国家科学图书馆青年人才领域前沿项目

专业领域知识组织模式研究
研究成果

项目名称：专业领域知识组织模式研究

成果类型：研究报告

成果版本：定稿

提交时间：2010-7-11

撰写人：刘峥、翟爽、鲁宁、景丽

2010 年 7 月

目录

1 引言

1.1 研究背景

随着网络信息技术的发展，数字化资源成为了学术信息中的主流，数字化的出版物是学术交流的主要方式，在科研教育过程中产生的各种原生资源也在逐步成为学术交流中重要介质。科学研究模式呈现出数字驱动、分布、合作、跨领域研究的特点[1]。因而科研人员对信息环境和信息服务提出了新的要求，所需的信息资源不仅包括文献信息资源，还包括综合科技信息资源，如科学家信息、科研项目信息、实验数据、科学仪器设备等综合的科技知识资源；不仅包括文献型资源，还包括视频资源、科学数据等多格式多类型资源。所需的信息组织与服务也不再是一个游离于科学研究之外的独立过程，而是参与到科学研究过程中，科研人员既利用数字信息资源，同时也在不断地生产新的数字信息资源。在这样的背景下，如何对特定的专业领域内数量巨大的各种类型的综合性信息资源进行描述、组织、集成和建设，成为了亟待解决的问题。

欧美等发达国家一向注重数字化及信息基础设施建设，在专业领域的知识组织的研究和建设上也启动了一批先导性的项目，以探讨如何进行专业领域的知识资源组织。比如欧盟 Renardus 项目以 DDC 分类法为基础，通过各种知识组织体系与 DC 资源类型的映射，建立欧洲学术主题门户信息集成访问服务；美国国会图书馆和 OCLC 分别以国会图书馆分类法和杜威分类法为基础，试验对信息资源进行自动的学科主题分类，以支持对大规模信息对象的主题搜索和按学科主题组织发现机制；美国康奈尔大学构建了虚拟生命科学图书馆系统，面向基因组学计划和新生命科学计划，集成涉及生命科学的科学数据、实验测试平台、研究项目、人物、数据库、电子期刊等等信息，提供集成的领域知识发现服务；华盛顿大学的健康科学图书馆（HealthLinks）提供与健康相关的信息，包括医学研究、医疗和护理等信息的集成服务。

而国内在知识组织的研究上相对薄弱，主要是一些探索性的理论研究和对传统知识组织方法的改造研究，如尝试用 SKOS 语法对中国分类主题词表进行描

[1] Julia Gelfand. The change collections for eScience: what it means for libraries. Workshop on the digital collection development and sharing, 2009

述；讨论采用元数据方法进行网络信息资源组织，将知识组织分成语义、逻辑等几个层次。

1.2 研究范畴和内容

项目中所指的综合科技信息资源，主要是根据 ARL 和中科院国家科学图书馆最新调研来确定。根据 ARL 的《当前数字学术交流的模式》调研显示，数字化学术信息资源主要包括电子期刊、评论、预印本和工作报告、百科全书、字典等注释性内容、数据、blog、讨论组、专业和学术性论坛[2]。根据中国科学院科研人员需求调查和分析，在新的科学研究形态和新的开放信息环境下，中国科学院科研人员在科学研究过程中所需的资源，除了正式出版的文献外，还包括了开放获取学术资源、机构自主知识资产、科学教育资源，以及支持科研成果转化和应用所需的社会、经济、技术、规范与政策信息等。

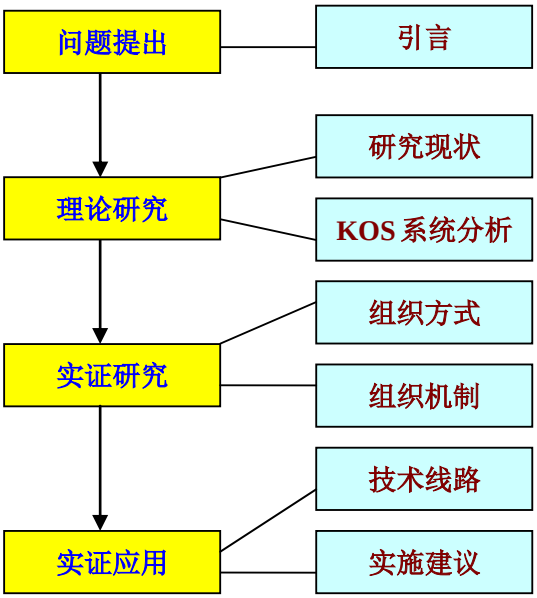
这些新形式的资源以不同的形态散布在网络当中，有的经过了信息加工，有着规范的元数据描述；有的以数据库、网站的形式存在；有的以原生态的形式，没有经过任何信息处理。为了使科研人员能够高效地发现、获取专业领域内的各种所需的信息资源，必须对这些信息资源本身进行描述和组织，同时也必须能效地揭示资源之间的相关关系。因而专业领域内综合科技信息知识组织的关键点是如何通过知识组织模型将不同来源异构的信息资源，有效集成，并揭示资源之间的隐含关系。

本项研究的目的在于探讨专业领域内综合科技信息资源的知识组织模式建设方法和建设思路，包括所采用的知识组织体系架构、知识组织的内容范围、实施路线等。

项目研究将分成如下几部分展开：首先，项目以 Taxonomy warehouse、专业领域信息检索指南、DAML ontology library 等为线索，调研各个专业领域内已有的知识组织体系，分析专业领域知识组织体系的现状和特点，形成专业领域知识组织体系数据集；其次，项目将选取具有代表性的专业领域内综合科技信息服务的项目进行剖析，分析其知识组织的方法、采用的知识组织体系。最后，论文将根据中国科学院的现状，总结中国科学院国家科学图书馆在综合科技资源组织上

[2] Maron.N L, Smith K. K. Current Models of Digital Scholarly Communication [EB/OL]. [2009-08-01]. http://www.arl.org/bm~doc/current-models-report.pdf

可能采取的途径和方法。



1.3 研究意义和方法

专业领域中综合科技信息资源的组织，是图书馆面向 eScience 环境，实现以数据为中心的综合性资源服务的开端，它的研究能为图书馆在数字化环境下拓展新的服务内容和领域，也能够为实现综合科技资源关联计算、信息检测和分析打下基础，因而此项研究具有重要的现实意义。

专业领域知识组织研究，是综合科技信息资源建设的一个关键环节，涉及综合科技信息资源采集、利用、利用和管理各个方面，它的研究有助于构建专业领域的综合科技信息知识体系，深化信息资源管理理论的内容，因此具有理论意义。

除了采取文献调研的方法外，项目研究还采取了系统分析、案例分析的方法，来探索专业领域综合科技信息组织模式。

2 专业领域知识组织的现状和需求

2.1 专业领域综合科技信息知识组织的特点和需求

e-Science 的发展，推动了科研过程、科研活动中各项信息的数字化。科研人员在面向专业领域工作和学习时候，除了需要代表科研成果的文献信息资源外，还需要获取同领域科研人员信息、科研项目信息、实验数据、科学仪器设备等科研活动中涉及的各项要素信息。在此背景下，图书馆和信息服务机构为了满足用户的需求，开始尝试面向专业领域进行知识组织，以提供更好的信息服务。

与过往的信息组织不同，专业领域知识组织呈现出数据异构、格式多样、资源拥有者多元、面向应用具体的特点。从信息的类型上，既有期刊文献、会议文献，又有科学数据、开放课件、即时学术交流信息；从信息的格式上，不仅包括文本型信息资源，还包括视频声频、动态图像、图片等多格式的资源；从信息被加工的程度上，有的含有标准的元数据描述，有的仅是结构化的数据，也有的被词表、分类、叙词表所标引。

理想上，用户希望通过一个信息检索的入口获取到不同来源不同类型的信息，并将这些信息之间千丝万缕的关系从多角度揭示出来。因而对专业领域的综合科技信息资源的知识组织模式需要具备：

（1）能实现异构数据的交互。不同的数据源有不同的运行环境，有着相应的硬件设备、操作系统和网络协议等；所采用的数据模型也不同，有结构化数据（如数据库）、半结构化数据（如 HTML、XML）和非结构化（如文本、图片）。知识组织体系模型要对各种异构数据进行有效集成，给出用户一个统一透明的访问界面，支持各种异构数据的关联、显示。

（2）能容纳各类型带有不同知识组织方式的资源。专业领域知识组织的资源，即包括了图书馆已经能较好进行组织的期刊、会议论文、图书，涵盖了科研过程中不断产生、尚需要规范组织的科学数据，还有来自人事、教务、科研基金的信息，并不断有根据专业领域科研的需求，需要纳入的信息资源。这要求专业领域知识组织模式要很强的容纳性，即要容纳各类型的资源，也要尽可能尊重资源已带有知识组织方式、揭示资源的角度。

（3）能揭示多重的知识关系。由于专业领域资源众多、需求专精的特点，

专业领域知识组织不再满足于单一角度、单一途径的知识揭示，需要尽可能从多个分面、多个角度来展示资源，并能反映资源之间的丰富关系。

（4）能随着资源的发展，进行动态扩展。专业领域中的资源和资源类型是不断发展和丰富，要求知识组织的架构不能是一成不变的，要能随着资源的发展，不断扩展，以实现不断更新的资源与知识组织体系的映射。同时，也需要满足资源因应用场景的不同，资源所处在知识组织体系中位置和其他资源间的关系能够动态的扩展。

