PATSTAT专利数据库数据集成策略研究*

杨冠灿，张静，望俊成
（中国科学技术信息研究所，北京 100038）

摘要：PATSTAT（Worldwide Patent Statistical Database，全球专利统计数据库）是一款由EPO欧洲专利局开发的、面向统计决策的专利数据仓库。该数据库在专利数据集成方面取得一系列的研究进展，其在语义映射、语义匹配方面的经验对于解决现阶段专利数据语义异构性问题具有较强的参考意义。通过逆向归纳的方法，文章阐述PATSTAT数据库在专利家族、优先权、摘要、标题、发明人、地址信息以及专利权人信息方面的集成策略，并对PATSTAT数据库在专利数据集成经验进行总结。

关键词：PATSTAT数据库；专利数据；数据集成；语义异构
中图分类号：G358；G353
DOI：10.3772/j.issn.1673-2286.2015.09.002

1 问题提出

长期以来，专利数据被广泛应用于科技评价活动中。专利数据由于涵盖的信息全面、规范且易于使用，所以受到了理论界与学术界的广泛青睐[1]。近年来，在数据提供商以及各国专利管理机构的共同努力下，一大批专利数据库被开发出来，其中既包括免费的，也包括商用的，这些数据库的存在使得我们可以便利的使用专利数据资源[2]。由于专利信息涉及的范围广、来源多，其蕴含的具体内容、涉及的时间空间跨度、信息的组织形式都存在很大的差异，使得不同数据来源的专利信息存在极大的异构性。

这种异构性表现在多个方面，既有数据的异构性，即多种专利信息数据库以不同的数据格式存储、展现；又有语义的异构性，即不同数据来源的同一专利信息类型、字段、值的表示方式存在差异[3]，表现为概念模式型语义异构以及数值型语义异构，概念模式型语义异构包括：表—表异构，属性—属性异构，值—属性异构，值—表异构，属性—表异构；而数值型语义异构则包括同一名称的数据具有不同表达形式，不同的数据粒度，不同的描述内容等。之所以存在这些语义层面的异构，根源在于数据源的加工者对于应用场景预期的差异。于是，对于这种质量上参差不齐，内容上千差万别，形式上多种多样的复杂专利信息资源，如何从应用场景出发对异构数据进行集成，提高专利信息分析的效率与精度是目前专利数据加工过程中面临的难题。

近年来，国内学术界与实务界开始关注专利数据集成问题，形成了一系列面向数据集成的专利数据集成方案，如《专利知识挖掘关键技术研究》[4]、《异构专利数据源集成方案设计与实现》[5]、《中文专利信息资源深加工方案设计与实证研究》[6]、《德温特专利信息清洗与标注模型研究》[7]等。然而，目前针对专利信息集成的研究主要还是集中于解决数据异构层面，即如何

* 本研究得到国家科技支撑计划“面向科技创新的专利信息加工与服务关键技术研究与应用示范”（编号：2013BAH21B00），国家自然科学基金青年基金项目“基于指数随机图模型的专利引用关系形成影响因素及机理研究”（编号：71403256）以及国家社会科学基金青年项目“专利信息的生命特征揭示和老化规律研究”（编号：12CTQ025）资助。

将多源、异构的专利数据集中存储到一个平台上供分析使用，而这种单纯针对数据异构进行集成所存在的问题在于：缺乏对于应用场景思考，主要是通过数据库工具或者ETL工具实现对异构数据的标准化存储，在数据库设计环节对于专利数据存在的语义异构性问题关注较浅，导致最终形成的专利数据产品往往可供分析的维度与深度不够。面对专利数据复杂性的挑战，亟需在深入理解数据结构的基础上，针对特定应用场景，通过数据集成为科研人员提供更为准确的专利数据以及可供分析的维度。

PATSTAT（Worldwide Patent Statistical Database，全球专利统计数据库）是一款在经济合作与发展组织（OECD）的倡导下，由EPO欧洲专利局主导开发的，面向统计决策的专利数据仓库产品[8]。该产品的主要特点包括：（1）面向统计决策分析；（2）多源数据的集成；（3）体现数据仓库特征；（4）资源共享与协同创新。PATSTAT在数据集成方面，尤其是语义集成方面的经验值得数据加工人员学习与借鉴。

