

医学科学数据管理与共享平台原型系统建设^{*}

孙小康 吴思竹 张 泽 何晓琳 修晓蕾 钱 庆

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 构建面向公众的医学科学数据管理与共享平台原型系统, 阐述系统设计情况, 包括总体框架、标准规范体系、功能层和应用层设计, 以及系统实现过程, 包括服务器配量、开发流程及效果, 实现对多源异构医学数据的有效组织、管理和利用。

〔关键词〕 医学科学数据; 数据管理; 数据共享; 系统建设

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.08.004

Building of Medical Science Data Management and Sharing Platform Prototype System SUN Xiao-kang, WU Si-zhu, ZHANG Ze, HE Xiao-lin, XIU Xiao-lei, QIAN Qing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China

〔Abstract〕 To build the medical science data management and sharing platform prototype system that faces the public, the paper elaborates on design of the system, including the general framework, standard specification system and design of the function layer and application layer, as well as the system implementation process which includes server configuration and development procedure and effect, to realize the effective organization, management and utilization of multi-source heterogeneous medical data.

〔Keywords〕 Medical science data; Data management; Data sharing; System building

1 引言

科学数据是推动科学发展的必要组成部分, 是

数字化科研时代的重要战略资源^[1]。继实验科学、理论科学、计算科学等范式后, 科学研究已进入一个崭新的领域——数据密集型科学的第 4 范式^[2]。第 4 范式的科学研究为科学发展创造有利条件的同时也带来了诸多挑战。就医学领域而言随着计算机科学和信息技术的快速发展, 信息化的普遍应用, 在医学研究、医学实验、医疗服务、健康保健和卫生管理等过程中产生的医学科学数据呈爆发式增长, 海量的医学科学数据为临床诊疗、药物研发、卫生监测、公众健康、政策制定提供可靠依据, 但也面临着数据管理、组织和利用等方面的困扰^[3]。虽然国内外已经建设一批医学科学数据管理与共享

〔修回日期〕 2018-01-16

〔作者简介〕 孙小康, 硕士研究生; 通讯作者: 钱庆, 研究员。

〔基金项目〕 国家社会科学基金青年项目“基于 R2RML 的 RDB 到 RDF 的转换模式研究与实现”(项目编号: 13CTQ009); 北京协和医学院协和青年基金“面向精准医学的数据汇交与共享政策研究”(项目编号: 2017330018)。

平台来解决上述困扰,但这些平台绝大多数是面向科研人员或特定的科研机构,无法服务于广大的网络数据用户。而且对多源异构数据的组织,偏向于数据集层面,缺乏对具体内容和结构的详细描述和规范。服务形式也侧重于原始数据层面,鲜见对于语义方面的探索,难以发挥数据的最大价值^[4]。

基于此,本研究将在借鉴已有科学数据管理平台的基础上,综合运用 Web 技术、结构化数据库、语义以及可视化技术等,构建面向公众的医学科学数据管理与共享平台原型系统 (Scientific Data Management and Sharing Platform for Medicine, SDMSPM)。

该系统适用人群为所有需要进行生物医学科学数据提交、管理、组织或共享的用户。通过优化数据上传方式、多层面数据规范处理、分类整合存储以及安全合理的数据共享,满足用户多途径数据汇交、多来源异构数据规范以及数据分类管理与共享等方面的需求,方便公众数据管理与共享的同时,有效促进数据价值的最大化。

2 系统设计

2.1 总体框架 (图 1)