2.2 现有专业领域知识组织系统的现状

知识组织系统是任何用来定义并组织和表述真实世界物体的术语和符号系统，在具体应用中往往泛指为语义工具，如大型数据库中使用的叙词表、检索引擎内部使用的分类表和自动扩检表、网站导航浏览用的等级体系结构、语义网用的实用分类系统[3]。因而专业领域内综合科技信息资源组织模式研究，先从专业领域内的知识组织系统现状入手，从专业领域知识组织系统的类型种类和使用情况等角度进行分析，以便为构建专业领域知识组织模式奠定基础。

为了能充分地了解专用领域的知识组织系统的现状，项目组以 Taxonomy warehouse 为主体，对科学技术等与中科院相关的学科领域进行了调研和分析。通过整理和分析，共获得有效的知识组织系统 165 个（具体详见附表 1）。

（1）从学科领域的角度来看，各学科知识组织系统的建设处于不均衡的状态（详见图 1）。具有大量规范知识组织体系的学科主要生物领域、环境科学领域，其次是计算机、地理学、通信、化学，而在力学、仪器、机械等工程领域基本没有成型可供利用的知识组织系统。从这个统计可以看出，各学科在知识组织上处理的方法有很大的不同，体现了不同学科的发展特点。生物、环境、地理学等相关学科的研究内容可以根据研究对象的特点进行定义和划分，如生物学可以根据研究对象进一步分为动物学、植物学、微生物学、古生物学等；计算机、通信、电子学等属于新兴学科本身注重在计算机和网络的应用，在术语规范、计算机处理上进行了大量的努力，如计算机领域出现了大量分类体系，用于网络信息分类；而在一些经典学科领域，除了化学十分重视术语、主题的规范外，在相对

[3] 曾蕾. 从情报检索语言到网络环境下的知识组织系统和语义工具. 上海图书馆, 2007

较长的学科发展史上仅有少量的知识组织系统；而对于力学、仪器、机械等属于应用学科，涉及的内容关联性较大，更倾向于形成具体的产品和专利，在这些学科领域就很难形成通用的知识组织系统。

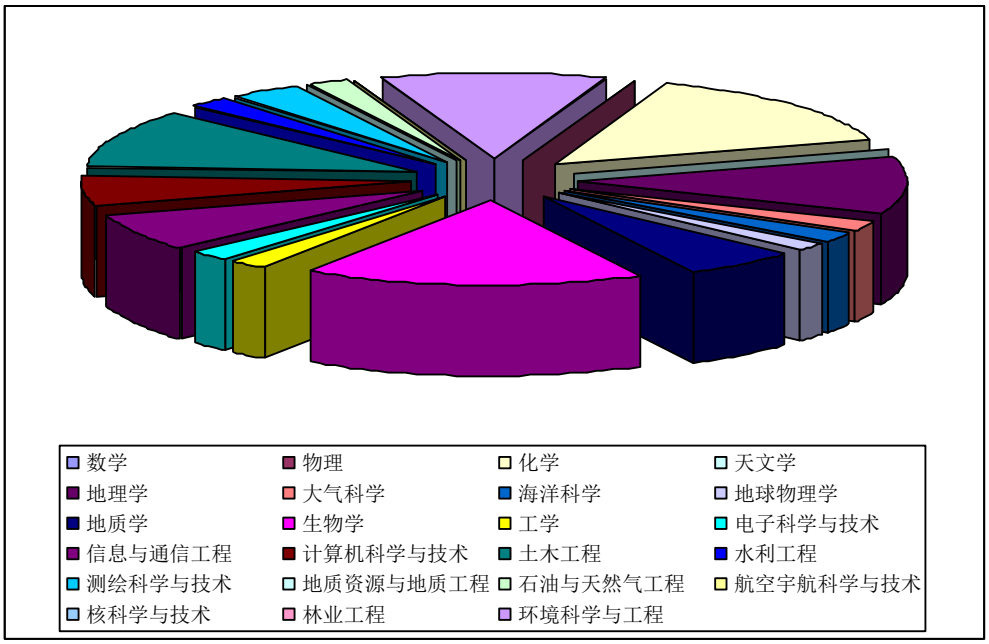


图 1 知识组织系统的学科分布图

（2）根据语言的规范性和结构化程度，Gail Hodge、Zeng 和 SaLaba 对各类知识组织系统进行了总结，他们认为知识组织系统大致可以分为三类：词单，一般对术语进行定义，为线性结构，如规范档、词典、字典、地名词表等；分类与大致归类，揭示术语间的等级关系，一般为树状结构，如分类法；关联组织，呈现网状结构，揭示整体-部分、因果等关系，如叙词表、本体[4][5]。

从语言的规范性和结构化程度对各学科的知识组织系统进行划分后分析发现，从总体来看在分析整理的 165 个知识组织系统中，数量最多的是分类表，其次是术语表和叙词表，仅有极少量的本体。

而对于具体的学科，除了计算机领域，分类体系占到了其学科中知识组织系统总数的 70%外，其他各学科出现频次较高的知识组织系统是术语表和叙词表（详见图 2）。各个学科基本上都对学科相关的标准术语进行规范和定义，以便文档的索引和检索，在具体应用上略有差异。在传统学科领域，如化学、地理学、地质学，术语词典占绝大多数，而在生物、环境领域研究内容能根据研究对象的

[4] Gail Hodge. Systems of knowledge organization for digital libraries, DLF, 2000
[5] Linda Hill, Olha Buchel, Greg Janee, 曾蕾. 在数字图书馆结构中融入知识组织系统. 现代图书情报技术, 2004 (1)

特点进行细分，叙词表的使用则相对较多；而对于计算机、通讯电子等新兴学科较常使用的是分类表。

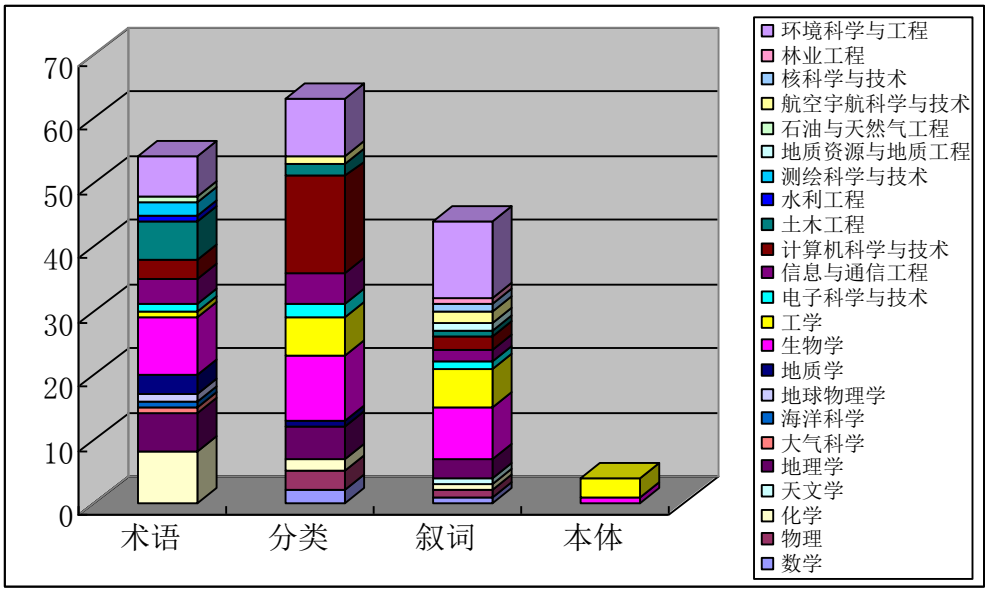


图 2 各学科不同类型的知识组织系统分布图

（3）从表现形式来看，目前专业领域知识组织系统主要有三种表现形式：印本形式的知识组织系统；数字化的知识组织系统；用语义网技术表示的知识组织系统。

在分析整理的 165 个知识组织系统中，绝大多数的知识组织系统都是数字化的，它们能够通过 web 方式用 html 网页提供浏览和查询功能。值得注意的是，一些专业领域的知识组织系统是和对应的电子资源集成在一起，也就是说，其建立目的是为了帮助特定的电子资源进行名称规范、提高检索的效率，如地理领域的 GEOLEX 是美国地质学会为美国地质学地图数据库的搜索而建立的专门术语词典。

在语义网框架下发展出来一系列的语义描述语言，包括描述结构的 XML、表达语义的 RDF 和表示本体的 OWL 等，其目的是实现机器可理解的信息描述。知识组织系统在语义网的环境下，充分利用了这一特点，用语义网技术来表示知识组织系统中的概念、特性、限制条件等，以便于计算机可读，也便于知识定义可被再利用。在分析整理的 165 个知识组织系统中，有少量用语义网技术表示的知识组织系统——本体，主要出现在工学和生物学领域，他们主要是对一个较小的领域知识进行表述的词和术语，并按照等级结构形成了类目。

此次调研中，也有少量印本形成的知识组织系统，他们多以 PDF 的形式存放在网络上，供大家下载和使用。

（4）从知识组织系统构建者的角度来看，调研收集到的知识组织体系的构建者有专门知识组织服务提供商，如 Intellisophic、Data Harmony，他们主要针对特定专门市场，如制药、教育、政府等，根据客户的要求构建专门的知识组织系统，并进行完善。在收集中发现了数量不小的此类型的知识组织体系，因商务的考虑，这些体系不能够获取，并复用。其次是数据库商，如 CSA、Gale，他们主要针对自己的数据库产品编制专门的叙词表和分类体系，因而这些知识组织体系无法脱离数据库产品本身，也难以被复用。专业领域的学协会，如美国光学协会、美国物理学会、国际纯化学与应用化学联盟，他们编制本领域的词表、分类法、叙词表。但值得注意的是，有些学协会已经建立的知识组织系统的目的是用于本学协会出版刊物的分类，如美国计算机协会（ACM）。