2 专利数据集成的概念与挑战

2.1 数据异构挑战

专利数据集成面临数据异构的挑战。不同来源的专利数据库往往由不同数据格式保存，数据集成过程中需要先统一转化为以关系型数据为基础的数据仓库。随着数据加工规范的逐步标准化，加之一些国家部门和商业机构的共同努力，基础层面（著录项、法律状态、过程信息）的数据异构正在逐步缩小。例如，EPO通过自己多年的数据加工实践，逐步形成了面向DOCDB（改进专利文献著录项数据质量）、REFI（协调全球专利引文数据）、MCD（协调各国国际分类数据）、EPR（利用专利申请过程信息补充提高数据质量）在内的多个数据库，这些数据库均遵循较为统一的数据加工规范，为解决数据异构问题奠定了良好的基础[14]。

2.2 语义异构挑战

专利数据集成中面临语义异构的挑战。从数据库系统实现角度理解，语义异构实质上就是要解决中介模式和源模式（或者任意两个模式）之间存在的差异[9]。通常情况下，在数据标准化之后，可以通过关系型数据库规范方式来对数据进行逻辑、物理建模，然而，由于多源专利信息的语义异构问题，使得数据库的建设必须回到最初预设的应用场景中，统一不同数据来源之间的深层次语义差异。

目前很多专注于数据异构集成的研究或多或少的涉及到了语义异构问题。例如，多源专利数据的组织问题，很多学者采用了数据库设计中通行的做法：设立虚拟技术标识符的方法或者利用DERWENT唯一标识符的方法来实现[7, 15]，这些做法在数据更新时以及数据之间存在差异时又会产生新的问题；另外在专利语义异构集成中面临的困难是：缺乏语义集成的依据[16]，实践中数据加工人员对于专利优先权、家族数据不敢触碰的，但优先权、家族数据对于识别技术发展战略又是至关重要的信息，这种割裂限制了数据分析的深度；又如多源专利数据的选择问题，很多情况下，由于数据加工人员缺乏对整体专利数据的理解（不同数据来源的专利信息存在数据范围、更新时效、权威性、应用目的上巨大的差异）[17]，在进行专利数据的选择过程中，不可避免会伴随着一些限制。在专利信息逐步走向更深层次的应用过程中，如何解决专利数据语义异构集成会是实践过程中的瓶颈。

3 PATSTAT数据库在专利数据集成上的策略

3.1 专利家族与优先权信息集成

3.1.1 专利家族和优先权

专利家族和优先权密不可分。《OECD专利统计手册》对专利家族的定义“专利家族是以一个或多个相互关联的共同优先权为基础，并在多个国家申请的一系列专利文献集合”[1]。目前，国际上同时存在多种专利家族体系，主要有DERWENT专利家族、INPADOC专利家族以及简单专利家族等。这些专利家族概念体系“并非由法律所规定，而是由各个数据库提供商出于自己便利的考虑而决定”[18]，这导致专利家族的概念变得难以理解。在Martinez和Catalina两位学者最新研究的基础上[19]，结合PATSTAT在专利家族方面的实践，本文认为理解专利家族概念需要理解专利家族概念的范畴问题和专利家族成员之间的关联规则。

PATSTAT数据库之所以能够对专利家族数据和专利优先权数据进行集成，其关键就是在理论上很好的回答了上述两个问题。目前，EPO自身包含了两种主流的专利家族分类体系：INPADOC扩展专利家族和DOCDB简单专利家族。INPADOC的专利家族的定义是“通过一份优先权文献直接或间接联系起来的一系列专利文献被称为一项INPADOC扩展专利家族”[20]。可见，INPADOC专利家族的概念范畴更宽泛的，既包括直接优先权关系也包含间接的优先权关系，这暗示了该家族所识别出的专利文献可能不仅包括一项技术，还会包含在其发展演化过程中产生的衍生技术。相对而言，DOCDB简单专利家族的范畴就比较窄，仅包含了由直接优先权关系组织在一起的专利家族成员，对应的技术内容也会窄一些，可以简单理解为一项技术本身，不包含技术的发展演化过程。

