基于约束规则的科技基础性数据质量审查模型研究与实现

张肖霞[1] 杜 平[2] 陈 杭[1] 鲁玉佳[1] 张金区[1]
（1.华南师范大学计算机学院，广东广州 510631；2.广东科学技术职业学院广州学院，广东广州 510653）

摘要：针对科技基础性工作项目汇交数据质量审查人工效率低、易出错的现状，设计了基于自定义约束规则的项目汇交数据质量审查模型。模型由构建器、规则库和判断器3个部分构成，其中构建器主要是用于约束规则的配置；规则库由一系列规则集构成，每个规则集都从数据的完整性、一致性和约束性3个方面进行约束规则的定义，完成定义的规则集构成一个审查模板；判断器主要是将审查模板应用于一个数据集进行审查并输出审查意见。实践证明，该模型能够满足科技基础性工作项目汇交数据质量审查的工作要求，提高了科技基础性工作项目汇交数据质量审查的质量和效率，同时也为其他类似数据质量审查工作提供了技术参考。

关键词：科技基础性工作；项目汇交；数据质量审查；约束规则；质量审查模型
中图分类号：TP306 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.05.008

Research and Implementation of Quality Inspection Model for Basic Data of Science and Technology Based on Custom Constraint Rules

ZHANG Xiaoxia[1], DU Ping[2], CHEN Hang[1], LU Yujia[1], ZHANG Jinqu[1]
(1. School of Computer Science, North China Normal University, Guangzhou 510631; 2. Guangzhou School, Guangdong Polytechnic of Science and Technology, Guangzhou 510551)

Abstract: Considering the situations of the low efficiency and fallibility in manually data quality inspection for the data from basic work of science and technology, a data quality inspection model on basic work of science and technology was designed based on custom constraint rules. The model consists of constructor, rule database and a judge determiner. The constructor is mainly used for the building of custom constraint rules. The rule database is composed of a series of rule collections. Each rule collection, namely as an inspection template, can be defined from integrality, consistency and restriction. The data will be checked based on an inspection template selected by the judge determiner with results exported. Study showed that the data quality review model can meet the work requirements in data review of basic work of science and technology. It improves the quality and efficiency of data quality review, and provides a technical reference for other similar data quality audits.

Keywords: basic work of science and technology, project remit, data quality review, constraint rule, quality audits model.

作者简介：张肖霞（1993—），女，华南师范大学计算机学院硕士研究生，研究方向：空间信息处理；**杜平**（1982—），女，广东科学技术职业学院广州学院教师、工程师，研究方向：嵌入式系统及软件工程；**陈杭**（1991—），男，华南师范大学计算机学院硕士研究生，研究方向：深度学习；**鲁玉佳**（1994—），女，华南师范大学计算机学院硕士研究生，研究方向：深度学习；**张金区**（1980—）：男，华南师范大学计算机学院副教授，研究方向：空间信息技术应用（通讯作者）。

基金项目：科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”（2013FY110900）；广东省科技计划项目“基于O2O模式的新一代科普作品研发”（2014A070711020）。
收稿时间：2017 年 7 月 14 日。

1 引言

据不完全统计，自 1999 年，我国启动科技基础性工作专项到“十一五”末，已经在气象、地球科学、生物学、农业、林业、医学、环境、材料等领域设置了 500 多个项目，投资总经费达 10 多亿元。通过这些项目，采集产生了一批重要的科学数据、文字资料、图集典籍、科学规范、标准物质、样本样品等。然而，由于缺乏国家层面的基础性工作数据资料的集成整编环境，绝大部分已结题的基础性工作数据资料仍然散落在各项目或课题承担单位中，并没有得到有效的集成、整编与挖掘，甚至有些数据资料濒临丢失，影响了基础性工作本质目标的实现。“科技基础性工作数据资料集成与规范化整编”项目的目标之一即为实现我国 1999—2010 年立项的基础性工作项目数据资料的分类集成与规范化整编，构建基础性工作数据资料集成服务平台，保障长期、持续地对我国基础性工作数据资料提供集成与共享服务。那么，如何保障项目汇交数据的质量，实现基础性工作数据资料的完整性、规范性、正确性和一致性，切实满足基础科学研究、重大公益性研究、战略高技术研究与产业关键性技术研发的基本需求，是当前最为关键的工作。

