图书馆资源组织中的数据关联特征研究*

常　娥，孟祥保

摘　要　深刻描述和揭示馆藏资源与数据自身及其之间各种复杂语义关系是图书馆资源组织的重要问题。文章分析了馆藏资源数据化转变问题，并重新界定了馆藏资源的类型，探索从资源外部到内部所存在的各种数据关联的特征、层次和基本结构，重点研究书籍、文章，以及其他学术文献与其内部原始数据、模型、算法等知识项的关联问题。

关键词　数据关联　知识关联　资源组织　馆藏资源　数据化

引用本文格式　常娥，孟祥保. 图书馆资源组织中的数据关联特征研究[J]. 图书馆论坛，2016（2）：49-56.

The Characteristics of Data Association in Library Resources Organization

CHANG E，MENG Xiang-bao

Abstract　It is very important to deeply reveal and describe a variety of complex semantic relations between collections and data，and their semantic correlations for library resource organization. This paper discusses the collection datazation and tries to redefine the types of library collections. Furthermore，it analyses the characteristics，layers and basic rules of data association of library collections，especially focusing on the association of articles，books，as well as other academic literature and its internal raw data，models，algorithms.

Keywords　data association；knowledge association；resource organization；library collection；datazation

随着数字图书馆的迅速发展，海量多源异构数据增长迅猛，图书馆不遗余力地从系统层面、元数据层面乃至资源层面整合馆藏资源，但数据融合与关联的范围、深度和效果仍不尽如人意。同时，网络数字资源和用户获取信息途径的发展与更新，使得图书馆的在线利用率仍处在一个较低水平[1]。面对馆藏资源组织与利用的困境，图书馆一直希望摆脱信息孤岛的束缚，实现馆藏资源与网络资源的关联，为搜索引擎广泛抓取并索引，以提升图书馆资源的可见度和可获取性。这一困境的解决有赖于图书馆突破传统资源组织的理念和方法，将各种载体形式的资源转化为万维网上相互连通的“数据”，以更加细粒度、关联化的方式组织馆藏资源[2]。因此，深刻揭示和描述馆藏资源与数据自身及其之间各种复杂语义关系成为图书馆资源组织的重要问题。

随着关联数据技术的快速应用与发展，图书馆界掀起了馆藏资源数据关联化的研究热潮[3-6]，然而已有项目大多从书目数据或本体模型方面研究馆藏数据之间的关联关系，鲜有嵌入资源数据化过程，从资源与数据的不同层面出发，深入比较和分析资源与数据自身及其之间各种复杂语义

* 本文系国家社会科学基金青年项目“图书馆资源组织中的数据关联机制研究”（项目编号：14CTQ003）研究成果之一

关系。邱均平等学者研究知识单元之间的关联关系，产生了大量的研究成果[7]，然而他们主要从知识计量角度进行研究，将粗粒度的资源与细粒度的知识统一视作知识单元，未加区分。本文将通过分析馆藏资源数据化转变问题，重新界定馆藏资源的类型，系统探索从资源外部到内部所存在的各种数据关联的特征、层次和基本结构，以弥补现有研究的不足，为后续研究提供思考方向。

1 馆藏资源的数据化

1.1 从数字化到数据化

1996 年，美国麻省理工学院教授尼古拉斯·尼葛洛庞帝(Nicholas Negroponte)在《数字化生存》(Being Digital)这部未来学著作中，充分展示了数字化技术给人类社会带来的巨大变革。数字化与网络化、信息化息息相关。所谓数字化是指把一切信息转化为计算机可以识别的比特[8]。比特是数字化信息的基本单位，被尼葛洛庞帝教授称为“信息的 DNA”。20 世纪 90 年代末，网络技术和计算机技术发展迅速，网络数字资源变得极为丰富，与此同时电子商务、电子政务、数字校园以及数字图书馆等项目建设得如火如荼，完成了海量信息资源的数字化加工。十多年过去了，这部未来学著作描述的情景已全部实现，数字化已成为人们生活、工作、教育和娱乐的基本方式。如今人类又来到了数据化(Datazation)时代，数据化是“人类在信息传播、人际交往乃至日常生活的过程中，为了便于沟通、传播与保存，将一切客观存在均处理为数据，进而使得整个人类社会成为一个庞大的数据库”“数据化是人类社会在数字时代的必然发展趋势”[9]。数字化带来了数据化，但数字化无法取代数据化，两者虽只有一字之差，却划分出两个截然不同的技术时代。

