GY

中华人民共和国广播电视和网络视听行业标准

GY/T XXX—XXXX

广播电视和网络视听收视大数据清洗规范

Specification for big data cleaning of radio & TV and network audiovisual viewership

（报批稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家广播电视总局　　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这些专利的责任。

本文件由全国广播电影电视标准化技术委员会（SAC/TC 239）归口。

本文件起草单位：国家广播电视总局广播电视规划院、国家广播电视总局广播影视信息网络中心、中央广播电视总台、央视国际网络有限公司、央视频融媒体发展有限公司、北京广播电视台、上海广播电视台、湖南广播电视台、浙江广播电视集团、江苏省广播电视总台、广东广播电视台、辽宁广播电视台、广西广播电视台、四川广播电视台、山西广播电视台、黑龙江广播电视台、中国传媒大学、中移（杭州）信息技术有限公司、咪咕文化科技有限公司、爱上电视传媒（北京）有限公司、未来电视有限公司、北京歌华有线电视网络股份有限公司、东方有线网络有限公司、华数传媒网络有限公司、广东有线广播电视网络股份有限公司、广东南方新媒体股份有限公司、百视通网络电视技术发展有限责任公司、北京快手科技有限公司、深圳广播电影电视集团、中国广电重庆网络股份有限公司、贵州省广播电视信息网络股份有限公司、贵州多彩新媒体股份有限公司、陕西广信新媒体有限责任公司、陕西广电网络传媒（集团）股份有限公司、湖南快乐阳光互动娱乐传媒有限公司、优酷信息技术（北京）有限公司、北京爱奇艺科技有限公司、腾讯云计算（北京）有限责任公司、秒针信息技术有限公司、上海成思信息科技有限公司。

本文件主要起草人：郑冠雯、李忠炤、谢垚、王幸、王志豪、苏畅、黄卓伟、杨质祺、李义彪、展祎萌、李凯东、陈珊珊、高譞、徐展、孙鹏飞、袁航宇、张冰、董哲颖、周玲、陈文浩、杨罡、芦磊、李婧雯、张莉、林春昭、刘玲、孙婧、许慧芬、段瑞忠、韩宇、肖红江、陈勇华、李鸣扬、钱江奇、吴兴荣、潘红梅、肖勇峰、王峰、李鸣、肖云、高明青、董原、顾荣宁、周青文、张玮、诸葛海标、汪昊辰、方辉、杨磊、冉大为、王庆宝、李正文、陈卫、郑炜、陆堃、董强强、高海漩、靳国卫、李新、刘超、廖玮、赵士原、杨敖、苟明宇、魏雪平、张沈阳、刘彦鹏、范斐、田方、梁琦、冯艳玲、田魁、佘一夫、黄立超、王金龙、刘杰、张国栋、葛承志、徐永太、曾亮、刘沛、张丽、刘向东、胡春磊、高燕肖、李百峰。

广播电视和网络视听收视大数据清洗规范

1 范围

本文件规定了广播电视和网络视听收视大数据的清洗流程和清洗规则。

本文件适用于有线电视、卫星直播、IPTV、OTT TV、互联网视听等收视数据的清洗。

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 2260—2007 中华人民共和国行政区划代码

3 术语和定义

下列术语和定义适用于本文件。

3.1

大数据清洗 big data cleaning

对原始数据进行处理和转换，以去除无效、不完整、重复或不一致的数据，提高数据的质量和准确性。

3.2

行为记录 behavior record

描述用户收视行为的相关信息，包含与该行为相关的所有数据。

3.3

字段 field

描述用户收视行为记录中某一特征的数据项。

3.4

单次播放标识 single play identifier

标识用户单次启动播放器后，一个内容对应所有收视行为的唯一性编码。

注：该标识在播放器初始化时设定，播放过程中切换码率、暂停等交互操作不会令该标识发生变化。

4 缩略语

下列缩略语适用于本文件。

ID 用户终端的唯一编码（Identifier）

IPTV 互联网协议电视（Internet Protocol Television）

OTT TV 互联网电视（Over The Top TV）

UTF-8 8位Unicode字符集转换格式（Unicode Transformation Format-8bit Unicode）

5 清洗流程

收视大数据清洗流程包含规范性清洗、有效性清洗、完整性清洗、唯一性清洗、合理性清洗5个规则，见图1。其中，规范性清洗用于规范数据格式，有效性清洗用于检验并剔除无效数据，完整性清洗

用于对行为记录进行合并和补全，唯一性清洗用于处理重复和重叠行为记录，合理性清洗用于处理异常流入行为记录。各步骤清洗规则见第6章。

应加载当次清洗处理时所需的全量数据进行收视大数据的清洗，并按图1依次执行各清洗步骤。前序步骤剔除的行为记录或字段，不再参与后续步骤的清洗。对于清洗过程中剔除的行为记录或字段，宜标记为异常后保存并记录剔除原因。