目前，对科技基础性工作专项项目数据汇交的审查工作主要采取人工逐项审查核对的方式。这种方法不仅费时费力，而且容易受到人为疏忽或经验水平有限而导致的审查错误。因此，基于科技基础性工作项目汇交数据的构成和特点分析，对不同的数据类别建立合适的审查模型，实现对汇交数据的计算机辅助审查，不仅提高效率，而且提高数据审查质量。从已有的研究来看，还没有专门针对科技基础性工作专项项目汇交数据质量审查的案例。但是，对于信息系统中数据质量的研究，历来受到建设者的高度重视。数据质量是进行数据分析和应用的基础，数据质量已经成为当前进行大数据价值挖掘的主要障碍[1]。在国内信息系统的建设中，通常将数据质量用正确性、准确性、不矛盾性、一致性、完整性和集成性等 6 个方面进行描述[2]。国际货币基金组织于 2001 年开发的《数据质量评估框架》列出了影响数据质量的 5 个方面，即诚信、方法的健全性、准确性和可靠性、适用性及可获得性，同时还定义了一套保证数据质量的制度前提[3]。欧洲统计系统建立的数据质量评估框架从统计机构环境、统计程序和统计产出 3 个方面对统计数据质量展开评估，开发了数据质量报告标准、质量报告手册和自我评估检查单等系列数据质量管理工具[4]。从上述可以看出，数据质量问题已经受到国内外的广泛重视。针对数据质量的不同方面，一系列数据质量评价的方法和系统相继开发实现，既有专门针对结构化数据进行质量检查的研究，也有专门针对空间数据进行质量检查的研究，还有专门针对特定行业数据质量检查的研究[4−7]。其中，基于规则引擎的数据质量检查，是常用的方法之一。王兴等[8]建立了基于规则引擎的多元大气信息数据质量检查方法，杨家芳[9]建立了基于规则引擎的基本农田划定内业数据质量检查方法，都取得了良好的效果。面对近年来大数据的兴起，研究确定了“Quality−in−Use”数据质量评价模型。该评价模型主要用于大数据分析时对输入数据的质量评价[10]。这些数据检查和分析评价的方法，大都是面向数据生产者服务。对于一些数据共享组织管理机构，通常是通过制定规范进行约束的。

科技基础性数据涉及学科广，类型复杂，从目前项目单位汇交数据看，普遍存在一些文档组织不规范、数据缺失、数据内容项不完整、文件打不开以及一些数值超限等问题。这些问题不仅增加了人工审核的难度，而且对科技基础性数据深层次的应用挖掘带来障碍。所以，建立面向科技基础性项目汇交数据的质量审查模型，开发相应的软件系统，对提高数据管理者的工作效率和促进科技基础性数据的应用挖掘具有重要意义。

2 科技基础性工作项目汇交流程分析

2.1 汇交数据构成及特点分析

科技基础性工作项目汇交数据主要来源于我国启动科技基础性工作专项以来立项的各类项目

所产生的数据。从学科来讲，包含气象、地球科学、生物、农业、林业、医学、环境、材料等多个领域；从数据存储格式上，有矢量数据、栅格数据、表格数据、文本数据、文档数据等；从表现形式上，有数据、图集、志书、典籍、标本资源、标准规范、论文专著或研究报告等。从对科技基础性工作项目汇交数据的构成分析可以得出项目汇交数据具有以下特点。

（1）多样性：主要指科技报告类型多样、数据类型多样、学科领域众多、科技数据提交加工环节多样等特点，使得提交上来的科技数据资源呈现多样化。

（2）异构性：科技基础性数据涉及专业广泛，领域众多，不同的专业领域对于科技基础性数据的记录形式各不相同。

（3）复杂性：不同专业领域的科技基础性数据形式不同，科学考察项目需要记录的数据有项目观测、监测、实验、调查和考察数据及相关的图件、报告等。图集、志书、典籍项目需要记录的数据有图集、志书、典籍及其支撑这些资源的数据等。标准规范项目需要记录的数据有标准规范文本及其支撑标准规范研制的基准、支撑、测试数据等。

