数据加工管理及数据加工日常工作中的文件管理探索

专利文献部 马媛媛 张红生 张旭

摘要:针对当前数据加工管理及数据加工日常工作中所面临的文件管理问题,本文提出一种应用MD5进行文件质量管理的方法,在进行存储、共享和传输时,可保证数据文件的安全性和完整性,确保无中途修改。与传统人工管理修改文件相比,该方法可方便地生成质量管理文档,避免人工误操作的同时,还可以检查被病毒恶意篡改的文件。

关键词:数据加工 数据加工管理 文件管理 质量管理

引言

目前，数据加工管理处的文件管理模式为将需要存档的文件分别进行人工命名后，存储于规定的共享空间中相应的文件目录下，以便供各位同事阅读及拷贝。这种存储模式在为大家提供方便的同时，也表现出了各种各样的弊端。以质检组日常工作中的

质检工作为例，按期完成的检测单一般是由相应的质检员各自命名后，上传至共享空间中的指定路径下，供复核拷贝等操作，由于该过程中可能出现在共享空间内打开文件，改动后不慎保存等误操作，不可避免地造成对文件的不可逆操作，影响后续他人对该文件的读取结果。再者，例如检测单汇总人员在对各个检测单进行汇总时，将质检员误发送的以往批次的某检测单当成当前检测批次的检测单进行了误汇总，以致汇总结果错误，却又得不到及时发现，就会造成重复老公，无形中增加工作量。

类似的，这类问题存在于数据加工工作中。例如，在将加工单正常存储于局域网内服务器时，加工单一般是同步自动保存到本地机，但有时本地机上却无任何记录（且加工人员也未及时另存至另一目录），此时若从服务器上找不到相应的记录，则需要加工人员手动地对该条记录重新进行加工。考虑到时间延续性，如果加工人员对该条记录所对应的文献印象模糊，本次重新标引则几乎相当于重新做一件新案，十分浪费时间，不利于提高加工效率。因此，希望保存各个时期的加工文件并对其进行及时备份，以便回溯到数据加工的各个阶段。待质检完成后，再根据具体情况决定是否需要将这批数据及时删除以节约硬盘空间。另外，在试验阶段规则不断变化，保存各个阶段重新加工后的数据，有利于后续工作中随时取用各个阶段的加工类目不同的加工数据。但这其中也存在数据加工管理中类似的问题。

此外，做课题研究以及日常工作中也经常生成不同版本的文件，如目前正在进行的案例库整理工作，若误修改共享空间中的最终版本，又由于最终版本时间较长，本地机上未必会有相应数据，就会造成大家工作中很多的不方便。

综上所述，可以看出目前的文件管理主要是依靠人工、形成文件库，在统一的共享空间内进行统一的规划存储。这种方法至少存在以下几点不足：管理文档不方便，需要人工逐个对比；更新后，如果文件的大小无变化仅依赖文件的修改日期来控制文件是不可靠的；当文件感染病毒后，使用手工对比的方式不容易发现问题。因此，如何对这些日常文件加以科学规范化的管理及高效利用是当前数据加工以及数据加工管理工作中的重点。针对上述问题，考虑对文件进行质量管理。针对上述的具体问题，本文提出一种应用 MD5 进行文件质量管理的方法：本文提出的方法使用软件自动生成文件质量管理文档，在进行存储、共享和传输时，可保证数据文件的安

全性和完整性，确保无中途修改。

一、MD5 特点分析

MD5(Message Digest Algorithm 5，消息摘要算法第五版）是计算机安全领域广泛使用的一种散列函数[1, 2]，用提供消息的完整性保护。该算法的文件号为 RFC 1321，在 90 年代初初期由 MIT Laboratory for Computer Science 和 RSA Data Security Inc. 的 Rivest 开发出来，经 MD2、MD3 和 MD4 发展而来。其作用为让大容量信息在数字签名软件签署私人密钥前被“压缩”成一种保密的格式（把任意长度的字节串变换成一定长的大整数），其中 MD2、MD4、MD5 都需要获得一个随机长度的信息并产生一个 128 位的信息摘要。但 MD2 的设计与 MD4 和 MD5 不同，MD2 是为 8 位机器做优化设计的，而 MD4、MD5 是面向 32 位的电脑。MD5 在 MD4 的基础上增加了“安全－带子”(safety-belts) 的概念。虽然 MD5 比 MD4 稍微慢一些，但更为安全。MD5 最广泛被用于各种软件的密码认证和钥匙识别上。该哈希函数 H(M) 有如下特征：

