DA

中华人民共和国档案行业标准

DA/T 82—2019

基于文档型非关系型数据库的档案数据存储规范

Specification for storage of archival data based on document-oriented non-relational database

2019-12-16发布　　2020-05-01实施

国家档案局　发布

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

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。

本标准由国家档案局档案科学技术研究所提出。

本标准由国家档案局归口。

本标准起草单位:国家档案局档案科学技术研究所。

本标准起草人:聂曼影、陶光毅、张淑霞、晏杰、王熹、刘春阳、陈吉。

引　言

目前，我国档案部门大都采用关系型数据库存储和管理档案数据。关系型数据库通常将数据存储在二维表中，高度结构化，对数据的规范性要求高。关系型数据库能够很好地管理和存储结构化数据，使用简单、功能强大。

非关系型数据库突破了关系型数据库严格的表结构，解决了关系型数据库模型简单、不易表达复杂嵌套数据结构的问题，存储的数据对象包括非结构化数据、半结构化数据和结构化数据。视处理数据对象而言，目前非关系型数据库主要有键值存储、列存储、文档型和图形四大类。

本标准基于文档型数据库具备的功能编制。文档型数据库允许创建不同类型的字段，存储任意格式的数据。在文档型数据库中，文档是处理信息的基本单位。文档型数据库存储每个对象的所有信息，并且每一个被存储的对象可与任一其它对象不同。采用文档型数据库，能够简单地将不同类型、不同格式的档案数据装入或映射到数据库。

从业务角度看，档案数据主要来自于各类业务系统产生的数据和传统载体档案数字化副本。每条记录一次写入、多次访问、几乎不可更改，适宜采用文档型数据库存储。

文档型数据库与关系型数据库不相互排斥，它们之间可以相互交换数据，从而相互补充、扩展。为了解决大规模档案数据集合多重数据种类带来的挑战，优化档案数据存储，推动档案数据科学管理，鉴于关系型数据库和文档型数据库各自的特点和优势，档案部门可以将文档型数据库、关系型数据库共同作为档案数据存储和管理的工具。

基于文档型非关系型数据库的档案数据存储规范

1 范围

本标准规定了使用文档型数据库存储档案数据的总体要求，提出了使用文档型数据库存储和管理档案数据的基本功能和实施方法。

本标准适用于各级各类档案馆、机关、团体、企业事业单位对档案数据的存储。

2 术语和定义

下列术语和定义适用于本文件。

2.1

数据 data

适合传输、诠释或者手动、自动处理，以一种规范的方式表达的事实或指令。

[GB/T 34840.3－2017，表 A.1]

注：档案数据包括电子档案的内容数据、传统载体档案数字化副本的内容数据以及两者的元数据（含目录数据）。

2.2

元数据 metadata

描述电子档案的内容、背景、结构及其管理过程的数据。

[DA/T 58－2014，定义 2.16]

2.3

内容数据 content data

电子档案和传统载体档案数字化副本固有的信息。

2.4

数据库 database

按照概念结构组织的数据的汇集，它描述这些数据的特征及与数据对应的实体间的关系并支持一个或多个应用领域。

[GB/T 5271.1－2000，定义 01.08.05]

2.5

关系型数据库 relational database

数据按关系模型来组织的数据库。

[GB/T 5271.17－2000，定义 17.04.05]

注：关系模型指用二维表的形式表示实体和实体间联系的数据模型，而一个关系型数据库就是由二维表及其之间的联系所组成的一个数据组织。

2.6

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。

注：NoSQL 泛指非关系型数据库。非关系型数据库去掉了关系型数据库的关系型特性，数据之间无关系，没有关系型数据库严格的表结构，随时可以存储自定义的数据格式。视处理数据对象而言，目前非关系型数据库主要有键值存储、列存储、文档型和图形四大类。文档型数据库允许创建不同类型的字段，存储任意格式的数据。