大数据时代，馆藏资源正经历着从数字化到数据化的重要转变。馆藏资源数字化是把纸质文献变成计算机可读取的比特流数据，简单而言就是把书籍内容整体数字化，把纸上的东西搬到计算机显示器上阅读和编辑。馆藏资源数据化则是把计算机可识别文本内容中的字、词、句、段落、图像以及各种概念、公式、数据等知识单元进行分割和提取，并建立知识单元之间的各种关联，从而进行无穷无尽的检索、分析与挖掘。数字化是馆藏资源描述方式的电子化，资源的载体形态发生了转变；而数据化则是对馆藏资源内容的重新拆分与组合，资源内容的组织粒度发生了根本转变。例如，日本对马克思、恩格斯文献的数据化处理不仅仅停留在数字化、文本化方面，而是将批注与其所解读的文本相链接，以实现视觉化展示和理解为目的[10]。

1.2 数据化的粒度控制

馆藏资源数据化的首要任务是将资源内容拆分为一个个可独立封装的知识单元。什么是合适的知识单元呢？知识单元的大小如何控制？在馆藏资源数据化过程中，将不可避免地遇到知识细分的粒度控制问题，这是目前困扰并阻碍知识组织、知识管理以及知识计量等领域深入发展的难题。学者们提出了粒度规范[11]、多级粒度(模糊等价)[12]、粗糙集[13]、数学期望[14]等诸多知识单元的粒度控制方法，以实现粗粒度知识单元向不同层次细粒度知识单元转化的目的。但将文献资源细分到什么程度才算最优，学界还未达成共识，并且对于知识单元的颗粒度也没有十分明确的规定，它既可以是一个词语、也可以是一小段文本或一幅图像，甚至是一篇文章[15]。

本文认为，知识单元的颗粒度与其包含的文本字符数关系不大，反映的是知识单元分解与组合的层次。例如，一个专业术语和一个数学定义各自包含的文本字符数可能不同，但颗粒度一致，属于同一层次的知识单元。文献资源通常由词语、句子、段落、小节以及篇章等不同层次的内容所构成。一本著作由若干篇章组成，包含大量的知识点；一篇论文由若干小节组成，同样拥有许多相关知识点，著作和论文是典型的粗粒度知识单元。为了规范文献资源转换为数据单元的颗粒度，本文将最小、独立且不可再分割的基本知识单元定义为元知识项，一般为独立的概念、定义、公式、定理、模型、方法、事实、数据、

图表、一段叙述或一组图解操作等。在馆藏资源数据化转换中，主要将文献资源统一解构为元知识项。

2 馆藏资源的类型

长期以来，图书馆为组织和描述各种馆藏资源，促进馆藏资源的检索与利用，创建并保存了各种类型的信息资源，如机读目录数据格式、各种元数据方案、主题表、分类表，以及规范文档等，这些作为图书馆相对独立的行业标准与最佳实践理应被视为“资源”看待[16]。

目前，伴随着馆藏资源数据化进程的发展，各国图书馆已将馆藏书目数据、各种规范文档等制作成关联数据集进行发布，从而使其在国际数据交换生态系统中的重要作用逐渐显现。例如，在开放数据网络中，国际虚拟规范档(VIAF)数据集已与维基百科、国际标准标识符(ISNI)、社会网络与档案文本(SNAC)等数据集进行了深度关联与融合，被认为是万维网利用图书馆资源数据的门槛。因此，图书馆收集、整理和发布于网络中的数据集是馆藏资源的另一种重要形式。

