# 机构名称规范数据的语义模型构建*

□曾建勋　贾君枝

**摘要**　机构名称数据是科研成果数据库、会议论文数据库、企业业务数据库中必不可少的构成。针对当前机构档规模小、质量低、使用范围有限等缺点，为构建一个有机的机构实体关联网，对机构名称数据呈现的特点进行了分析，从用户需求角度明确机构名称实体对象，设计了机构名称的属性及其机构实体间的关系，在此基础上引入 Schema 词汇表对其进行语义描述，从而确立了机构名称的数据模型。

**关键词**　机构名称　规范数据　语义模型
**分类号**　G254
**DOI**　10.16603/j.issn1002－1027.2019.01.006

## 1　导言

语义网技术的发展旨在使网络资源更好被机器所理解与交换，最大程度地实现资源开放、共享与互联，以满足用户多层面的应用需求。以 RDF、OWL、SKOS、SPARQL 语言、关联数据等核心技术的发展进一步提供了数据描述、使用词汇推理的应用环境，确保网络上存在着大量可用的标准格式的数据，这些数据集不仅容易获取，而且数据集之间建立关联，既可以被语义网工具所管理及获取，又为实现数据的跨组织及跨平台的集成应用提供了可能。机构名称是指政党组织、政府机关、军队警察、学术机构、群众团体、宗教团体、企事业单位等群体及其隶属部门的名称。机构名称数据是科研成果数据库、会议论文数据库、企业业务数据库中必不可少的构成。但由于机构名称数量庞大，名称变化复杂，名称形式多样，这样为机构名称的识别及其聚类带来极大的困难。为进一步提高名称识别效率，有效地区分同名、异名现象，提高用户检索效率，准确详细地获取机构名称的相关信息，需要对机构名称进行进一步规范化处理。因此机构名称规范档应运而生。机构名称规范档是机构名称规范记录的集合，它根据一定的规范控制规则将名称相关信息按照统一的标目形式展示，达到规范控制的目的。

从当前国内外规范文档建设看，呈现出机构规范档规模小、质量低、使用范围有限等特点，很难有效发挥其机构名称识别及数据聚类作用。如何借助于语义网技术，在现有规范库的基础上，运用现有丰富的本体词汇及其网络资源，采用实体描述方法，将机构名称作为实体对待，尽可能将不同属性特征的机构名称进行详细描述，并建立各个名称实体之间的关系，形成一个有机的机构实体关联网，采用 RDF 格式来表示其属性及关系，并运用命名空间方式定义。这样一方面是将外部资源作为名称规范档的来源库，以丰富其描述内容；另一方面开放现有资源库，建立与外部资源的连接，提高其影响力及应用范围。从而既符合语义网发展趋势，也为数据搜索、语义理解等应用做出重要贡献。

基于此，本文将对机构名称数据呈现的特点进行了分析，从用户需求角度明确机构名称实体对象，设计了机构名称的属性及其机构实体间的关系，在此基础上引入 Schema 词汇表对其进行语义描述，从而确立了机构名称的数据模型。

## 2　机构名称数据呈现的特点

### 2.1　机构名称数据数量大、价值高

机构是企业、事业单位、机关、社会团体及其他

* 国家社科基金项目“机构规范文档结构及构建方式研究”(15BTQ015)和国家社会科学基金重点项目“基于关联数据的中文名称规范档语义描述及数据聚合研究”(15ATQ004)的研究成果。
通讯作者：曾建勋，ORCID：0000-0002-0432-9618，邮箱：zeng@istic.ac.cn。

依法成立的单位通称，新机构的不断涌现，传统机构的淘汰、更名、重组与合并，累积了大量的机构数据，既存在于专门的机构库中，如各类型的机构规范档、国际虚拟规范文档（VIAF）、国际标准名称识别符（ISNI）库等；又出现于涉及机构数据的成果库、业务库、网络百科全书等。VIAF 所收集的来源数据中，机构名称有 510 万条，合并生成的 VIAF 记录有 3770650 条[1]。国际标准组织认可的用以识别创造性作品贡献者的全球标准号 ISNI，目前包含机构数据 565282 个[2]，每个机构都有一个唯一识别符。Wikidata 数据中机构名称涉及 846626 个（统计其 Organization 子类下所有实例数所得）[3]。这些丰富的结构化数据为机构识别及评价起到了重要作用。建立这些机构库之间的关联，充分利用现有成果，构建一个机构数据网，将有效地提高数据应用价值。

