基于领域知识关联的集成服务系统研究
——以植物学领域为例

朱艳华 胡良霖
（中国科学院计算机网络信息中心，北京 100190）

摘 要：在对知识关联的概念和常见应用形式概述的基础上，探索科学数据系统引入知识关联需要解决的两个关键性问题，即如何集成实体对象的知识节点和如何确定实体对象的关联标识。然后以植物学领域为切入点，分析知识关联模型在植物数据整合和增值服务方面发挥的作用，构建植物学知识关联的集成服务实验系统，从而为知识关联技术在实现异构数据库系统互操作方面做出了有益探索。

关键词：知识关联；科学数据；植物学领域；数据整合；增值服务；集成服务系统
中图分类号：TP311 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2013.01.010

Research of Integrated Service System Based on Domain Knowledge Connection by Botany Field

Zhu Yanhua, Hu Lianglin
(Computer Network Information Center of CAS, Beijing 100190)

Abstract: Knowledge connection technology is a useful exploration to achieve interoperability of heterogeneous database systems. Building a scientific and rational connection model is critical for computer information processing and data content interoperability. This paper outlines knowledge connection's concept and its common application forms. To introduce knowledge connection in scientific databases system, we need to solve two key issues. The first is how to integrate all attribute data of the same entity object; the second is how to determine its unique identifier. The paper mainly analyzes the role of knowledge connection and tries to develop integrated service experiment system in the botany field.

Keywords: knowledge connection, scientific data, botany field, data integration, value-added service, integrated service system

1 引言

随着科学数据的快速积累和信息服务技术的不断成熟，科学数据服务系统从孤立封闭的状态逐步走向开放融合的阶段，构建合理的数据服务系统要考虑其实用性、互操作性和持续性。同时，科学数据资源具有多元且复杂的异构性，不但不同学科数据库之间的数据格式、数据结构、操作平台和应用系统存在差异，而且同一学科数据库内部也存在命名方式、数据结构模型方面的不统一[1]。目前还没有适用的解决方案能够对各个层次的异构数据进行合理的处理，实现科学数据之间的有效连接和整合。而知识关联技术在实现异构数据库系统集成方面进行了有益的探索。

知识关联是指构成知识系统的知识节点与节点之间的联系，即使各相关节点间形成意义系统的联

第一作者简介：朱艳华（1982- ），女，工程师，硕士，主要研究方向：数据库技术与标准规范，数据应用服务。
收稿日期：2012年10月24日。

系。这种联系表现为一种拓扑结构形式存在的网络结构[2]。在知识关联中知识与知识之间以某一中介为纽带，建立起具有参考价值的关联关系。概括地说，每个知识关联网络都由各类知识单元以及这些单元之间的关联关系构成。知识关联能够整合和揭示知识节点之间的联系，是知识管理、知识发现和知识创新的基础。

科学知识具有继承性、累积性和连续性。任何新学科或新技术都是在原有学科或技术的基础上分化、衍生出来的，都是对原有学科或技术的继承和创新。同时，科学也具有统一性原则，各个学科之间都是彼此联系、相互交叉和相互渗透的[3]。建立学科之间的知识关联具有现实的基础和可能，构建科学合理的关联模型是计算机处理信息和解决数据内容整合的关键，科学数据引入知识关联的意义亦在于此。科学数据开展知识关联方面的研究需要明确自身的目标定位，最初可以从某一学科领域知识入手，深入考察该领域知识关联的内容范围、功能作用和应用模式，藉此为知识关联在科学数据应用环境中的全面实施探索道路。

本文拟以植物学领域为切入点，在科学数据库系统中选取与植物有关的数据资源，构建植物知识关联模型，这些知识节点不仅包括植物学内容，而且涉及到与植物相关的其他领域；深入挖掘知识关联在植物数据整合和增值服务等方面发挥的功用，并尝试开发基于植物领域知识关联的集成服务实验系统。