随着信息环境的改变，馆藏资源的类型在不断发展之中扩大。图书馆馆藏资源从最初的纸质文献资料，扩展成集电子资源、网络资源于一体的综合馆藏资源体系。实体馆藏、虚拟馆藏、电子馆藏、数字馆藏、复合馆藏等相关概念也反映了馆藏资源类型的变化。在开放互联的数据网络时代，馆藏资源的类型又有了新的变化，将更丰富多样。本文将不同载体形式的图书、期刊、报纸、手稿、研究数据等文献资料称为实际馆藏，将建立在实际馆藏资源之上的各种描述称为衍生馆藏。其中，衍生馆藏又可划分成为 3 大类型：

(1)资源数据集，是指图书馆建立的各种文献资料的直接描述。资源数据集是图书馆在开放数据网络中的主体资源，目前主要是发布为关联数据集的书目记录，例如瑞典国家图书馆的书目数据集、OCLC 的 WordCat 关联数据集等，未来会以 FRBR 等本体模型映射的资源数据集为主。相比而言，以 FRBR 模型为核心的资源数据集构建过程复杂，需要深入文献资源内部，将其解构为各种知识单元。其中，文献中零散数值数据的提取尤其需要注意：需连同其关键上下文或标题等信息一并提取，使其成为知识单元。数值数据只有在特定的语境中才能被理解和再使用，脱离了语境的数值数据毫无意义[17]。

(2)元数据元素集，由各种资源描述方案中的元数据词汇构成。元数据词汇是为了描述实际馆藏中资源实体的特征及其关系而定义的类别与属性，包括本体模型中定义的各种属性词汇。以图书情报机构为首的各行各业制定了各种资源描述方案，由此而引发了元数据互操作问题，这一直是知识组织领域研究的热点内容。在开放数据网络中，图书馆若将各种元数据词汇集作为资源进行发布，并采用 RDF 模型进行架构，为每一词汇赋予可被参引的 URI，那么将极大促进元数据词汇的重用和共享，建立起资源描述的共通基础。从这一层面来说，将元数据元素集纳入馆藏资源建设范畴，拓展了图书馆在开放数字环境中的优势。

(3)取值词汇集，指各种词表和规范文档，用来表示资源描述记录中某些元素的规范化取值。每个特定主题领域都有约定俗成的概念术语，充分利用这些概念术语将有助于资源描述的规范和统一，保证词汇在语义层面的一致性。图书馆历来十分重视书目的规范控制工作，积累了丰富的经验，包括各种标题表、分类表、叙词表、主题规范档和名称规范档等的制定。借助关联数据技术将取值词汇集转换为 RDF 模式，并为每一词汇赋予唯一的 URI 标识，可在开放数据网络中充分展示图书馆的规范控制优势。国内外图书馆界早已意识到了这一点，纷纷将取值词汇集作为书目数据之外的又一重要馆藏资源，积极将其发布为关联数据集。

3 馆藏资源组织中的数据关联特征

厘清馆藏资源与数据自身及其之间各种复杂的关联关系是图书馆资源组织中的重要问题。本文将从资源与资源、数据与数据，以及资源与数

据 3 个层面，全面、深刻地揭示馆藏资源组织中存在的数据关联关系及其特征。

3.1 资源与资源的关联

资源与资源的关联主要体现在两个方面：外部特征关联和内容特征关联。外部特征关联主要指以作者、引文、标题、机构、期刊、载体类型、发表时间、语种等信息建立资源之间的关联；内容特征关联主要指以反映资源主题内容的主题词、关键词、摘要、分类号等信息建立资源之间的关联。

利用外部特征关联文献资源，存在两种模式：(1)“文献—外部特征—文献”模式，即以文献资源作为网络节点，以外部特征作为节点连线，关联生成的网络称为文献资源网；(2)“外部特征—文献—外部特征”模式，即以外部特征作为网络节点，以文献资源作为节点连线，关联生成的网络称为非文献资源网。经常使用的外部特征主要包含作者、引文、期刊和机构。并非所有的外部特征都可用来关联文献资源，只有满足一对多关系的外部特征才能用来组织关联网络，也就是说，要么一篇文献具有多个同类型外部特征，要么一种外部特征对应于多篇文献。例如，资源的作者、引文和机构三种特征都满足上述两种形式的一对多关系，既可生成文献资源网，又可生成非文献资源网，但期刊特征只满足第二种形式的一对多关系，只能生成非文献资源网。

