

# 反应堆关键设备状态数据标准化研究

李书剑, 庞天枫, 杨泰波, 刘才学

(中国核动力研究设计院, 成都 610213)

**摘要:** 对应用于国内外核电站故障诊断产品所采集的反应堆关键设备状态数据进行了数据标准化研究, 梳理并分析了现有反应堆关键设备状态数据资源现状; 针对现状中存在的问题提出了数据标准化的思路与方法; 采用了标准化规范制定和标准化约束等数据标准化方法, 从数据文件格式、数据文件命名方法、文件编码方式、数据类型、数据解析规则、数据存储与读取方法等方面进行了数据标准化的研究与实践, 创新性地提出了“八字节法”标准化数据存储与读取方法; 经实际测试及应用, 实现了反应堆关键设备状态数据的标准化, 提高了数据的规范性与统一性。

**关键词:** 数据标准化; 核电系统; 故障诊断; 状态数据; 数据解析

## Research on State Data Standardization of Critical Reactor Equipment

LI Shujian, PANG Tianfeng, YANG Taibo, LIU Caixue

(Nuclear Power Institute of China, Chengdu 610213, China)

**Abstract:** Data standardization research is carried out for the state data of critical reactor equipment collected by fault diagnosis products at domestic and international nuclear power plants, and the state data resources of existing critical reactor equipment are sorted out and analyzed; The ideas and methods of data standardization are put forward in view of existing problems; Data standardization methods such as standardization specification formulation and standardization constraint are adopted. The engineering practice of data standardization is carried out from the aspects of data file format, data file naming method, encoding method, data type, analysis rules, data storage and reading methods, and an 8-Byte standardized data storage and reading methods are innovatively proposed; Through the actual test and application, the data standardization for the state data of critical reactor equipment is realized, and the data normalization and uniformity are improved.

**Keywords:** data standardization; nuclear power system; fault diagnosis; state data; data analysis

## 0 引言

故障诊断是保障核电站安全、稳定性运行的重要环节。在反应堆关键设备故障诊断专业领域, 中国核动力研究设计院自主设计、研发的相关产品种类丰富, 并在核电市场具有较广的覆盖面, 主要产品包括: 1) 松动部件监测系统(LPMS, loose part monitoring system), 主要用于反应堆一回路内可能出现的松动件、脱落件和遗留的金属零件等的在线监测<sup>[1]</sup>; 2) 堆内构件振动监测系统(VMS, vibration monitoring system), 主要用于周期性监测堆内构件的振动, 防止堆芯吊篮压簧弹簧过度松弛和支撑严重劣化<sup>[2]</sup>; 3) 主管道和波动管 LBB (leak before break) 泄漏监测系统, 简称“LBB”, 主要用于监测反应堆一回路主管道和波动管早期破前泄漏<sup>[3]</sup>; 4) 反应堆冷却剂泵监测系统(CPMS, coolant pump monitoring system), 主要用于主泵等旋转机械状态监测与分析<sup>[4]</sup>。上述产品主要采集振动、

运行工况等状态数据, 采集到的数据统称为“反应堆关键设备状态数据”。目前, 上述产品已应用于国内 46 台、海外 4 台核电机组, 并且建立了基于大数据和互联网的反应堆远程智能诊断平台 (PRID, platform of reactor remote intelligent diagnosis), 每月收集超过 10 万条监测系统数据, 而 PRID 依靠这些数据进行反应堆关键设备的状态分析、故障诊断及预警, 已成功进行了 60 余次故障告警<sup>[5]</sup>。

由于不同核电机组的设备型号及结构, 以及核电业主的实际需求存在较大差异, 再加上软硬件版本提升和技术更新换代, 上述产品应用在不同核电机组的版本不尽相同, 甚至同一机组在不同运行周期的产品版本也可能存在差异, 这直接导致了不同产品所产生的数据偏离了预先制定的统一标准规范, 数据资源出现“熵增”现象。另外, 随着数据的积累量越来越大, 数据的利用和管理效率也逐渐降低。而在数据使用后端, 依赖于数据驱动的机器学习、深度学习等技术在设备故障诊断领域的深入应用也对数据质量提

收稿日期: 2023-01-08; 修回日期: 2023-02-15。

作者简介: 李书剑(1996-), 男, 四川成都人, 硕士, 研究实习员, 主要从事反应堆关键设备故障诊断方向的研究。

引用格式: 李书剑, 庞天枫, 杨泰波, 等. 反应堆关键设备状态数据标准化研究[J]. 计算机测量与控制, 2023, 31(6): 198-203, 211.