2　植物学知识的关联形式

知识关联揭示了大量知识单元之间存在的序化联系，以及隐藏的、最终可用的关联关系。有学者认为知识关联具有相互性、传递性、普遍性、多重性、隐含性、积累性和动态性、可创造性、层次性和结构性等特征[4]。常见的知识关联类型有分类词表、主题词典、文献或数据引用以及语义网中的本体等。其中，分类词表和主题词典是早期的知识关联和分类组织形式；文献或数据的引用关联揭示了图书文献或数据之间的引证与被引证关系，开辟了知识组织的另一途径；本体作为语义网实现的关键技术之一，是对共享概念模型的明确的形式化规范说明[5]，领域本体描述了特定领域中的概念和概念之间的各种关系。

知识管理研究机构kmpro首席分析师王振宇认为知识关联在知识管理中有6种常见的应用形式，即类别关联、关键词关联、诊断/推理关联、聚类关联、行为关联和属性关联[6]。其中，类别关联是最为常见的一种关联方式，即属于同一个知识分类中的知识之间的关联。以植物学领域为例，物种分类体系中的界、门、纲、目、科、属、种揭示了物种之间的等级分类关系。关键词关联是以知识内容中的关键词作为关联纽带，有相同关键词的知识进入关联体系中。如研究植物的论文通过关键词聚合相似研究内容的文献。诊断/推理关联是以一个问题为核心，将解决该问题的知识层层推理出来。如在查找植物地区分布的过程中，通过行政区划表将较小的区县范围扩展到较大的省市，查找到更广泛的植物地区分布信息。聚类关联是通过对定量知识的分析，聚类出相关性较强的内容。如研究植物物种信息时，可以聚类物种的引种保育、化学成分和功能用途等相关性很高的内容。行为关联通过对知识使用者的行为进行分析，发现这些行为之间的关联性和连续性，从而推理出用户所进行这些行为是运用知识间的关联性。属性关联是以知识与知识之间的同一个属性为中介将知识关联起来。如根据植物的地区分布或花果期可以集成同一地区的物种或同花果期的物种等。

3　构建植物学知识关联集成服务系统

3.1　植物学资源范围

植物资源是科学数据资源体系的重要组成部分，包括中国植物物种信息数据库、东北植物与生境数据库、西双版纳热带植物园植物引种与保育数据库、中国热带亚热带植物学基础数据库和中国植物图谱数据库等。这些数据库整合了与植物相关的各类信息，如植物物种的基本信息、图片视频、引种保育、野外生长、染色体和研究文献等内容。科学数据资源体系还包括与植物相关的其他领域数据，如植物化学成分数据库收集了从植物所含有或者提取得到的化合物，包含化合物名称、结构、分子式和含量等信息；重要物种DNA条形码数据库整合了植物等重要类群的DNA条码数据，同时采集了与这些条码密切相关的样品采集信息、物种鉴定信息、条码引物信息、PCR扩增信息和Trace File信息等内容。

植物领域知识繁多而且分散，为保证植物集成服务系统的科学性和合理性，在系统构建之前，

必须确定资源内容收集指导原则。在对系统功能需求和数据资源获取易操作性等综合考虑的基础上，我们明确了植物内容收集的全面性原则和专指性原则。全面性指的是系统收录植物知识的完整程度，我们梳理了整个科学数据库系统中所有与植物相关的资源，不仅考虑植物领域学科知识，也涉及跨领域的知识扩展。专指性是指所收集的知识针对领域研究的核心和重点，所选择的知识节点是植物领域研究者关注的内容。

3.2 植物学知识节点

在确定了收集范围和收集原则后，我们就着手设计植物知识关联模型，关联模型是开发集成服务系统的关键。通过该模型，我们梳理了植物知识的脉络，并分类分层地组织所获取的知识节点；以植物物种为最小记录，确定的植物知识节点包括：植物物种基本信息、图片视频、引种、保育、野外采集、化学成分、染色体、DNA条码、植物功用和研究文献等，其关联模块、知识节点、来源数据库和详细内容见表1所示。值得说明的是，这种整合不是物理集中，而是基于知识的逻辑集中。