在众多资源外部特征关联网络中，利用文献之间的引用关系而建立起来的“文献—引文—文献”网络(简称引文网络)最为特殊。文献计量学领域进行了大量、细致而深入的研究，证明了它在尝试定义、解释科学结构方面，具有突出和卓有成效的作用[18]。在非文献资源网络中，除使用“文献”关联网络节点外，还可使用其他关系建立节点连线，例如合作关系、引用关系等。以作者关联网络为例，根据网络节点连线类型的不同，可细分为作者合作网、作者共被引网、作者文献耦合网、作者关键词耦合网以及作者期刊耦合网 5 种典型网络[19]。非文献资源网络除了具有挖掘学科领域核心作者和核心机构等功能外，还可从另一个侧面揭示学科领域的知识结构，但不同类型的非文献资源网络揭示学科结构的准确性与精确性、学科结构发现的角度，以及其与文献资源网络之间的替代性等问题，还有待进一步深入研究。

值得注意的是，资源的外部特征存在不稳定性。主要表现为：(1)当资源的载体形态由纸质转变为数字方式时，某些外部特征将随之改变；(2)当资源的知识粒度变小时，某些外部特征将随之改变。换言之，具有显著区分能力的外部特征会随着资源的载体形态和知识粒度的改变而变化。例如，页码数、尺寸等特征是纸质资源的重要标识项，然而当资源的载体转为数字形式时，页码数、尺寸等特征将完全消失，再如纸质资源需要标识作者、机构、出版社、出版时间等信息，数字资源则需另外标识网络平台、数据库商等信息。此外，当资源的知识粒度细化到元知识项时，除原始创作者，如公理、定理等核心学科概念的创立者外，出版、发行等外部信息则无需标识。

相比于复杂、多变的外部特征，资源的内容特征相对稳定，不会随资源载体形态和资源组织粒度的改变而改变。一篇文献可解构为多个知识单元，可利用知识单元共现原理关联文献资源。例如，若以关键词为知识单元，可生成“文献—关键词—文献”网络，即以文献资源为网络节点，为含有相同关键词的两个网络节点建立链接而形成的网络。文献节点连线可以单个关键词同现为基础，也可以一组关键词同现为基础，从而生成不同规模的文献资源关联网络。“文献—关键词—文献”与“文献—引文—文献”这两种网络形成的机理既有联系又有区别，前者是已有文献内容关联的客观呈现，后者是作者主动筛选文献内容建立关联的结果。两种网络都可看成是由知识单元同现关系所构成，在文献内容的关联与学科知识结构揭示上是相通的，然而它们在多大程度上功能相通，又有何差异，值得进一步深入研究。

3.2 数据与数据的关联

数据关联主要指知识单元之间的语义关系，又称为知识关联。语义关系最早在语言学、逻辑学、心理学和计算机领域中被定义和研究[20]。语

义关系是指两个或两个以上概念或实体之间有意义的关联，一般表现为“概念1—关系—概念2”三元组的形式。任何一个概念都不是独立的，需要借助其他概念进行定义，因此概念的含义包含了与其它概念的相互关系。不同于资源层面的关联，数据层面的关联是广泛、普遍存在的。总体来说，数据关联存在相互性、传递性、普遍性、多重性、隐含性、动态性、可创造性、层次性和结构性等特点[21]。

语义关系是概念之间内在关联性的体现，比语法更能体现概念间的联系。早期图书情报领域并未将语义关系作为研究重点，有关语义的研究主要集中在概念和术语构建上，建立了简单知识组织系统。分类表和叙词表是其中的典型代表，它们的语义关系简单、直接和明确。分类表中的语义关系包含等级关系、等同关系、相关关系和并列关系4种类型，没有进一步细分。叙词表中的语义关系包含等级关系、等同关系和相关关系3种类型，其中等同关系又细分为“用”“代”2种关系，等级关系则细分为“属”“分”“族”3种关系。