出了更高的要求<sup>[6-7]</sup>。

通过对核电故障诊断产品所采集的状态数据资源进行全面梳理和细致分析, 总结得到了目前的资源现状, 主要包括:

1) 文件格式多样: 包括“.dat”、“.tdms”、“.csv”等 7 种文件格式;

2) 解析规则众多: 海内外 50 个核电机组产品所产生的数据对应了 20 种不同的解析规则;

3) 数据解析复杂: 数据没有统一的解析规范, 且编码方式、数据类型等存在差异。

上述现状导致了 PRID 在进行数据解析时耗时较长以及数据解析容易出错, 十分不利于数据的统一管理和后续产品的开发与迭代。

本文将首先对核电故障诊断产品所采集的状态数据资源现状进行全面梳理和细致分析, 然后基于数据标准化的相关概念, 提出反应堆关键设备状态数据标准化的思路与方法; 在此基础上, 结合实际数据情况, 进行工程应用实践, 并评估数据标准化的效果, 最后进行总结并提出思考, 为未来核电市场产品供货、远程智能诊断平台拓展、健康管理平台开发等提供数据侧相关参考。

## 1 反应堆关键设备状态数据资源现状分析

### 1.1 整体情况

核电故障诊断产品所采集的状态数据资源主要来源于四大类: LPMS 产品应用于 50 个核电机组, VMS 产品应用于 26 个核电机组, CPMS 产品应用于 5 个机组, 而 LBB 产品应用于 2 个机组。在产品前端, 数据采集主要通过基于 LabVIEW 软件的程序实现, LabVIEW 是一款图形化的软件开发工具<sup>[8]</sup>, 在数据采集、硬件控制、数据交互等方面具有很大的优势<sup>[9]</sup>。通过 LabVIEW 采集得到的数据文件定期传递至 PRID, 存储在专门的数据服务器中, 并进行后续的处理与分析工作。

通过对数据资源进行全面梳理和细致分析, 发现目前在数据的采集、读取、存储、使用等方面存在较多不统一、不规范等问题, 具体表现在文件格式、文件命名、文件编码方式、文件数据类型以及文件解析规则等方面, 本节将依次进行分析。

### 1.2 文件格式不统一

文件格式是指数据在计算机中存储而使用的对信息的特殊编码方式, 每一类信息都可以有一种或多种文件格式保存在电脑存储中, 每一种文件格式通常会有一种或多种扩展名(文件后缀)用于识别, 不同的扩展名对应不同的打开和使用方式。

根据梳理, 目前反应堆关键设备状态数据资源文件格式种类多样, 具体如表 1 所示, 其中, 产品数据包括 7 种格式, 其中大部分为“.dat”格式, “.dat”格式文件通过二进制进行读写, 具有较快的读写速度和较小的文件体积, 适合高频数据采集以及大批量数据存储<sup>[10]</sup>。另外少量其他

格式也为二进制文件(如“.cbv”、“.DAN”等), “.tdms”与“.csv”格式需要辅助软件进行读取。

表 1 数据文件格式

类型	格式	机组数量	备注
LPMS	.dat	38	二进制文件
	.DAN	4	二进制文件
	.sig	4	二进制文件
	.tdms	4	结构化数据格式 <sup>[11]</sup>
VMS	.dat	18	二进制文件
	.cbv	4	二进制文件
	无	4	无后缀, 实为二进制文件
CPMS	.csv	4	逗号分隔符文件
	.dat	1	二进制文件
LBB	.dat	2	二进制文件

多文件格式存在的数据管理问题主要包括: 1) 文件格式种类较多, 若进行统一管理, 则需要对多种格式进行适配, 增加了软件的复杂程度; 2) 数据隔离程度高, 不同类型格式文件较难互相转化。

### 1.3 数据文件命名方式多样化

文件名是数据文件的标识符, 在同一个地址下具有唯一性。目前故障诊断产品数据文件命名方式种类多样, 具体情况如表 2 所示。由表可知, 大部分文件的文件名只包含采集日期、时间信息, 部分文件将传感器采集通道名作为文件名(如“VRA01”), 部分文件名中存在空格, 容易影响文件名读取。

