ICS 编号
CCS 编号

团　体　标　准

T/CHES XXX—20XX

非结构化水文资料数据库结构标准

Standard for structure of database for unstructured hydrologic data

（报批稿）

20XX-XX-XX 发布　　20XX-XX-XX 实施

中国水利学会　**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。

本文件由山东省水文局提出。

本文件由中国水利学会归口。

本文件起草单位：山东省水文局、山东国基光晔信息科技有限公司

本文件主要起草人：余国倩、张建新、陶光毅、封得华、赵天宇、王效忠、刘春阳、花基尧、郭增、李硕、刘冰、周希海、沈丽娟、王娟、池宸星、张鑫、王慧。

引　言

随着水文事业的发展和信息技术的进步，以及水文行业数据采集能力的不断提升，可收集的数据日益增多，其数据资源呈多源异构、分布广泛和动态增长的态势，利用价值高，且需要长期保存。从数据格式上看，水文资料（水文数据）包括结构化数据和非结构化数据。一般采用关系型数据库存储和管理水文资料结构化数据，而水文资料非结构化数据存储在文件系统，采用关系型数据库进行管理。非关系型数据库突破了关系型数据库严格的库表结构，可存储多种类型、多种格式的水文资料非结构化数据。

非关系型数据库与关系型数据库互不排斥，可以相互交换数据，从而实现相互补充、扩展。水文部门可根据实际需求，充分利用非关系型数据库和关系型数据库各自的特点和优势，实现统一存储和管理水文资料非结构化数据和结构化数据。

非结构化水文资料数据库结构标准

1 范围

本文件规定了非结构化水文资料数据库结构设计及采用非关系型数据库存储非结构化水文资料的技术要求。

本文件适用于非结构化水文资料数据库的设计、建设和应用。

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 32908—2016 非结构化数据访问接口规范

GB/T 50095—2014 水文基本术语与符号标准

SL 21—2015 降水量观测规范

SL 478 水利信息数据库表结构及标识符编制规范

SL 502—2010 水文测站代码编制导则

DA/T 58—2014 电子档案管理基本术语

3 术语和定义

GB/T 50095—2014 界定的以及下列术语和定义适用于本文件。

3.1

水文资料（水文数据） hydrologic data

各种水文要素的测量、调查、记录及其整理分析成果的总称。

[来源：GB/T 50095—2014，5.1.1]

3.2

关系数据库 relational database

用关系数据模型来描述的数据库，其最大特点是采用二维表保存数据。

[来源：GB/T 50095—2014，5.3.1.5]

3.3

非关系型数据库 non-relational database

数据不按关系模型来组织的数据库，其特点是去掉关系型数据库的关系型特性。

注：NoSQL泛指非关系型数据库。

3.4

结构化数据 structured data

能够用二维表结构进行逻辑表达的数据，且严格遵循数据格式与长度规范。

3.5

非结构化数据 unstructured data

没有明确结构约束的数据，如文本、图像、音频、视频等。

[来源：GB/T 32908-2016，2.1]

3.6

文件格式 file format

电子文件在计算机等电子设备中组织和存储的编码方式。

[来源：DA/T 58-2014，2.22]

示例：文本格式 pdf、doc、xls、ppt、txt、wps、xml、html 等；图像格式 JPG、Tiff、GIF、PNG、BMP 等；图形格式 DWG、DXF、IGS 等；音频格式 wav、mp3、mid 等；视频格式 avi、wmv、flv、mpeg、rm 等。