随着信息环境的变化和知识组织研究的深入，语义关系在知识组织、检索、推理与挖掘中的作用越来越重要，简单的知识组织系统正朝着能描述更细致、丰富、深入和全面语义关系的复杂知识系统转变。本体(Ontology)是复杂知识组织系统的典型代表，其语义关系设置取决于领域本体的构建目标和具体应用，因而不同领域本体的语义关系在数量和种类上差异较大。例如，CYC定义了成千上万种语义关系，而UMLS却仅定义了54种语义关系。

复杂知识组织系统中的语义关系是对等级关系、等同关系、相关关系和并列关系进行更为细致、具体而深入的表述。一般而言，复杂知识组织系统将等级关系细化为属种关系(ISA，AKO等)、整体部分关系(a-part-of，a-member-of等)、事物与属性关系(property of等)以及实例关系(an-instance-of等)等；相关关系细化为空间相关(located-at，near-to等)、功能相关(affect，interference，prevent等)、时间相关(before，after等)、概念相关(method-of，value-of等)等；等同关系细化为同义关系(same as)、反义关系(different-to)、近似关系(similar-to)、等价关系(equivalent-to)等。然而，无论语义关系在复杂知识组织系统中被如何细化，最终都可归为等级关系、等同关系、相关关系和并列关系这4种基本类型。

除了语义关联外，还可利用共现关系建立知识单元之间的联系。根据不同的共现方式，可构建相同或不同知识单元的各种知识关联网络。例如以关键词为例，可构建“关键词—文献—关键词”(简称关键词共现网络，或共词网)、“关键词–作者–关键词”、“关键词—机构—关键词”等网络。这些网络都以关键词为节点，以来源于同一篇文献、同一个作者、同一个机构等为共现方式，建立节点连线，形成关键词网络。由于关键词是文献核心内容的集中表达，因此通过分析关键词网络，可发现隐藏在某篇具体文本语义关系网背后的知识网络，这对了解研究领域的知识结构具有非常重要的意义。通过同现关联构建的数据之间的关联关系有强弱之分，可利用同现频次进行测度。

此外，“关键词—文献—关键词”与“文献—关键词—文献”这两种网络较易混淆。前者是关键词共现网，以关键词为网络节点，体现了数据层面的知识关联。后者是文献资源网，以文献为网络节点，体现了资源层面的知识关联。由于网络节点和知识粒度的差异，这两种知识网络虽有联系，但本质却不同。

3.3 资源与数据的关联

James P. McCusker等学者认为，关联科学面临的一个挑战就是如何在关联科学云里充分描述一条信息的来源，信息来源的表述是科学应用中信任关联数据的关键。换言之，知识网络无法完全取代文献资源本身，需要建立文献资源与知识单元之间的关联，这与计算机、人工智能等领域的知识组织研究略有不同。

图书情报学早期的文献标引组织研究就致力于从结构上揭示文献与知识单元的关联关系，提出了各种引用次序，例如阮冈纳赞的P、M、E、

S、T引用次序，我国叙词表“主题因素—通用因素—空间因素—时间因素—文献类型因素”的引用次序等。然而囿于纸质文献人工标引的局限性，这些文献标引方式无法充分揭示文献的内容特征，一般仅给出分类号和有限主题词，未严格按照规定的引用次序组织主题词，最终仅形成较为简单的词袋式、集合列表式的关联结构，即一篇文献关联到一组标引词集合，是一种松散的关联结构。

文献资源的数字化为深入文献内容的全面标引提供了可能。学者们更加致力于按照某种框架结构组织知识单元，从而在语义上尽可能还原文献内容的本来面貌，而不是罗列一组关键词或主题词。由于不同学科文献资源结构框架差异较大，即使是同一学科，不同类型文献的内容框架也不尽相同，因此学者们开始关注如何通过调研大量的文献内容结构框架，总结规律，以建立通用的文献结构框架，然后按照这个框架逐块提取知识单元，构建资源与知识单元的结构关联。