表 2 数据文件命名方式

类型	格式	数据文件命名示例
LPMS	.dat	20220302070745.dat
	.DAN	LPMS_180612200252_3.DAN
	.sig	s000015.sig
	.tdms	SM1_20200526_000309686.tdms
VMS	.dat	AC20211101145757.dat
	.cbv	CBV_HY2_20181103_233635.cbv
	无	VRA01(无后缀)
CPMS	.csv	20210611 RCP VIBRATION-1.csv
	.dat	FQ03ZB20210820161117000.dat
LBB	.dat	20210507145443.dat

随着数据积累, 为便于规范化管理和查询, 应当具备通过文件名初步识别该数据文件所属电站、机组、类型、采集数据时间等关键信息, 其中福清 3 号机组 CPMS 数据文件名较为规范(FQ03ZB20210820161117000.dat), 能够有效通过文件名识别关键信息, 但是该文件名分隔较少, 难以阅读。

### 1.4 数据文件编码方式不统一

编码是信息从一种形式或格式转换为另一种形式的过程, 使用预先规定的方法将汉字、数字或其他对象编成数码, 计算机在处理文件时, 将可以正常阅读或使用的十进

制数字、英文字母、汉字等转化为二进制编码。对故障诊断产品的数据文件进行读取分析时发现,数据文件存在编、解码方式的不统一。

二进制数据文件在计算机中进行编码存储时分为大端编码(big-endian,亦称高尾端)和小端编码(little-endian,亦称低尾端)两种模式,大端编码将低位数据存放在高位地址,小端编码将低位数据存放在低位地址<sup>[12]</sup>。使用大端解码的方式读取小端编码的数据文件,将会出现解析错误。在现有的数据文件中,大部分数据文件都采用大端编码的模式,少部分采用小端编码的模式。除大、小端编码方式以外,数据文件中还存在字符的编码问题。目前数据资源中存在数字字符(如“05”)、英文字符(如机组简标“HYH”)、特殊字符(如“/”)和中文字符(如“正常”),其中,数字、英文、特殊字符采用 ASCII 编码方式<sup>[13]</sup>,中文字符采用 GBK 规范<sup>[14]</sup>。

文件编码方式的不统一导致在文件读取时必须事先确定文件编码规则,否则将导致数据解析出错。

### 1.5 文件数据类型多样化

数据类型涉及到数据文件所存储的具体内容,根据针对上述所有类型数据文件的分析,可以总结得到目前数据资源所包含的 12 种数据类型,如表 3 所示。在目前已有的数据资源中,使用最多的数据类型是单精度浮点型(single)、短整型(short)和整型(int8),另外,“结构型”数据仅存在于“.tdms”格式文件中。

表 3 数据类型

序号	数据类型	字节长度
1	布尔型/bool	1
2	整型/int8	1
3	无符号整型/uint8	1
4	短整型/short/int16	2
5	无符号短整型/uint16	2
6	长整形/long/int32	4
7	无符号长整型/unit32	4
8	单精度浮点型/float32/single	4
9	双精度浮点型/float64/double	8
10	字符型(字母、数字、特殊符号)	不定
11	字符型(汉字)	不定(偶数)
12	结构型(由上述数据组成)	不定

### 1.6 文件解析规则复杂

由于在产品研发阶段考虑到数据文件的高速采集和大量存储等情况,大部分产品的数采系统将数据以二进制形式存储,这就导致了数据文件不便于直接阅读,必须通过数据解析的方式将二进制数据转换成十进制数据。

数据解析规则是将二进制数据转化为可供正常阅读的十进制数据的关键,缺少数据规则将只能通过按位尝试等“硬破解”方法进行数据读取(如早年国外产品服务机构提

供的数据),当遇到数据存在加密等情况时,没有解析规则将永远无法提取出有效数据。目前,数据解析规则复杂是反应堆关键设备状态数据资源中最为突出的问题,通过梳理,故障诊断产品数据资源目前具有 20 余项解析规则,如表 4 所示。

