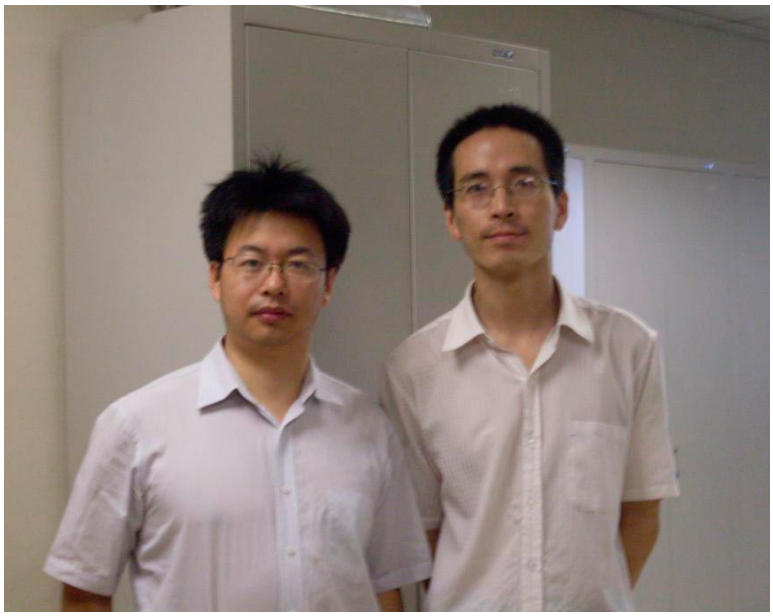
非专利文献数据加工系统流程和协作机制研究

专利检索咨询中心 颜平辉 孙亮



摘要: 本文分析了非专利文献数据加工系统流程管理的复杂性,设计了非专利文献数据加工系统数据存储逻辑结构和加工流程,研究了流程串行化、任务流转、数据传递机制,实现了数据加工流水线作业,任务实时流转,数据安全访问,多人协同工作,为工作顺利开展奠定了技术基础。

关键词: 非专利文献数据加工系统 流程管理
协同工作

一、研究背景

非专利文献数据深加工是我局信息资源建设的重要内容。非专利文献涉及各个技术领域，数据量巨大。目前以加工医药领域文献为主，从中文期刊中选取在相关专利审查工作中引用比率较高的 822 种期刊，主要标引内容包括化合物信息、方剂信息和同义词。

非专利文献数据加工系统是专利

检索咨询中心为非专利文献数据加工顺利开展自主研发的业务管理系统。系统需要解决如下基本问题：标引工作涉及数据量较大，选取期刊原始文献超过350万篇,需要有效管理数据;加工流程包括原始数据准备、数据筛选、标引、校对、质检等主要环节,需要各环节紧密衔接,数据流转顺畅;加工人员涉及多种角色，包括系统管理员、筛选员、校对员、结构绘制人员、质检员，每类角色同时有多人在工作，要保证数据并发访问正确性,保留各工作环节在每条数据上产生的修改“痕迹”。以上问题如果不能合理解决，数据加工效率和质量将受到严重影响。因此，开展非专利文献数据加工系统流程和协作机制研究是非常有意义的。

本文在数据集中存储的前提下,研究合适的数据流程管理存储结构、数据流转机制、界面组织方式、权限控制和协作机制，从而实现大数据量下多人并发访问数据，各加工环节无缝衔接,数据实时流转,权限按需控制。