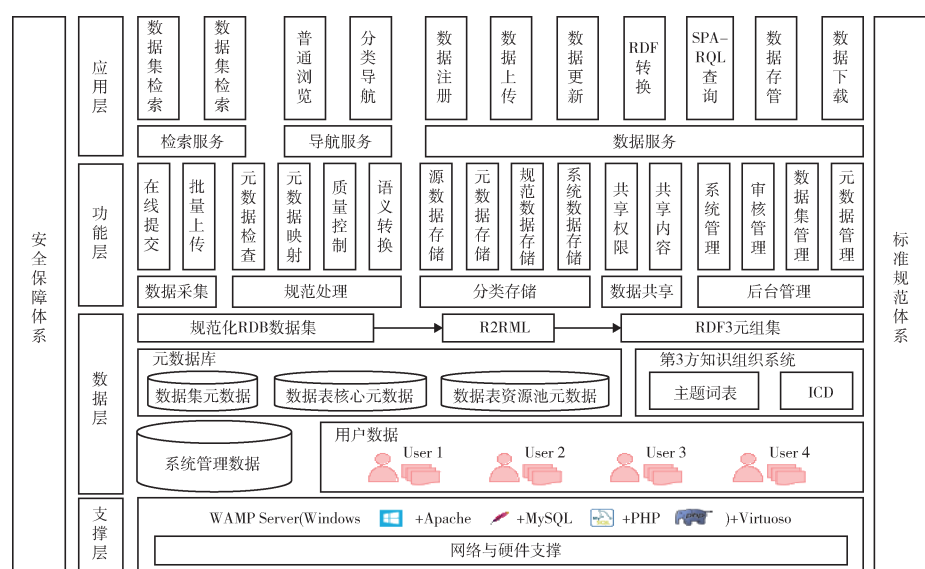


图 1 医院科学数据管理与共享平台总体框架

2.1.1 概述 平台总体框架共包括 4 层。平台设立完善的标准体系规范 (如数据提交规范、元数据规范、数据分类标准等) 以及数据备份等安全保障措施,为每一层的顺利运行提供保障。

2.1.2 支撑层 包括各种网络及硬件设备,用于存储关系型数据的 MySQL 数据库、存储 RDF 3 元组的 Virtuoso 数据库;用于搭建服务器运行环境的 WampServer (Windows + Apache + MySQL + PHP) 框架。在此基础上为简化系统开发,采用了一个快速、兼容且简单的轻量级国产 PHP 开发框架——ThinkPHP Framework。

2.1.3 数据层 存储用于维护用户个人信息的系统管理数据、用户提交的来源数据、数据集和数据

表中的元数据、第 3 方知识组织体系,通过元数据进行规范后再整合形成的整合数据,以及使用 R2RML 映射语言转换得到的 RDF 3 元组数据等。

2.1.4 功能层 是以平台为视角展示其所能提供的功能和服务。遵循科学数据生命周期进行数据管理功能设计,包括数据采集、规范处理、分类存储、数据共享以及其他一些后台管理功能 (系统管理、审核管理、数据集管理、元数据管理)。

2.1.5 应用层 是平台所能实现功能对用户展示的窗口,是用户与平台进行人机交互的桥梁。综合运用 HTML5、CSS、JavaScript 以及 Bootstrap 等技术为用户提供一个简单易懂的图形化用户界面,从而极大地简化用户进行数据检索、浏览导航以及数据

服务的过程。

2.2 标准规范体系

2.2.1 数据集分类与编码标准 对数据集进行分类与编码使得对同一领域的数据集进行集中管理, 不仅便于用户快速查找, 也为数据的整合和挖掘创造条件。通过对国家人口健康科学数据共享平台数据资源分类与编码、中图法 R 类(医药、卫生)、CNKI 医药卫生科技类分类体系以及数据堂的数据分类进行调研分析, 发现国家人口与健康科学数据共享平台的《数据资源分类与编码方案》^[5] 覆盖面广、类目层级设置适中, 符合本平台的数据组织要求, 因而在本平台中采用该分类体系对数据进行组织。该分类方案共包括两层: 7 大类和 34 亚类, 且每一亚类都赋予唯一的编码。

2.2.2 元数据标准规范 元数据是关于数据的数据, 是数字信息组织和处理的基本工具, 在网络信息资源组织与利用中发挥着日益重要的作用^[6]。元数据可以对平台内的大量数据进行规范并且可以作为数据交换的一种标准协议。在本平台中元数据的标准规范设计共分为 3 个部分: (1) 数据集元数据。是对数据集信息的基本描述, 可以帮助用户快速了解数据集的内容和状态。参考《国家科技基础条件平台资源元数据核心元数据》(征集稿)^[7] 的基础上, 结合本平台数据描述要求, 共制定 12 项元数据用于数据集的描述, 包括数据集名称、数据集描述、关键词、版权信息、提交人、1 级分类、2 级分类、访问权限、审核状态、创建时间、修改时间、审核意见。(2) 数据表核心元数据。数据表元数据是对用户上传的数据表中各字段的规范, 通过将不同来源的数据表中的字段进行统一, 实现多源异构数据的整合和关联。其中数据表核心元数据是根据平台数据需求所预设的元数据, 用户上传数据时选择与其数据对应的元数据项, 与之建立关联, 便于后期实现不同数据项之间的映射。平台会对数据表核心元数据进行动态扩展, 不断满足用户需求。(3) 数据表资源池元数据。如果数据表核心元数据集不能满足用户需求, 用户可以自定义元数据项。定义时需要提供元数据的唯一标识符(编码)、