表 4 数据解析规则

类型	格式	解析规则	适用核电机组
LPMS	.dat	10 通道规则	昌江等
		13 通道规则 1	宁德等
		13 通道规则 2	阳江等
		13 通道规则 3	福清等
		13 通道规则 4	红沿河等
		13 通道规则 5	岭澳等
		13 通道规则 6	田湾等
		16 通道规则	恰希玛等
	.DAN	20 通道规则	田湾等
.sig	21 通道规则	防城港等	
.tdms	16 通道解析规则	三门等	
VMS	.dat	VMS 规则 1	大亚湾等
		VMS 规则 2	福清等
		VMS 规则 3	卡拉奇等
		VMS 规则 4	昌江等
	.cbv	CBV 规则	三门等
无	无后缀规则	田湾等	
CPMS	.csv	可直接读取	三门等
	.dat	27 通道规则	福清等
LBB	.dat	LBB 规则	福清等

## 2 数据标准化的思路与方法

为解决数据资源现状中存在的问题,考虑应用数据标准化的相关方法,数据标准化用于解决数据不规范、不统一等问题,也称为标准化数据治理,包括标准定义、标准发布、标准维护等功能。数据标准化的目的在于尽可能对处于相对混乱状态的数据资源实现规范化管理,最大程度提升数据的使用便利程度和使用效率,为后续挖掘数据价值奠定基础<sup>[15]</sup>。针对核电故障诊断产品所采集的状态数据,数据标准化应当按照数据类型统一、解析规则少、解析快速方便、参数标准化度高、文件空间利用率高、容错性强、可移植性强、可扩展性强等原则进行设计,但在实际执行过程中会遇到较多问题,需要根据实际情况进行调整,数据标准化一旦执行将禁止轻易更改,因此需要对数据标准化方法进行充分测试<sup>[16]</sup>。对反应堆关键设备状态数据进行数据标准化将有助于:

1) 针对存量数据:对当前存量的数据资源,进行规范化整理,使数据便于调取和使用;

2) 针对增量数据:对未来增量的数据资源,数据产生和存储按照标准化执行,使数据统一。

本文提出的数据标准化思路如图 1 所示,主要流程即

对照数据资源现状中存在的问题, 进行标准化规范的定义与发布, 通过标准化规范的约束, 对原始较为混乱的数据资源进行强制化、规范化管理, 最后形成完整的数据标准化流程体系。本文主要根据上述思路, 从文件格式、文件命名、文件编码方式、文件数据类型、文件解析规则、数据存储与读取等方面入手开展数据标准化的实践。

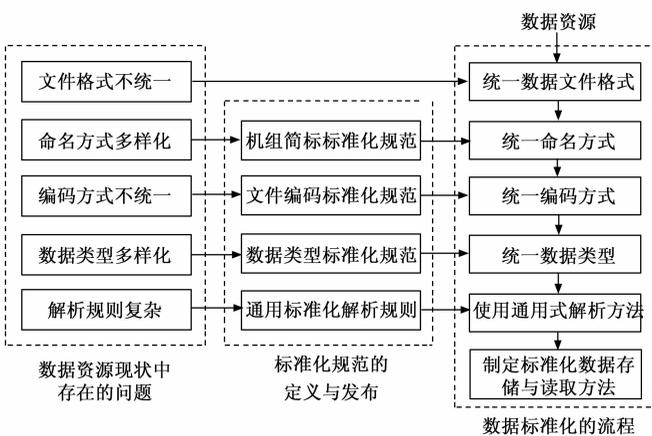


图1 数据标准化的流程

### 3 数据标准化的实践与效果

#### 3.1 文件格式标准化

经过对现有数据资源的梳理, 以及查阅相关资料, 包括对特殊格式数据(如“.tdms”)进行调研, 结合目前状态数据文件的实际情况(以“.dat”数据格式为主), 拟采用“.dat”文件格式作为数据存储文件格式。

使用“.dat”格式的主要优点包括:

1) 使用便利度高: 与现有数据、各类系统软件适配, 现有的LabVIEW数据读取、分析软件可以照常使用, 使用方便简单, 非标准化“.dat”数据文件格式可以较为简单地转化为标准化“.dat”数据;

2) 软件适配度高: “.dat”文件格式采用二进制方式存储数据, 占用内存小, 编码方式简单, 可以使用任何软件进行读取, 适用于LabVIEW、C、Java、Python、MATLAB等众多编程语言, 不存在知识产权限制;

3) 数据安全性高: 可以通过适当的加密方式对“.dat”中的二进制数据进行加密, 提高数据的安全性<sup>[17]</sup>。

目前, 使用“.dat”格式存在以下问题:

(1) 高度依赖于解析规则的准确性: 解析规则出现遗漏将直接导致数据解析出错, 另外, 数据规则过多也容易导致数据解析出现混乱, 因此需要优化解析规则, 确定较优的数据存储方式;

(2) 数据所包含的信息不够全面: 文件中仅包含数据内容, 不包含对数据的解释信息, 必须依赖解析规则中的解释字段才能识别解析数据的意义, 因此, 需要尽可能地将解释字段也包含在数据文件中。

#### 3.2 文件命名标准化

##### 3.2.1 机组筒标规范

现有的数据文件中, 部分文件存在“核电机组筒标”这一字段, 根据解析出来的数据, 该数据长度不固定, 如“TW05”、“HYH05”, 在字符前还有部分非打印ASCII码字符, 如文件结束传递标识等。为进行统一, 现对核电机组筒标做如下标准化规定:

核电机组筒标统一由5位字符组成, 通用格式可以表示为:“XXXYY”, 其中, “XXX”表示电站筒标, 为三位大写英文字母, “YY”为两位数字, 代表机组编号。

核电机组筒标的命名方式借鉴了国际航空运输协会制定的全球机场代码命名规则, 电站筒标由三位字母组成, 为核电基地拼音首字母大写, 当核电基地为三字时, 取每个字的拼音首字母, 当核电基地为二字时, 取第一字拼音首字母和第二字拼音前两位, 如红沿河(HYH)和海阳(HYA), 当出现冲突时, 新机组名称应当更换为其他字母。新命名规则的优点是能快速识别当前电站和机组, 且可使用的字母范围较广。由此对全国所有在运行及规划机组进行命名, 制定机组筒标标准规范, 如表5所示。

表5 机组筒标标准规范(节选)

核电基地	机组名称	电站机组筒标
苍南	三澳1号机组	SAO01
	三澳2号机组	SAO02
大亚湾	大亚湾1号机组	DYW01
	大亚湾2号机组	DYW02
漳州	漳州核电1号	ZZH01
	漳州核电2号	ZZH02
秦山	秦山一期1号	QSY01
	秦山二期1号	QSE01
海阳	山东海阳1号	HYA01
	山东海阳2号	HYA02

##### 3.2.2 文件命名规范

统一机组筒标后, 可以对文件命名格式进行规范, 文件命名格式可以表示为: XXXYY\_TTTT\_YYYYMMDDhhmmss\_AA.xxx, 具体表示方法为“机组筒标\_类型\_日期时间\_编码.后缀名”, 其中, 各字段对应含义如表6所示。值得注意的是, 文件命名规范规定了每个字段的长度, 所以在“类型”字段中, 所有的产品均采用了四位字母的编码方式, 并以产品英文缩写为主, 若字符不够四位(如“VMS”), 则使用第一个实词的非元音字母进行补位(如“VBMS”)<sup>[18]</sup>。

各个字段以下划线“\_”分隔, 便于阅读, 文件名长度固定, 包含电站机组、数据类型、日期时间等关键信息。

#### 3.3 文件编码方式标准化

1) 数据编码规范: 统一使用低位有效、大端编码方式进行数据存储, 即将二进制数据的高位存放在内存的低地址中, 将二进制数据的低位存放在内存的高地址中, 同时在

表 6 文件名命名格式规范

字段	含义	示例
XXXYY	核电机组简标	DYW01
TTTT	数据类型: LPMS、VBMS、CPMS、LKBB	LPMS
YYYYMMDD hhmmss	数据采集的日期时间	20210305 123654
AA	编码方式: 省略:原始文件编码、 B8:“八字节法”编码	B8
.xxx	文件后缀:.dat	.dat

每个内存空间中先存放字节的低位数据。对于目前采用小端编码方式存储的数据,采用交换字节顺序等方式将其转化为大端编码模式。在后续数据读取时,同样按照低位有效、大端编码的方式进行读取。

2) 字符编码规范:目前数据资源中存在英文字母字符、数字字符、特殊符号字符和汉字字符,其中前三种字符属于 ASCII 编码,汉字属于 GBK 编码规范,在数据标准化中,将统一使用 UTF-8 的编码方式。UTF-8 编码是一种使用广泛、包含字符数量极多的一种编码规范,使用 UTF-8 编码可以将所有语言体系下各种类型的字符都涵盖<sup>[19]</sup>。