（4）保密性：部分科学基础性数据涉及国家机密，具有保密性特点。

2.2　项目汇交数据审查的主要流程

为了有效监督和管理科技基础性工作专项项目的执行，促进项目汇交科学数据的共享与服务，科技部专门出台了《科技基础性工作专项项目科学数据汇交管理办法》，明确规定了项目承担单位负责项目科学数据的整理和汇交，包括：（1）组织编制项目数据汇交方案；（2）按照汇交方案组织整理项目数据，并按时完成汇交；（3）确保项目数据的完整性和质量。科学数据管理机构负责项目科学数据的接收、保存、管理、共享与服务。其对项目数据汇交数据审查的主要流程如图 1 所示。

数据汇交管理机构主要基于项目承担单位编制的数据汇交方案，对项目基本信息与元数据、数据实体、数据文档、论文专著及辅助软件等进行规范性、完整性和一致性的审查。

项目数据汇交方案：包含项目编号、项目名

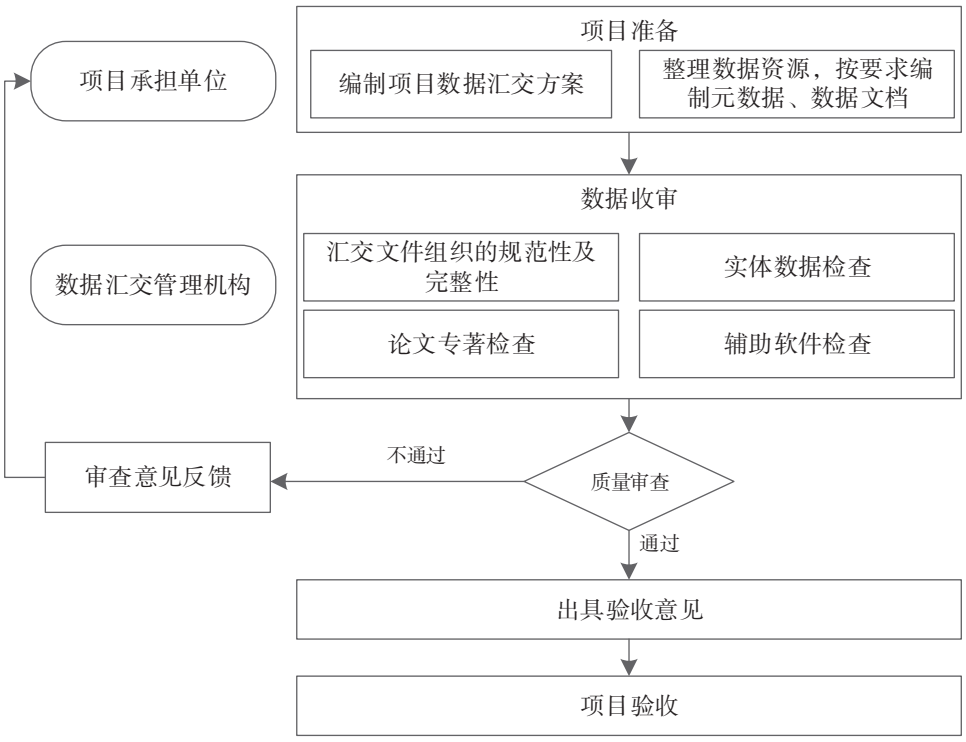

图 1　项目汇交数据审查的主要流程

称、项目负责人、项目承担单位等基本信息，还包括项目计划任务书规定的任务和考核指标及调整情况、汇交的资源内容、资源质量控制等相关说明。汇交方案是进行汇交数据审查的基础和依据，如果汇交的文档中没有汇交方案文件，则直接反馈错误信息。

项目基本信息与元数据：项目基本信息和元数据中主要包含项目编号、项目名称、所属类型、第一承担单位、项目依托部门、成果类型、项目起止时间、项目负责人和数据汇交联络人基本信息、成果介绍、资源描述摘要、关键词、资源质量描述等。主要审查这些描述项是否有漏填及不一致现象。

数据实体：汇交的数据实体有 4 种格式类型，分别是矢量数据、栅格数据、表格数据、文本及其他类型数据，每种数据描述表的字段不同，针对不同的字段进行自定义约束审查。依据《自然科技资源共性描述规范》[11]，结合科技基础性工作专项项目的特点，形成对植物种质资源、动物种质资源、微生物菌种资源、人类遗传资源、生物标本资源、岩矿化石资源、实验材料资源、标准物质八大类标本资源描述信息的规定。每种资源的描述规范表中含有 5 个字段，分别是序号、描述符、数据类型、数据限制、备注说明。