在文档型数据库中，文档是处理数据的基本单位。

2.7

结构化数据　structured data

能够用统一的结构加以表示的数据，或者说能用二维表结构进行逻辑表达的数据。

2.8

半结构化数据　semi-structured data

具有结构性，但结构变化大，且难以用结构化数据的处理方法将其放进二维表的数据。

示例：XML 文档内容，每项都被一对标记封起来，如＜title＞＜/title＞，表面上看是结构化数据，但＜title＞＜/title＞之间的数据却是千变万化，这是典型的半结构化数据。

2.9

非结构化数据　unstructured data

不能用数字或者统一的结构表示的数据，或没有固定结构的数据。这些数据不能用二维表存放。

2.10

文档型数据库的拆分　document-oriented database splitting

一个数据库被分成若干个数据完整的子数据库。

2.11

文档型数据库的合并　document-oriented database merging

若干个子数据库合成一个数据完整的数据库。

3　总则

3.1　适用性

文档型数据库应适用于大规模、多种类、多种格式档案数据的存储、备份和管理。

3.2　规范性

文档型数据库结构应满足档案数据存储的要求。使用文档型数据库存储档案数据应将内容数据和元数据按照规范结构装入或映射到数据库中，保障两者之间的可靠联系。

3.3　开放性

文档型数据库结构应具有开放性，能与搜索引擎及相关工具软件集成，能与其他类型数据库和系统交换与共享档案数据和相关信息。

3.4　互补性

基于文档型数据库的档案数据存储与基于关系型数据库的档案数据存储和管理可形成互补。文档型数据库是内容数据的存储方法之一，元数据的存储可采用文档型数据库和关系型数据库，日志等的存储和管理宜采用关系型数据库，建立由关系型数据库和文档型数据库组成的数据库系统。

4　文档型数据库字段

4.1　字段要求

文档型数据库应能创建不同类型的字段，存储任意格式的档案数据，并能根据用途和需求变化对字段进行增加、删除和修改。

4.2 字段类型

文档型数据库字段类型宜包括字符串、数值、日期、时间、文本、二进制等，并符合以下规定：

a) 字符串字段：用于存储定长字符串和变长字符串；
b) 数值字段：用于存储整数和实数；
c) 日期字段：用于存储日期类型数据；
d) 时间字段：用于存储时间类型数据；
e) 文本字段：用于存储文本中的句子和段落；
f) 二进制字段：用于存储任何类型的字符，包括 ASCII 字符，以及图像、视频、音频等二进制数据。

5 文档型数据库存储

5.1 存储对象

文档型数据库的存储对象包括内容数据等非结构化数据、XML 文件等半结构化数据以及元数据等。

5.2 存储方式

文档型数据库存储方式包括：

a) 将内容数据和元数据全部装入数据库，按照与元数据的匹配关联关系，内容数据存储在文档型数据库的二进制字段中。
b) 将元数据装入数据库，将内容数据映射到数据库。
c) 同一数据库可存储多种格式的内容数据，不同记录(行)的内容数据的格式可不同。
d) 同一条记录可存储一个或多个内容数据，同一条记录(行)中多个内容数据的格式可不同。

5.3 存储介质

5.3.1 文档型数据库可存储在磁盘、固态硬盘、光盘等存储介质上。

5.3.2 文档型数据库可在不同的存储介质之间转移。

5.4 存储信息管理

文档型数据库存储档案数据的过程中，应显示以下信息：

——存储档案数据进度和存储完成的信息；
——存储过程中出现的问题或错误的信息。

5.5 存储安全

存储档案数据的文档型数据库应保存在安全的存储系统中和存储介质上，防止被非授权改动数据库存储位置的设置，并保障档案数据不被非授权修改、访问、删除、复制和破坏，对授权修改、访问、删除和复制要做审计跟踪。涉密档案数据在线存储应加密，涉密档案数据导出和备份应采取数字签名等安全措施。