4 原则和要求

4.1 原则

4.1.1 非结构化水文资料数据库结构设计应遵循完整性、一致性、准确性、实用性和规范化、可扩展的原则。

4.1.2 非结构化水文资料数据库结构应满足大规模、多种类型、多种格式的水文资料非结构化数据存储的要求。

4.1.3 非结构化水文资料数据库结构设计中，应对所管理的数据库按学科或业务需求进行分类和编码。

4.1.4 非结构化水文资料数据库结构应具有开放性和包容性，能与搜索引擎集成。

4.2 要求

4.2.1 非结构化水文资料数据库宜采用非关系型数据库。

4.2.2 非结构化水文资料数据库应由记录组成，记录由字段组成，字段存储水文资料非结构化数据和结构化数据。

4.2.3 非结构化水文资料数据库结构内容应包括数据库名、数据库标识、字段描述。

4.2.4 非结构化水文资料数据库名应使用简明扼要的文字表达该数据库所描述的内容，命名准确、无歧义。

4.2.5 非结构化水文资料数据库标识由英文字母、数字和下划线“_”组成，首字符应为大写英文字母。

4.2.6 字段描述应包括字段号、字段名、字段标识、字段类型及长度、是否允许空值、计量单位、是否索引等，并应符合以下规定：

a） 字段号采用数字表示；

b) 字段名采用中文字符表征字段的名称，命名准确、无歧义；

c) 字段标识由英文字母、数字和下划线“_”组成，首字符为大写英文字母；

d） 字段类型及长度描述该字段的数据类型和数据长度；

e） 是否允许空值描述该字段是否允许空值；

f） 计量单位描述该字段数据的计量单位；

g) 是否索引描述该字段数据是否进行索引。

5 字段类型

5.1 非结构化水文资料数据库应能创建不同类型的字段，存储多种类型、多种格式的水文资料非结构化数据和结构化数据，并根据用途和需求变化对字段进行增加、删除和修改。

5.2 字段类型主要有字符、数值、日期、时间、文本、二进制等类型，使用规则按SL478执行。

a） 字符字段用于存储定长字符串和变长字符串，其类型长度描述为：C（d）；

b） 数值字段用于存储整数和实数，其类型长度描述为：N（D[，d]）；

c） 日期字段：Date，用于存储日期类型数据，其格式为：YYYY-MM-DD（年-月-日）；
d） 时间字段：Time，用于存储时间类型数据，其格式为：hh:mm:ss（时：分：秒）；
e） 文本字段用于存储自由文本中的句子和段落，能够存储任意数量的段落、任意数量和任意长度的句子，包括从文本格式水文资料非结构化数据中抽取的文字内容，其描述格式为：Text；
f） 二进制字段用于存储文档、图像、音频、视频等二进制数据，其描述格式为：B。

6 记录

6.1 非结构化水文资料数据库的记录可由任意多个字段组成，字段类型应按 5.2 规定执行。每条记录的长度无限制。
6.2 一条记录可有多个字符、数值、日期、时间的字段。
6.3 一条记录可有多个文本字段，存储多个自由文本，包括从文本格式水文资料非结构化数据中抽取的文字信息。
6.4 一条记录可有多个二进制字段，存储多个水文资料非结构化数据，一条记录中多个水文资料非结构化数据的格式可不同。
6.5 同一非结构化水文资料数据库可存储多种格式的非结构化数据，不同记录的水文资料非结构化数据的格式可不同。
6.6 每条记录在数据库中应有唯一的记录号，记录号应自动生成。
6.7 每个文本格式水文资料非结构化数据装入非结构化水文资料数据库时，应生成字符、文本、二进制和数值等 4 个字段的数据，存储在一条记录中，并符合以下规定：
a) 水文资料非结构化数据名称存储在字符字段；
b) 从文本格式水文资料非结构化数据中抽取的文字信息存储在文本字段；
c) 水文资料非结构化数据存储在二进制字段；
d) 水文资料非结构化数据容量存储在数值字段。
6.8 每个非文本格式水文资料非结构化数据装入非结构化水文资料数据库应生成字符、二进制和数值等 3 个字段的数据，存储在一条记录中，并符合以下规定：
a) 水文资料非结构化数据名称存储在字符字段；
b) 水文资料非结构化数据存储在二进制字段；
c) 水文资料非结构化数据容量存储在数值字段。

7 非结构化水文资料数据库字段定义

7.1 文本格式数据的字段定义

7.1.1 存储文本格式水文资料非结构化数据的数据库字段类型应包括字符、数值、文本和二进制字段。
7.1.2 应按照记录的字段内容建立字段。
示例：水文测站考证簿数据库字段定义详见表 1。

表 1 水文测站考证簿数据库字段定义

字段号	字段名	字段标识	类型及长度	是否允许空值	计量单位	是否索引
1	考证簿名	FILEN	C（255）	否		是
2	电子文件容量	FILEC	N（7）		KB	是

3	文字信息	FILETEXT	Text			是
4	电子文件	FILE	B			否

说明：各字段存储内容为：

a) 考证簿名：数字（化）考证簿的名称；

b) 电子文件容量：数字（化）考证簿容量；

c) 文字信息：从数字（化）考证簿中抽取的文字信息；

d) 电子文件：数字（化）考证簿。

7.2 非文本格式数据的字段定义

7.2.1 存储非文本格式水文资料非结构化数据的数据库字段类型应包括字符、数值和二进制字段。

7.2.2 应按照记录的字段内容建立字段。

示例：流域图数据库字段定义详见表 2。

表 2 流域图数据库字段定义

字段号	字段名	字段标识	类型及长度	是否允许空值	计量单位	是否索引
1	流域图名	FILEN	C（255）	否		是
2	电子文件容量	FILEC	N（7）		KB	是
3	电子文件	FILE	B			否