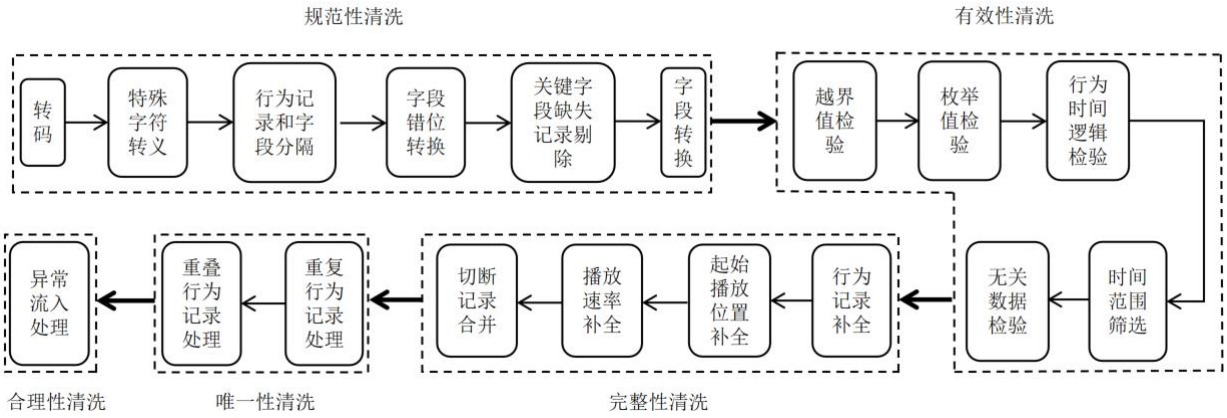

图 1　清洗流程

6　清洗规则

6.1　规范性清洗规则

6.1.1　转码

对于非UTF-8格式的数据，应统一转码为UTF-8格式。

应对转码失败的字段进行剔除。

6.1.2　特殊字符转义

应对后续数据处理会造成影响的特殊字符进行转义。

6.1.3　行为记录和字段分隔

应使用行为记录分隔符和字段分隔符将数据分隔为行为记录和字段。

应对分隔失败的行为记录或字段进行剔除。

6.1.4　字段错位转换

应对字段错位的记录进行转换，使其字段顺序满足处理要求。

转换失败的行为记录应剔除。

6.1.5　关键字段缺失记录剔除

对于行为主体（如终端ID）、行为时间（如播放时间）、行为对象（如节目信息、频道信息等）3项中的其中1项字段完全缺失的行为记录，应进行剔除。

6.1.6　字段转换

6.1.6.1　时间和时长字段

应对时间和时长字段进行转换，转换后的时间字段格式应保留年、月、日、时、分、秒；转换后的时长字段格式应精确到秒。

转换失败的行为记录应剔除。

6.1.6.2　频道字段

应对电视频道的字段进行转换，频道编号格式应符合附录A的规定。

转换失败的频道字段宜保留，并标记转换失败的行为记录。

6.1.6.3 地域字段

应对地域字段进行转换，地域字段格式应符合GB/T 2260—2007的规定，数据格式应为6位十进制数字的国家行政区划码，宜按全国、省（自治区、直辖市）、县（自治县、市）、乡（民族乡、镇）的次序进行数据整理。

转换失败的地域字段宜保留，并标记转换失败的行为记录。

6.1.6.4 节目字段

宜参照GY/T 361—2022附录A中的“节目名称命名及示例”，对节目字段进行命名并转换成统一格式，宜按平台、频道/节目集/页面、内容板块、节目、时间的次序进行数据整理。

转换失败的节目字段宜保留，并标记转换失败的行为记录。

6.2 有效性清洗规则

6.2.1 越界值检验

应剔除超出阈值范围的数值型字段。

注：如时长字段中出现小于0的值。

6.2.2 枚举值检验

应剔除收视行为记录中超出枚举范围的枚举型字段。

6.2.3 行为时间逻辑检验

应对包含开始时间字段、结束时间字段的行为记录进行检验，剔除开始时间大于等于结束时间的行为记录。

6.2.4 时间范围筛选

应剔除开始时间、结束时间均不属于当次清洗处理时间范围的行为记录。

6.2.5 无关数据检验

宜参照GY/T 350.2—2021或GD/J 074—2018的中所述数据元素集的要求，剔除无关的行为记录或字段。

6.3 完整性清洗规则

6.3.1 行为记录补全

对于采集心跳数据，未采集行为开始或结束时间的数据源，宜将同一内容（或同一单次播放标识）、同一用户、相同行为类型且相邻（间隔不多于2次心跳时间视为相邻）的心跳数据相连，组成一条收视行为记录。