6 档案数据导入

6.1 导入对象

导入对象包括内容数据等非结构化数据、XML 文件等半结构化数据以及元数据等。

6.2 导入方法

6.2.1 在线录入数据

在线录入是借助录入表单，将内容数据和元数据直接录入到文档型数据库，一次建立一条记录。

6.2.2 批量导入数据

批量导入是按照内容数据与元数据的匹配关联关系，通过导入程序，将内容数据和元数据批量导入文档型数据库。

6.2.3 导入 XML 文件

将 XML 文件和内容数据通过导入程序导入文档型数据库。

6.2.4 导入数据库数据

将关系型数据库中的数据，通过导入程序导入文档型数据库，并将挂接在关系型数据库的内容数据装入或映射到文档型数据库；或将一个文档型数据库中的数据，通过导入程序导入另一个文档型数据库。

6.2.5 导入其他系统的数据

从其他系统导出的数据按照原有的数据类型和文件格式导入文档型数据库。其他系统包括：

——电子档案管理系统；

——档案管理的其他相关系统；

——形成电子文件的业务系统，等等。

6.3 档案数据导入质量控制

6.3.1 对导入文档型数据库的档案数据的字段类型和文件格式进行设置，并在导入文档型数据库时按照设置自动进行检查，对不合格的字段类型和文件格式显示提示信息。

6.3.2 记录档案数据导入文档型数据库的数量、时间、载体、处理人员、格式转换等处理过程相关信息。

6.3.3 自动监测和过滤错误数据，进行数据完整性校验，日志文件自动记录档案数据导入文档型数据库产生的错误，对错误显示提示信息。

6.3.4 对内容数据、元数据重复导入文档型数据库进行识别。

6.3.5 允许不同归档单位的相同档案数据导入文档型数据库。

6.3.6 设置记录档案数据导入文档型数据库过程的日志，确保能够跟踪、审计、检索（参见附录 A）、统计分析（参见附录 B）。

6.4 档案数据导入的维护

档案数据导入的维护包括：

a) 增加、删除、撤回/恢复、批量修改文档型数据库中的记录；

b) 增加、删除、修改文档型数据库中记录的内容等。

7 文档型数据库的拆分与合并

7.1 文档型数据库的拆分

拆分以记录为单位。拆分的方法包括：

a) 建立若干个结构相同的子数据库，确定分配给各子数据库的原数据库的起始记录号和终止记录号，从原数据库中提取相应的记录，按顺序将记录导入各子数据库。
b) 建立若干个结构相同的子数据库，确定各子数据库的最大容量，计算各记录的容量和分配给各子数据库的容量（分别小于各子数据库的最大容量），从原数据库中提取相应的记录，按顺序将记录导入各子数据库。
c) 建立若干个结构相同的子数据库，利用检索（参见附录 A）和统计分析（参见附录 B），从原数据库中提取相应的记录，按照记录的数量或数据库容量要求导入各子数据库。
d) 建立若干个结构相同、但与原数据库结构不同的子数据库，利用字段读写访问限制功能，从原数据库中提取部分字段的记录，按照记录的数量或容量要求导入各子数据库。

7.2 文档型数据库的合并

合并以记录为单位。合并的方法包括：
a) 确定各子数据库合并记录的记录号，从各子数据库中提取相应的记录，按顺序将记录导入其中一个子数据库或新建的与子数据库结构相同的数据库。
b) 确定各子数据库合并记录的容量，从各子数据库中提取相应的记录，按顺序将记录导入其中一个子数据库或新建的与子数据库结构相同的数据库。
c) 利用检索（参见附录 A）和统计分析（参见附录 B），从各子数据库中提取相应的记录，按照记录的数量或容量要求导入其中一个子数据库或新建的与子数据库结构相同的数据库。
d) 新建一个与子数据库结构不同的数据库，利用字段读写访问限制功能，从各子数据库中提取部分字段的记录，按照记录的数量或容量要求导入新建的数据库。

8 档案数据导出

8.1 导出对象

导出对象包括内容数据等非结构化数据、XML 文件等半结构化数据以及元数据等。

8.2 导出形式

导出包括以下形式：
a) 从文档型数据库导出全部数据；
b) 从文档型数据库按记录（行）导出部分数据；按字段（列）导出部分数据；利用检索（参见附录 A）和统计分析（参见附录 B）导出部分数据；
c) 从文档型数据库导出的元数据存储到 excel、csv、xml 等文件；
d) 从文档型数据库导出的内容数据存储到文件系统。