说明：各字段存储内容为：

a) 流域图名：数字（化）流域图的名称；

b) 电子文件容量：数字（化）流域图容量；

c) 电子文件：数字（化）流域图。

7.3 非结构化数据及结构化数据的字段定义

7.3.1 存储水文资料非结构化数据及结构化数据的数据库字段可包括字符、数值、日期、时间、文本和二进制字段。

7.3.2 应按照记录的字段内容建立字段。

示例 1：水文测站测验资料数据库字段定义详见表 3。

表 3 水文测站测验资料数据库字段定义

字段号	字段名	字段标识	类型及长度	是否允许空值	计量单位	是否索引
1	档号	ARCNO	C（30）	否		是
2	流水号	SEQNO	C（4）			是
3	年度	YR	N（4）			是
4	站码	STCD	C（8）			是
5	站名	STNM	C（24）			是
6	题名	FILETITLE	C（255）			是
7	页数	PAGES	N（4）			是
8	备注	NOTE	C（255）			是
9	电子文件名	FILEN	C（255）			是
10	电子文件容量	FILEC	N（7）		KB	是
11	文字信息	FILETEXT	Text			是
12	电子文件	FILE	B			否

说明：各字段存储内容为：

a） 档号：水文测站测验资料的档号；

b） 流水号：卷内水文测站测验资料的流水号；

c） 年度：水文测站测验资料形成的年度；

d） 站码：水文测站代码应符合 SL 502—2010 的规定，8 位（数字和字母）；

e） 站名：水文测站的中文名称；

f） 题名：水文测站测验资料的名称；

g） 页数：每份水文测站测验资料的页数；

h） 备注：水文测站测验资料内容的说明；

i） 电子文件名：数字（化）水文测站测验资料的名称；

j） 电子文件容量：数字（化）水文测站测验资料容量；

k） 文字信息：从数字（化）水文测站测验资料中抽取的文字信息；

l） 电子文件：数字（化）水文测站测验资料。

注：可根据实际需求，增加或减少字段。

示例 2：水资源资料数据库字段定义详见表 4。

表 4 水资源资料数据库字段定义

字段号	字段名	字段标识	类型及长度	是否允许空值	计量单位	是否索引
1	档号	ARCNO	C（30）	否		是
2	流水号	SEQNO	C（4）			是
3	年度	YR	N（4）			是
4	题名	FILETITLE	C（255）			是
5	页数	PAGES	N（4）			是
6	备注	NOTE	C（255）			是
7	电子文件名	FILEN	C（255）			是
8	电子文件容量	FILEC	N（7）		KB	是
9	文字信息	FILETEXT	Text			是
10	电子文件	FILE	B			否

说明：各字段存储内容为：

a) 档号：水资源资料的档号；

b) 流水号：卷内水资源资料的流水号；

c) 年度：水资源资料形成的年度；

d) 题名：水资源资料的名称；

e) 页数：每份水资源资料的页数；

f) 备注：水资源资料内容的说明。

g) 电子文件名：数字（化）水资源资料的名称；

h) 电子文件容量：数字（化）水资源资料容量；

i) 文字信息：从数字（化）水资源资料中抽取的文字信息；

j) 电子文件：数字（化）水资源资料。

注：可根据实际需求，增加或减少字段。

示例 3：降水自记纸图像文件数据库字段定义详见表 5。

表 5 降水自记纸图像文件数据库字段定义

字段号	字段名	字段标识	类型及长度	是否允许空值	计量单位	是否索引
1	站码	STCD	C（8）	否		是
2	站名	STNM	C（24）			是
3	起始日	BGDY	Date			是
4	起时间	BGTM	Time			是
5	终止日	ENDDY	Date			是
6	止时间	ENDTM	Time			是
7	自然虹吸水量	NSWV	N（6，1）		mm	是
8	查得未虹吸水量	CNSWV	N（6，1）		mm	是
9	查得底水量	CBWV	N（6，1）		mm	是
10	查得日降水量	CDP	N（6，1）		mm	是
11	虹吸订正量	SC	N（6，1）		mm	是
12	虹吸订正后日降水量	SCDP	N（6，1）		mm	是
13	时钟误差	CE	N（2）		min	是
14	备注	NOTE	C（100）			是
15	正面图像文件名	FIFN	C（30）			是
16	正面图像文件容量	FIFC	N（7）		KB	是
17	正面图像文件	FIF	B			否
18	背面图像文件名	BIFN	C（30）			是
19	背面图像文件容量	BIFC	N（7）		KB	是
20	背面图像文件	BIF	B			否