中文名称、英文名称、数据类型、数据长度、描述、元数据类型等信息。用户定义的元数据将会进入资源池元数据集中进行保存, 管理员定期对其进行检查, 将同义异名的元数据进行规范, 将规范后的元数据根据需求添加到数据表核心元数据集中。

2.3 功能层

2.3.1 数据采集 平台获取用户数据的过程, 包括在线提交和批量上传两种方式。(1) 在线提交。当数据量不大或需要对已上传的数据进行增补时可选择该方式。用户根据之前选择以及定义的元数据, 逐项填写, 完成提交。(2) 批量上传。当数据量较大时可以通过文件进行批量上传。用户需要先下载平台提供的上传模板, 按照其方式整理数据。目前平台支持的文件格式有 EXCEL、CSV、JSON 和 TXT。对于用户上传的文件系统自动进行解析, 然后添加到相应数据表中。

2.3.2 规范处理 对用户上传的数据进行标准化的过程, 也是后续进行数据整合与数据关联的基础。(1) 元数据检查。用户完成元数据定义以及数据提交后管理员会进行审核, 主要检查元数据定义是否符合要求、是否有现成的元数据以及用户提交的数据项是否与填写的元数据项对应。(2) 元数据映射。通过元数据检查后平台便可以建立各个表项的元数据映射关系。映射到同一个元数据下的相关字段可以建立相关关系, 从而将各个孤立的数据联结在一起形成一个交互的关联网络, 为数据挖掘和知识获取提供可靠来源。(3) 数据质量控制。以上两个方面只是针对元数据层面的规范, 而数据质量控制则更多地涉及原始数据层面, 包括数据格式能否正确解析、字段类型和长度是否满足要求以及数据分类是否准确等。(4) 语义转换。存储在关系型数据库(Relational Database, RDB) 中的数据虽然具有良好的可扩展性、高效存储性、优化的查询执行以及可靠性等众多优点^[8], 但其缺乏语义, 不利于数据的互操作以及全球共享。而用资源描述框架(Resource Description Framework, RDF) 表示的数据能够被计算机理解、处理和推理, 有利于全球数据的互联互通。语义网之父 Tim Berners - Lee 制定

的开放数据 5 星评级模型中将以 RDF 格式发布的开放数据定义为 4 星^[9]。SDMPSM 倡导数据有意义的共享,因此除支持原始数据浏览和下载外还支持 RDB 数据到 RDF 格式的转换。用户上传数据时如果选择同意进行 RDF 转换,后台将会完成数据的转换工作。RDB2RDF 映射语言选择 W3C 推荐的 R2RML,它既可以实现 RDB 到 RDF 的缺省映射,也可以根据数据库模式和目标词汇定制 R2RML 映射。本平台目前采用的是缺省映射模式,即结果 RDF 图结构和词汇直接反映数据库的结构和模式元素的名称,且结构或目标词汇都未被改变^[10]。

2.3.3 分类存储 平台中数据的存储形式可以分为 4 类:源数据存储、元数据存储、规范数据存储以及系统数据存储。其中源数据指用户上传的原始数据;元数据包括数据集元数据、数据表核心元数据和数据表资源池元数据;规范数据既包括经过元数据规范后,将同一类目的数据进行整合形成的整合数据集,还包括利用 R2RML 规则进行转换后形成的 RDF 3 元组数据;而系统管理数据指用于维护系统正常运行的其他各类数据,如个人注册信息、访问日志、各类数据之间的关联信息等。这 4 类数据在数据库中分类存储,保持物理上的隔离,但逻辑上高度集中。这既有利于数据的管理,也能提高系统相应速度。