8.3 档案数据导出质量控制

8.3.1 对导出文档型数据库的档案数据的文件格式进行设置，并在导出文档型数据库时按照设置自动进行检查，对不合格的文件格式显示提示信息。

8.3.2 记录档案数据导出文档型数据库的数量、时间、载体、处理人员、格式转换等处理过程相关信息。

8.3.3 自动监测和过滤错误数据，建立数据导出的检验机制，日志文件自动记录档案数据导出文档型数据库产生的错误，对错误显示提示信息。

8.3.4 设置记录档案数据导出文档型数据库过程的日志，确保能够跟踪、审计、检索、（参见附录 A）、统计分析（参见附录 B）。

9 文档型数据库的备份、还原与恢复

9.1 文档型数据库的备份

9.1.1 备份的内容

文档型数据库备份的内容应包括数据库数据、数据库结构和数据库定义文件。应按照数据库结构(字段)备份数据库数据。

9.1.2 整体备份

当文档型数据库的容量小于备份介质的容量时,无需对文档型数据库做任何处理,直接对数据库整体进行复制备份。

9.1.3 拆分备份

当文档型数据库的容量大于备份介质的容量时,应按照 7.1 所述,将数据库拆分成容量小于备份介质容量的若干个子数据库,然后将各子数据库分别备份到备份介质上,保留原有的访问控制策略,并保证原数据库的完整性。

9.1.4 增量与差异备份

对文档型数据库中新增的档案数据进行增量备份,对被修改的档案数据进行差异备份。

9.1.5 在线备份和离线备份

在线备份是将数据库数据、数据库结构和数据库定义文件备份到在线存储介质上。

离线备份是将数据库数据、数据库结构和数据库定义文件备份到离线存储介质上。

9.1.6 自动备份

自动备份通过软件的控制方式将数据库数据、数据库结构和数据库定义文件有规律的进行备份。

9.2 文档型数据库的还原

9.2.1 整体还原

整体还原是将整体备份的数据库数据还原到原数据库系统。

9.2.2 合并还原

合并还原是将拆分备份的多个子数据库数据还原到原数据库系统,采用的方法包括:

a) 在新建数据库系统合并全部子数据库数据,然后将合并形成的数据库数据还原到原数据库系统。

b) 在原数据库系统中合并还原全部子数据库数据。

9.3 文档型数据库的恢复

9.3.1 正常恢复

用备份的数据库数据覆盖原数据库数据。

9.3.2 异常恢复

在异常情况下，用备份的数据库数据、数据库结构、数据库定义文件和日志文件进行恢复，并进行数据完整性校验，以确保数据的完整性。

10 文档型数据库管理与数据管理

10.1 数据库管理

数据库管理包括：

a) 数据库结构的定义、设计、复制、导入、导出。
b) 数据库的新建、删除、修改；数据库字段的新建、删除、修改、排序。
c) 数据库管理权限的新建、删除、转移。
d) 数据库管理员授权用户访问数据库。数据库授权访问包括但不限于：授权用户访问全部数据库或部分数据库；授权用户访问数据库的全部字段或部分字段；授权用户访问包含特定内容的记录或不包含特定内容的记录。
e) 数据库存储位置的设置；存储数据库的存储介质的选择。
f) 数据库的备份、复制、转移存储、迁移、还原与恢复。

10.2 数据管理

数据管理包括：

a) 档案数据导入和导出数据库；
b) 数据库拆分与数据库合并；
c) 数据库记录的增加、删除、修改；
d) 数据库数据的恢复、还原及内容覆盖的检测提醒。

10.3 用户管理

用户管理包括：

a) 用户的新增、修改、删除、激活、锁定等；用户信息管理和维护；
b) 用户的分组、分类与权限控制与管理；
c) 记录用户访问数据库，包括但不限于：
   ——登录系统的时间、次数；
   ——访问数据库和档案数据的时间、次数；
   ——访问数据库的名称和档案数据的内容；
   ——用户登录的 IP 地址。