基于表1中确定的16个关联节点，我们设计了植物知识关联图（图1）。图1中每个节点对应一个来源数据库，每个来源数据库都有物种拉丁名信息，因此，各知识节点之间通过物种拉丁名进行关联。针对同一个物种拉丁名之间的差异，我们采取人工方式进行判断和处理。

3.3 系统平台的构建

开发集成服务系统平台主要包括3个方面工作：首先从来源数据库中选择系统实现所需要的样例数据；其次设计系统平台页面呈现风格和服务功能；最后由程序员开发平台系统。

作为一个实验系统平台，首要任务是确定样例

表1 植物知识关联节点

关联模块	知识节点	来源数据库	详细内容
植物基本信息	基本信息	物种基本信息库	种拉丁名、种中文名、属拉丁名、属中文名、科拉丁名、科中文名、种别名、国外分布、国内分布、形态特征、生长环境、用途、参考文献等
植物图片视频	线描图	物种图谱数据库	物种线描图
	照片	物种图片库	物种照片
	视频	物种视频库	物种视频
植物引种	引种	引种植物数据库	引种材料、引种数量、引种时间、引种人、引种地点、海拔、坡向、坡度、坡位、母岩、地貌、性状、采集母株、生长状况、Totalnum、记录人等
植物保育	保育	保育植物数据库	现成活数量、国内分布、引种时间、引种地点、引种人、海拔、坡度、坡向、坡位、地貌、性状、原始株数、死亡记录次数、繁殖记录次数、移植记录次数、植物类型等
植物野外采集	野外生长	野外采集数据库	采集时间、采集地点、采集人、采集样品类型、生活型、生态环境、种子颜色、叶子颜色、果期、分布、单位面积产量、油脂含量初检、标本鉴定人等
植物化石	化石信息	植物化石标本数据库	名称信息、采集信息、时代产地、参考文献等
植物化学成分	化学成分	植物化学成分数据库	化合物名称、结构、分子式、含量等
植物染色体	染色体	植物染色体数据库	分布/采集地、材料来源、研究部位、倍性、核类型、臂指数、最长染色体/最短染色体、臂比>2染色体的百分比、核型不对系数、染色体相对长度组成、刊登刊物、作者、卷期/页码等
植物DNA条码	DNA条码	重要生物类群DNA条码数据库	样本编号、matK、rbcL、trnH−psbA等
植物功用	药用	药用植物数据库	中文名、药性描述
	饲用	饲用植物图谱数据库	形态特征、地理分布、生物生态特征、饲用价值等
	有毒	有毒植物图谱数据库	形态特征、分布、生境、毒性、毒理、科中文名、科拉丁名、本科概述等
	油脂	油脂植物数据	命名人、种下等级类型、种下等级加词、种下等级命名人、描述文件、油脂含量等
植物文献	研究文章	研究论文数据库	物种名称、文献标题、页码等