数据文档：主要包含数据集/图集内容特征、学科及行业范围、精度、存储管理、质量控制、共享及使用方法、知识产权等说明信息。标准规范编制说明主要有工作简况、主要起草过程、重大意见分歧的处理依据及结果等。这部分主要是进行内容的描述，主要依靠人工进行审查。

论文专著及辅助软件工具：论文专著主要指与项目数据直接相关、在数据引用时需要使用的专著或论文。辅助软件工具则是对汇交的数据进行查看和处理的专用工具。此部分主要从文件是否存在、是否有关联性、是否能正确打开等方面进行审查。

3 基于约束规则的数据审查模型设计

由于科技基础性工作项目汇交数据包含气象、地球科学、生物学、农业、林业、医学、环境、材料等学科领域，计算机辅助审查只能从数据的共性层面建立规则来构建审查模型，对于具体数据内容的真实性、可靠性还必须依靠人工进行判断。

3.1 科技基础性工作项目数据汇交审查内容

审查的方式有系统自动审查和人工审查两种形式。主要包括以下几个方面的审查内容。

（1）完整性审查。汇交数据的完整性审查主要从 3 个方面进行审查：一是从文件组织上看汇交的数据文件是否遵循项目科学数据汇交的统一规范，“汇交规范”规定了每个专项项目汇交数据时的文件构成和组织方法，如有遗漏，则完整性审查不通过。二是基于各专项项目提交的数据汇交方案来审查，在汇交方案中列明本项目的数据组成情况。模型将通过对汇交方案的解析实现对数据完整性的审查。三是从数据文件构成的完整性上进行审查，例如一个矢量数据的 shape 文件，至少由 .shp、.dbf、.shx 3 个文件组成，如果缺少一个那么完整性审查将不能通过。

（2）一致性审查。主要指对汇交数据中文档的一致性、内容的一致性等内容审查。

（3）约束性审查。主要是对数据内容的约束性审查，约束性审查主要是对二维表格数据、二维表中每一列属性进行判断，审查每一行的值是否在约束范围内。

3.2 数据审查模型框架

为了灵活实现对不同学科领域的数据审查，本文探讨基于自定义约束规则的数据审查模型，模型框架如图 2 所示

数据审查模型主要由构建器、规则库和判断器构成。构建器主要是用于创建约束规则的工具，约束规则由判断条件和值域构成。规则库存储了用户进行数据审查时创建的各类规则集。判断器则将这些规则集应用于一个待审查的项目汇交数据集，并对是否满足规则的情况进行输出。

3.3 自定义审查规则集的构成

根据科技基础性工作项目汇交数据的内容和特点，从完整性、一致性和约束性 3 个方面进

行约束规则的定义，审查规则包括数据文档存在性审查、文件组织和命名规范审查、数据质量审查、数据文档审查、论文和辅助软件审查。其中，数据文档存在性审查是指文档是否存放在指定的路径位置上。文件组织审查指文件的存放路径是否符合规范的统一约定，命名规范审查指文件的命名是否符合要求。数据质量审查和数据文档审查模块包括数据项内容审查、行数据审查、列数据审查、多表审查等。一个数据审查规则集的构成如图 3 所示。

数据项审查是指对某一数据表中的某一个数据项进行审查，包括非空审查、数据类型审查、正则表达式审查、数据范围审查等。在数据项审查中，非空审查通过设置数据项能否为空的约束条件来审查数据项内容是否满足约束规则。数据类型审查主要审查所采用的数据类型必须是指定的某一数据类型或满足预先设定的几种类型中的某一类型。正则表达式审查是由于采用单个字符串描述或者匹配一系列某个句法规则的字符串，也就是用一个“字符串”来描述一个特征，因此主要审查某一个“字符串”是否符合这个特征。如审查电话号码、邮箱、日期是否满足规格。数据范围审查包括常规的数值范围审查和数据项内容是否在自定义的范围之内，是一种约束性的审查，如审查某一物质的PH值必须在 3 ～ 7，审查植物种植的气候带必须为热带、亚热带、温带、寒温带、寒带、其他这 6 项中的一项等。