（5）从利用的角度来看，在整理收集的 165 个知识组织系统中，有 42%知识组织系统需要购买，42%的知识组织系统可以浏览检索，16%的可以免费下载。其中传统学科领域知识组织系统可浏览检索和下载的比较较高；其次生物、环境领域；而在新兴科学，如计算机、电子、通信，虽然有大量成型的知识组织系统，但多需要购买。

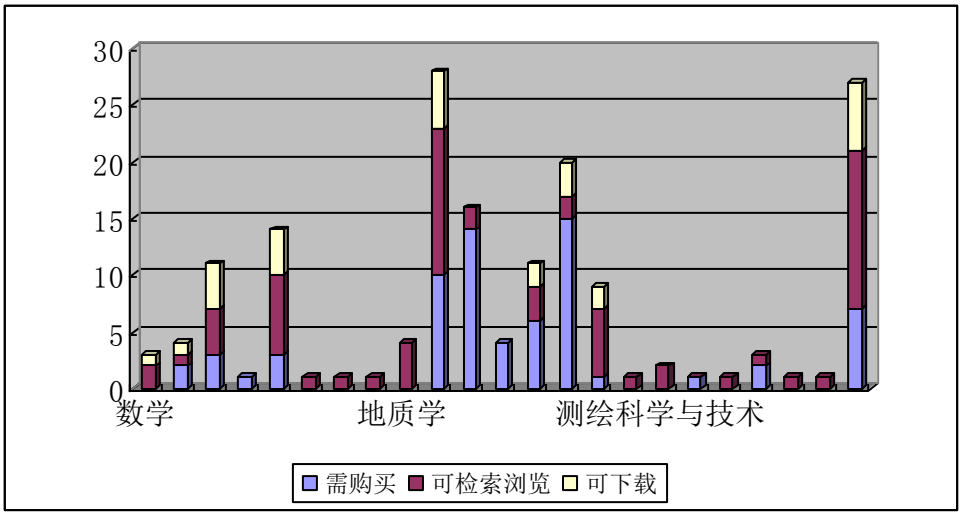


图 3 各学科不同利用形式的知识组织系统分布图

小结：从专业领域知识组织系统的调研可以发现，不同的学科领域在使用知识组织系统时，有较大的差异性。传统学科领域，多有成型的术语定义、规范文档等，多可免费浏览和下载使用；在电子、通讯、计算机领域，虽有大量成型的

分类表，但多为商业目的构建，需要购买；而在生物、环境、地理领域，由于学科本身的特点，其含有的知识组织系统最为丰富，并多可以免费获取再利用。

2.3 现有知识组织系统应用于综合科技知识组织的局限

通过上一节专业领域知识组织系统的调研，结合专业领域知识组织的要求，分析发现，专业领域现有的知识组织系统直接应用到专业领域环境中综合科技资源的组织存在如下的问题：

（1） 更新频率

现有的知识组织体系相对稳定，更新频率慢，无法跟上专业领域科学研究的发展速度。以收集到物理学领域知识组织系统为例，物理学领域具有代表性的4个知识组织系统，分别是INSPEC叙词表和分类表、美国物理学会的物理航空分类框架、美国光学学会的光学分类和索引框架，均采取年度更新的方式。而在收集到化学领域11个知识组织系统，除了3个明确提出时年度更新、1个是常年不定期更新外，另外的7个都没有明确提出其更新周期。也就是说，这7个知识组织系统在建设过程中没有考虑更新机制，建设完成后其包含的内容和知识组织结构也就固定了，是无法随着专业领域的发展而继续扩展。

（2） 知识组织对象

现有成型的知识组织系统从知识组织的对象角度可以分为两类：一部分是针对文献，包括期刊、档案、标准等，来进行组织；另一大部分是为了支持专业领域范围内用语的规范，如术语表、叙词表。现有知识组织系统与专业领域要涵盖的人员信息、项目、科学数据、科学数据，以及他们在科研过程中封装形成的复合数字对象差别很大。因而现有知识组织系统在直接应用于专业领域知识组织显得力不从心。

（3） 知识组织特点

专业领域现有的知识组织系统，多是采取先组式的方式，基本上都是静态的、列举式的结构，不能展示信息内容之间的关系。知识组织系统内容描述的颗粒度大，无法准确和细致标引专业领域内知识。

（4） 知识组织加工

专业领域不同来源的综合科技资源，在获取组织的时候有的已经带有自身的

分类、主题标引，有的则是结构化数据，如带有元数据描述，也有的是非结构化数据。要采用现有的知识组织体系重新来对他们进行组织，就必须进行不同来源数据的标引。即使采用机器自动标引，工作量大，准确性差，同时也无法利用已有的知识关系。

（5） 本体复用

虽在一些专业领域已建有一些成熟的本体，但这些本体主要面向特定的应用场景，在揭示问题的角度上具有针对性，移植到其他的应用场景，会存在语义二义性等问题。另外，本体如何进行规模化应用是目前本体研究中正在研究解决的一个问题，如欧盟第六框架的 NeOn 项目，希望通过方法和工具的研制来解决本体规模化应用的问题[6]。

2.4 小结

本章从专业领域综合科技资源组织的需求和特点出发，系统地考察专业领域内现有的知识组织系统是否能满足专业领域综合科技信息知识组织的需求。通过分析发现：

专业领域综合科技知识组织，具有资源类型和格式多样、数据异构、资源拥有者多元、面向应用领域专深的特点。它要求知识组织模型能够容纳性强，能解决异构数据的交互、能揭示多重数据关系、能随着资源内容发展而动态扩展。

专业领域现在已建好的知识组织系统，因专业领域研究对象、研究方法的差异，存在较大的学科差异。传统学科领域，多有成型的术语定义、规范文档等，多可免费浏览和下载使用；在电子、通讯、计算机领域，虽有大量成型的分类表，但多为商业目的构建，需要购买；而在生物、环境、地理领域，由于学科本身的特点，其含有的知识组织系统最为丰富，并多可以免费获取再利用。

将现有的专业领域知识组织系统应用于综合科技信息资源的组织，发现现有专业领域知识组织系统在更新频率和方式、面对知识组织对象、知识组织特点、本体复用上都存在局限性，直接采用现有成型的专业知识组织系统难以满足专业领域综合科技资源组织的需求。

[6] NeOn 项目. http://www.neon-project.org/nw/About_NeOn. [EB/OL]. [2009-10-21].

3 专业领域知识组织模式分析

本章将重点从专业领域知识环境和系统采用的知识组织模型，来分析专业领域面向综合科技信息资源组织的方法。

3.1 专业领域知识组织的典型案例分析

为了深入的分析专业领域的知识组织的方法，首先选择了一些专业领域知识组织的典型项目进行研究。

3.1.1 MathDL 项目

MathDL 是由美国国家科学基金赞助并由美国数学学会直接管理，其建立的目标是弘扬高等数学教育，为数学协会会员提供交流并发表文章的平台；支持开发网上数学教学的改革；普及并提高网上数学教学的科学教育质量；为数学教育者们自身提高提供相应的网上资讯；支持并协助全美数学协会宣传其负责的项目等[7]。

MathDL 是在早期美国国家科学基金赞助的数学门户和数学科学数字图书馆的基础上合并形成，沿用了原数学科学数字图书馆的名字，以便将分散的资源重新整合，使用户一站到位、统一检索。

MathDL 是在对其所拥有资源充分了解的基础上，将众多资源合理组织，按内容归类，并用 HTML 格式对各数据库及资料进行有条理的链接。经过整理后，MathDL 的主要栏目包括：

- Loci：在线电子期刊，是继承早期的三种电子杂志：《The Journal of Online Mathematics and its Applications》、《Digital Classroom Resources》和《Convergence》。
- 数学学会写作奖：提供获得数学学会写作奖得主的简短个人信息和文章的 PDF 链接。
- 检索：列出所有与 MathDL 合作伙伴及其数据库的资料。
- 数学新闻：数学界最新发展趋势及动态
- 数学历史上今日：专门讲述一些数学史上的名人趣事

[7] MathDL.http://mathdl.maa.org/.[EB/OL]. [2009-07-13].