二、数据存储逻辑结构

数据存储逻辑结构以单篇文献为对象，记录了该文献在数据加工过程中标引各阶段相关数据和流程管理信息，如图1所示。

标引各阶段相关数据包括原文及提取信息、各阶段标引文档、标引单提取信息和化学结构信息。原文提取信息主要包括文章题名、作者、摘要、关键词、正文、期刊名、卷、期、出版日期等信息，通过OCR扫描识别获得。由于OCR正文识别错误较严重,不能依据该正文信息进行标引，只能将这些信息用于任务分配和数据筛选。因此，在确定要标引某篇文献时需要上传原文，数据标引依据原文进行。各阶段标引文档包括标引文档、校对文档、校返确认文档、质检文档、通过确认文档、提交文档，每个处理阶段即数据会发生变化的环节都会单独保存文档。在多人共同处理同一标引单的情况下，为了使后续修改不会覆盖前一阶段处理结果，保存了各自的工作痕迹，且他人不能修改，为实现明晰的责任认定机制奠定了基础。同时，如果后续处理环节错误，可以很方便回溯到前面任一处理环节，提取正确数据。标引数据以Word文档(*.doc)保存在数据库中，优点在于标引人员可以利用Word强大的编辑功能，比如Word中的审阅功能，不仅可保留修改痕迹,还便于修改确认,用于标引员和校对员之间，标引员和质检员之间信息沟通。同时，Word文档真实存储并保留了标引时的数据和原始样式，如果只在数据库保存Word文档提取数据信息，对于特殊

符号，如上下标，数据库无法正常保存和查询。此外，Word标引单也便于记录质检和修改情况。由于Word文档内信息不能直接用于检索，还需要对Word文档信息进行提取并存于数据库表中，一方面用于加工过程中数据检索，另一方面，可以检查提取信息，发现有系统性错误时进行批量修改，并进一步修改Word文档。数据库存储了多个Word文档，只要保存文档，系统都会自动提取信息，从而保证查询数据是最新的。Word文档提取信息中包含了对化学结构图形图像数据的引用，因此，数据库中还存储了化学结构图形图像数据。以CN为主键,结构图形以Mol文件存储，用于基于结构式的化学结构式检索，图像以Gif文件存储,用于数据展示。

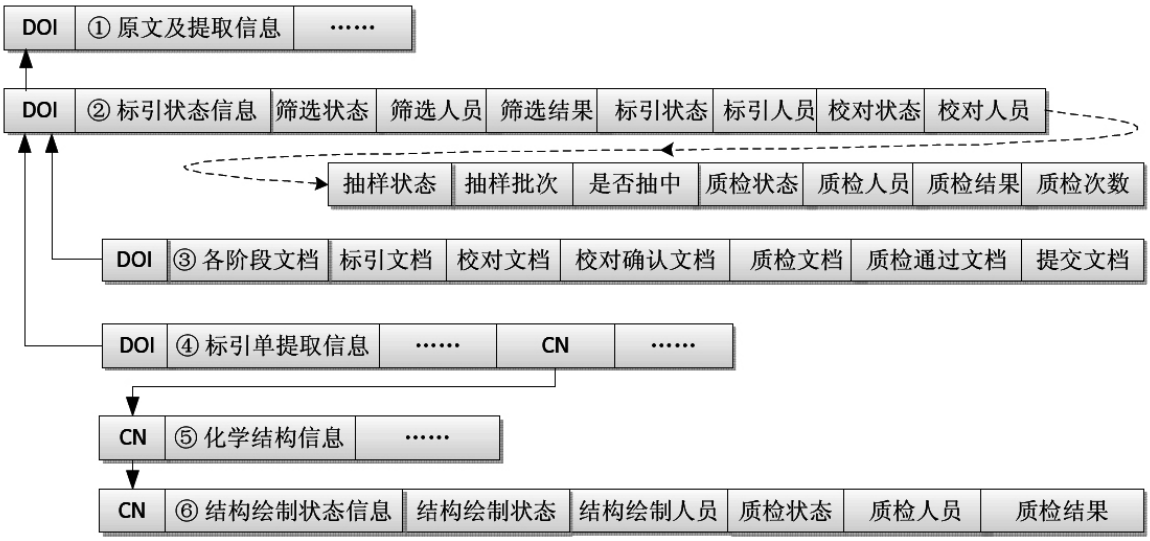


图1 数据存储逻辑结构

流程管理信息包括数据标引状态信息和结构绘制状态信息。数据标引状态信息包括筛选、标引、校对、质检各处理阶段流程管理信息；结构绘制状态信息包括结构绘制和结构质检两阶段状态信息。每个阶段包括处理状态、负责人、处理时间、处理结果四方面信息。处理状态用于描述文献的加工进度，各处理阶段状态类型如表1所示。此外，可通过处理状态、负责人信息、可编辑人员结合判定数据访问权限。