3) 字节对齐规范:统一使用 1 字节对齐模式,需要在编译、解析软件中进行说明。当使用其他字节对齐方式时,可能会出现空字节等情况<sup>[20]</sup>。

### 3.4 文件数据类型标准化

数据是数据文件最重要的内容,数据类型标准化对已有的数据类型进行统一编码,规定对应的字节数、位数和计算规则,在后续数据的存储和读取中都统一使用该标准规则,将数据资源中的 12 种类型缩减到 7 种,并给每一种数据类型赋予固定的“编码”,制定的数据类型标准规范如表 7 所示。

表 7 数据类型规范

编码	类型	标识	含义	字节	备注
1	布尔型	bool	布尔型	1	
2	整型	int8	有符号单字节整型	1	二进制运算
3		int16	有符号双字节整型	2	
4		int32	有符号四字节整型	4	
5	浮点型	single	单精度浮点数	4	IEEE754 标准 <sup>[21]</sup>
6		double	双精度浮点数	8	
7	字符型	string	字符串	N	UTF-8 编码

### 3.5 解析规则标准化

为了将解析规则标准化,首先制定标准化通用解析规则,当前阶段,准化通用解析规则(包括各类标准化规范)都使用 Excel 表格形式建立,后续逐步在数据库中建立。标准化通用解析规则如表 8 所示,其中,“循环采集次数”指某些产品会进行特定次数的平滑计算,该次数为前期指定。

将 VMS 某产品的数据解析规则转化成通用解析规则,如表 9 所示。

表 8 通用解析规则表

属性	通道	单位	数据类型	起始位置	单个数据长度	数据数量	循环采集次数	大端编码
属性名称 1		单位 1	编码 1	字节	字节	数量 1	数量 1	编码
属性名称 2	通道名称 1	单位 2	编码 2	字节	字节	数量 2	数量 2	
	通道名称 2							
	.....							
	通道名称 N							
属性名称 3	通道名称 1		编码 N	字节	字节	数量 N	数量 N	
.....	.....							
属性名称 N	.....							
文件尾								

表 9 解析规则标准化后的 VMS 某产品数据解析规则(节选)

属性	通道	单位	数据类型	起始位置/字节	单个数据长度/字节	数据数量	循环采集次数	大端编码
核电机组简标			7	0	9	1	1	1
采样长度			5	9	4	1	1	
采样频率		Hz	5	13	4	1	1	
平滑次数			5	17	4	1	1	
启堆日期			5	21	4	1	1	
循环周期			5	25	4	1	1	
功率		%	5	29	4	1	1	
压力		MPa	5	33	4	1	1	
冷段温度		℃	5	37	4	1	1	

通用解析规则的优点在于:

1) 规则表形式结构标准固定: 规则表的结构一经确定就不需更改, 只需要往里面填入对应字段, 且填写内容除属性、通道、单位外, 全部为数字, 便于软件进行识别;

2) 规则表适配度高, 解析规则得到优化: 根据目前的实践情况, 通用解析规则表适配于松脱、振动、主泵的所有“.dat”文件格式, 也适配于“.sig”、“.cbv”等非标准格式;

3) 阅读简单, 数据清晰: 通用解析规则表在填入内容以后, 内容非常清晰, 因为较多需要重复填写的内容都可以赋空值, 由程序自行运算, 因而相较于原始解析规则, 可阅读性得到了较大提升。

### 3.6 数据存储与读取标准化

#### 3.6.1 “八字节法”标准化数据存储方法

数据存储使用“八字节法 (8-Byte, B8)”标准化数据存储方法, 对应“八字节法 (B8)”标准化读取方法。“八字节法 (B8)”标准化数据存储方法是一种为故障诊断产品数据定制的数据存储方法, 规定数据文件由以下三部分内容构成: 1) 头文件; 2) 头文件结束标志; 3) 数据。

1) 头文件: 头文件部分按照每八字节存储一个“属性”下的一个“通道”的“单位”和“数据”相关信息, 故称为“八字节法”。“八字节法”头文件定义规则如表 10 所示。

表 10 “八字节法”头文件定义规则

字节位置	数据类型	含义	数据范围	
属性	1	单字节整型	“属性”字段的字节长度	0 ~ 127
通道	2	单字节整型	“通道”字段的长字节度	0 ~ 127
单位	3	单字节整型	“单位”字段的字节长度	0 ~ 127
数据	4	单字节整型	数据类型编码	1~7
	5~8	四字节整型	当前通道下的数据数量	1~2 <sup>31</sup> -1