1. H 可以作用于一个任意长度的文件；

2. 可以产生固定长度的输出；

3. 容易计算出 h=H(M)；

4. 对于任意给定的 h，很难找到 M 满足 H(M)=h；

5. 任意给定的 M，找到 M′ 满足 H(M)=H(M′) 在计算上具有不可行性；

6. 任意数据对 (x，y)，满足 H(x)=H(y) 在计算上是不可行的。

MD5 计算程序以只读方式读取文件内容，计算生成 MD5 值，同时显示该文件名称。可以同时计算多个文件的 MD5 值，选定需要的 MD5 值，复制，粘贴到选定的文件中。该 MD5 值可以检验所拷贝文件的版本，以便确认在拷贝和使用过程中保证文件的完整性和安全性。

日常的数据加工和数据加工管理工作中涉及到的文件不外乎以下几种：（1）word 文档，其文件名后缀是 doc；（2）文本文件，其文件名后缀是 txt；（3）excel 文件，其文件名后缀是 xls 等；（4）配置和日志文件，其文件名后缀是 CFG、INF、LOG 等；（5）临时文件；（6）PDF 和 caj 文件等。在日常文件的质量管理中，需要管理的主要是前三类文件，具有各自的文件名称，文件目录位置、修改日期、文件大小和文件版本号等属性。其中，文件名称、目录位置和内容，这三者是文件质量管理的关键。目录遍历方法可以确定文件名称和目录位置；MD5 根据文件内容产生文件的信息摘要特征。目录遍历方法是十分成熟的技术，本文不予赘述，在编程环

境中通过调用库函数实现[3]。

二、文件质量管理文档的生成

针对本文所介绍的基于MD5的文件质量管理方法，分别以数据加工管理处和数据加工日常工作中的内容为例进行分析验证。

文件质量管理文档生成的具体步骤如下：

首先根据选定的文件，生成文件的MD5值；将文件的名称，文件目录位置和MD5值按照统一格式存储在一个文件中，作为上述文件的质量管理文件，然后应用该质量管理文件检验测所对应的存储文件。

先以数据加工管理处dataprocess共享空间中日常工作内容为例进行分析。数据加工质检工作流程如下：质量检测人员完成所分配的检测任务后，上传至“\\10.50.1.13\dataprocess\数据加工管理处 项目管理\数据深加工检测共享\开发公司专利文献深加工\检测结果（供处内人员粘贴结果使用）\2011年度各月检测单和错误统计\2011年12月5日批次数据”子文件夹内。

在规定的完成日期，按照复核表所规定的复核对象，从该路径下，拷贝相应检测人员的检测单,进行复核。复核期满后，各检测人员根据复核人员的复核意见及讨论结果，对检测单进行修改，发送给汇总人员进行汇总后，发布本批次数据错误和特殊建议单，供后续本批次讨论会使用等。

由于上述过程中可能出现在共享空间内打开文件，改动后误保存等误操作，不可避免地造成对文件的不可逆操作，影响后续他人对该文件的读取结果。再者，例如检测单汇总人员在对各个检测单进行汇总时，将质检员误发送的以往批次的某检测单当成当前检测批次的检测单进行了误汇总，以致汇总结果错误，却又得不到及时发现，就会造成重复劳动，无形中增加工作量。

以“\\10.50.1.13\dataprocess\数据加工管理处 项目管理\数据深加工检测共享\开发公司专利文献深加工\检测结果（供处内人员粘贴结果使用）\2011年度各月检测单和错误统计\2011年12月5日批次数据”子文件夹内的检测结果为例，分别给出改动前和改动其中一部分文件后各自的对应结果。