三、流程组织

（一）角色与功能

系统流程包括申请、筛选、标引、校对、结构、质检、重做七个处理阶段，每个阶段同时也代表一类角色的职责和相关功能。筛选员通常承担了

表 1 流程状态说明

处理流程	状态	说明	可编辑人员
筛选	Undo	还未筛选数据	筛选员
	Finish	筛选数据结束，可进行标引	筛选员
标引	Undo	未标引	标引员
	Doing	正在标引	标引员
	Finish	标引工作初步完成，可指派校对	标引员
	Preaf	已指派校对	校对员
	PreafConfirm	已对校对返回数据进行确认	标引员
	Submit	标引数据已提交，可抽样质检	质检员
校对	Undo	未校对	校对员
	Doing	正校对	校对员
	Finish	校对已结束，标引员可进行校返确认	标引员
质检	Undo	未质检	质检员
	Doing	正质检	质检员
	Finish	质检完成，若有错，标引员可进行错误确认	质检员
	Submit	质检提交，质检过程结束	标引员
结构绘制	Undo	未绘制	结构绘制人员
	Doing	已绘制	结构绘制人员
	Submit	已提交质检	结构质检人员
结构质检	Undo	未质检	结构质检人员
	Doing	正质检	结构质检人员
	Submit	质检结束	结构绘制人员

申请阶段的工作，按需申请任务。结构处理阶段由结构绘制人员和结构质检员共同完成，由于两角色工作联系紧密，操作同质性强，使用相同的辅助工具，因此将其组织在一起，复用软件功能。重做阶段主要处理局质检不合格数据，所有角色都会参与该工作，让其独立成为一个阶段。

将各加工阶段工作进一步细化，拆分为一个个功能点。功能点通常具有单一职责特性，如找出某状态的任务列表、完成一个处理步骤、实现一类数据的状态转换。一个功能点与一个窗口界面相对应。再将功能点按角色分类，形成该角色的功能组，如图 2 所示。

图 2 角色职责与功能

（二）流程编排

流程编排将各功能点按照一定顺序合理组织形成工作流，通过执行工作流上功能点功能，从而完成数据加工任务，流程如图 3 所示。非专利数据加工系统是一个并行处理系统，多阶段工作同时开展，多人同时处理同一阶段数据。流程编排需要解决两方面的问题，一是流程串行化，二是各角色之间的协作。

流程串行化的目的是防止多人同时修改同一数据，导致数据不一致，发生错误。实现的方法如下：首先定义规则，只允许对最新处理阶段的文档进行编辑，一篇标引文献在数据加工过程中只存在一个最新处理阶段，该信息可通过状态信息表或具体窗口界面推断获得，根据最新处理阶段信息可实现只有一种状态数据可编辑。再获取状态信息表中最新处理阶段的人员信息，只允许该人对数据进行编辑，由于一篇文献在某一处理阶段只有一人负责处理，因此就保证了任一时刻只有一人对一篇文献进行编辑，防止了数据访问冲突，实现了流程串行化。图中，虽然存在化学结构绘制与质检流程与主流存在并行的情况，但该流程与主流程间不存在数据访问冲突，所以也是串行的。

（三）界面组织

界面组织主要任务是合理组织数据加工流程，按角色组织工作任务、展示任务进度、统计加工数量。各角色职责界面集中了整个加工流程中和该角色相关的职责和功能，按照加工阶段和各阶段状态变化顺序组织，形成功能组。以标引员工作界面为例，如图 4 所示。其中标引任务、标完确