- 个性化图书馆

而 MathDL 资源整合中关键是检索，MathDL 用 MathResources Inc.所开发的内容管理系统取代了原来 Lucidea Corporation 的软件来进行检索处理及链接。同时 MathDL 增加了高级检索功能，用户可就具体学科（几何、代数等）、作者、资源类型（新闻、教学参考、科研基金）、技术类型（flash 模拟演示、java 程序）以及发表时间进行限制进行针对性检索。用户检索后，系统会标示检索结果的位置[8]。

同时为了加强对合作伙伴资源的检索，数学门户合作伙伴完成了一个数学教育专业分类词汇表，以进行便于进行元数据处理。但由于元数据处理需要大量人力资金及时间，MathDL 合作伙伴尚未完成界定元素和统一词汇。

3.1.2 STERNA 项目[9][10]

STERNA（Semantic Web-based Thematic European Reference Network Application）是以语义网为基础的欧洲参考网络专题应用的最佳实践项目，与 12 个欧洲自然史和生物多样性机构合作，在 2008 年 6 年到 2010 年 11 月获得 eContentplus 项目的 1500 万欧元支持。它的目标是为了在动态和多语言的知识系统中建立最佳实践，将众多具有不同结构的数据模型的分布资源整合起来。

STERNA 是以鸟类和各种鸟类相关的信息为核心，将各种关于鸟、鸟类物种和他们习惯的多媒体资源，包括科学数据、文章、图片、视频和声频文件，汇集和显示在同一个信息空间，以支持欧洲数字图书馆，整合在自然科学、生物多样性及保护领域带有语义的丰富数字资源。

（1）STERNA 的系统架构

STERNA 基础架构是通过元数据的 RDF 模型和用 SKOS 格式描述的网络参考模型对成员网站的内容进行分布式语义检索。整个系统分为三部分：最基础是语义检索，用于实现对不同成员网站的分散异构数据库的检索；网络工具集，用于内容拥有的机构整合和丰富现有的内容和元数据，并链接到用 SKOS 的参考模

[8] 刘燕权，王晓燕. 敞开数学知识王国之门：美国数学数字图书馆——MathDL. 数字图书馆论坛, 2009(6):74-77

[9] Hans Nederbragt. An architecture for networked collections overview of the RNA architecture[EB/OL]. [2009-09-13]. http://www.sterna-net.eu/images/stories/documents/sterna_del_4.2_architecture_networked_collections_revised_20091009-1.pdf

[10] Andrea Mulrenin. STERNA annual report 1 [EB/OL]. [2009-09-13]. http://www.sterna-net.eu/images/stories/documents/sterna_del1.1.2_annual_report1.pdf

型；API 层，用于每个成员网站实现 STERNA 系统的检索功能。

拥有内容的机构依靠一系列的网络工具和数据挖掘程序，通过添加或者自动抽取元数据来丰富自身的内容，同时也通过网络工具根据参考模型来链接和整合其他内容提供者的资源和内容条目。用户也能利用这一系列工具，他们即是软件又是服务，来进行检索查询，并能够通过如分面导航功能，来丰富自己的查询。

STERNA 分布式图书馆的基本功能如图 4 所示：

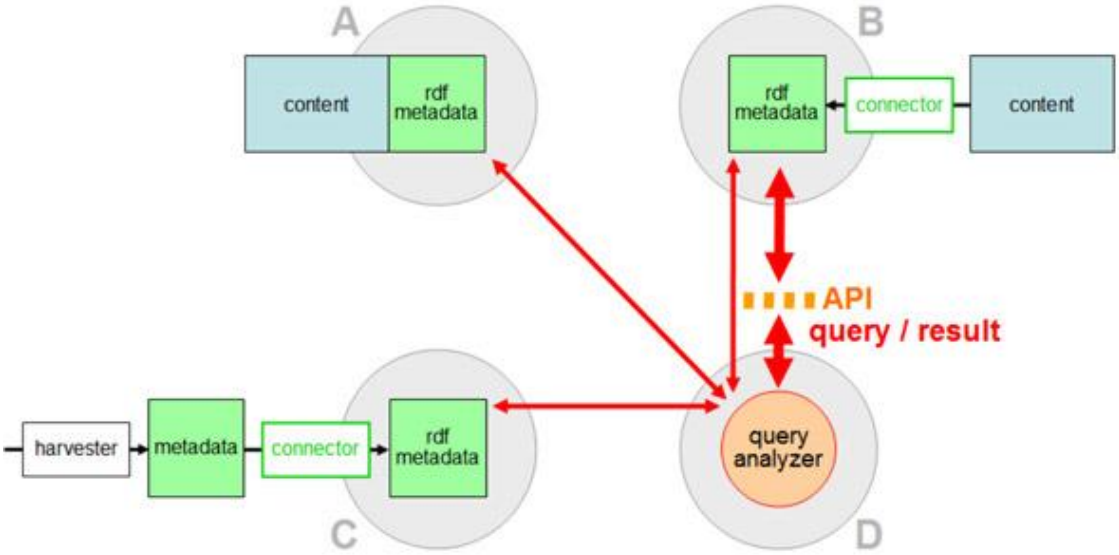


图 4 STERNA 系统的功能图

分布式查询使得联合工作成为可能，它意味着可以由一个成员网站（如网站 D）来提供必要的技术和数据，用来支持分布式查询，而能从其他网络成员处（如网站 A、B、C）分享他们的数据。

用户可以从任何一个成员网站（如网站 A、B、C）发起查询，此查询都会传送到联盟网站（如网站 D），即安装了查询分析的网站，由它来智能将查询传送到相关成员网站。查询分析将收集和整合形成结果表单，最终将结果反馈给最初发起查询的网站。

STERNA 系统中，对于内容提供者的网站有技术要求：用 RDF 方式存储成员网站；参考模型用 SKOS 方式的表示；高级功能通过联邦网站实施，如检索功能；采取软件即服务的方式。

（2）STERNA 信息组织

整个 STERNA 项目最具有意义和值得借鉴是采用语义建构“动态”的知识组织（如图 5）。

- 元数据是整个知识组织中的基础组成部分，项目制订了元数据生成和受控词表示的规范。元数据采用 RDF 元数据模型，RDF 的三元组结构易于链接，使一个物体即可以作为主语，也可以作为宾语；也易于机器处理，利用 URI 来链接相关资源。

- 受控词表示在 STERNA 系统中被称为“参考结构”，采用 SKOS 进行描述。参考结构中包含各种的不同组织描述资源的方式，即有低结构化，如词表和术语，也有高结构化能揭示词和词间关系，如叙词表、本体、知识组织系统。为了在语义网络环境下充分利用现有的知识组织系统，使其能被机器处理和整合到发现层，STERNA 项目采用了 SKOS 方式对 ITIS（Integrated Taxonomic Information System）、ISO3166 国别名称和代码、自建的标准化列表，如机构结构、人员结构进行描述。

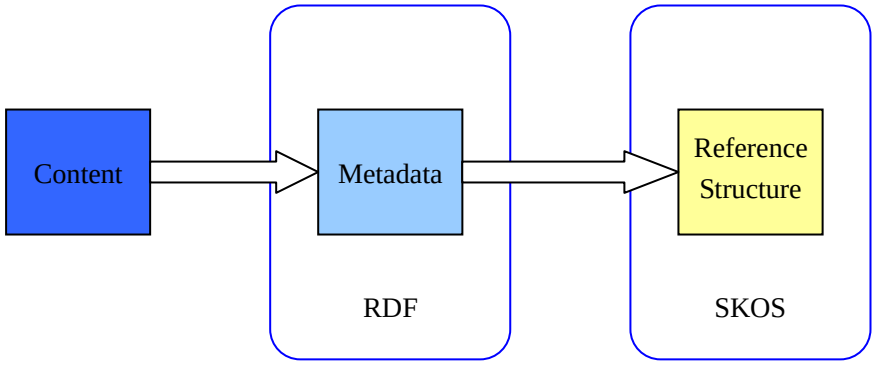


图 5 STERNA 系统的组织结构图

- RNA（Reference Network Architecture）参考网络架构，在系统中内容条目使用了多个参考模型，它们通过等级结构将内容相连接，同时内容条目又通过各自间的元数据相连，从而形成了参考网络架构。参考网络架构中内容条目之间的关系最基本是通过参考模型形成的等级结构，同时还存在虚拟的等级关系，RDF 的三元组则形成了属性链接的关系，而在 RNA 之外，内容条目可以通过超链与系统外的文本、网页链接。RNA 被用于创建一个发现层，来检索不同语言的各种数字资源，同时 RNA 工具被用来帮助创建和维护发现层及产生结构化文本内容。

3.1.3 VIVO 项目

VIVO web 是为了构建大型社交网络型的科研脸谱网，旨在方便科学家寻找

同行,改进研究,形成合作，通过类似 Facebook (脸谱网)交流方式，把全美各地的生物医学研究人员联系起来。Vivo web 在 2009 年获得美国国立卫生研究院 1220 万美元为期两年的研发基金(Grant)，用于资助康奈尔大学，印第安纳大学和佛罗里达大学等七所学校和机构用于开发 VIVO Web。其中康奈尔大学将主持这个多方参与的 VIVO 平台建设中的技术功能部分的开发和完善，佛罗里达大学将致力于开发保持每个站点的数据流；印地安那大学将开发社交网络工具，使研究人员能够发现具有相似兴趣的人和技术。其他四个机构-Scripps 研究所，Juniper 公司，佛罗里达州 Ponce 医学院，华盛顿圣路易斯大学和纽约市的魏尔一康奈尔医学院，将作为项目的实施地点。

VIVO 项目最早开始于 2003 年，是为了解决康奈尔大学校内跨学科寻找合作者的困难，实现超越行政区划，创造单点汇集式的学术交流访问平台。最初是在 Cornell 大学的生物科学领域内进行，在获得好评后逐步推广到 Cornell 大学的所有学科，此后获得了美国国立卫生研究院的支持在全美推广。

（1）VIVO 的系统架构

VIVO 系统的核心是一个网络本体编辑器和一个内容管理系统，系统采用同一套工具对底层的本体和填充本体的用户可见的内容进行编辑。

此外，作为系统的补充还开发了自动抓紧程序，便于 Vivo 系统从现有的网站和出版物中通过定制化的自动抓取和人工编辑获得数据。VIVO 能从商业数据库中自动抽取出版物信息，从中央管理数据库中获取研究基金信息并链接相应资助者、基金和管理部门，从 Cornell 人事管理数据库中获取人员信息。