假设，所述共享空间dataprocess中的文件目录结构如图1所示。

改动前后的质量检测文档分别如表1和表2所示。

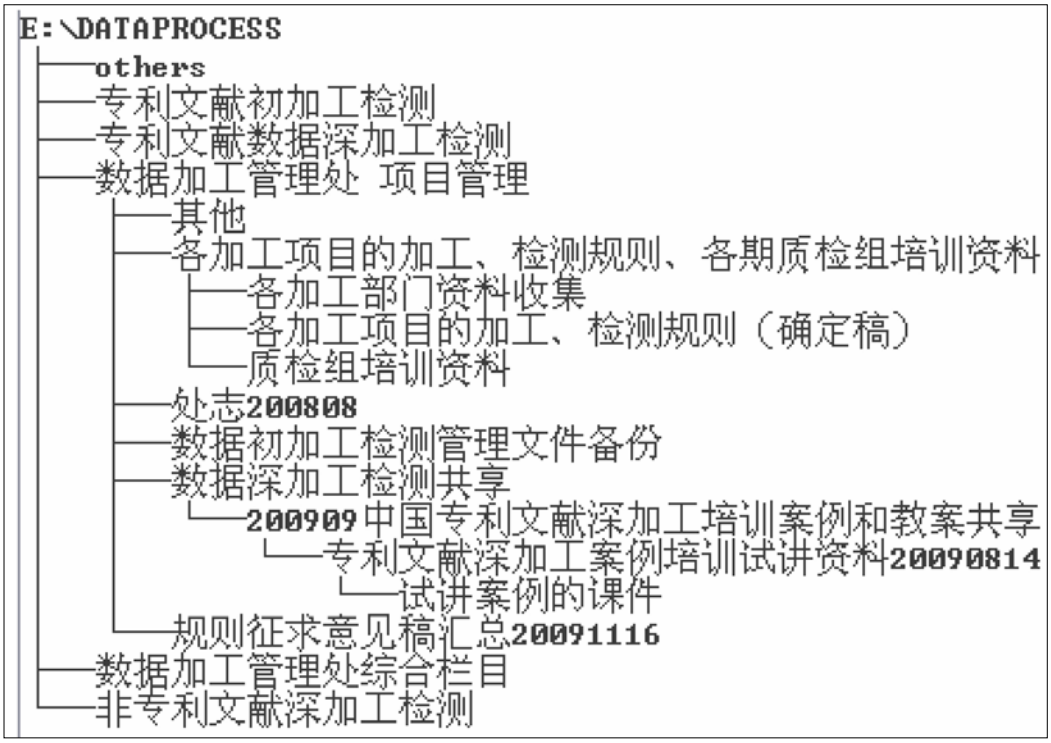

图 1 共享空间 dataprocess 中的文件目录结构

表 1 20111205 检测批次各检测单质量管理文档（修改前）

检测批次:20111205	文件目录位置：\\10.50.1.13\dataprocess\数据加工管理处 项目管理\数据深加工检测共享\开发公司专利文献深加工\检测结果（供处内人员粘贴结果使用）\2011 年度各月检测单和错误统计\2011 年 12 月 5 日批次数据			
检测代码	文件名称（*.xls）	文件内容（MD5）	备注	
			文件大小(字节)	修改时间
200803	200803-30111205	3960CD3AE27254950CA03017EA5440D7	1307648	2011 年 12 月 14 日，16:29:38
200904	200904-20111205	4D554510E7C93D9227C7292F3E2053EE	1091584	2011 年 12 月 15 日，14:31:42
200907	200907-20111205	99278BB57A3AA01FDDCC61B76B46FF3D	576000	2011 年 12 月 16 日，10:02:54
201101	201101-20111205	033D5BE8784A61256E69CF5053E0D726	1071616	2011 年 12 月 15 日，13:42:16
……	……	……	……	……
200107	201107-20111205	527B0EE203D010D03DD8C1F9A9957E99	318464	2011 年 12 月 16 日，9:56:22