6.3.2 起始播放位置补全

对于缺失起始播放位置的非直播收视行为记录，宜将起始播放位置记为节目开始时间。

6.3.3 播放速率补全

对于缺失播放速率的非直播收视行为记录，宜将播放速率记为1倍速。

6.3.4 切断记录合并

对于由于数据采集系统原因导致一条行为记录被切断为两条或多条的，应进行合并。

6.4 唯一性清洗规则

6.4.1 重复行为记录处理

对于属于同一用户的两条或多条行为记录，如果所有字段完全一致，应保留其中一条，剔除其余行为记录。

6.4.2 重叠行为记录处理

对于存在时间重叠（不含开始时间、结束时间完全一致的情况）且属于同一用户的两条行为记录：

——如属于同一频道或内容，应将两条行为记录合并为一条，取第一条行为记录（即两条行为记录中开始时间较早的）的开始时间作为开始时间，第二条行为记录（即两条行为记录中结束时间较晚的）的结束时间作为结束时间；

——如属于不同内容且开始时间不一致，应剔除第一条行为记录的时间重叠部分；

——如属于不同内容且开始时间一致，应保留行为记录中结束时间较晚的，剔除其余记录。

6.5 合理性清洗规则

应剔除在特定时间点开始的异常行为记录，如批量重启或采集故障等原因产生的行为记录。

附 录 A
（规范性）
频道编号

A.1 电视频道编号（不含付费电视频道）

对于电视频道（不含付费电视频道），其频道编号为10位，构成形式为：7位播出机构许可证编号+3位播出机构内频道序号，见图A.1。

播出机构许可证编号见《地级以上播出机构及频道频率名录》《县级广播电视播出机构名录》。

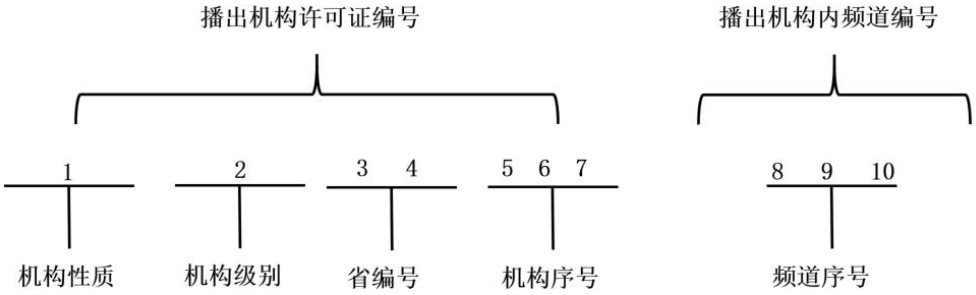

图 A.1 电视频道（不含付费电视频道）编号构成形式

电视频道（不含付费电视频道）编码说明如下：

a) 第 1 位：机构性质，1 广播，2 电视，3 广播电视，4 教育；
b) 第 2 位：机构级别，1 中央，2 省，3 地市，4 县；
c) 第 3～4 位：省编号，如：00 中央、01 北京、02 天津；
d) 第 5～7 位：机构序号；
e) 第 8～10 位：频道序号，根据《地级以上播出机构及频道频率名录》《县级广播电视播出机构名录》“节目设置”中的频道依次进行排序。

示例 1：中央电视台综合频道的频道编号为 2100001001。
示例 2：广东广播电视台珠江频道的频道编号为 3219001002。
示例 3：青岛市电视台体育休闲频道的频道编号为 2315003005。

A.2 付费电视频道编号

对于付费电视频道，其频道编号为8位，为行政主管部门颁发的该频道许可证编号，构成形式见图A.2。

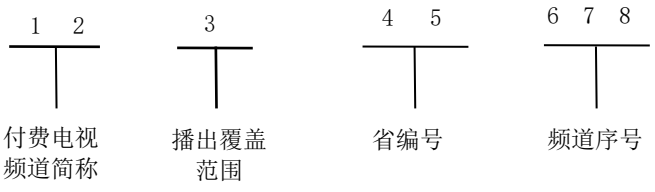

图 A.2 付费电视频道编号构成形式

付费电视频道编码的说明如下：

a) 第 1～2 位：付费电视频道简称，固定为 PT；
b) 第 3 位：播出覆盖范围，0 省内覆盖，1 全国覆盖；
c) 第 4～5 位：省编号，如：00 中央、01 北京、02 天津；
d) 第 6～8 位：频道序号。

示例：杭州市广播电视台开办的求索纪录的频道编号为PT111108。

[来源：GD/J 075—2018，附录B，有修改]

参 考 文 献

[1] GY/T 350.2—2021 网络视听收视大数据技术规范 第2部分：数据元素集
[2] GY/T 361—2022 电视播出节目信息即时传输技术规范
[3] GD/J 074—2018 电视收视数据元素集规范
[4] GD/J 075—2018 电视收视数据交换接口规范
[5] 国家广播电视总局．地级以上播出机构及频道频率名录．http://www.nrta.gov.cn/col/col69/index.html.
[6] 国家广播电视总局．县级广播电视播出机构名录．http://www.nrta.gov.cn/col/col69/index.html.