### 2.2 名称形式多样化

机构名称是对现实存在的实体的符号化表示。尽管存在的实体具有唯一性，但由于语言的丰富性及其时空的变迁，有多种多样的形式与实体对应。既有正式名称，又存在多个变异名称形式，如曾用名、全称与简称、译名等并列名称等多种变异形式。而这些不同的名称形式之间如果不建立同一关系，则会导致输入不同的名称查询形式，得到的结果不一样，不能全面完整地获得一个机构实体所有信息，从而影响到机构名称数据的应用，如实现机构名称聚类检索及准确地科研评价。科研评价包括知识点评价和科研产出评价。评价中，由于机构重名、不同语种之间机构名称缺乏对应、机构变迁等问题存在，围绕机构层面的评价如研究群体、科研布局及影响力、科研合作等内容开展的研究，由于缺乏各类型机构名称之间的语义关系定义，仅仅依赖于名称形式上匹配，会出现漏统计、错误统计等问题，从而影响到科研评价质量。

### 2.3 机构名称标识符的不统一

机构名称数据出现在期刊、论文、专利等科研成果数据库或机构名录数据库，由于数据库结构及其语种差异，随着数据库集成应用与发展，出现了跨库、跨语言、跨领域的操作，指代同一机构的不同数据库之间存在着数据整合的需求，这很大程度上需要通过构建唯一的机构名称标识符建立数据间的关联。但当前各大数据库使用的名称标识符并不统一，有表示 MARC 数据的规范记录号、国际标准名称识别符（INSI）、Ringgold 号、全国组织机构代码管理中心给定的机构代码等，而且各标识符之间并未建立有效的关联，从而为数据整合造成了困难。如果明确给定每一个机构名称实体的 URL 地址，运用命名空间方式定义，同时将机构名称实体的 URL 与其他名称标识符之间建立外部等同关系，这样为数据之间的引用及其整合重用提供了便利。如汤森路透已与 Ringgold 完成超过 40 万条机构记录的无缝链接，实现机构名称的快速消歧和机构层级关系的准确关联，以便于维护与提高其数据质量[4]。

### 2.4 中文机构英译名称准确规范性有待提高

中文机构英译名称是中文名称的其他语种表达方式，属于并列名称。其目前在中英文数据库都有描述，尤其在外文数据库中成为主要的描述方式。科研成果检索中，通常会涉及中英文数据库的跨库集成检索，由于作者机构英译名称存在表达不规范、拼写错误、翻译不准确等问题，再加上中文机构名称与英文机构名称并未建立对应关系，很容易造成漏检、误检。从 Web of Science 数据库中下载高引用量前 1000 条中国作者发表论文数据[5]，由于有多个合作者及同属一个机构的数据，去重后共涉及 1589 个机构名称，结果发现机构名称全部采用名称缩写形式，如 Peking Univ（北京大学），其中 50 个机构名称存在名称拼写形式不一致（如前后次序颠倒、字母丢失、翻译不一致等），144 个机构名称详略度描述不统一，存在指代不明确问题，如“Tsinghua Univ”“Tsinghua Univ，Ctr High Energy Phys”两个机构名称之间有从属关系，从而为机构名称的识别及消歧增加了困难。

## 3 机构名称规范数据的语义模型构建

如何明确清晰地表达机构名称规范数据，构建各类机构名称数据之间关系，是解决名称多样化、来源多样性、数据孤立所带来问题的关键，也是保证机构名称规范数据在检索与评价应用质量的关键。语义模型的构建旨在明确用户所关注的对象，并从实体分析角度定义实体间的关系、实体对象所具有的属性，为数据的描述做充分准备。