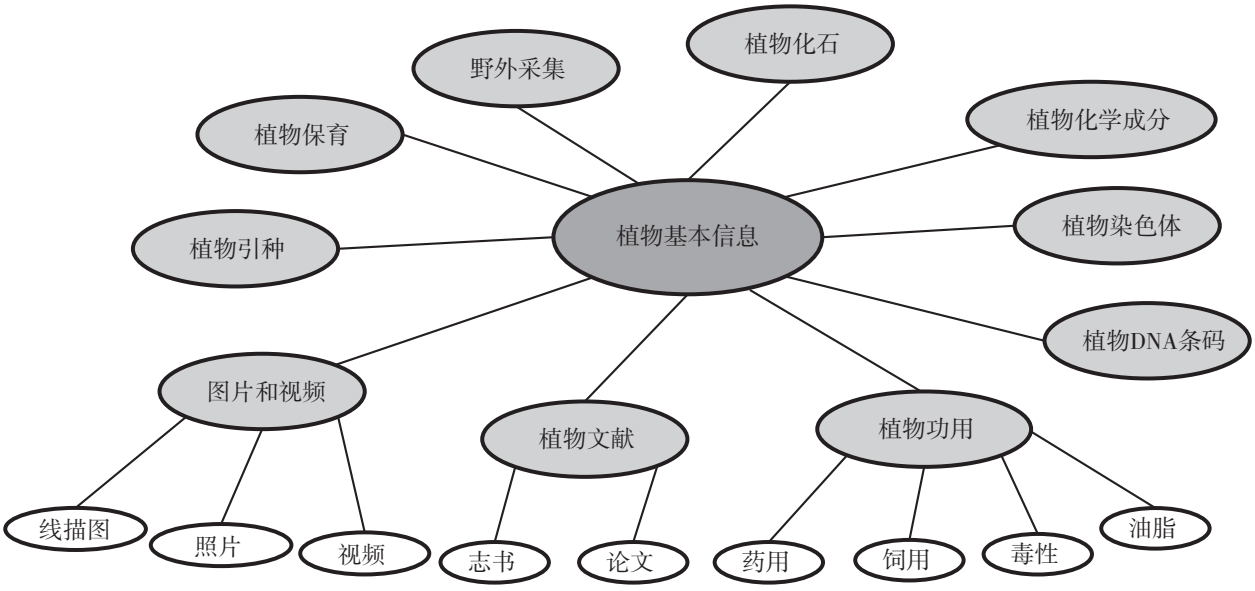

图1　植物知识关联模型图

数据。我们选择那些在知识节点中出现较多的数据作为样例数据。为了实现平台服务功能，知识关联图的每个节点都要包含尽可能多的数据内容。我们主要通过物种拉丁名查找在来源数据库中重复较多的数据记录。物种拉丁名是每一个知识节点来源数据库都包含的信息，而且在很多情况下作为唯一标识符，因此依据物种拉丁名，在这些选定的数据库中确定重合度高的物种作为系统样例数据。

平台页面主要包括3类：首页面、概览页面和细览页面。其中，首页面包括检索功能区域、内容简介和知识关联图。概览页面主要展示检索结果的概要信息，当检索结果不止有一条记录时，系统将返回一个检索列表，列表中主要显示物种基本信息。细览页面展示每个节点的全部信息，即基本信息、详细信息、更多信息、关联信息和知识关联图信息。基本信息显示物种基本内容；详细信息显示每个知识节点的核心信息；更多信息跳转至该节点来源数据库记录的详细页面，关联信息显示与该物种有关的其他物种信息；知识关联图以拓扑图形式展示与该节点相关的其他节点入口。

确定了样例数据和页面设计方案后，程序员就可以开发系统平台了。当然，在开发的过程中，我们还要根据具体数据情况调整知识关联节点。系统开发完成后，我们还需要制定规范的系统测试方案，修改和完善平台功能，保证平台稳定有效地展示和运行。

4　植物学知识关联集成服务系统优势

相比一般的数据服务平台，我们构建的集成服务系统在提高检索效率和增值服务方面具有显著优势。其优势主要表现在以下几个方面。

（1）实现智能语义检索。在植物学领域构建知识关联模型可以实现服务平台数据查询的语义理解和扩展，具有智能检索功能。现阶段的查询请求主要是通过将查询语句解析成一个个单词，然后进行关键词匹配，再把匹配的结果按照一定算法进行过滤和排序，最后提供给用户，这个过程几乎没有语义分析。引入知识关联模型后，用户基于自然语言的查询请求就可以翻译成植物学领域相关概念组成的查询语句，并根据检索提问的不同，对查询请求进行知识扩展，如通过植物物种异名和俗名等字段获取检索关键词的同义扩展；通过植物分类等级确定的物种分类关系而进行的属性扩展；利用行政地理分布字典的等级区划推断物种分布地信息等[7]。用户输入检索词后，经过有关字段的扩展和推理，得到语义丰富的关键词列表，用户再根据这些扩展后的关键词组，进一步明确检索需求，提高检全率，在一定程度上实现了智能语义检索。