在 vivo web 项目中，vivo 还将构建网络交流工具，加强和完善数据管理。

（2）VIVO 知识组织

VIVO 系统采用了实体关系的模型来组织和表现知识内容。在最初设计本体模型，考虑到系统组织的内容类型多是根据机构和组织，而不是根据科学概念和术语，VIVO 没有选择生命科学的学科本体[11]，而是在先进知识技术（Advanced Knowledge Technologies，AKT）支撑门户本体的基础上修改完成，AKT 本体能提供一个关于大学研究活动描述有用的类，如教育员工、政府组织、出版物。

VIVO 存储的信息是作为一个类型实例，按照明确的类型对象关系来链接，

[11] Medha Devare, Jon Corson-Rikert, Brian Caruso, Brain Lowe, Kathy chiang, Janet McCue. VIVO connecting people, creating a virtual life sciences community. D-lib magazine, 2007, 13(7/8) . http://www.dlib.org/dlib/july07/devare/07devare.html

如老师和课程的关系是他是否任教。每个实体都具有有限的数值属性，却具有相对较多标识性从上位类继承来的数值属性（实体关系显示见图 6）。数值属性可以是文本、数值和日期。在编辑任何一个类型的信息，与其相关和来自其上位类的数值属性都可以被填写或编辑。数据属性超越了传统的实体关系结构，提供另一层次的灵活性，通过简单的接口使单一的文本值保持在数据库中，比通常采用添加新的对象实体可用性更强。他们通过对细微的数据元素和外部的系统识别符，极大地促进 VIVO 与来源数据保持一致。标签和外部系统的数据分类能够在 VIVO 中存档，完整性检查，更新，正式这一处理保证了 VIVO 系统的可持续性发展。

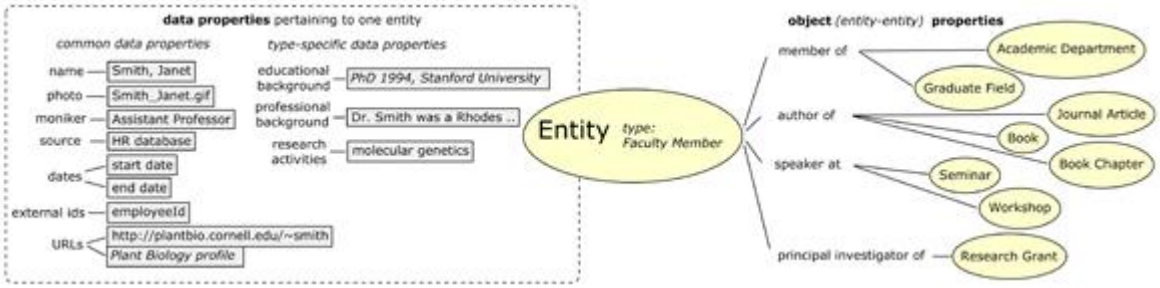


图 6 VIVO 实体关系图

在 VIVO 中的类型和属性关系，即内置的本体，是规定什么是实体以及实体间的关系，是独立于在数据库内容和网站显示的内容。这使 VIVO 在不同领域使用同一内容，可以拷贝相似的应用；或者在内容完全不同的情况下，作为一个空的结构框架来设置类和属性。

3.1.4 FOS 项目

联合国联农组织为了提高渔业信息服务的能力，发起了 FOS 项目，旨在通过创建、整合和利用本题，来加强渔业信息系统的信息整合和语义互操作能力。

FOS 项目整合了资源包括：

OneFish：一个渔业项目的门户，采用等级主题树方式对信息进行组织，大约有超过 1800 个主题，主题含有简短的摘要、标识符、相关联的材料，如文件、网站、元数据。

AGROVOC 叙词表：包含大约 2000 个渔业的叙词和 16000 相关的扩展词。

ASFA 叙词表：超过 6000 个叙词。

FIGIS：一个整合渔业信息的全球化网络，其采用参考表来组织资源，主要

包括水生物种、地理对象、水生资源、海洋渔业、渔业技术，大约有 300 个顶级概念，向下分成 4 级，共含有 30000 个资源对象，并能支持多语种互操作。

（1） FOS 的知识组织

FOS 希望设计出一个全面的本体参考模型，满足：是（部分是）以领域为基础本体，能分享规范的 KOS；足够的灵活性，能在同一背景下包括不同的观点或者视角；聚焦在渔业领域的核心推理框架[12]。

最终，FOS 建立一个多层级的本体仓库来整合资源（如下图），主要包括三层内容：

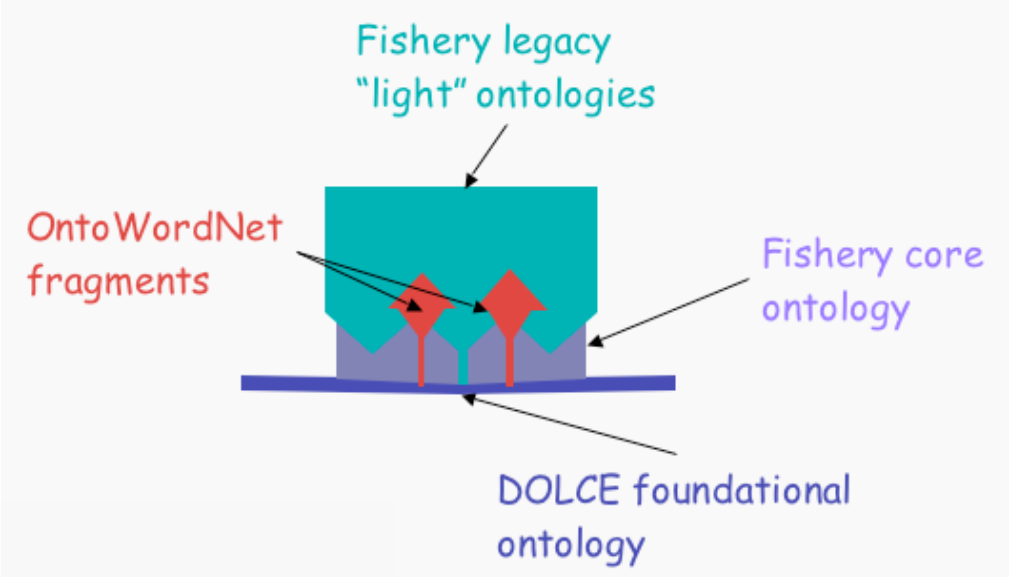


图 7 FOS 本体仓库模型

顶层本体或称为基础本体，用来表示通用一般性的概念。FOS 采用了 WonderWeb European 项目建立 DOCLCE 本体模型作为基础，DOCLCE 分为 3 个基本的大类：持久性和临时性；品质和品质属性；抽象概念，并采用了乘法的方式来扩展下位类。

核心本体是在顶层本体的基础上，结合本体描述与情景的原理（根据情景的作用、任务、参数、状态来具体化说明）来构建了核心本体。将 FOS 需要整合的资源按照 ODP 的本体描述模型转化成术语数据库，提取术语数据库顶层概念，保留 TDB 框架，专家精简以及采用其他本体设计模型[13]。最终，在 ASFA 的 1600

[12] Aldo gangemi, Frehiwot Fisseha, Ian Pattman, Johannes Keizer. Building an integrated formal ontology for semantic interoperability in the fishery domain. [EB/OL]. [2010-3-13]. ftp://ftp.fao.org/docrep/fao/008/af242e/af242e00.pdf

[13] Aldo Gangemi. Reusing semi-structured terminologies for ontology building. [EB/OL]. [2010-3-13]. http://wonderweb.semanticweb.org/deliverables/documents/D16.pdf

个顶级类，AGROVOC 的 83 个顶级类，FIGIS 的约 400 个顶级类中选取了 10%，作为核心本体。

领域本体，是在核心本体设计和检验后，将术语数据库中遗留的数据转化为本体数据模型，被称为“原形本体”。将 COF(核心本体)与 OntoWordNet 和“原形本体”进行映射、模块化、合并，形成领域本体。

（2）FOS 知识组织内容框架来源

通过对 FIGIS 渔业概念框架的讨论、精练、正规化，建立出初步的核心本体。同时将其他来源的叙词表的上级概念与 FIGIS 概念框架进行匹配。形成发展核心本体的材料；翻译 FIGIS 参考表：分类、个体和本地关系，转换为正式的公理。形成发展领域本体的材料；复用 oneFish 的主题树，设计最初的 IFO 仓库的架构，此架构要与核心本体匹配。形成设计本体仓库的材料；从 AGROVOC 和 ASFA 的 BT/NT 上位词下位词关系中抽取叙词。形成发展核心和领域本体的材料；从 AGROVOC 和 ASFA 中扩展 RT 相关词关系，同时非 IS-A 关系也在此展开。形成发展核心和领域本体的材料；复用现存的文献：oneFish 主题文摘、AGROVIC 和 ASFA 叙词的范围注释、FIGIS 的术语表。形成本体文档的资料；复用所有资源的 UF 代关系和多语种的等价关系。形成本体的概念资料。

3.1.5 NeOn 项目中的应用项目 FSDAS

NeOn 项目是由欧盟第六框架计划资助，14 个欧盟研究机构参与，旨在分散机构中的使用本体以进行大规模语义应用，通过发展工具和相应方法，产生经济有效方式处理方式解决整个过程应用问题，尽可能促进新一代语义应用。

其核心思想是网络化本体，通过发展一套整合的方法来进化网络化的本体和相关元数据。通过 9 个场景来实现现有本体的重构、映射、模块化、本体化，并将非本体资源进行整合。