行数据审查是对数据表中行与行数据项之间关系的审查，包括行数据项之间的对应关系、限制约束关系。如项目编号字段与项目名称字段是一一对应关系，一个项目编号有且仅有一个项目名称。

列数据审查指的是对同一字段的数据项与数据项之间关系的审查，包括对比审查、累计值审查、四则运算审查等。如表格数据详细描述表中“数据记录数”字段需要运用四则运算统计表格记录的整列数据总量。

多表审查是对两个及其以上数据表中数据项关系的审查，也叫数据项动态联合审查。如表格数据详细描述表中描述字段必须包含被描述数据表的所有字段。

4　数据质量审查系统开发与实现

4.1　系统工作流程

依据上述数据质量审查模型的设计方案，梳理数据质量审查系统的工作流程，如图 4 所示。

系统在应用上，首先读取项目数据包，然后从规则库中选择审查模板，依据审查模板定义的规则进行逐项检查。在检查过程中，首先检查是否存在PDF格式的数据汇交方案。其次审查Dataset的内容，检查Dataset文件夹存放的数据实体和数据说明文档，以数据资源唯一的标识号作为下一级文件夹的名称，每个文件夹中又存放着Data、Document、Thumbnail 3 个文件夹，它们分别用来存放数据实体、数据说明文档和数据缩略图。此部分审查主要是针对文件的组织和命名是否符合规范。接着用自定义约束规则审查模型对数据质量进行审查，检查数据的完整性、一

