



中华人民共和国国家标准

GB/T 32909—2016

非结构化数据表示规范

Unstructured data expression specification

2016-08-29 发布

2017-03-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会 发布

目 次

前言	III
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 缩略语	1
5 非结构化数据 DDL 表示框架	2
5.1 概述	2
5.2 整体框架	2
5.3 基本抽象描述符	3
5.4 基本抽象描述方案	5
5.5 根元素	7
6 数据类型	8
6.1 基本数据类型	8
6.1.1 数值数据类型	8
6.1.2 字符串数据类型	9
6.1.3 日期数据类型	9
6.2 扩展数据类型	10
6.2.1 标量	10
6.2.2 向量	11
6.2.3 矩阵	11
7 基本属性描述	13
7.1 基本元素	13
7.1.1 唯一标识符描述符	13
7.1.2 时间描述方案	13
7.1.3 人物描述符	13
7.1.4 地点描述方案	14
7.2 创建信息描述方案	14
7.3 修改信息描述方案	15
7.4 格式与大小描述方案	15
7.5 数据定位描述符	16
8 语义特征描述	16
8.1 对象描述方案	16
8.2 事件描述方案	17
8.3 概念描述方案	17
8.4 数据编码描述方案	17
8.5 数据语义描述方案	18

9	底层特征描述	19
9.1	图像类型	19
9.1.1	颜色相关描述符	19
9.1.2	纹理特征	25
9.1.3	形状特征	26
9.1.4	局部特征	27
9.2	音频类型	28
9.2.1	标量序列	28
9.2.2	低级音频描述符	30
9.2.3	静音描述符	32
9.2.4	MFCC 特征描述符	32
9.3	视频类型	33
9.3.1	时间域特征描述类型	33
9.3.2	运动特征	35
9.3.3	镜头和关键帧描述符	38
9.4	文本类型	39
10	原始数据描述	40
11	关联描述符	41
11.1	语义特征-底层特征关联描述符	41
11.2	基本属性-原始数据关联描述符	41
11.3	语义特征-原始数据关联描述符	42
11.4	底层特征-原始数据关联描述符	42
附录 A (规范性附录)	Unsigned 类型语法定义	43
附录 B (资料性附录)	非结构化数据表示 XML Schema 定义	47

前 言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位:北京航空航天大学、中国电子技术标准化研究院、清华大学、浙江大学、中国人民大学。

本标准主要起草人:马殿富、郎波、黄雷、高林、吴新松、刘贤刚、卫凤林、李莹、张展新、王朝坤、王建民、陈俊、庄越挺、陈岭、邵健、孙建伶、杜小勇、陈晋川。

非结构化数据表示规范

1 范围

本标准规定了文本、音频、视频、图像等非结构化数据的统一数据表示。
本标准适用于非结构化管理系统中非结构化数据的表示。

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。

GB/T 18793—2002 信息技术 可扩展置标语言(XML)1.0

XML Schema 可扩展置标语言模式(XML Schema, W3C Recommendation, 2 May 2001)

3 术语和定义

下列术语和定义适用于本文件。

3.1

特征 feature

数据的特性。

3.2

逆向文件频率 inverse document frequency; IDF

一个词语普遍重要性的度量,由总文件数目除以包含该词语的文件数目,将得到的商取对数获得。

3.3

词频 term frequency; TF

某一给定的词语在文件中出现的次数。

4 缩略语

下列缩略语适用于本文件。

D:描述符(Descriptor)

DDL:描述定义语言(Description Definition Language)

DS:描述方案(Description Scheme)

IDF:逆向文件频率(Inverse Document Frequency)

GPS:全球定位系统(Global Positioning System)

MFCC:梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient)

SIFT:尺度不变特征转换(Scale-Invariant Feature Transform)

TF :词频(Term Frequency)

TF-IDF:词频-逆向文件频率(Term Frequency-Inverse Document Frequency)

UML:统一建模语言(Unified Modeling Language)