作为 NeOn 实践案例，联合国粮农组织（FAO）采用 NeOn 技术方法搭建了以本体驱动的鱼产品消耗评估系统（Fish Stock Depletion Assessment System，FSDAS）。联合国粮农组织的渔业和水产部管理和维护着多个渔业和水产信息和知识组织系统，尽管他们含有的数据多是结构化数据，却难以互操作。在 NeOn 项目的帮助下，FAO 基于联合国的资源创建了一个渔业本体化网络，包括了 FIGIS 的渔业时间序列参考表、AGROVOC、ASFA、渔业和水产的分面表、FAO

地理本体。

FSDAS 采用本体网络化的方法以集成来自 FAO 不同信息系统的数据。根据渔业本体生命周期管理的要求，分别从本体工程师和本体编辑人员两个角度，归纳出 FSDAS 的一般性要求。

根据来自 FIGIS 的数据创建本体，即形成了 6 个本体模型：地域、渔业区域、生物物种、渔业商品、船只类型和载重、工具类型。通过对渔业领域范畴的分析，将对同一实体的多重知识组织方式进行分析，来扩展和修订本体模型；参考表作为本体化的网络的基础，由于已建立的单个本体在覆盖范围上不存在重复，采用数据连接的方式，即以连接数据到参考表，链接数据到半结构化文件、链接数据是从数据或本体中被推理获取，形成本体化网络[14]。

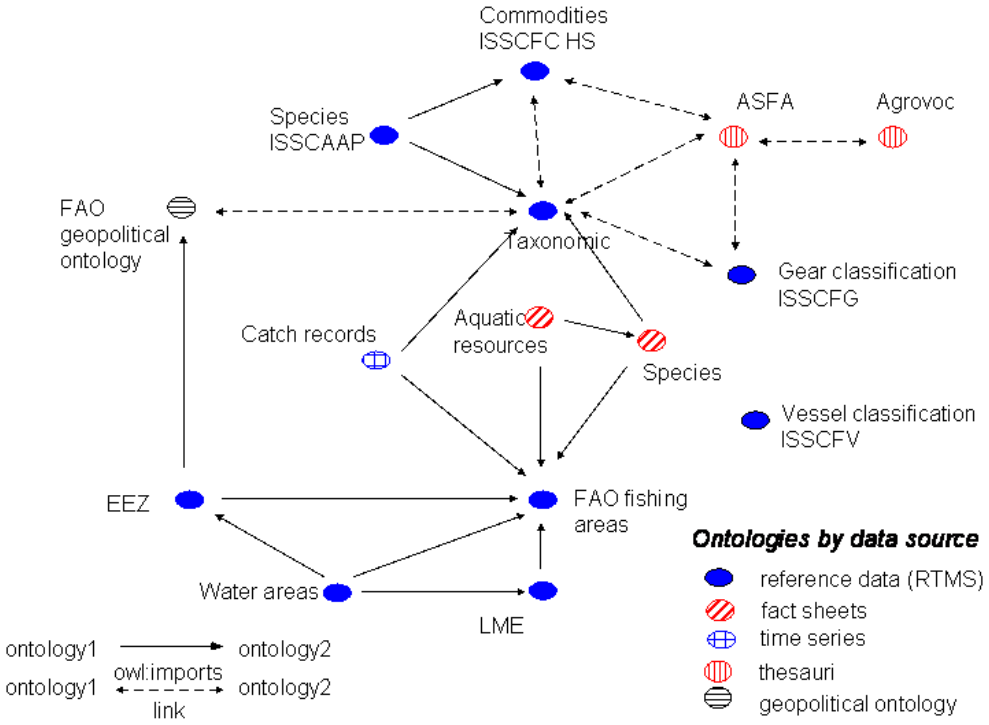


图 8 FSDAS 的网络化本体

3.2 专业领域知识组织方式

从上一节典型案例的分析，专业领域综合科技信息的知识组织，并不是要替代原来专业领域各种复杂的研究公开信息的网站，而是要提供一个知识发现和导航层，以帮助用户以学科领域为基础，实现将跨系统的人员、工具、机构、科学

14 Caterina Caracciolo. Second Network of Fisheries Ontologies [EB/OL]. [2010-3-13]. http://www.neon-project.org/nw/images/7/75/NeOn_2010_D724.pdf

数据等的关联。根据不同的用户需求和应用场景，知识发现和导航层的构建采用了不同知识组织处理方法，归纳起来主要有：

3.2.1 以元数据互操作为基础的组织

专业领域往往包含多种资源类型，带有多种的元数据描述模式，所以基于资源类型的元数据的互操作是常被使用的方法。根据都柏林元数据互操作的层次模型设计出来的应用框架，为平衡不同层次上元数据的语义互操作，定义了四个层次的互操作。第一层次规范了元素名称，第二层要求“概念”具有形式语义，第三层规定了编码模式的规范，即语法规范，第四层是整个元数据级别的规范（包括各类语义和语法约束）[15]。

专业领域在知识组织上主要采取的元数据互操作，主要是用于支持跨库检索和检索结果的呈现，一般的做法：在异构元数据描述中需要使用统一检索界面；集成或合并补充可能重叠的元数据模式或标准的描述；根据用户特定的兴趣、视角或需求，使一个隐含和完整的元数据描述产生不同的视图。

此类元数据的互操作，常通过建立一个辞典索引作为概念和计算相等的等级关系的识别条件。它能用来合并不同知识组织编码体系，如词表、数据辞典和交叉关系，进行数据交换的中央程序。使用者使用一个术语来查询专业领域的资源时，首先通过辞典转化为不同语言形式或不同知识编码体系中的专门词汇，向带有不同元数据的不同集合进行查询，然后将查询结果合并、去重后，进行数据的呈现。

美国国家数字图书馆在建设各个专业数字图书馆时，基本都采取此方法，将其所拥有资源在充分分析了解的基础上，将众多资源合理组织，按内容归类，并用 HTML 格式对各数据库及资料进行有条理的连接。在资源检索的设置上，采用统一的界面对其组织的资源进行检索，并对相应的元数据进行了规范。

3.2.2 以知识组织系统规范和映射为基础的组织

专业领域往往包含多种资源类型，每种资源类型由一个或多个资源集合组成，一个特定的资源集合和特定的元数据标准及知识组织体系相连。为了便于用户使用一组熟悉的术语或者同一知识组织体系来搜索或浏览不同类型的资源，最常见的方法是对资源集合采用的知识组织体系进行规范和映射。

15 Baker T. Interoperability Levels for Dublin Core Metadata[EB/OL]. [2009-10-27]. http://dublincore.org/documents/interoperability-levels/index.shtml.

知识组织系统的映射，常通过选择一个知识组织系统作为数据交换的中央程序，将不同知识组织体系向它进行关联和映射，以实现概念和计算相等的等级关系的识别。当使用者使用一个术语来查询专业领域的资源或者使用一个分类框架来浏览时，作为中间处理程序的知识组织系统就会按照映射关系，到不同的知识组织系统中，对所组织的资源集合进行查询，

美国国家医学图书馆建立 UMLS（Unified Medical Language System）关于生物医学和健康的知识组织体系，是采用知识组织体系映射中概念映射的典型例子。UMLS 中的 Metatheasurus 数据库是一个多用途多语言的词汇库，词汇来源于 100 多个术语表、分类表、叙词表及不同版本的概念和属性，通过不同知识组织体系中概念的映射，实现同一概念的不同名称和形式的连接，同时通过语义网络能提供Metatheasurus数据库中所有概念的统一分类和一系列概念之间的关系。

OCLC 推出的术语服务，将规范文档、主题词表、网络分类、分类表等，将源词表和目标词表用结构化的 MARC21 规范格式进行编码，将所有选用词和非选用词进行直接映射和同现词匹配。实现在多个受控词表之间有选择地建立映射，帮助图书馆、博物馆、档案馆等馆藏建立相互兼容的元数据。

知识组织体系映射能将采用不同知识组织体系的资源集合映射到同一个知识组织框架之下，提高知识检索的精准性，映射成果也可为整个网络信息组织所共享，但此方法的成本高、时间比较长，并需要不断维护和更新。因而在专业领域信息组织时，采用此方法主要利用已有建设好的映射表，以节省系统建设成本。

3.2.3 以本体为基础的语义组织

在语义网的发展和驱动下，本体的开发和应用获得长足进步。本体能够正式定义一个领域概念及其关系的结构化表示，并能够被机器所理解，它在语义网的环境下能对知识的分享和交换起到重要的作用。

以本体为基础的专业领域组织，是在用户呈现页面和不同来源数据之间搭建一个本体层，作为数据交换的中央程序。本体可以从实体关系出发，跳出原有资源的描述框架，来揭示和描述资源和资源之间的关系。当使用者使用一个术语来查询专业领域的资源时，搜索引擎检索会按照一定因素形成的索引来提供查询结果，并不反映查询结果与其他资源的关系，检索就到此停止了。而以本体为中央处理程序，无论是查询还是浏览，查询到的结果又变成了一个进一步导航的开始，

可以减少检索交互，同时能发现相似内容。

从上一节的典型案例分析中，VIVO、STERNA、FOS、NeOn 的 FSDAS 在搭建以本体基础的导航发现层时采用了不同的方式。VIVO 是使用了一个通用型的科研本体，STERNA 是创建了一个语义网，FOS 是形成了一个本体仓库，FSDAS 是形成了一个本体化网络。但无论采取何种方式，以本体为基础的专业领域知识组织都具有更强的灵活性，它将用户呈现数据、来源数据进行了分离，减少了来源数据进行改造和加工的处理，同时本体的语义关系丰富了数据的内涵、便于用户从多角度来获取资源。