说明：各字段存储内容为：

a） 站码：同表 3 站码字段；

b） 站名：同表 3 站名字段；

c） 起始日：降水量发生时段的起始日期；

d） 起时间：降水量发生时段的起始时刻；

e） 终止日：降水量发生时段的终止日期；

f） 止时间：降水量发生时段的终止时刻；

g） 自然虹吸水量：自然虹吸水量（由储水瓶内雨量换算得到），精确到 0.1mm；

h） 查得未虹吸水量：自记纸上查得未虹吸水量，精确到 0.1mm；

i） 查得底水量：自记纸上查得底水量，精确到 0.1mm；

j） 查得日降水量：自记纸上查得日降水量，精确到 0.1mm；

k） 虹吸订正量：虹吸订正量=自然虹吸水量+查得未虹吸水量-查得底水量-查得日降水量，精确到 0.1mm；

l） 虹吸订正后日降水量：虹吸订正后的日降水量=查得日降水量+虹吸订正量，精确到 0.1mm；

m） 时钟误差：精确到 1min；

n） 备注：记录信息说明，包括未按 SL 21 规定更换自记纸的降水迹线的信息、缺降水自记纸正面和背面填写的信息、降水迹线异常情况；

o） 正面图像文件名：降水自记纸正面图像文件的名称；

p） 正面图像文件容量：降水自记纸正面图像文件容量；

q） 正面图像文件：降水自记纸正面图像文件；

r） 背面图像文件名：降水自记纸背面图像文件的名称；

s） 背面图像文件容量：降水自记纸背面图像文件容量；

t） 背面图像文件：降水自记纸背面图像文件。

注：可根据实际需求，增加或减少字段。

8 非结构化水文资料数据库设计

8.1 库名和库标识编制

8.1.1 编制数据库名，应符合数据库存储的内容。

8.1.2 应根据数据库名编制数据库标识。

8.2 存储位置

8.2.1 应选择数据库和索引文件的存储位置。

8.2.2 数据库和索引文件的存储位置可不同。

8.2.3 各数据库的存储位置可不同。

8.2.4 各索引文件的存储位置可不同。

8.3 中文分词

8.3.1 选择中文自动分词方法。

8.3.2 各数据库的中文自动分词方法可不同。

8.4 字段设计

8.4.1 存储文本格式水文资料非结构化数据的数据库字段类型应包括字符、数值、文本和二进制字段，按照记录的字段内容建立字段。

8.4.2 存储非文本格式水文资料非结构化数据的数据库字段类型应包括字符、数值和二进制字段，按照记录的字段内容建立字段。

8.4.3 存储水文资料非结构化及结构化数据的数据库字段类型可包括字符、数值、日期、时间、文本和二进制字段，按照记录的字段内容建立字段。

9 非结构化水文资料数据库的拆分与合并

9.1 拆分

9.1.1 一个非结构化水文资料数据库可被拆分成若干个子数据库。

9.1.2 非结构化水文资料数据库拆分应以记录为单位。

9.1.3 非结构化水文资料数据库拆分的方法可包括：

a) 建立若干个结构相同的子数据库，根据数据库的记录号从数据库中提取记录，将相应的记录导入各子数据库；

b) 建立若干个结构相同的子数据库，按照记录的容量从数据库中提取记录，将相应的记录导入各子数据库；

c) 建立若干个结构相同的子数据库，根据检索结果从数据库中提取记录，将相应的记录导入各子数据库。

9.2 合并

9.2.1 若干个非结构化水文资料数据库可合并成一个数据库。

9.2.2 非结构化水文资料数据库合并应以记录为单位。

9.2.3 非结构化水文资料数据库合并的方法可包括：

a) 根据各数据库中的记录号从各数据库中提取记录，将相应的记录导入其中一个数据库或新建的数据库；

b) 按照记录的容量从各数据库中提取记录，将相应的记录导入其中一个数据库或新建的数据库；

c) 根据检索结果从各数据库中提取记录，将相应的记录导入其中一个数据库或新建的数据库。

10 水文资料文字信息索引

10.1 宜对非结构化水文资料数据库中的所有文字信息进行索引。

10.2 对非结构化水文资料数据库中的文字信息索引宜包括：

a) 对字符字段每个字、词、词干、整个字段内容进行索引；

示例：观测项目字段的文字内容包括“流量 水位 降水”，对流、量、水、位、降、水、流量、水位、降水等字词进行索引，其中单字“水”出现了两次，都索引，对整个字段内容“流量 水位 降水”进行索引；