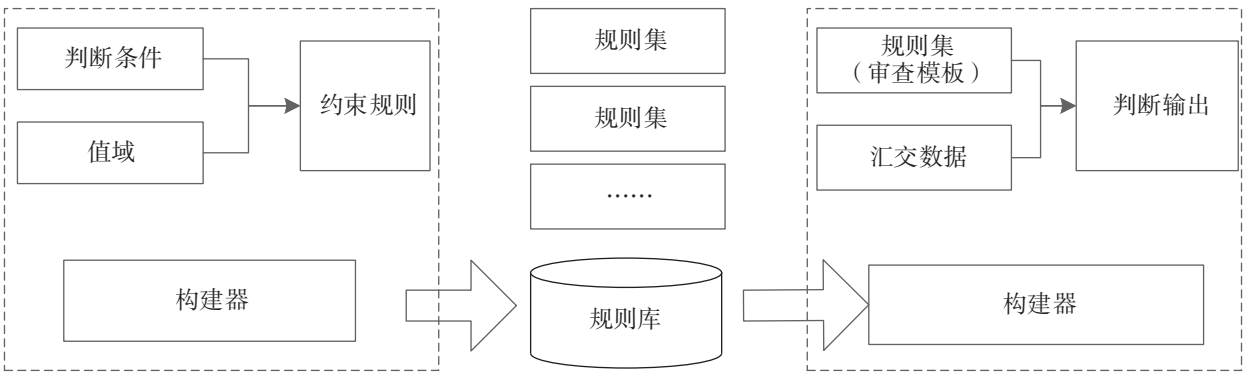

图 2　数据审查模型的框架构成

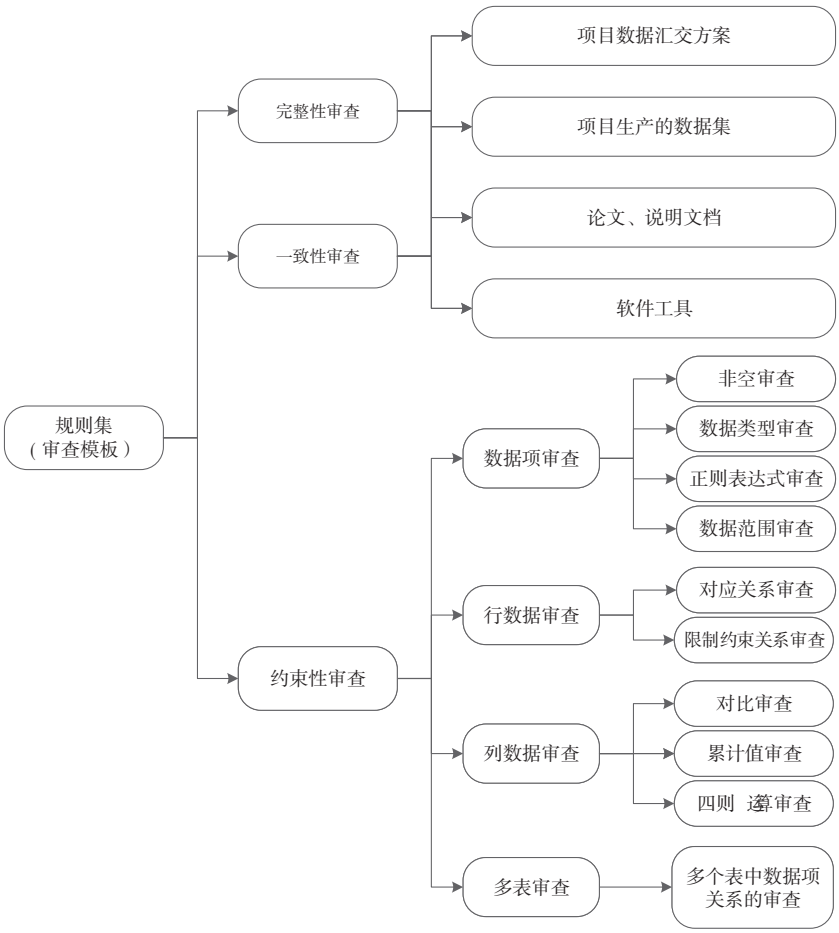

图 3　数据审查规则集构成

致性等。再次对数据说明文档进行审查，重点对科学数据/图集说明文档、标准规范编制说明、八大类标本资源规范描述表进行审查。最后对Paper部分和Software部分进行审查。

4.2　审查模板与自定义规则配置

规则集通过采用基于XML的模板文件进行存储，一个规则集就是一个审查模板。XML文件是一种可扩展标记语言，其具有可扩展性、交互性好、跨平台的特点，还具有结构性强、易于处理、灵活性好等优点，易于进行自定义审查规则的存储[12]。用户可以对不同的数据集创建不同的审查模板。当审查要求有变动时，只需添加或修改审查规则文件，在审查时进行相应的配置即可，便于灵活的数据审核。

自定义数据审查规则的配置方法是根据待审查数据集的不同而进行设计的，需要分别对项目数据汇交方案、项目基本信息与元数据、数据资源实体、数据说明文档、辅助软件工具、专著论文等数据进行创建。为了便于用户操作，系统开发了自定义审查规则的配置界面（图 5）供用户使用。

4.3　审查日志与审查结果反馈

为了方便管理和记录每一个项目汇交数据的质量审查情况，该系统还增加了用户管理和审查日志的功能，每次审查数据的结果都会被记录在审查日志中。而对同一数据集的审查则根据时间轴来记录每次的审查情况，方便用户追溯数据资源的审查和修改记录情况。

每一次的审查结果都记录着对汇交数据资源审查评价的信息，包括审查的数据是否正确，数据错误的原因等。数据审查结束后，系统会自动生成一个审查结果的报告文档。审查人员可以在此文档的基础上，继续添加人工审核的意见。最

后，将审查结果文档反馈给汇交单位，供汇交单位进行数据集修改完善。

5　结论

本文首先分析了科技基础性项目汇交数据的构成及特点，进而梳理了项目汇交数据审查的主要流程。按照科技基础性项目汇交规范，设计了基于自定义约束规则的数据质量审查模型，模型从数据的完整性、一致性和约束性3个方面进行约束规则的定义，能够对项目数据汇交方案、项目