3.3 专业领域知识组织的参与机制

要实现专业领域知识组织，搭建一个知识的发现和导航层，受到一个重要因素的影响，即专业领域知识组织的资源对象其所有者与搭建专业领域知识组织平台间的关系，它决定了是数据如何与知识发现层建立关系，也影响了专业领域知识模型的设计。

从典型案例的分析可以得出，资源与专业领域平台的所有者的关系，大致可以分为三种类型：

（1）专业领域知识组织平台的创建者与专业领域资源的所有者隶属于同一机构，或者专业领域知识组织平台的创建者获得使用的合法授权。在这种情况下，专业领域知识组织模型的设计，具有了较大的自主权。如 3.1 节中提到的 VIVO 系统、FOS 项目、FSDAS 项目都是此类型。FOS 项目和 FSDAS 项目的来源数据都是联合国粮农组织所收集，只是处于不同的信息系统当中，虽然是结构化数据，但难以进行互操作。在这种情况下，FOS 和 FSDAS 项目对资源的来源数据库采用的数据库、数据格式、采用的知识组织系统十分了解，并容易获取相关数据，为构建本体仓库、本体化网络打下坚实基础。VIVO 系统中综合组织的数据，在隶属关系上，分为两部分：一部分是 cornell 大学跨部门所拥有的资源，如人事部门的人事信息、教务部门的课程信息；另一部分是从订购数据库，如 Biomed 中抽取的论文信息。因而 VIVO 在以本体作为资源发现层的时候，只考虑不同资源类型之间的关系，不试图去统一数据的描述格式，容许相同资源可以出现在多条目下，并以不同的数据描述显现，这样不依靠来源数据库的知识组织系统、数

据描述格式，减少数据的依赖性，避免数据处理难度，增强了灵活性。而在 Vivo 项目的延伸项目 vivo web 中，这一思想更进一步得到了发挥。VIVO web 的合作伙伴除了采用 VIVO 中已建立的科研本体外，可以创建本地化的本体来满足特定的需求。VIVO 建设团队试图通过此来保持一致，并尊重本体需求。

（2）专业领域知识组织的资源所有者之间是合作伙伴关系。3.1 节中的 STERNA 项目是典型代表。联盟的形式使得在构建专业领域知识模型时，一方面能对拥有的数据结构、采用的知识组织方式、数据存储等有全面的了解；另一方面又面临多头来源、多种类型和形式的处理。在这种情况下，知识组织模型的建设，一般都会采用知识组织系统的映射或者在原来的各系统之上搭建一个中间层来解决数据异构、知识组织体系不一致的问题。STERNA 项目在对参与项目的机构数据上搭建了一个 RDF 层，而不去统一不同机构的知识组织系统或者元数据格式，通过不同成员网站的分散异构数据库的语义检索和用 SKOS 格式描述的网络参考网络，来发现资源和揭示资源间的关系。

（3）专业领域知识组织对象是供用户进行公开使用的，实施专业领域知识组织的机构仅是作为一个用户身份，来调用资源和服务。在这种情况下，一般都采用了元搜索的方式，对不同资源的元数据格式进行分析，形成视图，通过对异构数据进行检索查询来完成知识的组织。因缺乏对来源数据的数据存储、数据关系的认识，这种情况下很难深入揭示资源间的关系。

虽然资源提供者参与专业领域资源组织的方式的不同，但总的来说，来源数据与知识组织模型关联，数据转化的主要有以下几种情况：

（1） 采取了活链接（非通用性方法）

在用户查询的时候，数据才从来源数据中收集。这种情况主要适用于：一是来源数据库处于一个动态变化状态，数据还在不停发生变化；二是作为原创性的来源数据库，需要提供验证服务，来源数据库也并非合作伙伴或者隶属于同一机构。

（2） 表单写入（非通用性方法）

在来源数据的历史数据需要更新，并被作为一种原创资源。知识组织系统需要根据一定的时间或者通过来源数据库的一个诱因的促发，写入就能自动进行。这种情况适用于来源数据库有一部处于动态变化状态，并且是一种定期变化。

（3） 一次性写入（非通用性方法）

一次性写入，是历史数据或者原创的数据，不需要在经过任何编辑或者是转换的数据已经能满足知识组织模型的要求。这种情况适用于：来源数据库不在更新维护；数据库所有者同意将数据整体转入新的专业领域知识模型下；在数据写入过程，数据库进行大量手工一次性纠错；数据库包含了许多解释，依靠查询很难转入，比如多等级的推理。

（4） 通过标准化工具链接（通用性方法）

通过一个能适用于每个数据库的标准化工具，如 OAI 工具，常常被用来检索档案数据，TAPIR 常用于自然史领域数据的收割。通用标准化工具，要求来源数据库的数据管理系统能支持适用标准化的数据交换。

（5） 以 excel 作为通用链接器（通用性方法）

以 Excel 作为通用链接器，主要是考虑输入表单数据到知识组织模型的框架中。其工作模式是根据特定的资源类型创建一个 excel 模板。数据被手工写入到数据表中，即可以是从头开始输入，或者从数据库中进行拷贝粘贴到对应的栏目。由于通过此方法，链接将会保留，关系数据库的转入可以采取此办法。此方法适用于：来源数据库不会再进一步发展；来源数据库是由相对简单的关系表构建而成；所含有的数据不是原创型数据。

3.4 专业领域知识组织模式的发展特点

从上述国外专业领域知识组织的项目分析，可以看出专业领域知识组织发展呈现如下的发展特点：

从规范数据格式、数据内容到构建信息框架。专业领域信息组织最初从不同资源类型和资源集合的元数据格式和知识组织系统入手，围绕元数据格式规范、知识组织系统映射来实现，其目的是为了提供集成检索，提高检索的效率，并能够依靠知识组织系统映射进行扩展；而随着语义网和本体的发展，专业领域的知识组织能超出资源类型和资源集合知识组织的限制，根据用户需求来构建以本体为基础的信息组织层，在新的信息框架呈现数据间关系，进行知识的推理。

从封闭到动态的信息组织架构。专业领域信息组织最初多根据资源类型聚合不同来源的综合科技信息，其知识组织的框架和元数据格式是固定不变的，而本

体的架构使资源集合的元数据、知识组织体系、本体分离，知识组织架构可以动态进行内容扩展，如 STERNA 知识组织框架中的参考网络可以根据纳入的资源进行调整，从而动态进行知识架构和聚合。

专业领域的知识组织模式显示出多角度的融合。专业领域知识组织最初仅是在知识组织系统、元数据上进行操作，本体的引入使其跳出了原有的框架，来构建知识发现层，同时又能兼容元数据和知识组织系统映射的成果，来帮助提高知识检索和呈现。

4 专业领域知识组织模式设计和构建

4.1 专业领域知识组织模式设计

根据上一章，专业领域知识组织典型案例的分析，归纳出专业领域知识组织系统建设的特点：专业领域知识组织模型是作为一个知识信息发现和导航的中间件，并不试图取代专业领域各类知识资源的来源信息系统；专业领域知识组织模型在反映和揭示专业领域的各类型资源时，也并不试图重新建立各类资源的描述框架，而是从多角度反映资源与资源、资源与实体关系的角度来揭示；专业领域知识组织模型并不是要重建一套知识组织系统，取代已有的知识组织系统，也不是要将专业领域的综合科技资源采用的知识组织系统进行统一，而是要通过工具、重用现有的知识组织体系，搭建信息框架来反映资源间的关系。遵循专业领域知识组织系统建设主流的发展态势是：构建信息框架，支持动态的信息集成扩展，尽可能从多角度来展示资源间的关系。专业领域知识组织模型设计要考虑三个层次的问题：

来源数据的获取和存储，专业领域知识组织所获取的资源，根据 3.3 参与机制分析可以得出，来源主要来自同一机构不同信息系统、参与机构以合作的方式提供、从公开提供服务的系统获取，从专业领域知识组织模型建立连接的方式可以分为通用方法，如标准工具的连接，和非通用方法，如一次性转换、定时转换、活连接。采用搭建信息框架，不再对来源数据进行再加工和处理的原则，又要保证与来源数据的保持一致，以支持数据的更新。来源数据应该独立于专业领域知识组织模型存储，因此可采用 RDF 资源描述框架来对来源数据进行表述。根据 RDF 的三元组结构，RDF 可以将所描述的事物、所描述事物的属性、所描述事物属性的值按照主-谓-宾三层来描述，便于了发现知识的内容。来源数据的 RDF 转换，可以提供一个插件或者工具，在来源数据系统基础上搭建一个插件，也可以采取在专业领域知识组织系统的主体上，具体采取何种方式，取决于专业领域知识组织系统的参与构建机制。

知识组织参考架构：专业领域知识组织模型作为一个中间件，来支持导航和知识发现。知识组织模型是要容纳多个资源集合，一个资源集合是由多种类型的内容条目，并具有一定的等级关系。在同一个资源集合进行知识组织时，并不是

要随机将各种资源内容类型容纳进去，而只是将满足一定结构功能的内容类型包括进去，即反映资源与实体关系的结构，因此这一层可以采用以 owl 或 SKOS 表示，来构建一个语义化的资源关系框架。

以 SKOS 构建一个通用知识组织参考结构为例，可以采用概念、扩展概念属性、概念的对象记录和概念相关文本描述的结构，并以概念为节点形成等级结构：概念结构，即采用 SKOS 的概念属性来描述，如选用词(skos:prefLabel)、同义词(skos:altLabel)、变化词(skos:hiddenLabel)、概念定义(skos:definition)等；概念属性结构，当含有比 SKOS 更丰富的概念属性，即在 SKOS 已定义的属性基础上，如一些概念属性很重要，则在 SKOS 结构的基础上进行扩展；概念的对象记录描述结构；概念下对象记录含有文本、图像。除了采用 SKOS 为概念词作为结构组织外，知识组织参考架构也可采用本体形成来反映。