### 3.1 明确机构实体对象

国际图书馆协会和机构联合会（IFLA）研究组于 2008 年提出及推动发展的规范数据功能需求

(FRAD)，基于用户需求任务，明确将机构作为实体对象对待[6]。事实上，将客观存在的机构作为实体，并用 URI 进行命名，既确保实体的唯一性、可获取性，又有利于将其携带的各类型数据集中呈现。URI 不仅是名称，也是资源获取方式。为实体建立 URI 时，首先要定义 URI 的命名规则，好的 URI 能够更加直观地识别出该实体并方便与其他实体连接。一般 URI 的命名规则是＜基地址＞/＜实体类型名称＞/＜实体 ID＞，基地址可根据服务器名或机构名来定，如中信所国家工程技术数字图书馆的机构名称规范数据，确定基地址为 http://www.istic.ac.cn/，放在 Authority 目录下，实体类型为 Organization，因此 URI 为 http://www.istic.ac.cn/authority/Organization/ID。访问该实体时，根据 303 重定向机制，服务器会定位到与之相关的信息资源中，如 http://www.istic.ac.cn/doc/ID，返回结果时，服务器根据内容协商机制选择返回格式，如 http://www.istic.ac.cn/doc/ID.htmL 或 http://www.istic.ac.cn/doc/ID.rdf[7]。

### 3.2 机构实体的属性选取

机构实体属性揭示了机构的各种特征，属性的获取及其选择源于用户的需求。对于机构查询者言，需要了解机构的名称、地理位置、功能、创建时间及其所属领域等相关信息；对于统计及评价用户言，机构名称的识别及归一是关注核心。当前各类型的机构数据所定义的属性如表 1，其一定程度上为机构实体的属性选取提供了便利。名称规范档对机构名称的各种形式进行了定义，维基数据着重于机构属性全面详细地展示，规范数据功能需求从用户角度定义了规范名称应具有的属性，这些数据可以作为借鉴。

**表 1 机构实体的属性列表**

| 名称规范档 | 维基数据 | 规范数据功能需求 |
|---|---|---|
| 名称规范形式 | 机构名称标识符 | 名称类型 |
| 名称变异形式(不同语言、不同拼写、缩写、全称) | 机构类型 | 名称字符串 |
| | 地理位置 | 相关联的地点 |
| | 网站 | 相关联的日期 |
| | 邮编 | 语言 |
| | | 地点 |
| | | 活动领域 |
| | | 历史 |

综合考量各类用户，充分利用现有机构数据库数据，以确保所表达的属性尽可能满足用户所需。我们选取机构实体的九个属性：机构名称标识符、机构规范名称、机构变异名称、时间、地点、类型、活动领域、评论、描述。每种属性下都包含若干子属性。

机构名称标识符需要定义一套字符串体系以唯一标识机构实体，通过定义机构名称的唯一识别符实现机构识别；机构规范名称是大众所知的、首选或惯用的名称，尽可能选自名称规范档、INSI、VIAF 数据源、全国组织机构管理中心注册时登记的名称。机构并列名称是指不同语言形式的名称。机构变异名称是指表示同一机构名称的不同形式，包含全称、简称、曾用名称、其他名称等，其他名称主要指写法(异体字、音译、大小写、标点符号)不同或者词序(倒序、正序)不同的名称。类型是对机构实体进行按类组织，根据功能、性质或其他属性对机构进行的不同层次划分，明确各机构名称所属的范围。需构建范畴树，以此作为机构分类的依据。机构类型的选择，尽可能遵循标准化词表，以确保该数据的共享性及可重用性。活动领域是指组织机构从事的业务领域，包含组织的能力、责任、职能范围等；图像是组织机构的图像，通常与组织机构图像 URL 地址建立链接；评论是对组织机构所做出的各层面的评价与评论，以作为用户查看机构信息的依据；描述介绍机构实体的基本信息，包括对机构规模(如职工数量)、发展历史等方面的描述；时间是与机构实体生命周期中的事件相关的时间，包括机构创建、撤销、机构名称变更时间；地点是指机构实体所处的地理位置及其虚拟位置，用地名、邮编、网址表示。如果与相关 GPS 定位系统相连，可以显示机构实体的经纬度。包括街道地址、邮编、电话号码、电子邮件及与团体有关的网站地址。全国组织机构管理中心的组织机构代码共享平台提供了地址、电话号码、邮编等详细信息，可以作为中文机构构建的参考数据源。

