、存储周期和访问权限，以支持安全事件溯源。

7 元数据信息质量评估方法

元数据信息的完整性、准确性、一致性是影响元数据质量最重要的三个因素。另外，还可以依据业务应用进行量化评估，如数据查准率指标等，业务应用的量化指标需联合业务应用场景进行设计。

7.1 完整性量化评估

对完整性进行量化的最直接方法就是计算非空字段的个数，计算公式如下：

$$Qcomp = \frac{\sum_{i=1}^{N} P(i)}{N} \quad (1)$$

公式中若第 i 个字段为空，则 P(i)为 0，非空则为 1。N 为元数据规定盘点或自动挖掘的字段总个数。Qcomp 可被称为简单完整度，因为公式中每个字段对于衡量完整性具有等同的意义。
Qcomp 值域为[0,1]，Qcomp 值越接近 1，表示元数据的完整性越高。

7.2 准确性量化评估

准确性是指元数据提供的内容正确、客观地反映被描述资源的程度。准确度是对准确性的定量计算值，其测量方法是计算用户从元数据记录中获取到的信息与同一个用户从资源自身获取到的信息之间的语义距离。该距离越短，表明元数据提供的内容与资源自身内容越吻合，元数据记录的准确性就越高。

语义距离的计算可借鉴信息检索领域用来计算两个文本之间相似度的向量空间模型。具体公式如下：

$$Qaccu = (\sum_{i=1}^{n} tfr_i * tfm_i) / \sqrt{\sum_{i=1}^{n} tfr_i^2 * \sum_{i=1}^{n} tfm_i^2} \quad (2)$$

其中，tfr_i 和 tfm_i 分别是第 i 个词在被描述资源的文本和元数据记录中出现的相对频次，n 为两个文本中不同词的总个数。两个向量之间的距离采用余弦函数来计算，从而得出元数据与被描述资源间的语义距离，即元数据的准确度。

Qaccu 值域为[0,+∞)，Qaccu 值越大，表示元数据采集或挖掘结果越准确，元数据信息量越大，区分度越高。依据以往的研究经验，通常 Qaccu 值大于 1 的，具有明显的区分信息，能够体现数据本身的特征信息；Qaccu 小于 0.01 的，不具备区分信息，难以体现数据本身的特征信息。

7.3 一致性量化评估

一致性核心考察的是元数据遵循元数据规范进行一致性取值的情况。常见的破坏一致性的情况有:

①元数据中包含了指定元数据规范并没有定义的字段;

②盘点字段里没有元数据规范规定的必备字段;

③某些字段没有遵循元数据规范中的取值限制;

具体量化计算公式如下所示：

$$Qcons = 1 - \frac{\sum_{i=1}^{N} brokeRule_i}{N} \quad (3)$$

其中Qcons为元数据的一致性值，$brokeRule_i$为盘点字段遵循第i条规则的情况，取值为1或0，如果盘点或挖掘的字段遵循第i条规则，$brokeRule_i$为0，否则为1。N为元数据盘点或挖掘过程中所采用的元数据规范中规则的条数。

Qcons值域为[0,1]，Qcons值越接近1，表示元数据的一致性越高。