多知识组织参考架构形成知识网络：知识组织参考架构是从一个角度，反映了资源条目与其他资源条目的等级关系，但从其他的角度，资源条目与其他资源的关系也可能形成其他的知识参考架构，因此最终形成了一个互联的参考网络来反映资源与资源的关系。知识网络，即可以表现为本体的网络，也可以表示为多个知识组织系统的网络。

4.2 中科院专业知识组织系统建设的需求

4.2.1 中科院构建专业领域知识组织系统的背景

进入本世纪以来，国家对科研信息化建设越加重视，中国科学院投入了大量资金支持科研信息化建设发展，分别启动重大项目计划、加快科研信息化基础设施建设、深入推广科研信息化应用、创建虚拟科研组织机构[16]。

中国科学院在“十五”之初开始制订了信息化发展规划，提出了打造数字科学院（digital CAS）的长远发展目标，以科研活动信息化（e-science）和科研管理信息化（Academy Resources planning，ARP）为主要内容实施信息化建设专项。中科院科学数据建设、中科院国家科学数字图书馆等从“十五”起得到了院信息化专项的强化支持，制定了科学数据库元数据标准规范，能为中科院提供数据的存储和长期保存服务；构建科学研究和国家创新体系服务的科技文献信息支持系统，能为用户提供开放、集成的数字图书馆服务。同时，中科院一些机构提出建

[16] 阎保平, 桂文庄, 罗泽. 我国科学研究信息化的发展与启示.科研信息化技术与应用, 2010(1):10-17

设虚拟科研信息组织，消除地域、组织的界限来进行科学研究，如中科院国家天文台提出了建设中国虚拟天文台计划，依托中科院网络中心成立的“计算化学虚拟实验室”、联合动物所、武汉病毒所、遥感所等与青海湖国家级自然保护区管理局成的“中国科学院青海湖国家级自然保护区联合科研基地”。

在这样背景下，中国科学院科研人员在 e-science 环境下，在获取科研信息所需时候，面临着需要从多个来源获取信息，如文献信息服务来自中科院文献情报系统、科学数据来自中科院科学数据库、课程信息来中科院研究生院教务管理系统，而项目信息、人员信息则是作为了存储在 ARP 中，作为管理信息，并不对外公布。其次，在学科交叉日益、科学协作的情况，中科院以研究所为单元进行组织，如何跨院中科院系统内的组织机构限制寻找合作伙伴、了解同类研究的进展也是科研人员面临的新问题。另外，互联网连接了科研信息化设施，不断产生出新类型的科研信息需求。同时在专业领域内去跟踪国际一流科研团队、了解科研进展，并去获得相关合作机会和基金项目支持，也是专业领域知识组织要解决的问题。

构建专业领域的知识组织模式，为中科院专业领域的科研人员、研究生、科研管理人员服务，为中科院系统外的科研人员、项目资助者和投资人等了解中科院的科研人员、科研项目、科研产出。以此专业领域知识模型搭建的专业领域信息环境并不是要取代现有的其他信息系统，而是为了提供一个更便捷和灵活的导航层来反映信息之间的关系。

它的创建能有效的服务专业领域相关各个方面：

（1）为中科院的研究生或要报考中科院研究生的学生来发现研究领域、研究项目、指导老师、招生信息

（2）为中科院或中科院外部的研究人员快速寻找相关的合作者，了解相似研究领域的活动

（3）也可以帮助工业伙伴和投资人了解中科院领域的项目成果、获得联系，促进成果转化，获得投资

（4）为管理人员和期刊编辑，寻找专家信息，获得专家对热点问题的评论提供了方便

4.2.2 中科院专业领域知识组织模式的需求

根据专业领域知识模式的三层设计，即在数据获取和存储上采用 RDF，知识组织参考架构上采取 owl 和 skos，最终由多知识组织架构形成知识网络。从中科院专业领域知识组织模式建设的背景和定位，逐层进行分析：

（1）中科院专业领域知识组织的资源从采集的角度，有来自本系统和授权购买的资源，如数字化文献、科学数据；也有需要通过采集、编辑产生，如一些项目信息。如使用同一机构数据或者提供数据方是专业领域知识组织的合作伙伴，可以采取分布存储的方式，通过一个将数据转换成 RDF 的链接工具来实现数据的转换，数据的转换形式可以根据具体情境，参见 3.3 节中的方法；同时由于部分数据需要通过用户参与、编辑，属于原创性的资源；并有部分资源数据不再更新、资源数据也能够被直接整合到专业领域知识组织的系统，在专业领域知识组织系统的本地需要建设一个存储库，支持将获取的数据转换成 RDF 格式，并进行存储。

（2）从知识组织参考架构的角度，由于是专业领域知识组织模型的定位是为了在专业领域提供综合资源服务，中科院从研究领域上涵盖了不同科学领域，因此专业需求上有着不同侧重，是无法建立一个通用知识组织参考架构来满足所有的需求。通用科研本体的架构可以解决科研基本要素，如人员、机构、科研产出、科研活动，但要面向特定应用的时候，特殊要求是无法满足。在面向特定专业领域建设时候，需要根据特殊的要求进行架构，因此在这个层次上需要解决产生和维护参考架构。

（3）从知识组织参考架构形成知识网络的角度，每一个专业领域知识组织参考架构是面向一个专业领域或者一个知识集合，但其中的内容条目，如一个科研人员，从学科交叉的角度，也会属于另外的一个专业领域，会具有另外一个知识参考架构，从而形成知识网络。因此从知识网络的角度，要支持链接内容条目和知识组织参考架构。

从构建专业领域知识组织系统，系统架构角度的要求：（1）支持存储和检索实体、元数据和参考模型；（2）展示实体、元数据和参考模型；（3）支持编辑实体、元数据和参考模型。

4.3 专业领域知识组织模式构建

目前中科院专业领域知识环境(SKE)的建设,采用了康奈尔大学开发VITRO系统,根据3.1.3的vivo系统典型案例分析,VITRO提供了一个网络本体编辑器和一个内容管理系统。利用VITRO建立一个通用科研本体,来表示科研人员、科研活动、科研产出、科研机构间的关系。通用科研本体按照专业领域知识组织模式的设计,即属于知识组织参考架构。在数据获取和存储层,VITRO能支持RDF格式的数据导入,但并没有提供数据的获取、关联和转换成RDF;在知识组织参考模型架构形成知识网络,VITRO目前已经能支持构建多个本体,同时在VIVO web项目中正在进一步研发,支持本地化本体与核心科研本体互操作,来支持数据交互。

从这个角度分析,中科院专业领域知识模式以VITRO为基础进行构建时,目前需要在工具集上进一步建设:

(1)生产和维护内容:支持在VITRO系统上构建一个工具,来支持数据录入,包括元数据、记录数据,甚至是记录数据所带有的文本文件、图像、视频、有声材料。元数据和记录数据能够以RDF的方式存储,以文本为基础的内容能以XML和XHTML的方式存储。

(2)链接外部数据:建立一个工具作为基础,来支持建立与外部数据库的链接,来自外部的数据能被存储在RDF发现层。

4.4 灵长类动物知识组织模式的片段实验

昆明动物所想从灵长类动物出发,来描述物种状态、形态描述、生存状态等,并揭示与灵长类动物相关的科研人员、科研机构、科研项目。虽然中科院专业领域知识环境已经建成了科研本体,但是它是以科研人员为核心,没有针对灵长类动物的参考知识模型,因此选择灵长类动物作为对象,采用专业领域知识环境模式的构建思路来进行实验。

来源数据分析:

1 灵长类动物的描述:物种描述信息来自ARKIV地球上的生物图像网站,能提供物种属种信息、形态描述、生存状态、生物学特点、保持现状与需求、参考文献。

2 科研人员的描述,人员描述信息来自中科院昆明动物所网站,能提供人员

基本信息、简历、研究项目、代表论著。

3 研究机构的描述，取自机构文档库，能获得地址、负责人、邮编、概括。

4 灵长类动物的资源，可以取自 ARKIV 地球上的生物图像网站，包含图片和影像。

模式设计：

1 数据存取和获取：从目前可获取的信息，人员和研究机构是结构化数据，数据属于专业领域知识环境的构建者，这部分信息可以直接获取，转换成 RDF 格式，可以 excel 通用工具转换的方法；灵长类动物描述、图片和视频属于公共服务的网站，这部分信息是原创性信息，可采取灵活链接图片和视频、一次性导入动物描述。

2 知识组织参考模型：确认物种、资源、人员间的关系，按照标识、属性、可选属性类型、交互性来构建数据架构，参见表 1。在数据格式的基础上，在 VITRO 平台上编辑形成，形成一个灵长类动物描述本体。

表 1 物种的数据格式片段

标识	属性	可选的属性类型	交互性	描述
taxon concept	rna:taxonConcept	rna_taxonConcept		
taxon name	rna:taxonName	rna_taxonName		
type	dcterms:type	rna_objectType		物种的种类。参考实体类型表
owner	rna:owner	rna_person rna_organisation		
expert	rna:expert	rna_person rna_organisation		
collection	rna:collection	rna_collection		物种是资源的一部分，参考资源类型表

3 在知识组织领域环境上，构建显示的条目，选择关联的灵长类动物本体和中科院领域知识环境的通用本体，形成知识化网络。