2.3.4 数据共享 数据汇交与管理的最终目的是科学合理地共享。本平台中数据共享具有严格的权限设置,用户在上传数据可以为其定义相应权限:公开、半公开、保密。公开的数据所有用户均可下载;对于半公开数据用户只能查看数据的描述信息,不能浏览全部或者下载;保密数据只有数据提交者能进行查看和下载,其他用户在平台中搜索不到该数据的相关记录。而共享的内容既包括原始数据,也包括转换后的 RDF 3 元组。为了方便用户使用,对于原始数据平台提供多种下载格式,如 EXCEL、CSV、JSON 等。

2.3.5 后台管理 是管理员对于平台各种功能和数据的管理和监控,包括系统管理、审核管理、数据集管理和元数据管理 4 部分。其中系统管理包括用户、角色权限和日志管理;审核管理包括数据

集、数据内容审核;数据集管理包括源数据集、分类及 RDF 3 元组管理;元数据管理包括数据集元数据、数据表核心元数据和资源池元数据管理。

2.4 应用层

2.4.1 检索服务 平台提供检索应用,满足用户对特定数据的需求,提高数据查找准确度,缩短数据获取时间。检索的数据内容包括数据集和数据表两个层面。检索机制上除使用字面匹配外,为提高查全率平台用各种词表归并同义词及下位词,从而使检索结果更加全面。

2.4.2 导航服务 平台提供普通浏览和分类导航两种数据查看方式。普通浏览主要适用于那些没有特定主题需求、想快速了解平台数据收录情况的用户;而分类导航则是用户快速访问领域数据的向导。将上传的数据按照预先设定的分类体系进行归类,用户可以根据个人专业领域完成数据的快速定位。为准确地将数据进行分类,为平台前台以及后台提供同一种分类体系。前台分类体系仅适用分类体系中 1 级类目,便于用户进行浏览和查找。而后台粒度更细,以便对平台数据进行规范管理及进行后续数据的关联整合。

2.4.3 数据服务 (1) 数据注册。用户在提交数据时对数据集和数据表基本特征以及元数据的说明。注册后通过管理员审核才能进一步上传数据。(2) 数据上传。用户上传数据的方式分为两种,即在线提交和批量上传。在线提交只需按照平台的提示依次填写各个字段项即可,适用于数据量较少的情况;批量上传时需要按平台提供的模板对数据进行预处理,文件的格式也要满足平台规定,以免发生解析错误。(3) 数据更新。用户对数据进行增添、删除、修改等操作的统称。它是用户自主参与的数据维护与管理过程,以保证存储在个人数据库中的数据质量和完整性。在本平台中数据更新的对象不但有用户上传的原始数据,还包括注册时填写的数据描述信息以及元数据相关内容。(4) RDF 转换。为增加数据的语义,平台也提供 RDB2RDF 转换服务。整个转换工作在后台进行,利用 R2RML 规则的缺省映射模式。其他用户在浏

览相关数据时也可以查看其 RDF 3 元组形式, 支持下载。(5) SPARQL 查询。对于生成的 RDF 3 元组系统也提供了相应的查询功能, 即通过编写 SPARQL 语句完成相关查询任务。SPARQL 查询语言相当于关系型数据库中的 SQL, 是专门针对 RDF 3 元组的语义查询技术^[11]。查询的结果同样可以导出。(6) 数据存管平台。为用户提供一个安全可靠的存储环境, 通过定期检查及备份, 保障用户上传数据的安全性和完整性。不同于普通的云端保存, 本平台对数据的管理不局限于数据集层面, 还能摄入到各字段项, 为用户进行各种处理提供便捷。(7) 数据下载。为促进数据共享, 共同推动科技进步, 在提交者许可的情况下, 本平台鼓励数据广泛的传播。为方便用户使用, 对于公开数据, 提供多种文件格式可下载, 如 EXCEL、CSV、JSON、RDF。

3 系统实现

3.1 服务器配置

操作平台: Windows Server 2008 R2 Enterprise;
平台类型: 64 位操作平台; 处理器: Intel (R) Xeon (R) CPU E7 - 4820 @ 2.00GHz; 安装内存:

8.00GB。

3.2 开发流程

由于开发时间有限, 故平台在实现过程中运用快速迭代的开发模式, 这既保证每次交付的结果都是完整可用的, 也能及时得到使用反馈, 有助于快速调整产品方向, 避免浪费时间和精力。平台开发流程, 见图 2。具体开发时将平台分前台和后台两大模块, 使用 ThinkPHP 开发框架快速实现平台功能模块, 后台 Web 端页面以开源 Bootstrap 模板 ACE Admin 为基础。平台各功能模块实现流程基本类似 (ThinkPHP 框架模式): 基于模块/控制器/操作设计 (MVC), 平台通过入口文件绑定模块和控制器后, 对应用的所有请求都定向到应用入口文件, 系统会从 URL 参数中解析当前请求的模块、控制器和操作。在平台实现过程中, 每一个 MySQL 数据集都绑定特定模型文件 (平台中使用 ThinkPHP 中 M 变量) 访问及操作数据库, 其上通过控制器处理输入, 操作数据库或者请求查询, 将结果数据通过模板变量传递至 Web 前台, 使用 JavaScript 等前端脚本语言便可将其填充到前台页面或进行进一步处理 (如可视化展示、分页展示等)。

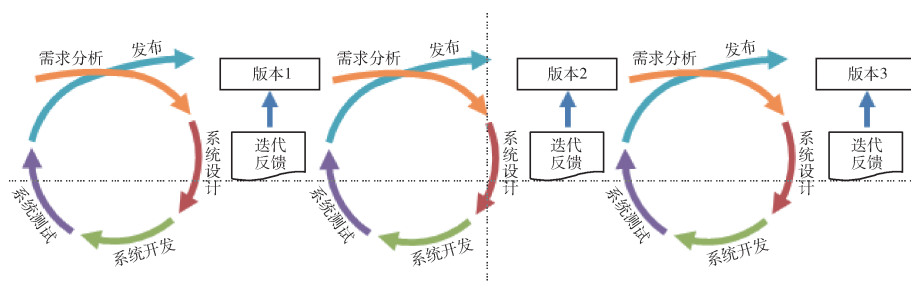


图 2 平台开发流程

3.3 实现效果

医学科学数据管理与共享平台原型系统构建较为完善的前后台功能。其中前台具体实现 6 大功能模块: 系统首页、数据列表、个人数据、数据服务、发布数据和用户中心; 后台实现 4 大功能模板: 系统管理、审核管理、元数

据管理和数据集管理。用户可以通过数据集定义——数据表定义——信息审核——数据上传——数据审核 5 个步骤上传数据。上传的数据也会根据用户所定义的权限, 展示数据集信息、元数据信息等部分或者全部内容, 方便用户了解平台中所收录信息的具体细节。平台首页实现效果, 见图 3。