举例而言，如图1所示，专利文献D1的优先权文献为P1，专利文献D2的优先权文献为P1和P2，专利文献D3的优先权文献为P1和P2，专利文献D4的优先权文献为P2和P3。根据INPADOC扩展专利家族原则，专利文献D1、D2、D3、D4为同一个扩展专利家族，因为，专利文献D1、D2、D3、D4之间通过直接和间接优先权链接关系被联系到了一起；而根据DOCDB简单专利家族原则，专利文献D2、D3属于一个简单专利家族，因为专利文献D2、D3拥有共同的优先权链接关系P1，P2，但专利文献D1、D4则不属于共同的简单专利家族成员。

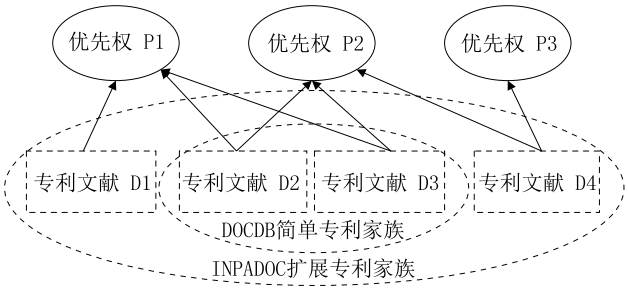

图1 INPADOC扩展专利家族和DOCDB简单专利家族概念范畴的比较示例

在定义了概念范畴之后，专利家族的建立还需要以一定的关联规则确定哪些专利属于同一专利家族。PATSTAT是以DOCDB数据的链接方法指针字段（LMI，Link Method Indicator）为基础，该指针又被称为链接类型（Linkage Type）。DOCDB在对优先权关系进行深入研究的基础上，对优先权链接关系类型进行了扩展，并将这4类链接关系类型作为DOCDB简单专利家族判断的主要依据，分别是：（1）巴黎公约优先权链接类型；（2）技术相似性链接类型（亦称为非公约优先权、知识优先权或技术关系）；（3）国内优先权链接类型（如续案、部分续案、临时专利申请、分案等）；（4）进入地区/国家阶段的专利合作条约（PCT）链接类型[21,22]，具体含义见表1。

表1 两个来源的申请人/专利权人信息示例

链接类型	含义
巴黎公约优先权链接类型	根据《巴黎公约》专利优先权国际保护规定，形成的首次专利申请与后续专利申请之间的链接关系。
进入地区/国家阶段的（PCT）链接类型	根据《专利合作条约（PCT）》规定，形成的PCT国际申请与进入到国家/地区阶段的专利申请之间的链接关系。
技术相似性链接类型	专利审查员在专利审查过程中通过人工判别方式形成的链接关系。人工判别的依据是专利审查员在审查过程中发现一些专利之间虽然没有直接或间接优先权链接关系，但发明人、申请人相同或相近的，且申请的专利在技术上具有高度的相似性。
国内优先权链接类型	有些国家的国内程序也会导致首次专利申请与后续专利申请之间形成某种链接关系。最常见的情形是续案、部分续案、临时专利申请（这三种情形多见于美国专利及商标局）或者是分案（许多专利局都存在）。

3.1.2 专利家族和优先权信息集成

基于专利家族概念范畴以及专利家族成员之间链接的关系问题的分析可知，专利家族与专利优先权之间有着紧密的联系，即专利家族是以概念范畴指引，以优先权链接关系为基础，辅助人工判断及其他因素的一套体系。从面向统计决策支持的应用场景考虑，PATSTAT数据库并没有预设一个专利家族概念范畴，而是希望通过数据集成，使科研人员能够自由地从不同的概念范畴视角来观察专利数据，这里不仅包括是INPADOC扩展专利家族，也包括DOCDB简单专利家族数据，更