对文献内容框架的研究最好莫过于研究学术文献的内容框架。学术文献题材属于议论文范畴，知识内容规范严谨，在结构上有规律可循。其中，自然科学类学术文献的内容结构最为规范，基本按照“研究问题—实验方法—实验数据—评价与分析”的框架来组织知识内容[22]。社会科学类学术文献的内容结构形式多样，但与自然科学文献相同的是，都围绕着某一问题展开论述或进行试验，因此社科类文献可归纳为“目的 / 意义—方法 / 过程—结果 / 结论”3 个核心部分。为了追求最小本体论承诺，有学者还提出了“假设—现象”这种最为简洁的知识结构[23]。此外，有些专业期刊要求作者在摘要中提供结构化的知识，主要包括 4 个部分：背景 / 目的 / 意义、方法 / 材料 / 过程、结果 / 结论、相关工作 / 未来工作。这一系列工作都为揭示资源与数据的关联关系奠定了基础。

结合文献标引引用次序原理，以及学术文献内容结构框架，本文认为应引入完整的学术元数据框架，以更好地建立资源与数据的关联。学术元数据框架应包含研究对象、背景、问题、意义、目标、方法、材料、过程、实验数据、结果、结论、讨论、评价等重要元数据项，将文献中析出的概念、原理、观点、规范、标准、设计、技巧、事实、数据、人物等知识单元，在此框架下与文献资源建立关联。学术元数据框架有别于传统资源描述框架(比如，MARC、DC 等)，它重在建立文献资源与内部知识单元的映射与关联，而非进行资源外部特征描述。

4 馆藏资源数据关联的基本结构

“数据—信息—知识”之间的组织与转化体现了图书情报学独特的价值，三者相互之间如何更为有效地组织与转化也一直是学界争论和研究的焦点[24]。与数据、信息和知识密切相关的另一个重要概念是文献。从定义上看，文献指记录有知识的一切载体。实际上，文献同时承载了数据、信息和知识，并记录及表达了数据、信息和知识的转化结果。但毕竟文献资源只是知识的载体，与知识本身存在巨大差异，因此知识与文献资源的组织与利用并不完全遵循同一规律和模型。本研究通过深入分析馆藏资源数据化转变问题以及馆藏资源类型的发展与变化，探索了从资源外部到内部所存在的各种数据关联的关系、特征与结构，下面将从数据关联的主体、关系类型和层次结构三个方面，进一步总结馆藏资源数据关联的基本结构。

4.1 数据关联的主体

数据关联的主体指馆藏资源知识网络中的各种知识节点。对于文献资源而言，知识关联网络的构建可通过外部特征或内容特征的关联实现，也可通过外部特征与内容特征的交叉关联实现。知识节点可以是文献资源本身，也可以是代表文献外部特征和内容特征的各种特征项。根据知识节点的不同，可将知识网络分为文献网络、学者网络、机构网络、期刊网络、关键词网络和交叉综合网络等类型。随着馆藏资源数据化加工与建设的全面推进，文献资源的外部特征和内容特征逐步融合，形成各种知识单元，主要包括作者、单位、参考文献、事件、生物、矿物、产品、设备、公式、算法、定理、概念术语、图表等。文

献资源作为实际馆藏，将被赋予 URI 参引，成为各种知识节点的来源项，在知识网络中被关联，因此在馆藏资源数据整体关联网络中，知识节点类型十分丰富。从知识单元的颗粒度来看，知识节点既包含细粒度的元知识项，又包含粗粒度的文献资源；从传统文献资源角度看，知识节点既包含文献的外部特征项，又包含文献的内容特征项。

4.2　关联关系的类型

无论馆藏资源知识网络中的节点类型是粗粒度的文献资源，还是细粒度的知识单元，知识节点之间的关联关系可归纳为以下 3 种类型：

(1)引文关联，指以文献之间引证与被引证的关系构建知识节点之间的关联。引文关联将文献之间的引证关系转化为文献主题内容的相互关联，以构建资源层面的关联。引文的本质是进行科学对话[25]，从而促使知识发生流动与关联。引文是学术文献特有的结构，虽然引文动机十分复杂，包含观点佐证，方法识别，提供背景资料，证明事实数据，表达敬意、赞赏，争论或否认他人观点等 10 余种[26]。从知识组织角度而言，这些引用动机都可看作是被引文献与施引文献在知识结构上存在的关联关系。由于引文的复杂性，认为所有引文同等重要，进而构建引文网络进行学术评价的做法，近年来一直饱受学术界诟病。然而对知识组织而言，引文关联无疑是建立知识关联的重要路径之一。若能借鉴已有引文动机的研究成果，将引文动机转变为语义关系，将对知识组织产生更大影响。