### 3.3 机构实体关系定义

机构实体关系包含机构实体之间的关系及机构实体与其他名称实体间的关系。实体关系越丰富，所形成的机构名称数据网络价值越高。

#### 3.3.1 机构名称实体之间关系

机构名称实体之间存在着同一、隶属、相继、相关等关系。

(1)同一关系。

同一关系是使用不同的指代形式表示同一实

体，包括不同名称之间及其名称标识符之间的关系。机构名称可分为规范名称、译名、变异名称等多种名称形式。规范名称是指某一机构各个名称中的首选名称，可选用全国组织机构管理中心注册登记的机构名称；变异名称是与规范名称表示同一机构实体的除规范名称、译名外的其他机构名称形式，包括并列名称、简称、曾用名等，并列名称是指同一机构实体应对不同公开身份所采用的其他名称，曾用名是该机构在变更前所使用的规范名称。如“北京大学”与“Peking University”为同一关系；“北京大学”与“北大”也是同一关系。表示同一机构名称标识符与其他标准标识符如 VIAF ID、ISNI、LC Authority ID 等属于同一关系。

(2)等级关系。

实为机构名称实体之间的从属关系，包括指主体机构与其下属机构间产生的行政隶属关系，也可指代不同机构类型之间的层级隶属关系。以“北京大学”为例，北京大学与其下属的数学科学学院、图书馆、北京大学附属中学、北京大学深圳研究生院皆为等级关系。机构名称类型学校包含小学、中学、大学等，它们之间属于此类关系。

(3)相继关系。

按照顺序或者时间依次产生的两个以上团队之间的关系，比如机构之间的合并或者分解，如 1862 年洋务运动期间的京师同文馆并入京师大学堂，则两个机构之间是相继关系。具体可分为两种关系，一种是前身关系，一种是后继关系，如京师大学堂是北京大学的前身关系，北京大学是京师大学堂的后继关系。

(4)相关关系。

由于机构间的合作，或者共同从事某一活动而产生的关联，比如高校之间的项目合作等。由于北京大学参与了办学，北京大学与北大培文九华实验中学为相关关系。

3.3.2 机构名称与其他名称实体间的关系

包含机构名称与个人名称、机构名称与事件名、机构名称与会议名之间。

(1)成员关系。

个人以不同形式在不同时期表现为某一机构内部的成员，包含创建者、员工、会员等类型。通常如“蔡元培”与“北京大学”。

(2)事件关联。

机构与某一事件的关联关系。机构内部发生的事件、关联的事件。“五四运动”与北京大学。

(3)会议关联。

机构与某一会议的关联关系。包括主办、组织及其其他关系。如北京大学主办了“连续制造研讨会”。

3.4 描述模型确立

依据上述定义的属性及其关系，每一实体对应为类，明确所引用的类，构建类与类之间的关系，这样以确保机构名称实体的描述基本遵循此模型。

为了提高重用性及互操作性，类名及属性名称的定义尽可能选用已有的成熟词汇集。目前 SKOS、FOAF、Dublin Core、RDA 等都曾被用来描述名称规范数据。由 VIAF 数据模型的演进情况来看，VIAF 将对规范记录的语义描述侧重点从描述概念、名称逐步转到对实体本身的描述。VIAF 起初几乎全部使用 SKOS，之后引入 FOAF 来增加事物描述的逼真性，后加入 RDA 作为 RDF 模型的一部分。2011 年，谷歌、Bing 和雅虎共同发起一个新项目Schema.org[8]，它是以微数据(Microdata)为基础的通用标记词汇集，除了一些原始数据类型如数字、文本外，Schema.org 包含了很多新的标记类型，包括创造性工作(Creative Work)、事件(Event)、组织(Organization)、人物(Person)、地点(Place)、产品(Product)、评论(Review)等。Schema.org 所包含的事物类型中，这些类型以一定的层次结构组合起来，每一类都有自己的属性，子类继承父类的属性。2014 年 VIAF 参考 Wikidata 的做法，主要使用 Schema.org 作为其核心 RDF 词表，描述机构时，以 Schema:Organization 作为实体的唯一类型。

结合所定义的机构名称属性和关系，主要采用 schema 词汇表表示，如果 Schema 词汇表不足以进行描述时，则采用 RDA 词表进行补充，如表 2 所示，对属性、关系的名称、说明、RDF 映射及其取值类型进行列表描述。

可以看出，所选用 Schema.org 是描述机构名称属性的核心词汇表。RDA 定义的机构与机构之间关系属性对其进行了有益的补充。现对其属性描述模型和关系描述模型进行描述，如图 1、图 2 所示。