重要的是帮助研究人员通过自定义概念范畴来实现自定义的专利家族。

在这种思想的指引下，PATSTAT将专利家族和优先权概念有机的统一到了一个逻辑体系，实现了三种家族数据层次之间自由沟通。研究人员可根据研究目的选择等同专利家族层次（Equivalent Patent Family-level）、简单专利家族层次（Simple Patent Family-level）以及INPADOC专利家族层次（INPADOC Patent Family-level）来观察专利数据。更重要的是，研究人员可以根据自己定义的专利家族概念范畴，通过优先权链接类型的组合重构符合自己研究要求的家族定义。

从数据集成的实践角度来看，PATSTAT考虑到研究人员研究目的的差异，因此，首先保留了INPADOC扩展专利家族数据集（TLS219_INPADOC_FAM）和DOCDB简单专利家族数据集（TLS218_DOCDB_FAM）。在此基础上，PATSTAT通过四张表中的TLS204_APPLN_PRIOR，TLS201_APPLN，TLS205_TECH_REL，TLS216_CONTN将四种优先权链接关系类型的数据展现出来。从表2中可以观察到，INPADOC扩展专利家族与DOCDB简单专利家族概念范畴的差异在于是否包含间接优先权，而在链接关系类型方面的最大差异在于INPADOC扩展专利家族不包含技术相似性链接，即INPADOC扩展专利家族的建立不需要人工干预，而DOCDB则是需要根据数据质量流程监控和审查员的人工判断才能完成的。

3.2 专利摘要和标题信息的集成

一项专利申请可能多次公开、公告，每次公开、公告的标题和摘要信息可能是一致的，也可能并不一致；同时，一项专利还可能在多个国家和地区提出申请，也会产生多种语言版本的标题与摘要信息。PATSTAT对专利摘要和题目信息建立选择规则：

（1）根据公告日期，选择最新的英文标题；

（2）选择最新公告中语言的标题；

（3）选择最新任意语言的标题。

通过上述规则对专利数据库中的题目和摘要信息进行筛选、去重，实现数据集成。

3.3 专利发明人和地址信息的集成

各国官方专利数据库在专利发明人和地址信息方面缺乏较为统一的规范，翻译为英文之后也会带来一些问题，再加上拼写错误、印刷错误等，使得专利发明人和地址信息的处理显得异常的困难[23]。以PATSTAT中美国专利发明人和地址信息为例，分析PATSTAT解决专利发明人和地址信息集成的思路。PATSTAT的美国专利数据源有两个：一是来源于EPO的以交换数据为基础的DOCDB数据，另一个是来源于USPTO提供的专利数据。DOCDB中的美国专利发明人和地址信息，来源权威、覆盖面广，缺点是存在一定的更新和修改时滞问题；USPTO中的美国专利发明人和地址信息在数据的权威性以及更新的时效方面都是最为权威的，但数据涵盖范围有限。因此，PATSTAT对专利发明人和地址信息进行集成的策略是：

（1）1976年1月1日之后的美国授权专利发明人和地址信息都是来源于USPTO专利授权数据。

（2）2005年9月29日之后的美国申请专利发明人和地址信息都来源于USPTO专利申请数据。

（3）1976年1月1日之前的美国授权信息，以及在2005年9月25日之前的美国申请专利信息中，包含的发明人和地址信息均来源于DOCDB数据库。

表2 INPADOC扩展专利家族、DOCDB简单专利家族的差异比较

	细分类别	INPADOC扩展专利家族	DOCDB简单专利家族	数据表	对应字段
专利家族范畴	直接优先权关系		√	TLS204_APPLN_PRIOR	prior_appln_id
	直接优先权关系、间接优先权关系	√		TLS204_APPLN_PRIOR	prior_appln_id
链接关系类型	巴黎公约优先权链接	√	√	TLS204_APPLN_PRIOR	prior_appln_id
	进入地区/国家阶段的PCT链接	√	√	TLS201_APPLN	internat_appln_id
	技术相似性链接		√	TLS205_TECH_REL	tech_rel_appln_id
	国内优先权链接	√	√	TLS216_CONTN	parent_appln_id

3.4 专利权人信息的集成

在数据集成过程中，数据匹配的任务就是要找出描述相同的现实世界的数据项。PATSTAT数据库在该项工作的实践方面也取得了较大的进展，较为突出是专利权人名称的清洗与协调。就专利权人名称清洗与协调而言，可分解为四个方面的任务：数据清洗、内外部数据补充、数据匹配以及筛选[23-24]。其中，数据清洗、补充是基础，匹配是数据集成工作的关键。本文着重介绍PATSTAT在专利权人字段在补充、匹配工作方面的进展。