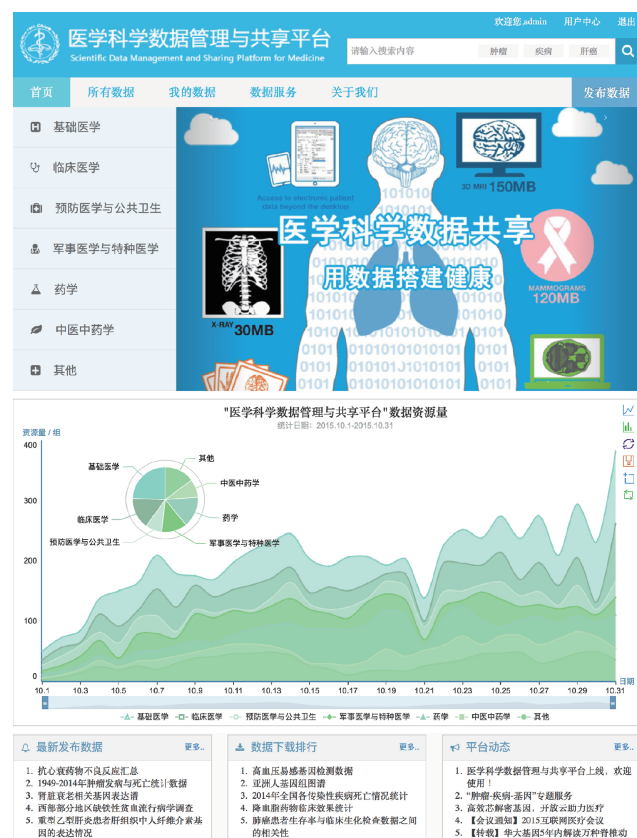


图 3 首页实现效果

4 结语

医学科学数据管理与共享平台通过规范化的数据上传流程以及灵活的元数据配置,对用户上传的数据内外部特征进行规范,有助于同类目数据的整合,为科学发现提供条件;数据分类管理不仅有利于不同层次数据的聚合,也提高数据查询和响应速度;多层面的规范处理在保证数据质量的同时也增加数据的语义和关联,有助于多源异构数据间的整合挖掘;分级共享在促进科学数据最大限度传播的同时,也保护提交者的专有权。

目前平台仍存在诸多尚待完善之处:(1)平台收集的医学科学数据主要来自开放资源,元数据格式多样,虽然平台已使用一套比较粗的元数据标准规范体系,采用人工规范和动态扩展元数据项相结合的方式,但这些元数据主要侧重于数据分类、数据集和部分数据表层面的标准,在后续研究中还需进一步深入研究数据表和数据记录层元数据规范。

(2)虽然使用元数据对多源异构的数据进行规范,但目前平台的主要功能还侧重于数据的管理与整合,缺乏不同数据之间的关联及挖掘。因此在后续工作中应引入相关数据挖掘算法,探索不同数据后潜藏的价值。(3)使用 R2RML 映射语言进行 RDB 到 RDF 转换时,仅能由管理员通过后台进行操作,用户无法在线编辑映射规则。仅使用 R2RML 语言中缺省映射模式,未能实现数据向相关标准本体的定制映射。此外映射过程中还存在空白节点的处理、父节点的表示等问题。因此在后续工作中需要不断优化映射规则并增强用户的交互性,使语义转换功能更加完善、便捷。(4)构建阶段也曾尝试使用非关系型数据库 mongoDB 以及 RDF 3 元组仓储工具 Virtuoso,在后续工作中将继续探索不同类型数据库在医学科学数据存储以及整合利用方面的特点,从而选择更加高效的仓储方式。

参考文献

- 顾立平. 科学数据开放获取的政策研究 [M]. 北京: 科技文献出版社, 2016: 45.
- 徐敏, 李广建. 第四范式视角下情报研究的展望 [J]. 情报理论与实践, 2017, 40 (2): 7-11.
- 代涛. 健康医疗大数据发展应用的思考 [J]. 医学信息学杂志, 2016, 37 (2): 1-8.
- 马秀平. 基于 RDF 的海量语义数据管理研究 [J]. 科技信息, 2010, 17 (22): 595-596.
- 国家人口健康科学数据共享平台. 数据资源分类与编码方案 [EB/OL]. [2017-05-03]. <http://www.ncmi.cn/UploadFile/7/0/b660d6db9e0b9e97a4421fbd765d07.pdf>.
- 曹锦丹, 李欣欣. 基于 DC 的医学信息资源元数据比较分析 [J]. 图书情报工作, 2003, 24 (7): 24-27.
- 中国国家标准化管理委员会. 国家科技基础条件平台资源元数据核心元数据 [EB/OL]. [2017-05-03]. <http://www.ncmi.cn/UploadFile/2/f/286b6c638d46592285b6e183e140b3f2.pdf>.
- W3C Incubator Group. A Survey of Current Approaches for Mapping of Relational Databases to RDF [EB/OL]. [2017-05-03]. https://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.
- Tim Berners-Lee. Linked Data [EB/OL]. [2017-05-03]. <https://www.w3.org/DesignIssues/LinkedData.html>.
- DAS S, SUNDARA S, CYGANIAK R (Eds.). R2RML: RDB to RDF Mapping Language (W3C Recommendation) [EB/OL]. [2017-05-03]. <http://www.w3.org/TR/r2rml>.
- 汤怡洁, 杨锐, 刘毅. 近年来 SPARQL 查询技术的研究热点及进展 [J]. 知识管理论坛, 2014, 5 (74): 1-6.