**表 2　机构相关的属性、关系及 RDF 映射**

| | 名称 | | 说明 | RDF 映射 | 取值类型 |
|---|---|---|---|---|---|
| 属性 | 机构规范名称 | | 常用或者正式名称 | schema:name | Text |
| | 其他语种规范名称 | | 并列名称 | schema:name,标明语种属性 | Text |
| | 其他名称 | | 机构的其他名称 | schema:additional name | Text |
| | 类型 | | 机构所属的范畴 | schema:genre | Text |
| | 活动领域 | | 组织活动的专业领域 | RDA:field of activity of the corporate body | Text |
| | 图像 | | 组织的图像 | schema:image | url |
| | 评论 | | 与组织相关的评论信息 | schema:review | Text |
| | 描述 | | 描述机构的其他信息 | schema:description | Text |
| | 时间 | 创建日期 | 机构创建时间 | schema:founding date | Date |
| | | 撤销日期 | 机构撤销时间 | schema:death date | Date |
| | | 开始日期 | 机构活动的开始时间 | schema:start date | Date |
| | | 结束日期 | 机构活动的结束时间 | schema:end date | Date |
| | 地点 | 电子邮件 | 组织联络的电子邮件地址 | schema:email | Text |
| | | 地点 | 组织的位置或邮编,组织相关的事件发生地 | schema:location | Place Postal Address |
| | | 电话 | 组织的联络电话 | schema:telephone | Text |
| | | 网址 | 组织主页 url 地址 | schema:url | url |
| | 机构与个人 | | 雇员 | schema:employee | Person |
| | | | 成员 | schema:member | Person |
| | | | 创建者 | schema:founder | Person |
| | 机构与事件 | | 与该机构相关的事件 | schema:event | Event |
| | 机构与会议 | | 机构组织的相关会议 | RDA:conference | Event |
| | 机构与机构 | | 等级关系 | subclass of | Organization |
| | | | 相继关系 | RDA:preceded by | Organization |
| | | | | RDA:succeeded by | Organization |
| | | | 相关关系 | RDA:associated institution | Organization |