专利权人名称清洗与协调工作可以初步分为两个层次：专利权人名称清洗（Patentee Name Harmonization）以及法律实体协调（Legal Entity Harmonization）[25]。专利权人名称清洗，主要是对于专利权人字段本身进行清洗，通过清洗统一专利权人名称的用词规范，实现对专利权人名称的标准化；法律实体协调则强调对专利权人名称所对应的历史信息、权属关系进行标准化。两者的差异具体体现在以下三个方面：工作任务的差异、遵循加工原则的差异、数据来源的差异，见表3。

PATSTAT在处理专利权人清洗与协调工作时，采取了更加务实的方案：对于专利权人名称清洗工作，PATSTAT采用了DOCDB的标准化专利权人数据集（DOC_STD_NAME），该数据集对DOCDB的专利权人名称进行了统一的数据清洗工作，该数据集主要采取的工作包括去除多余的字符、空格等。对于需要利用外部数据辅助进行的法律实体协调工作，PATSTAT吸收来自EURASTAT，OECD，KU Leuven等机构的研究成果，逐步完善目前的专利权人名称数据。目前，已经采纳的数据集包括OECD的HAN数据集以及KU Leuven的EEE-PAT数据集。

4 结语

PATSTAT作为专利信息服务领域最杰出的产品，其在专利数据加工、集成、设计方面的方法和经验具有独到之处，值得学习和借鉴。PATSTAT吸纳技术创新理论发展的最新成果，将《OECD专利统计手册》作为其理论依据[1]；吸纳OECD在专利引文[27]、专利家族[28]、专利权人清洗[29]、专利地理信息[30]、专利质量评价[31] 等方面的最新成果，形成了一系列的专利数据集[32]。

PATSTAT数据库公开了其在数据库设计与开发过程中的一些核心文档，便于参考，同时这些设计文档中

表3 专利权人名称清洗与法律实体协调差异比较

差异体现	专利权人名称清洗 (Patentee Name Harmonization)	法律实体协调 (Legal Entity Harmonization)
工作任务	（1）拼写变化，如“IBM” and “I.B.M.” （2）印刷错误，如“INTERNATIONAL BUSINESS MACHINES” and “INTERATIONAL BUSINESS MACHINES” （3）法律形式的统一，如如“IBM”，“IBMCORP.”，“IBM CORPORATION” （4）（地区、机构、部门）的统一，如“IBM” and “IBM JAPAN”； （5）缩略语的统一，如“IBM” and “INTERNATIONAL BUSINESS MACHINES”	（1）对属于同一法律实体的不同机构进行识别（业务部门、下属机构） （2）识别法律实体名称的变化情况 （3）识别企业兼并并购情况 （4）识别企业合资情况 （5）识别母子公司关系以及下属机构情况
加工原则	以准确性为基础，兼顾全面性，通常采用人工与算法结合方法进行质量控制	难以顾及全面性，主要是采用算法对准确性进行质量控制
数据来源	主要以专利数据为基础，如利用地址信息、非专利引文信息、专利权人合作、引用关系信息进行补充	需要借助外部数据源，如融资并购数据库中的专利权人名称信息、股票年报、地理信息等
方法选择	以规则匹配为主的方法 这里主要可以参考EURASTAT出版的技术报告，其中，对于专利权人部门（Sector Allocation, Name Harmonization）的介绍[25]	采用算法来进行匹配： 编辑（Levenshtein）距离算法[25] 块匹配（Block Matching）算法[23] 作者重名辨别（Torvik Smalheiser）算法[26]

涉及的数据库设计基本原则、元数据信息以及具体实施规则与代码，都能够对专利数据集成提供帮助。

参考文献

[1] OECD. OECD Patent Statistics Manual [M]. Paris: Organisation for Economic Co-operation and Development, 2009.

[2] Albrecht M A, Bosma R, van Dinter T, et al. Quality assurance in the EPO Patent Information Resource[J]. World Patent Information, 2010, 32(4): 279-286.