10.4 日志管理

日志管理包括分类、存储、备份、检索、查询和管理日志信息。

附 录 A
（资料性附录）
检索

A.1 中文自动分词与中文分词词典

中文索引和检索是对档案数据中的中文单字和词（或词组）进行索引和检索。为提高文档型数据库中档案数据的查准率和检索速度，实现对中文内容的统计分析，应对内容数据和元数据的中文信息进行中文自动分词，并根据不同的档案门类编写相应的中文分词词典。

A.2 索引

A.2.1 应具有对文档型数据库中的内容数据和元数据的文字信息进行全文索引的功能，包括：

——字符串字段每个字、词、词茎、整个字段内容进行索引；

——对文本字段每个字、词、词茎进行全文索引。

A.2.2 应具有对日期、时间和数值进行索引的功能。

A.2.3 索引与数据库可保存在同一位置，也可分别保存在不同位置。

A.3 检索

A.3.1 检索范围

检索范围包括内容数据和元数据的文字信息。

A.3.2 数据库一般检索

数据库一般检索包括：

——全文检索，是对所有字符串字段和文本字段的全部字符进行检索；

——字段检索，是对具体字段的数据和内容进行检索，包括对字符串字段、文本字段、数值字段、日期字段、时间字段的数据和内容进行检索；

——分类检索与模糊检索；

——布尔检索，是用逻辑"或"、逻辑"与"、逻辑"异或"、逻辑"非"等运算符进行检索，带运算符检索的效果如图 A.1 所示；

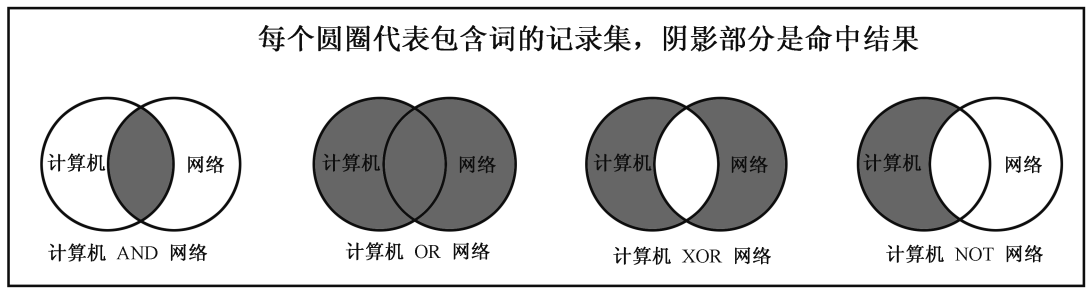

图 A.1 布尔检索效果图

——二次检索，是在前面（上次或更前面）检索结果的基础上进行检索；
——数值检索，是用运算符在指定数值范围进行检索；
——日期检索，是用运算符在指定日期范围进行检索；
——时间检索，是用运算符在指定时间范围进行检索；
——“时间戳”检索，是查找某个时间段新增或修改的记录。

A.3.3 数据库检索限制

数据库检索限制包括：
——检索和显示部分数据库；
——检索和显示部分字段；
——检索和显示包含特定内容的记录；
——对部分信息进行屏蔽和遮盖。

A.3.4 多库检索

多库检索包括：
——同构多库检索，对相同结构的多个文档型数据库同时进行检索；
——异构多库检索，对不同结构的多个文档型数据库同时进行检索。

A.3.5 数据库间接检索

文档型数据库之间在结构上没有相互关联的字段，每个数据库都是完全独立的数据库。文档型数据库之间的关联查找需采用间接检索。间接检索是用一个数据库检索所获得的结果作为检索项，对本数据库或其他数据库进行检索。

示例：病症数据库收藏关于各种疾病的症状、失调和治疗的数据，病历数据库收藏各种疾病诊断的病历。在病历数据库能查找到“睡眠失调”的病历，但不能找出与“睡眠失调”症状相关的记录。通过间接检索，先在病症数据库中查找“睡眠失调”，然后用命中记录之症状字段的内容自动作为检索词在病历数据库进行检索。