（2）提升数据增值服务。在植物学领域构建知识关联模型可以提升数据增值服务。通过领域知识关联中概念之间丰富的关联和其中包含的规则进行逻辑推理，深入分析和挖掘数据内部隐含的语义知识关系，由一个知识点扩展到相关知识单元，并最

终形成整个领域，甚至跨领域的知识网络，实现数据更高层次的增值服务。基于关联知识模型的系统平台能够对检索结果所隐含的知识关联进行有效分析，让用户基于一次查询就能快捷获取增值服务的体验。用户如果检索某物种的引种保育情况，平台还能提供与物种引种保育相关的各种信息，如集成同引种人物种、同引种地物种、同海拔物种和同性状物种等额外信息，为用户提供新的研究视角。

5 科学数据库系统知识关联的探索

1982年，中国科学院正式提出科学数据库及其应用系统建设项目。经过30年的持续发展，参与数据库建设的单位从最初的几家研究所扩展到院内62家研究所，几乎覆盖了中国科学院所有的研究领域；科学数据库工程已经建成为一个庞大的、资源类型多样的科学数据库群。“十一五”期间，中国科学院科技数据基础设施正式列入中国科学院信息化基本环境，进行重点建设，为科研活动提供综合性的数据应用环境[8]。在“十五”的基础上，全院从信息化战略资源高度，系统规划科学数据资源体系。基于院内有特色和长期积累的数据资源，通过严格质量控制与管理建成了具有完整性和权威性的2个参考型数据库；根据国家和院内部署的重大研究计划或项目，建设了4个专题数据库；面向特定的学科和应用领域，整合若干逻辑相关的数据库，建设了8个主题数据库；并从“十五”期间已支持且服务比较好、使用比较广泛的数据库中择优确定了37个重点专业数据库[9]。

面对这样一个来自不同建库单位，数据量庞大且存在着复杂异构的科学数据库体系，如何实现数据资源的集成和共享引起研究者的普遍关注。要解决科学数据跨学科集成，我们需要解决两个关键问题，一是如何实现同一个实体对象所有属性数据的集成，二是如何确定不同来源数据库的对象唯一标识，完成个体之间的有效关联。知识关联可以解决数据库跨学科集成问题，每个知识关联网络都是由各知识单元以及这些单元之间的关联关系构成，我们引入知识关联的意义亦在于此。

5.1 集成实体对象的知识节点

科学数据一般包括标识实体对象的数据和描述实体对象属性的数据。事实上，实体对象的标识属性和描述属性也是相对的，特别是在交叉学科和跨领域的科学数据中[10]。在某个学科领域作为描述对象属性的数据，在另一个相关学科领域可能是标识数据，而在某个学科领域作为对象个体标识的数据，在另一个相关学科领域则成为属性数据。如植物研究论文数据库，从文献角度看，论文是其实体对象，文献中研究的物种是论文的属性数据，表示文章的研究内容，而从植物学角度看，物种也是个实体对象，论文可以作为其属性数据。

在科学数据库系统内部构建知识关联模型，需要明确研究实体对象和内容范围，集成同一个对象的相关属性数据，并从中抽象出实体对象的知识节点，确定实体对象的唯一标识，建立对象之间的关联关系。例如在化学领域构建知识关联，以化合物为实体对象，关联植物化学成分数据库中的植物物种数据、药物数据库中化合物治疗的疾病数据等。知识关联实现了跨学科领域对象个体间的连接，数据整合不再局限于某一个学科，而是扩展到多个相关学科。利用知识关联模型图，我们就能从实体对象的一个知识单元找到另一个知识单元，而且在同一个节点上还能深入挖掘其相关信息，比如借助药用植物数据库中的物种信息可以集成具有相同药用价值的物种。