注：1. 按照实际要求本批次数据的检测时间为 2011 年 12 月 8 日至 2011 年 12 月 15 日下午 14:00 前；

2. 假设上表中的修改时间中记录的时间点仅为上传时间点。

若仅误改动 201107-20111205.xls，则重新计算结果如下所示：

表 2 20111205 检测批次各检测单质量管理文档（修改后）

检测批次：20111205	文件目录位置：\\10.50.1.13\dataprocess\ 数据加工管理处 项目管理 \ 数据深加工检测共享 \ 开发公司专利文献深加工 \ 检测结果（供处内人员粘贴结果使用）\2011 年度各月检测单和错误统计 \2011 年 12 月 5 日批次数据			
	文件名称（*.xls）	文件内容（MD5）	备注	
			文件大小（字节）	修改时间
200803	200803-30111205	3960CD3AE27254950CA03017EA5440D7	1307648	2011 年 12 月 14 日，16:29:38
200904	200904-20111205	4D554510E7C93D9227C7292F3E2053EE	1091584	2011 年 12 月 15 日，14:31:42
200907	200907-20111205	99278BB57A3AA01FDDCC61B76B46FF3D	576000	2011 年 12 月 16 日，10:02:54
201101	201101-20111205	033D5BE8784A61256E69CF5053E0D726	1071616	2011 年 12 月 15 日，13:42:16
……	……	……	……	……
200107	201107-20111205	80930206B52596CD902E4867E46F59A0	315392	2011 年 12 月 26 日，13:27:02

从上述表 1 和表 2 中所列的数据可以看出，如果对 200107-20111205 文件进行了改动，则相应的 MD5 值就发生了变化（仅改动文件名，不改动文件的情况下，MD5 值不会发生变化），再结合文件大小和修改日期，可确定文件的具体修改情况，根据对应参数，追溯相应的文件。针对具体情况，还可以增加 SHA1 和 CRC32 校验，进一步确保安全性。

由以上所描述的处内质检流程中也可以看出，在质检过程中本地机上也会保存几个版本的过程文件，如首次检测结果、复核后对检测单会进行一次修改、讨论会后对检测单的修改以及撰写检测分析报告时对检测单的修改等等。这些文件很容易混淆，如采用本文中提出的方法对各个时期的

文件进行管理,发送给汇总人员时(或汇总人员进行汇总时）可以查询相应参数，避免误发送，造成后续工作中的不便。

下面列举另外一个典型案例来进行说明。对“\\10.50.1.13\dataprocess\数据加工管理处 项目管理\数据深加工检测共享\2010年深加工（高层次）培训案例共享\中国专利文献深加工标准案例库201101”文件夹下的案例库相关内容来说，在近期整理案例库的过程中，出现如下情况，例如，误修改共享空间中的最终版本。由于并不是每个人都有定时从共享空间中将需要文件拷贝到本地机的习惯，通常，一般情况下都是需要进行某项工作的时候，再从共享空间中拷贝。对于这部分数据来说，由于最终版本时间较长，本地机上未必会有相应记录，则会造成大家工作中很多的不方便。下面分别申请号为200520121714的案例的不同版本的数据进行了校验，其对比结果如表3所示。

表3 申请号为200520121714的案例的对比结果

申请号		200520121714		
	类型	文件大小（字节）	修改时间	MD5
改后文件	分析表	50176	2011年12月23日，10:06:49	1A823E43DCF99341FAD58A4EF58B33CC
	加工单	91136	2011年12月23日，10:06:49	A0E406160FD5FA9B7B7DE70963425E68
核对文件	分析表	51712	2011年3月14日，16:38:01	E8C9CE0472AB679B3C68DF7105BF38AB
	加工单	91136	2010年11月5日，9:31:42	A0E406160FD5FA9B7B7DE70963425E68

从上表中可以看出，共享空间中的分析表和加工单在“2011年12月23日,10:06:49”被修改过，但是在接下来的工作中该改动被及时发现，指出共享内容未经许可不能修改后，相关人员对其进行了修改（按要求，应退回为改动前的内容）。但是，将按要求进行退回修改的文件与最终核对版本的对应的文件进行比较后，可以看出，申请号200520121714的案例“分析表”所对应的MD5值前后不一致，证明所述“分析表”还是被改动过，并未完全的退回到最终核对版本（即本次整理工作开始前供大家拷

贝修改的版本），这一点从改动后分析表的大小“50176”字节和改动前分析表的大小“51712”字节也能够看出。这就会给工作中带来很多麻烦，比如前期拷贝跟后期拷贝的版本不一致，前期已经发表的案例（已核对的案例）定稿的结果发生变化，后续案例入库的时候，如果误采用改动过的共享中的文件，可能会在造成结果不准确等结果（具体内容，还需要逐个对文件进行验证）。