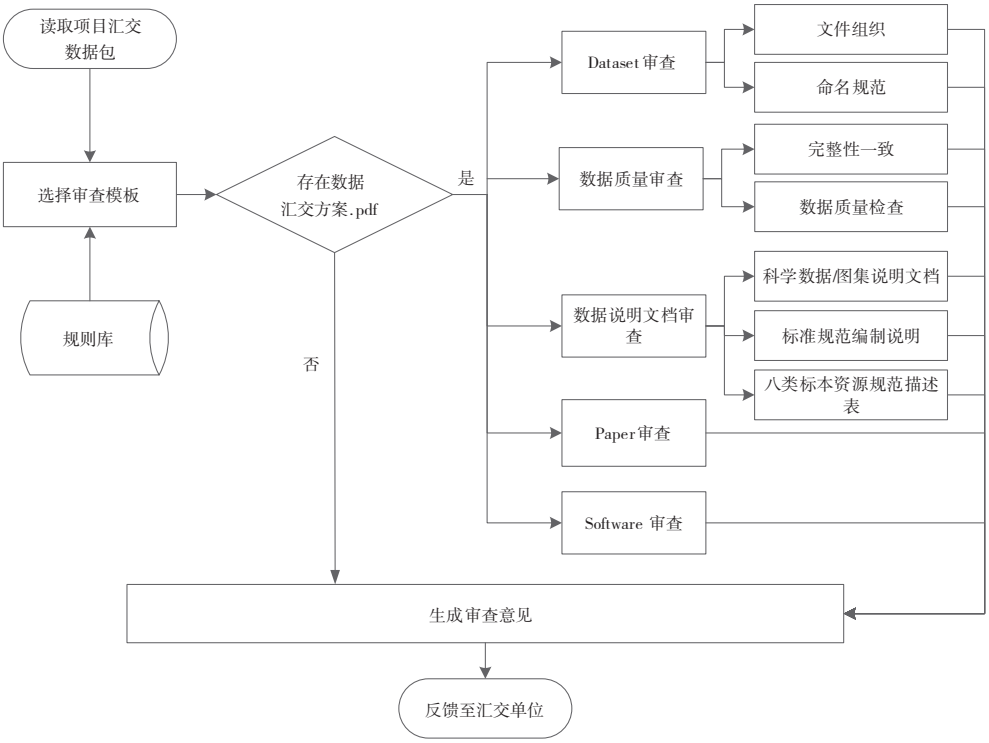

图4　数据质量审查系统应用流程

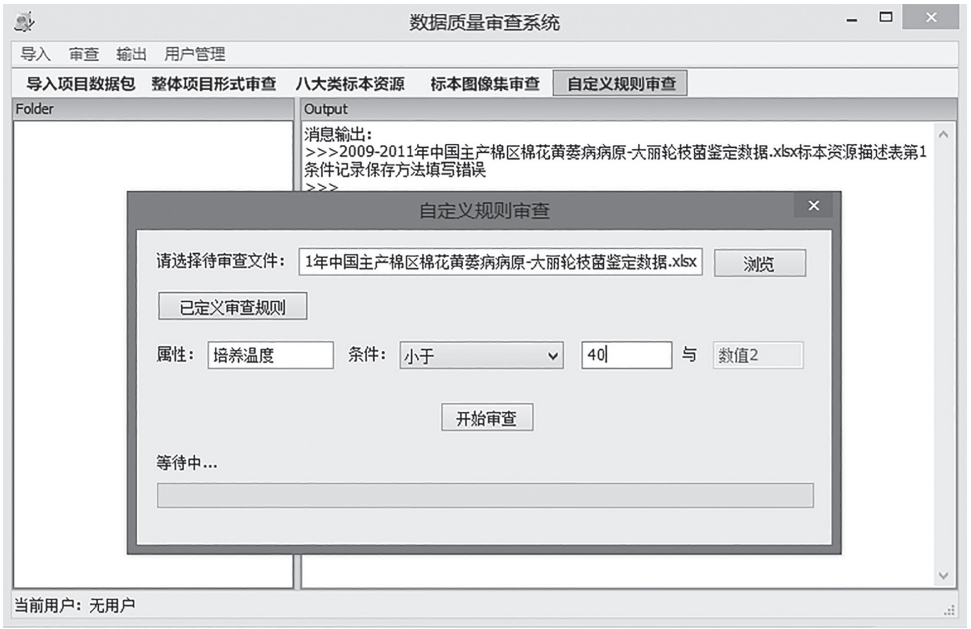

图5　自定义审查规则的配置界面

基本信息、数据实体、数据文档、论文专著和辅助软件工具等进行全面的审查。最后，基于本模型开发了科技基础性项目汇交数据质量审查系统。除了数据审查功能之外，还开发了用于约束规则配置的工具界面，添加了审查日志和审核意见导出等功能，方便对同一数据集的持续跟踪审查。

通过基于约束规则的科技基础性数据质量审查模型的研究与实现，能够提高科技基础性工作项目汇交数据质量审查的质量和效率，节约数据审核人员的时间，使数据审核人更集中精力去审核一些更深层次的数据质量问题。科技基础性项目数据汇交是一项长期持续性的工作。目前，数据质量审查模型还更多地侧重于形式方面的审查，随着数据汇交工作的开展而不断深入，数据质量审查模型将深入数据内容，建立经验模型，逐步实现数据质量的自动审查和意见反馈。

参考文献

[1] SADIQ Shazia, INDULSKA Marta. Open data: quality over quantity[J]. International Journal of Information Management, 2017, 37(3): 150−154.