[3] 方丽英,王普,闫健卓. 面向语义异构的信息集成系统查询处理方案[J]. 北京工业大学学报, 2007(8): 819-822.

[4] 翟东升. 专利知识挖掘关键技术研究[M]. 北京: 知识产权出版社, 2013.

[5] 王志,孙涌,张书奎,等. 基于本体的专利数据源集成的研究及应用[J]. 计算机技术与发展. 2009, 19(7): 87-90, 94.

[6] 张兆锋,桂婕,李颖,等. 中文专利信息资源深加工方案设计与实证研究[J]. 数字图书馆论坛,2014(7): 45-51.

[7] 翟东升,李倩,张杰,等. 德温特专利信息清洗与标注模型研究[J]. 情报杂志,2013(8): 150-154.

[8] EPO. EPO Worldwide Patent Statistical Database (PATSTAT) [EB/OL]. [2014-04-26]. http://www.epo.org/searching/subscription/raw/product-14-24.html.

[9] Doan A,Halevy A,Ives Z. 数据集成原理[M]. 北京: 机械工业出版社, 2014.

[10] Coffano M, Tarasconi G. CRIOS - Patstat Database: Sources, Contents and Access Rules[R]. Universit Bocconi: Organization and Strategy CRIOS, 2014.

[11] 胡开胜. 基于WEB元数据抽取的ETL资源整合模型研究与实现[D]. 长沙: 湖南师范大学, 2010.

[12] 储振华,孙涌,张书奎,等. 异构专利数据集成研究[J]. 计算机与现代化, 2009(5): 29-32.

[13] 翟东升,禾文汇. 异构专利数据源集成方案设计与实现[J]. 现代图书情报技术,2010(9): 67-73.

[14] 汤艳莉,杜萍. 浅析欧洲专利局专利信息资源的质量保障[J]. 中国发明与专利,2012(6): 94-97.

[15] 禾文汇. 基于BI的专利数据整合分析研究及实现[D].北京: 北京工业大学, 2011.

[16] 经济合作与发展组织. 弗拉斯卡蒂手册[M]. 北京: 科学技术文献出版社,2012.

[17] 甘绍宁,曾志华. 全球专利信息公共检索资源指南[M]. 北京: 知识产权出版社, 2015.

[18] Simmons E S. "Black sheep" in the patent family [J]. World Patent Information, 2009, 31(1): 11-18.

[19] Martinez C. Insight into Different Types of Patent Families[R]. OECD Science: Technology and Industry Working Papers, 2010.

[20] Lingua D G. INPADOC: 30 years of endeavours yet unmapped territories remain[J]. World Patent Information. 2005, 27(2): 105-111.

[21] EPO. Data Catalog V 5.01 Patstat. 2014 [EB/OL]. [2014-04-26]. http://www.epo.org/searching/subscription/raw/product-14-24.html.

[22] EPO. DOCDB User Documentation v. 2.4.4. 2014 [EB/OL]. [2014-04-26]. http://www.epo.org/searching/subscription/raw/product-14-7.html.

[23] Lissoni F, Coffano M, Maurino A, et al. APE-INV's "Name Game" algorithm challenge: A guideline for benchmark data analysis & reporting[R].Technical Report: Academic Patenting in Europe-APE-INV, 2010.

[24] Den Besten M, Lissoni F, Maurino A, et al. APE-INV Data Dissemination and Users' Feedback Project [EB/OL]. [2014-04-26]. http://www.esf-ape-inv.eu.

[25] Van Loody B, Du Plessis M, Magerman T. Data production methods for harmonised patent statistics: Assignee sector allocation[R]. Luxembourg: European Commission,2006.

[26] Ventura S L, Nugent R, Fuchs E R. Methods Matter: Rethinking Inventor Disambiguation with Classification & Labeled Inventor Records[C]. Academy of Management Proceedings. Academy of Management, 2013(1): 14537.

[27] Webb C, Dernis H, Harhoff D, et al. Analysing European and International Patent Citations: A Set of EPO Patent Database Building Blocks[R]. OECD Publishing, 2005.