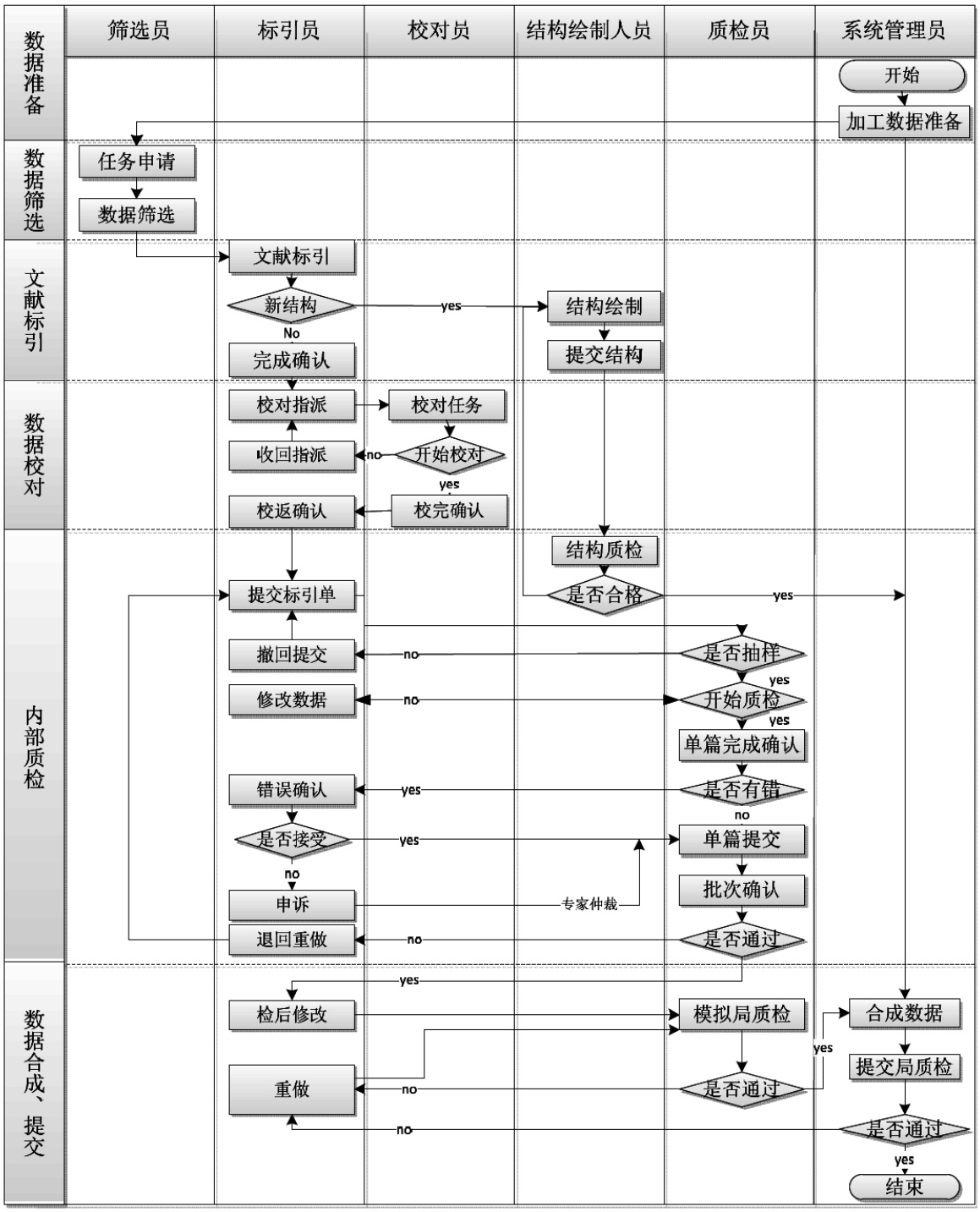


图 3 非专利文献数据流程和协作关系

认属标引阶段的工作；校对指派、校返确认在校对阶段与校对员之间的协作;标引提交、提交未检、质检错误、错误确认、重做提交、检后修改是质检阶段的工作，与质检员协作处理质检过程中发生的各种情况。