(2)共现关联，指利用文献中相同或不同类型知识单元的共同出现构建知识节点之间的关联。一般来说，共同出现的知识单元间一定存在某种关联。知识共现主要包含 2 种形式：不同知识单元在同一篇文献中出现和同一知识单元在不同文献中出现。如果知识单元出现在建立了引证关系的不同文献中，那么可建立引文关联。因此，引文关联可视为共现关联的一种特例，共现关联比引文关联建立的知识链接更为广泛。共现关联可体现在资源层面、数据层面，以及资源与数据层面的关联关系中。目前对共现关联的研究大多集中在知识单元二重共现上，将来可将共现扩展到三重或者更多，从多维度挖掘和揭示关联关系，并把知识关联的程度用共现频次进行测度。与引文关联一样，共现关联是建立知识关联的重要路径之一，但如何将共现关系转变为语义关联仍有待进一步研究。

(3)语义关联，指通过知识单元之间各种复杂的语义关系构建知识节点之间的关联，包括同义、反义、属种、实例、属性、空间、等价。知识单元之间的语义关联是普遍存在的，除了词表、领域本体等所包含的语义关系外，还有大量的语义关联隐含在各种文献资源中，有待挖掘、整理、组织与利用。语义关联是知识与知识之间最本质的关系，共现关联是语义关联的客观呈现，引文关联则是语义关联的人为主观选择。共现和引文是计算机自动构建语义关联的基础，在此基础上可借助语法、语用分析，或者模式识别、引文动机识别等方法，判断具体的语义关系。语义关联主要是数据层面的知识关联，资源层面的知识关联通常表现为引文或同现关联，而资源与数据间的关联通常也表现为同现关联，资源与数据层语义关联的构建有赖于学术元数据框架的实施与应用。

4.3　数据关联的层次结构

知识组织领域早期重在研究资源层面的知识关联，是一种简单的资源聚合；如今研究逐渐深入到文献资源内部，出现了数据层面的知识关联。知识单元间通过多重、多维、动态的关联关系交织在一起，形成庞大的网状馆藏资源知识关联体系，是一种深层次的资源聚合。错综复杂的馆藏资源数据关联网络可划分为 3 个层次：资源层、数据层和中间层。资源层体现了资源与资源之间的关联，主要以同现、引文为基础建立各种关联；数据层体现了知识单元之间的关联，主要以语义为基础建立各种关联；中间层体现了资源与数据的关联，主要以来源同现为基础，附加学术元数据框架建立各种语义关联。这 3 层关联关系的建立将极大增强馆藏资源数据关联网络的整体连通性，无论是从资源出发，还是从知识单元出来，都能迅速实现关联并发现用户所需知识或

文献资源。

5 结语

戴维·温伯格(David Weinberger)的“新数字秩序理论”包含了深邃的知识组织思想[27]，为馆藏资源的数据化组织提供了理论指导。温伯格意识到数字环境中实体秩序和传统理性秩序存在着巨大缺陷，从而提出希望建立个性化、多维、网状结构的知识秩序，即数字秩序。在数字秩序中，一切都是零碎、自然和无序的。这与本文所讨论的馆藏资源数据化组织理念一致，即将各种载体形式的馆藏资源解构为万维网上可标识、相互连通的“数据”，尽可能为同一资源实体挂满各种各样的数据标签，从而建立馆藏资源的数字秩序。其实，数字秩序的无序并不是真正的无序，只是知识之间存在着复杂、多维的语义关系，从而导致数字秩序处于一种凌乱的状态。

参考文献

[1] OCLC. Perceptions of Libraries，2010：Context and Community [R/OL].[2015- 06- 07]. http：//www.oclc.org/reports/2010perceptions/2010perceptions_all.pdf.