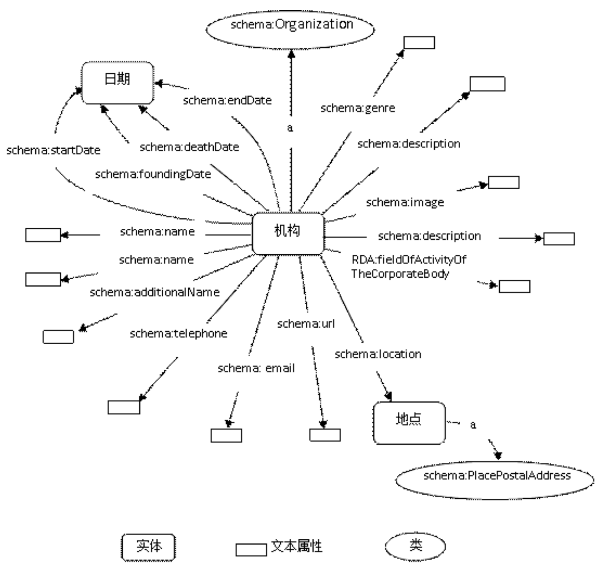


图 1　机构名称规范数据实体属性模型

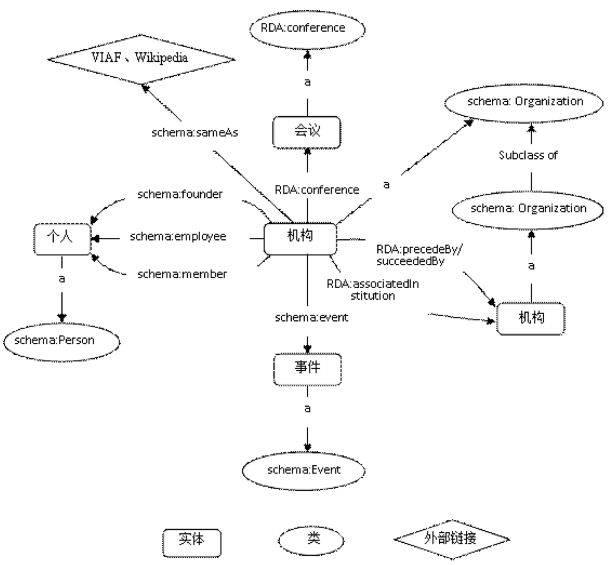


图 2　机构名称规范数据实体关系模型

## 4 结论

机构名称规范数据模型的构建是保证机构名称描述准确及其广泛应用的重要基础。借助现有丰富的本体词汇及其网络资源，采用实体描述方法，所形成的机构实体关联网将在很大程度上满足各类用户的需求，并有助于机构数据与其他资源建立联系，成为语义网的重要构成，有助于推动语义机器理解、问答系统等应用发展。随着机构名称数据数量增大，如何统一机构名称标识符，确定机构名称标准化描述内容及其方式将是我们未来关注的研究问题。

## 参考文献

1 OCLC developer network. VIAF (The Virtual International Authority File)[EB/OL]. [2016－06－24]. http://www.oclc.org/developer/develop/web-services/viaf.en.html.
2 ISNI. How ISNI works [EB/OL]. [2016－08－15]. http://www.isni.org/how-isni-works.
3 Wikimedia Foundation. Wikidata[EB/OL]. [2016－10－10]. https://www.wikidata.org/wiki/.
4 贤信，曾建勋. 科研实体唯一标识系统研究[J]. 图书情报工作，2015(12)：113－119.
5 Web of Science [EB/OL]. [2016－10－01]. http://apps.webofknowledge.com.
6 Functional Requirements for Authority Data. IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR) [EB/OL]. [2016－09－01]. http://www.ifla.org/publications/functional-requirements-for-authority-data.
7 贾君枝，石燕青. 中文个人名称规范文档的关联数据化研究[J]. 情报学报，2016，35(7)：696－703.
8 Schema 的组织结构[EB/OL]. [2016－07－11]. http://schema.org.cn/docs/schemas.html.

作者单位：曾建勋，中国科技信息研究所信息资源中心，北京，100038
贾君枝，中国人民大学信息资源管理学院，北京，100872

收稿日期：2017 年 8 月 14 日

# The Construction of Semantic Model of Organization Name Data

Zeng Jianxun　Jia Junzhi

**Abstract**：Organization name data are necessary components of database of research outputs, conference papers and business operation data. In view of the shortcomings of the current institutional file such as small scale, low quality and limited scope of use and in order to develop an institutional entity network, this paper analyzes the characteristics of the organization name data, defines the organization name entity from the perspective of user requirement and designs the properties and relation between entities. Then it establishes the data model by describing semantically in schema.org.

**Keywords**：Organization Name, Authority Data, Semantic Model

# 封面照片简介：浙江越秀外国语学院镜湖校区图书馆

浙江越秀外国语学院镜湖校区图书馆于 2014 年初动工兴建，2017 年 3 月投入使用，占地面积 4649.58 平方米，建筑总面积 18915.86 平方米，共五层，地下一层，地上四层。

图书馆建设立足于百年经典，力求成为传世之作。整体造型结构体现“厚重、大气、简洁、典雅”的建筑风格，建筑设计呈现中西合璧、外方内圆。在内部空间处理上，通过过厅、楼梯等设计手段和空间形态、色彩、光影的使用，使整个图书馆形成流畅的、有节奏的、虚实相结合的大气而又有趣味的整体空间。外墙采用清水混凝土装饰幕墙体系，与校园整体建筑风格相得益彰。

除拥有报刊及中外文图书借阅区外，图书馆还设有新书展示区、自助服务区、视听阅览室、教师研修室、教师指导室、文化展览空间、学术交流空间等多个特色功能性区域，将图书借阅功能、教学研究功能和休闲娱乐功能融为一体，体现图书馆人性化的服务理念。图书馆采用藏、借、阅、咨一体化开放式管理模式，且层层相通、室室相连，真正实现“全面开放，藏阅合一”的服务效果。同时，馆内舒适、便捷的休闲阅览桌椅，自助、高效的信息化设备，高速、稳定的 360°覆盖全馆的网络，以及可实现三种语言两路同声传译效果的多功能国际化报告厅，都充分体现图书馆“高定位、高配置、高标准”的特色。

图书馆独特的建筑设计、丰富的馆藏资源、先进的管理模式和高起点的服务层次和水平，不仅能为全面提高学校教学科研和学科建设水平提供有力保障，也是广大师生追求真理、探索新知的知识殿堂。