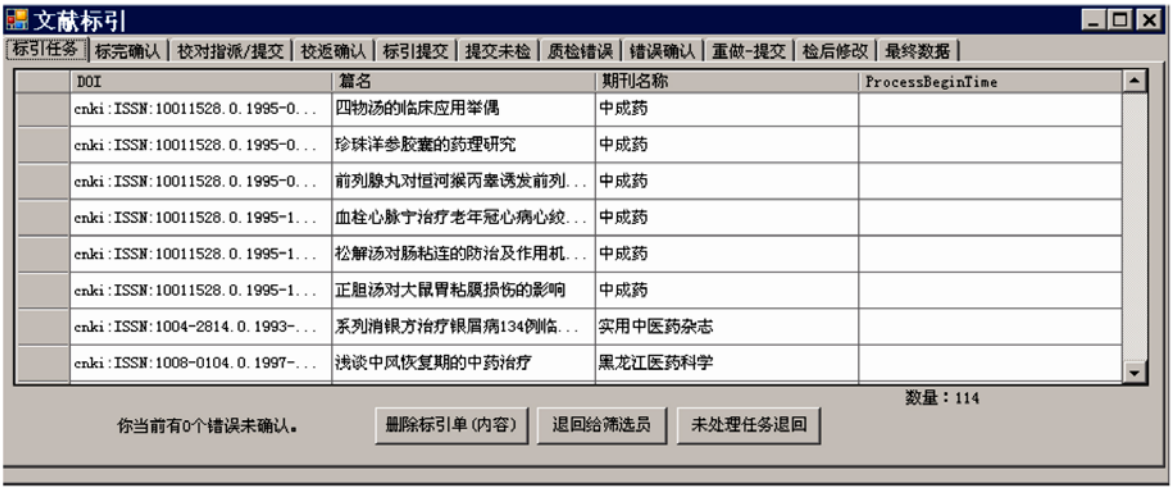


图 4 标引员界面

角色界面的每个功能页面通常包括数据列表、查看数据、修改数据、改变数据状态、任务提醒等功能。数据列表通过在用户登录 ID 和功能页面隐含处理阶段信息和阶段状态信息从标引状态信息表中查询获得；加工文档通过文献 DOI 和功能页面隐含的处理阶段信息和阶段状态信息从文档信息表中查询取得；能够查看和修改数据根据用户权限和文献处理阶段状态而定；数据状态改变通常包括两种方式，通常在保存数据时，会将未处理的数据自动变为正处理，通常正处理的数据变为完成状态由用户自行控制，以便反复修改数据；并在界面明显位置提示重要且迫切需要处理的任务，如错误确认、重做任务，提醒加工人员及时处理任务。

四、协同工作

角色之间协作主要内容包括处理任务转移、数据传递、数据访问控制

（一）任务流转

实现的基本机制是改变数据状态并对数据显示和数据访问权限进行控制。以数据提交为例，标引员在提交数据的那一刻，该数据的标引状态改变为“Submit”，由于标引状态改变为“Submit”，通过即时刷新数据，该数据将从未提交列表将中消失。同时，该数据会出现在质检员可抽样的数据列表中，因为该数据标引状态为“Submit”，但质检抽样状态为“Undo”，意味着数据提交但未进行质检抽样，处于可抽样阶段。通过这样实现了任务转移。

（二）数据传递

数据产生和查找机制是数据实时流转的基础，保证了通过DOI和数据处理阶段信息获取正确数据。在数据加工过程中，如图5所示，正常情况下一篇原始文献会发生5次实质性数据修改(①－⑤),1次格式调整(⑥),共产生6个Word文档。