A.4 词表与词表检索

A.4.1 词表

词表指主题词表或叙词表，是记录有关词的内容、概念以及相互关系的语言工具。在文本处理系统中，词表是提高检索质量的强有力的工具。典型的词表是根据意思或内容把每个词与词（或词组）之间的关系，按照树状（或其他形状）结构上下左右相互关联起来的数据库。词表应包括但不限于如表 A.1 所列的五种词：主词（控制词）、同义词、广义词、狭义词和相关词，它们既可以是中文的词，也可以是其他语种的词。

表 A.1 词表结构

名称	别名	角色名	标记	含义说明
主　词	受控词、优选词	用 或 Y	CT	表达某一特定概念的标准术语
同义词	非优选词	代 或 D	UF	与主词含义相同的其它称谓
广义词	上位词	属 或 S	BT	概念上包含主词的上一层术语
狭义词	下位词	分 或 F	NT	概念上比主词更狭的下一层术语
相关词	关联词	参 或 C	RT	概念上和主词有关联、并列的术语

A.4.2 词表检索

词表检索是用词表库中的词及词之间的逻辑关系对文档型数据库中的文字信息进行检索，是保障查全率、查准率，提高检索质量的强有力手段。

词表检索方法包括：

——用主词和同义词进行检索；

——用主词、同义词和广义词进行检索；

——用主词、同义词和狭义词进行检索；

——用主词、同义词和相关词进行检索。

示例：主词是自行车，同义词是单车、脚踏车和 bicycle。采用主词和同义词进行检索时，只输入“自行车”进行查找、就会自动查找含有“自行车、单车、脚踏车、bicycle”等四个词的记录。

A.5 检索结果排序和打印

A.5.1 对检索结果进行排序，包括单级排序和多级排序。多级排序是指检索结果的输出先按某一字段内容（如题目）排序，如遇有内容相同的多条记录，再按另一字段内容（如单位名称）排序。

A.5.2 对检索结果进行全部或部分打印、输出和下载。

附 录 B
（资料性附录）
统计分析

B.1 统计分析的对象

档案数据的统计分析对象是内容数据和元数据的文字信息。

B.2 统计分析的方法

B.2.1 按照字段类型进行统计分析

按照字段类型进行统计分析包括：

a) 字符串字段统计分析。对整个字段内容和字段中的词汇（包括词、单字、数字等）分别进行统计分析，给出整个字段内容相同和不同的记录数、字段所含词汇出现的频度和记录数。

示例：全宗名、归档部门、题名、发文单位、责任者等字段是字符串字段，利用统计分析，获得全宗名、归档部门、发文单位、责任者的数量和分布。

b) 文本字段统计分析是对文本字段中的词汇进行统计分析，给出不同词汇出现的频度和记录数。

c) 数值字段统计是对给定的记录，统计最大、最小、总和、平均等数值。

d) 数值字段统计分析是对给定的记录、数值范围和间隔，列出数值分布情况。

e) 日期字段统计分析是对给定的记录、日期范围和间隔，列出各日期段的分布情况。

f) 时间字段统计分析是按给定记录范围、时间范围和间隔，列出各时间段的分布情况。

B.2.2 档案数据内容的统计分析

按照档案数据内容进行统计分析，包括：

a) 内容数据和元数据的数量和容量的统计分析；

b) 档案全宗、门类、归档部门、责任者、题名等的统计分析；

c) 档案数据文件格式的数量和容量的统计分析；

d) 内容数据和元数据的保密、解密、开放、保管期限的统计分析；

e) 不同时间段各类数据和内容的统计分析；

f) 各类日志信息查询的统计分析。

参 考 文 献

[1] GB/T 5271.1—2000 信息技术 词汇 第1部分:基本术语

[2] GB/T 5271.17—2010 信息技术 词汇 第17部分:数据库

[3] GB/T 18894—2016 电子文件归档与电子档案管理规范

[4] GB/T 34840.3—2017 信息与文献 电子办公环境中文件管理原则与功能要求 第3部分:业务系统中文件管理指南与功能要求

[5] DA/T 58—2014 电子档案管理基本术语