下面，以检索中心机械电学等领域以往数据加工工作中的内容为例进行分析。

回顾从2008年至今的加工过程，到现阶段为止，各个阶段的加工项目如表4所示：

表4 各阶段深加工项目列表

阶段	深加工字段							
	标题	摘要	附图	关键词	同义词	范畴	IPC	材料组分
第一阶段	✓	✓	✓	✓	✓	✓	✓	✓
第二阶段	✓	✓	✓	✓	✓	✓	✓	✓
第三阶段	✓	✓	✓	✓	✓		✓	✓
第四阶段	✓	✓	✓	✓	✓		✓	✓
第五阶段	✓	✓	✓	✓	✓			✓
第六阶段	✓	✓	✓	✓	✓			✓
第七阶段			✓		✓		✓	✓

下面分阶段进行一下简单的介绍：

第一阶段，为最初的试验标引阶段，这一阶段要求的比较全面，依托于医药领域非专利文献的加工经验，分别从标题、摘要、附图、关键词、同义词、范畴分类、IPC分类以及材料组分等几个方面进行探索，其中标题的格式要求为“原标题/修改后标题”；摘要要求先给出原加工，然后深加工摘要另起一段标引内容规定为对原文摘要中所遗漏的技术方案信息进行补充；摘要附图要求对多个附图进行处理后，作为一幅图存储于非专利文献数据加工标引工作单的FIG标引项内；关键词项目要求保留原文关键词作为AKW，新标引的关键词作为KW项目；同义词按照要求进行标引；同时标引内部范畴分类号；对该条记录给出对应的IPC分类号；对于组分和配比均完整的材料组分，应对其进行标引，置于材料组分标引项目内。

第二阶段，在第一阶段加工的基础上进行改进。标题要求仅保留修改后的标题，原标题删除；摘要要求删掉原文摘要，重新对文中的技术方案进行标引；附图要求选取最能反应非专利文献技术主题的一幅图作为摘要附图；关键词类目依旧是原文关键词保留作为 AKW，新标引的关键词，作为 KW 项目;内部范畴分类号、同义词、IPC 分类以及材料组分项目标引要求与第一阶段相同。

第三阶段，标题要求仅保留修改后的标题，原标题删除；摘要要求删除原文摘要，重新对其按要求进行标引且多技术方案建议分段进行撰写。同时，取消内部范畴分类号的标引；其余类目与第二阶段的标引要求相同。这一阶段初步开始研究了多主题技术方案的标引方法，可看做多主题技术方案阶段 1。

第四阶段，这一阶段可以看做试验的过程，可看做多主题技术方案阶段 2，该阶段要求标引的项目，还是与第三阶段相同，不同之处在于不同的技术方案要求进行分段处理。标题和摘要中要求将不同技术方案放在不同的加工单内，即分别对各个技术方案进行单独标引。

第五阶段，标题要求仅标引修改后的标题；摘要的要求是删掉原文摘要，重新进行标引且多技术方案分段撰写；不同之处在于，此阶段的摘要附图要求单图。

第六阶段的摘要附图要求多图，其余要求与第五阶段相同。

第七阶段，标引项目仅包括附图、同义词、IPC 分类号和材料组分。

在试验阶段规则不断变化的时期，保存各个阶段重新加工后的数据，有利于后续工作中随时取用各个阶段标引项目不同的标引数据。在这一特殊阶段采用文件质量管理，可以避免误修改上一时期的版本，避免造成文件混乱，可有效保存各个阶段数据。

三、结论

本文提出一种应用 MD5 进行文件质量管理的方法，针对当前数据加工管理及数据加工日常工作中所面临的具体文件管理问题，确保在进行存储、共享等操作时数据文件的安全性和完整性，确保无中途修改。与传统人工管理修改文件相比，该方法可方便地生成质量管理文档，避免人工误操作的同时，还可以检查被病毒恶意篡改的文件。

（专利文献部　李明　审校）

参考文献

1. 蔡红柳、何新华.《信息安全技术及应用实验》. 科学出版社.2004年第69-72页。

2. Brandon Staggs. File check MD5 [EB/OL]. 2004:1[2009-11-01]. http://www.brandonstaggs.com/filecheckmd5.html. (2011年6月30日最后访问)。

3. Tom Archer, Nishant 5ivakumar. 周良忠译.《MFC应用程序在.NET框架下的扩展》. 北京电子工业出版社.2005年第96-98页。