5.2 确定实体对象的关联标识

要实现实体对象知识节点之间的有效关联，需要确定个体对象唯一标识。由于数据来源不同或采用不同的标识规则，科学数据库的个体异构是一个很普遍的现象。所谓科学数据的个体异构，指的是对同一个对象使用了不同的表述方法，使得在不同数据库中的相同个体无法确定相互间的关系[9]。以植物为例，物种作为一个实体对象，通常以物种名作为标识，但是在实际操作的过程中，物种名又存在拉丁学名、中文学名、异名和俗名等情况。现在通用的林奈双名命名体系中，植物拉丁名采用两个拉丁化的名字来命名。第一个名代表“属”名，第二个名代表“种加词”，属名和种加词组合起来构成了物种名。在种名的后面，再注上命名者的姓名。即使都使用物种拉丁名，也会因为不同分类体系中不同人名的拼写、属种分类的差异而造成的物种标识异构，导致植物数据之间无法直接采用物种拉丁名集成和共享。为此，我们需要建立不同标识转换的映射表，通过这个对照表，同一个对象的不同标识符号都会映射到表中确定的唯一标识，解决个体对象的异构问题。

将知识关联引入到科学数据库集成服务是我们探索科学数据特色应用服务的一种尝试，本文基于科学数据库系统中植物及其相关领域的数据资源，构建植物领域知识关联模型，并尝试开发基于植物领域知识关联的集成服务实验系统，为知识关联技术在科学数据应用环境的实施探索道路。

6　结语

构建领域知识关联模型对解决科学数据异构，实现多个数据库系统之间的互操作和多角度整合具有重要意义，知识关联模型在实现跨领域数据整合、智能检索、数据增值服务和数据文献关联等诸多方面都能发挥作用。本文探索了在科学数据库系统创建知识关联需要解决的两个关键问题，即挖掘集成实体对象的知识节点和确定实体对象的关联标识，并以植物学领域为切入点，基于系统内的植物相关资源内容，构建植物领域知识关联模型，尝试开发了集成服务实验系统。

植物学知识关联的集成服务系统是个开放的平台，随着植物领域研究的不断深入和拓展，必然会产生更多的研究数据，我们的平台也需要及时跟踪最新的领域数据资源，补充知识节点，更新关联模型，收录新的知识内容。

本文的研究为知识关联技术在科学数据应用环境中的实施途径作出了探索。知识关联模型对解决科学数据异构，实现多个数据库系统之间的互操作和多角度整合具有重要意义，知识关联模型在实现跨领域数据整合、智能检索、数据增值服务和数据文献关联等诸多方面都能发挥作用。

参考文献

[1] 刘炜,李大玲,夏翠娟.元数据与知识本体[J].图书馆杂志,2004(6):51.

[2] 文庭孝,刘晓英.知识关联的结构分析[J].图书馆,2011(2):1−7.

[3] 邱均平.信息计量学[M].武汉:武汉大学出版社,2007:318−319.

[4] 文庭孝,龚蛟腾.知识关联:内涵、特征与类型[J].图书馆,2011(4):32−35.

[5] Gruber T. Ontolingua: A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993,5(2):199−200.

[6] 王振宇.浅谈知识关联在知识管理中的应用[EB/OL].[2010−04−08]. http://www.kmpro.cn/html/kmyanjiuyuan/kmproheibanbao/10242.html.

[7] 戴维民等.语义网信息组织技术与方法[M].上海:学林出版社,2008:111−116.

[8] 中国科学院信息办.2010年中国科学院信息化资源报告[R].北京,2010.

[9] 中国科学院数据应用环境.[EB/OL].[2012−08−15]. www.csdb.cn.

[10] 陈维明.科学数据个体识别和跨学科集成[C]//科学数据库与信息技术论文集.北京:科学出版社,2012:10−17.