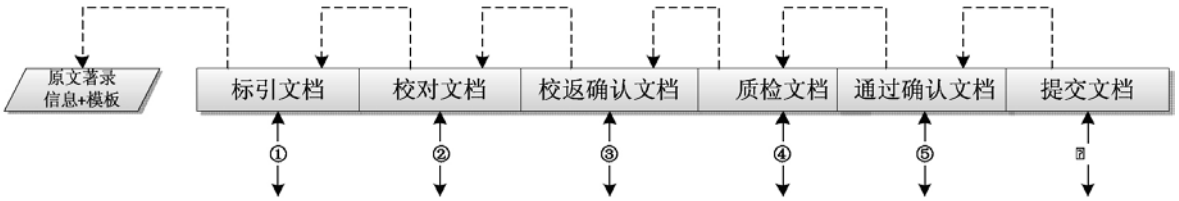


图5 标引文档生成和查找过程

数据查找的目的是找到当前处理阶段最近一次保存的文档，同时考虑数据处理的逻辑关系，如首次校对找不到引文档，首次校返确认找不到校对文档都是错误的。已知标引阶段信息，按如下方法查找数据：

1. 标引阶段：如果标引文档存在，直接取回；若不存在，说明是首次处理，将获取标引模板和原文著录信息，自动生成Word标引单。

2. 校对阶段：如果校对文档存在，直接取回；若不存在，说明是首次处理,将从标引文档获得数据拷贝，若标引文档为空，抛出异常。

3. 校返确认阶段：如果校返确认文档存在，直接取回；若不存在，说明是首次处理，将从校对文档获得数据拷贝，若校对文档为空，抛出异常。

4. 质检阶段：如果质检文档存在，直接取回；若不存在，说明是首次处理，将从校返确认文档获得数据拷贝，若校返确认文档为空，说明未进行较返确认，将从校对文档获得数据拷贝，若校对文档为空，说明未进行校对,将从标引文档获得数据拷贝，若标引文档为空，抛出异常。

5. 通过确认阶段：如果通过确认文档存在，直接取回；若不存在，按质检阶段的方法查找。

6. 通过确认阶段：如果提交文档存在，直接取回；若不存在，按通过确认阶段的方法查找，找到之后根据提交文档要求自动生成提交文档。

若未知标引阶段，则按通过确认阶段查找数据，必定能找到任何处理阶段的数据，好处在于简单灵活，只需ＤＯＩ则可查到任何文献最新处理结果信息，存在的问题是查找过程较长，性能上有一定损失。

数据保存，在找到文档基础上修改文档,分别存储于各阶段文档位置，若已经存在，则覆盖该文档。

（三）访问控制

系统数据一致性主要通过 以下

几方面进行控制：

DOI和各加工阶段负责人信息修改限制。在整个加工过程中，禁止任何方式对DOI进行修改；加工阶段负责人信息只允许通过系统修改，保证权限验证基础信息正确性。

保存检查。系统界面只提供了一个保存按钮，任何数据修改都必须通过该按钮进行保存操作，修改才能生效，系统在保存之前，都会对数据进行权限验证，防止非法数据存储。

在客户端，对可编辑文档数量作了限制。在同一时刻只允许存在一个可编辑文档，但可打开多个非编辑文档，以便查询。有效防止了可编辑文档互相覆盖，发生错误。

即时刷新数据状态，更新数据列表。功能切换和数据状态改变时立即刷新界面，保证数据列表里的文献都是属于该状态的。

五、结论和讨论

本文设计了非专利文献数据加工系统的数据存储逻辑结构，并在此基础上研究了加工流程的组织和串行化机制，任务流转和数据传递机制，实现了非专利数据加工流水线作业，任务实时流转，数据安全访问，多人协同工作，解决了非专利文献数据加工系中一个重要技术问题，为工作顺利开展奠定了技术基础。

但是，以下两方面内容还需研究，一是在筛选和加工规则有变化时会导致流程改变，因此，研究具备随需应变能力的流程管理技术非常重要。二是数据加工与局质检部门之间的协作关系并未纳入流程，需要进一步研究，通过部门间协作，实现非专利数据加工完整流程。

（专利检索咨询中心　杨晓春 审校）