[28] Dernis H, Khan M. Triadic Patent Families Methodology[R]. OECD Publishing, 2004.

[29] Ribeiro S P, Menghinello S, De Backer K. The OECD ORBIS Database: Responding to the Need for Firm-Level Micro-Data in the OECD[R]. OECD Publishing, 2010.

[30] Squicciarini M, Dernis H. A Cross-Country Characterisation of the Patenting Behaviour of Firms based on Matched Firm and Patent Data[R]. OECD Publishing, 2013.

[31] Squicciarini M, Dernis H, Criscuolo C. Measuring Patent Quality: Indicators of Technological and Economic Value[R]. OECD Publishing, 2013.

[32] OECD. OECD work on patent statistics [EB/OL]. [2014-04-26]. http://www.oecd.org/sti/inno/oecdpatentdatabases.htm.

作者简介

杨冠灿，男，1981年生，博士，中国科学技术信息研究所助理研究员，研究方向：专利数据、技术竞争情报，E-mail：yangc@istic.ac.cn。
张静，女，1975年生，博士，中国科学技术信息研究所副研究员，研究方向：科技政策与管理、信息分析等。
望俊成，男，1984年生，博士，中国科学技术信息研究所副研究员，研究方向：专利分析、信息老化、用户行为等。

Study on Patent Data Integration Tactics of PATSTAT Database

YANG GuanCan, ZHANG Jing, WANG JunCheng
（Institute of Scientific and Technical Information of China, Beijing 100038, China）

Abstract: PATSTAT (Worldwide Patent Statistical Database) is the world's leading patent warehouse product developed by the EPO to serve statistical decision supporting. The product has made a series of research progress at patent data integration, and has instruction and reference value to solve the issues on patent semantic-level heterogeneity. The paper adopts a reverse inductive method, summarizing the progresses of PATSTAT database. Finally, this article summarizes the patent data integration experiences of PATSTAT database, which can use for patent data processing and analysis in future.

Keywords: PATSTAT Database; Patent Data; Data Integration; Semantic Heterogeneity

（收稿日期：2015-07-20；编辑：王立学）

《数字图书馆论坛》2015年征稿启事

《数字图书馆论坛》是由科学技术部主管、中国科学技术信息研究所主办的专业性学术刊物（月刊），国际标准刊号ISSN：1673-2286，国内统一刊号CN：11-5359/G2。本刊是“中国科技核心期刊”统计源刊，是CSSCI扩展版来源期刊。

本刊是我国唯一一本以“数字图书馆”命名的刊物，一直关注国内外数字图书馆领域的相关研究和实践，设有特别关注、专家访谈、专题研究、技术前沿、应用案例、业界动态等栏目，报道主题涵盖信息检索、数字资源、知识组织、语义技术、开放获取、用户服务等，侧重反映数字图书馆领域在资源建设、技术应用和产品服务等方面的新趋势、新发展和新变革。

本刊注重稿件的学术水准、研究内容和研究特色，来稿需要满足以下基本要求：①未发表过、未一稿多投的原创性论文；②主题鲜明、数据可靠、文字通顺、引用规范；③来稿应包含以下项目：中文和英文的标题、作者姓名、单位、摘要和关键词，以及中图分类号、参考文献和作者联系方式。请登录本刊网站（http://www.DLF.net.cn）进行在线投稿。

本刊收到稿件后，会及时登记、编号，分至责任编辑。初审合格的稿件将送至相关领域的同行专家进行外审，周期为半个月左右。本刊会将评审意见通过E-mail通知作者，作者应在规定时间内将修改稿返回编辑部，并对修改意见作出逐条答复。修改后通过主编终审的稿件，本刊将寄送录用通知。文章在发表前，本刊会将编辑加工过的稿件清样通过E-mail发送给作者校对、修订。文章发表后，本刊将向作者寄送样刊并付稿酬。作者可登陆本刊网站查询稿件处理情况。

本刊既厚名家、更重新人。欢迎国内外作者赐稿。本刊特别期待相关专家就某一课题项目/主题提供系列专题稿件。本刊开放出版（网址：http://www.DLF.net.cn），也期待着相关专家在阅读或利用后提出宝贵意见和建议。