b) 对文本字段每个字、词、词干进行全文索引，包括从文本格式水文资料非结构化数据中抽取的文字信息；

c) 对数值、日期、时间进行索引；

d) 对数据库的记录号进行索引。

10.3 水文资料文字信息索引应存储在索引文件中。

11 其他规定

11.1 宜将水文资料非结构化数据装入非结构化水文资料数据库（参见附录 A），不宜将水文资料非结构化数据挂接在非结构化水文资料数据库上。

11.2 水文资料非结构化数据及结构化数据装入非结构化水文资料数据库时，当一个数据库装不下时，可续装入到下一个数据库，可根据数据的容量增加数据库的数量。

11.3 宜从文本格式水文资料非结构化数据中抽取文字信息装入非结构化水文资料数据库的文本字段。

11.4 应对非结构化水文资料数据库中的中文信息进行中文自动分词。宜使用基于词典的中文自动分词方法，并按照水文门类编制相应的中文分词词典。

11.5 非结构化水文资料数据库可存储和备份在磁盘、固态硬盘、光盘等存储介质上，并符合下列要求：

a) 当数据库容量小于存储介质的容量时，可直接将数据库存储和备份到存储介质上；

b) 当数据库容量超过存储介质的容量时，应将数据库拆分成容量小于存储介质容量的若干个子数据库，将各子数据库分别存储和备份到存储介质上。

11.6 对于需要长期保存的非结构化水文资料数据库，宜备份在档案级可录类蓝光光盘上。

11.7 非结构化水文资料数据库的水文资料非结构化数据存储与关系型数据库的水文资料结构化数据存储可形成互补，按照资源配置和使用的需求，构建由非结构化水文资料数据库和关系型数据库组成的异构数据库系统，将水文资料非结构化数据存储在非结构化水文资料数据库，将水文资料结构化数据存储在关系型数据库。

附录 A

（资料性）

水文资料非结构化数据及结构化数据导入

A.1 在线录入水文资料非结构化数据及结构化数据

A.1.1 应按照存储水文资料非结构化数据及结构化数据的数据库字段类型和记录的字段内容（参见 7.3）建立字段以及设计录入表单。

A.1.2 录入表单应由字段名和数据录入区组成，数据录入区与字段关联。

A.1.3 用录入表单，可将水文资料非结构化数据及结构化数据直接录入非结构化水文资料数据库。

A.2 批量导入文本格式水文资料非结构化数据

A.2.1 应按照存储文本格式水文资料非结构化数据的数据库字段类型和记录的字段内容（参见 7.1）建立字段。

A.2.2 可将多个相同或不同格式的文本格式水文资料非结构化数据放入一个或多个文件夹，采用数据导入程序将文件夹中的文本格式水文资料非结构化数据批量导入非结构化水文资料数据库，同时将从文本格式水文资料非结构化数据中抽取的文字信息批量导入非结构化水文资料数据库。

A.3 批量导入非文本格式水文资料非结构化数据

A.3.1 应按照存储非文本格式水文资料非结构化数据的数据库字段类型和记录的字段内容（参见 7.2）建立字段。

A.3.2 可将多个相同或不同格式的非文本格式水文资料非结构化数据放入一个或多个文件夹，采用数据导入程序将文件夹中的非文本格式水文资料非结构化数据批量导入非结构化水文资料数据库。

A.4 批量导入水文资料非结构化数据及结构化数据

A.4.1 应按照存储水文资料非结构化数据及结构化数据的数据库字段类型和记录的字段内容（参见 7.3）建立字段。

A.4.2 可将多个相同或不同格式的水文资料非结构化数据存放在任意位置，将水文资料结构化数据及水文资料非结构化数据的名称和存放位置录入 excel、csv 等文件，采用数据导入程序将水文资料非结构化数据和 excel、csv 等文件中的水文资料结构化数据批量导入非结构化水文资料数据库。

A.4.3 可将多个相同或不同格式的水文资料非结构化数据放入一个文件夹，将水文资料结构化数据录入 excel、csv 等文件，水文资料非结构化数据的名称与水文资料结构化数据中的一个或多个数据有匹配关联关系，采用数据导入程序将文件夹中的水文资料非结构化数据和 excel、csv 等文件中的水文资料结构化数据批量导入非结构化水文资料数据库。