“八字节法”限制每个“属性”、“通道”以及“单位”的字符串长度为 127 个字节, 一般情况下字符长度满足要求; 数据类型小于 127 种, 根据数据类型标准表的定义, 当前只定义了 7 种数据类型; 数据数量为 32 位整型, 数据范围非常大, 各种类型数据均满足要求。

2) 头文件结束标志: 头文件所有数据存储完毕以后, 存储一个全部为 1 (0xFF) 的八字节数据作为头文件结束标志。

3) 数据: 在头文件结束标志以后, 按照头文件的规定顺序存放数据。

#### 3.6.2 数据的读取方法

当数据文件按照“八字节法”进行存储时, 数据的读取将变得非常简单, 读取八字节格式的文件不需要任何的解析规则, 仅需要八字节格式编码的逆算法——解码算法即可。

八字节格式文件的读取方式如下:

1) 以流式数据方式打开“.dat”文件, 每次将八个字

节的内容读入内存, 其中, 第一个字节内容为“属性”字段字节长度, 第二个字节内容为“通道”字段字节长度, 第三个字节内容为“单位”字段字节长度, 第四个字节内容为数据类型, 第五至八字节内容为数据数量, 解析规则和表完全一致, 将读取并解析后的数据存放在内存中;

2) 当读取到八个全为 1 (0xFF) 的字节时, 头文件读取结束, 此时内存中的头文件解析数据即为该文件的“目录”;

3) 在头文件结束标志后, 按照“目录”中的“属性”字符长度、“通道”字符长度、“单位”字符长度、数据类型和数据数量, 依次读取对应的数据并解析。数据读取完成后, 同样存储在标准化数据结构体中。

#### 3.6.3 “八字节法”的优点

“八字节法”标准化数据存储方法具有如下优点:

1) 数据文件完全脱离于解析规则: 按照“八字节法”存储的数据文件不再依赖于复杂的解析规则, 通过算法解码可以直接读取文件的全部内容;

2) 数据解码方法简单: 通过简单的解码算法即可将文件内容解析出来, 算法简单, 易于理解, 通过编写简单的程序模块即可实现数据解析;

3) 数据文件结构固定: 所有的八字节文件格式固定包括“头文件”、“头文件结束标志”、“数据”三部分内容, 使文件各部分易于识别, 各部分内容长度根据需要动态变化;

4) 文件标准化程度高: 数据文件经过文件格式标准化 (均采用“.dat”存储格式)、文件命名标准化、数据类型标准化和编码方式标准化 (大端编码), 标准化程度高;

5) 易于错误排查: 根据八字节编码特性和头文件结束标志, 容易排查“.dat”文件是否存在字节错位、漏字节等情况。

### 3.7 数据标准化的实践效果

针对反应堆关键设备状态数据现有资源, 通过文件格式标准化、文件命名标准化、文件编码方式标准化、文件数据类型标准化、解析规则标准化以及数据存储与读取标准化等手段, 初步实现了数据标准化, 部分经标准化后的数据资源如图 2 所示。可以看出, 所有的数据文件都是“.dat”格式, 并且严格按照前述机组简标规范和文件命名规范进行命名, 所有的数据文件统一采用大端编码、字符采用 UTF-8 编码、1 字节对齐方式, 并且所有文件都按照“八字节法”标准化方法进行存储。

名称	类型	大小
CHA03_LPMS_20161209155100_B8.dat	DAT 文件	313 KB
CJI01_LPMS_20211216111656_B8.dat	DAT 文件	196 KB
CJI01_VBMS_20200409100703_B8.dat	DAT 文件	12,001 KB
CJI02_VBMS_20200629101703_B8.dat	DAT 文件	12,001 KB
FQI03_CPMS_20210820161117_B8.dat	DAT 文件	2,147 KB
FQI05_LKBB_20210507150208_B8.dat	DAT 文件	12,110 KB
FQI05_VBMS_20211101145757_B8.dat	DAT 文件	12,005 KB
HVH05_CPMS_20210726095232_B8.dat	DAT 文件	12,005 KB

图 2 经过标准化处理后的数据文件 (部分)

(下转第 211 页)