[2] 陈远, 罗琳, 沈祥兴. 信息系挽中的么据质童问题研究[J]. 中国图书馆学报, 2004(1): 48−50.

[3] 蒋萍, 田成诗. 全方位、立体性数据质量概念的建立与实施[J]. 统计研究, 2010, 27(12): 8−15.

[4] 许涤龙, 龙海跃. 欧盟数据质量评估框架及其对我国的启示[J]. 统计与决策, 2013(8): 4−7.

[5] TAGGARTA Jane, LIAWA Siaw−Teng, YU Hairong. Structured data quality reports to improve EHR data quality[J]. International Journal of Medical Informatics, 2015, 84(12): 1094−1098.

[6] PRESSER Karl, HINTERBERGER Hans, WEBER David, et al. A scope classification of data quality requirements for food composition data[J]. Food Chemistry, 2016, 193: 166−172.

[7] 徐启恒, 张新长, 张兴飞. GIS数据检查与质量控制系统的设计与实现[J]. 测绘通报, 2012(5): 38−40.

[8] 王兴, 朱定真, 苗春生. 基于规则引擎的多元大气信息数据质量检查方法[J]. 南京信息工程大学学报(自然科学版), 2011, 3(3): 238−243.

[9] 杨家芳. 基于规则引擎的基本农田划定内业数据质量检查方法研究[D], 杭州: 浙江大学, 2014.

[10] MERINO Jorge, CABALLERO Ismael, RIVAS Bibiano, et al. A data quality in use model for big data[J]. Future Generation Computer Systems, 2016, 63: 123−130.

[11] 曹一化, 刘旭, 许增泰, 等. 自然科技资源共性描述规范[M]. 北京: 中国科学技术出版社, 2006: 1−86

[12] HELLMANN D. The python standard library by example[M]. Indianapolis, Indiana: Addison Wesley, 2011: 1−10.

（上接第59页）

[3] 胡光晓.提升我国地层研究知名度展现我国地层工作最新成果:《中国岩石地层名称辞典》[J].科技成果管理与研究, 2015(8): 79−80.DOI: 10.3772/j.issn. 1673−6516.2015.08.029.

[4] 王训练, 徐均涛.古生物学研究的新成果: 中国古生物志与中国各门类化石编研[J].中国基础科学, 2002(5): 18−23.DOI: 10.3969/j.issn.1009−2412.2002.05.004.

[5] 吴小红.京族医药调查报告[J].中国民族医药杂志, 2016, 22(3): 57−59.DOI: 10.16041/j.cnki.cn15−1175. 2016.03.037.

[6] 徐福荣, 戴陆园, 韩龙植.21世纪初云南稻作地方品种图志[M].北京: 科学出版社, 2016.

[7] 张芳, 王思.中国农业古籍目录[M].北京: 北京图书馆出版社, 2003.

[8] 徐冠华.加强科技资源研究促进科技资源共享[J]. 中国科技资源导刊, 2008, 40(3): 3−5.DOI: 10.3772/j.issn.1674−1544.2008.03.001.

[9] 叶玉江.加强科技平台工作推进科技资源管理[J].中国科技资源导刊, 2015, 47(2): 1−6.DOI: 10.3772/j.issn. 1674−1544.2015.02.001.

[10] 国家科技基础条件平台中心.国家科技基础条件平台发展报告: 2011−2012[M].北京: 科学技术文献出版社, 2013.

[11] 王卷乐, 杨雅萍, 诸云强, 等. “973”计划资源环境领域数据汇交进展与数据分析[J].地球科学进展, 2009, 24(8): 947−953.DOI: 10.3321/j.issn: 1001−8166. 2009.08.013.

[12] 王建涛, 朱龙文.基于XML元数据描述的空间数据共享管理平台的实现与应用[J].测绘工程, 2007, 16(1): 12−15.DOI: 10.19349/j.cnki.issn1006−7949. 2007. 01. 003.