[2] 常娥. 图书馆书目数据组织模式发展研究[J]. 图书馆论坛，2014 (12)：14- 19.

[3] 白海燕，乔晓东. 基于本体和关联数据的书目组织语义化研究[J]. 现代图书馆情报技术，2010 (9)：18- 27.

[4] 欧石燕，胡珊，张帅. 本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J]. 图书情报工作，2014 (9)：5- 13.

[5] 刘炜，胡小菁，钱国富，等. RDA与关联数据[J]. 中国图书馆学报，2012 (5)：34- 42.

[6] 毕强，牟冬梅，王丽伟，等. 数字资源语义互联研究：体系结构设计[J]. 现代图书馆情报技术，2010 (9)：3- 7.

[7][15] 邱均平，文庭孝，宋艳辉，等. 知识计量学[M]. 北京：科学出版社，2014：77.

[8] 黄巍巍. 论高校数字图书馆建设[J]. 南华大学学报，2001 (3)：83- 84.

[9] 韩晗. “数据化”的社会与“大数据”的未来[J]. 中国图书评论，2014 (6)：26- 32.

[10] 大村泉. 马克思恩格斯文献在日本的典藏与数据化[J]. 谢海静，范大祺，译. 马克思主义与现实，2012 (5)：26- 29.

[11] 徐绪堪，房道伟，蒋勋，等. 知识组织中知识粒度化表示和规范化研究[J]. 图书情报知识，2014 (3)：101- 106，90.

[12] 刘平峰，余文艳，游怀杰. 基于模糊等价关系的文本多级粒度划分方法[J]. 情报学报，2012，31 (3)：589- 594.

[13] 李秀红. 粗糙集概念与运算的知识粒度表示[J]. 计算机工程与应用，2011，47 (11)：34- 36，45.

[14] Yao Y，Zhao L. A measurement theory view on the granularity of partitions [J]. Information Sciences，2012，213：1- 13.

[16] 范炜. 走向开放关联的图书馆数据[J]. 图书情报知识，2012 (5)：94- 102.

[17] Faniel I M. Putting research data into context：A scholarly approach to curating data for reuse [EB/OL]. [2015- 06- 07]. http：//www.oclc.org/content/dam/research/presentations/faniel/faniel- asist- 2014.pptx.

[18] 尤金·加菲尔德. 引文索引法的理论及应用[M]. 北京：北京图书馆出版社，2004：83.

[19] 邱均平，董克. 作者共现网络的科学研究结构揭示能力比较研究[J]. 中国图书馆学报，2014 (5)：15- 24.

[20] 王知津，郑悦萍. 信息组织中的语义关系概念及类型[J]. 图书馆工作与研究，2013 (11)：13- 19.

[21] 文庭孝，龚蛟腾，张蕊，等. 知识关联：内涵、特征与类型[J]. 图书馆，2011 (5)：32- 35.

[22] Liu，X. Z.，Guo，C，Zhang，L. Scholar metadata and knowledge generation with human artificial intelligence [J]. Journal of the Association for Information Science and Technology，2014 (6)：1187- 1201.

[23] 唐义，肖希明. 关联科学：一种全新的科研支撑方式[J]. 图书馆杂志，2013 (8)：4- 11.

[24] 化柏林，武夷山. 序化转化，双管齐下[J]. 情报学报，2012，31 (1)：卷首语.

[25] 叶继元. 引文的本质及其学术评价功能辨析[J]. 中国图书馆学报，2010 (5)：35- 39.

[26] 刘茜，王健，王剑，等. 引文位置时序变化研究及其认知解释[J]. 情报杂志，2013 (6)：166- 169，184.

[27] 文庭孝，刘璇. 戴维·温伯格的“新秩序理论”及对知识组织的启示[J]. 图书馆，2013 (5)：5- 7，11.

作者简介 常娥，女，博士，东南大学图书馆副研究馆员；孟祥保，男，硕士，东南大学图书馆馆员。

收稿日期 2015- 06- 30