

Terminology Services and Technology

JISC state of the art review

术语学服务和技术: JISC 的最新进展评论

Douglas Tudhope et al.

[Http://nkos.slis.kent.edu/](http://nkos.slis.kent.edu/) accessed Oct.17,2006

曾新红翻译 2007.1

要点:

TS 可以是 m2m 或交互的、面向用户的服务,可以应用于搜索 (search) 过程的所有阶段。服务包括: **将搜索术语转化为受控词表, 解疑服务, 提供浏览访问, 提供词表间的映射, 查询扩展 (query expansion), 查询重构 (query reformulation), 组合的搜索和浏览。** 这些服务可以作为最终用户界面的直接元素应用, 也可以根据应用环境支持场景之后的服务。在交互和自动服务组件之间的适当平衡需要谨慎的考虑。

投资回报 (return on investment) 应该在任何服务提供中考虑。有服务于不同目的的不同种类的词表, 它们有着不同的词表控制程度、语义关系丰富程度、形式化程度和编辑控制程度。交互的和自动的 TS 都有广泛的选择。有希望将 TS 用来增强现有的 JISC 项目和计划。

有时 TS 会被拿来与自由文本搜索相比较, 自由文本搜索有统计信息检索技术在自动标引和分级方面相助。但这些并不是互相排斥的选择, 我们有机会探索这两种方法的不同组合。**应该注意到 Web 搜索引擎已经引进了 TS 元素, 提供同义词和词汇扩展选项。** 因此, TS 不应该被看成是自由文本搜索的对立, 它可以增强自由文本搜索。

有许多现存的词表。可以发现有关词表的所有权、维护和许可的不同方案。谁将维护词表, 它基于什么被描述或在一个注册中可利用等问题需要调查, 因为这支持词表在 JISC IE 中的系统应用。这也关系到为词表的访问和维护建立商业模型。

映射是在异构环境中实现语义互操作的一个关键需求。虽然 schemas、框架和工具能够有所帮助, 但概念层次上的详细的映射工作是必要的, 需要结合人工工作和自动化的辅助。需要重点考虑对检索的影响。

自动分类和标引工具对于解决将 TS 应用于标引集合和知识库时的潜在资源开销是重要的。有些工具正在出现, 应该为 JISC 目的对它们进行调查。许多人正在讨论将人工和自动方法相结合。

在设计和评价 TS 时, 考虑人们如何进行信息搜索是重要的, 这可以降低设计错误的范围和提高服务能够实际应用的可能性。在正在进行的项目工作中应该尽可能实施用户研究 (user study)。

TS 不应该被看成是一种孤立的、自立的组件。TS 需要放在 JISC IE 的更宽泛的上下文中考虑, 并且需要和 eFramework 的其它组件整合。它们应该被看成是构成一组服务, 这组服务又可以和大范围的其它服务相结合。需要制定 TS 和其他工作流的业务规范, 作为 JISC IE 的一部分。

互操作性需要共同认可的标准和协议。在互操作的不同层次和类别中存在着不同的标准。有希望出现涉及术语学服务各个方面的广泛的标准集——永久标识符, 词表的表示, 程序访问协议, 知识库中的词表级元数据等。**这些标准是未来 TS 的基础设施, 但要等待国际上的认可是不切实际的; 国际共识将受到实践经验 (operational experience) 的影响。** TS

试点项目应该面向现有的潜在标准（在永久标识符，表示，程序访问协议方面的），并对它们进行评估和发展。

建议

该评论要求包括：“对该领域所需的进一步活动提出建议，JISC 在此工作中应该包含的内容的扩展（短期或长期的），包括与其它组织的合作，作为一种包含的可能形式。”下列建议按评论的相关章节列出，在章节中有更进一步的内容。

1. 介绍

评论的目的

- 术语学服务可以支持信息生命期的不同阶段
- JISC 应该在所有相关的 JISC 项目中强调主题访问和术语学服务，作为对现有项目的扩展或作为新项目。

术语学服务评论

- 论证术语学服务与 JISC 信息环境其他组件的整合（参见建议 4.3）

1.2.3 术语学工具和技术的组合

- 鼓励在开发术语学服务中进行学科间（inter-disciplinary）协作以及与存储机构和档案库的合作。
- 调查 TS 和非 TS（non-TS）搜索的不同组合

1.3.2 投资回报

- 调查使词表通过注册（Registry）对教育部门可用的方法，最初用于实验目的，但最终要以可持续、可维护和官方许可的方式。（参见建议 3.7）

2 用例 – 场景

- 用例应该在一个正在进行的基础上进行开发和精化，连同实际 TS 的案例研究、用户会议登录（usersession logging）、观察等。

3 词表的类型

- 根据环境需要提供对一系列不同词表的访问
- 考虑更宽泛的背景和投资回报是重要的

3.1 按结构分的词表

- 在开发词表和 TS 时要考虑分面方法（faceted approach）。

3.2 按目的分的词表

- 对词表的用途描述是词表注册的有用元素（参见建议 3.7）。

3.2.4 eLearning 目的

- eLearning 和数字图书馆领域之间的增强的交互改良（cross-fertilisation）
- 由标引员（编目员）、学生、教师进行用户行为研究。研究如何用不同的标引和检索工具支持有效的实践。

- 为兼容词表调查 VDEX 和 SKOS Core 表示之间的转换（参见建议 6.2）

3.2.5 eScience 目的

- 用词表描述搜索数据的用户实践研究

3.3 命名实体规范档（named entity authority）和解疑服务

- 调查机构名称和学术单位（IESR Agent 等）目录
- 研究在 OPAC 和学术 web 出版（LEAF, CiteSeer 等）上的可利用的名称规范档的覆盖面。
- 参加国际合作（如 LEAF, OCLC, SURFDARE）
- 开发一个示范性的 UK 名称规范文件原型，可能包括 BL 和大学（验证、职员、机构数据库），并在一个受限的应用中评估其使用。
- 着手解决 UK 服务和活动中的地方和地理名称的处理问题，并与相关的项目和术语学研究计划合作，开发标准和规范档。
- 支持 UK 机构积极参与科学学科中的国际命名标准研究计划，通过项目支持，帮助他们在 UK 实现。
- 应用名称抽取的方法，并研究它们相对于传统规范系统的长处和与传统规范系统的结合。建立和评估不同的名称解疑示范品（demonstrator）。
- 用一个名称规范档 Web Service 进行试验，例如，嵌入元数据生成工具。
- 开发或支持元数据增强服务（校正，丰富）：词表，分类法（scheme），映射，名称。

3.4 社会标记法（social tagging）和民间分类法（folksonomy）

- 将使用一个已建立的词表进行基于知识组织系统的受控标引，与用于某一特定领域搜索目的的自由（社会）标记法相结合进行试验，为了发现和检索目的进行优化。
- 对从标记到分面、受控词表和规范档的自动链接的潜力进行试验。
- 整合标记与现有的服务，例如知识库、OPAC、（RDN/Intute）主题网关、数字图书馆、KOS 创建和管理系统、博物馆展览以及编目、元数据增强服务等。
- 比较研究不同类型的用户参与：注释、推荐、个性化、信息重构、分类（categorization）、概念空间、概念图、主题图工具。这项工作可以传达一个将不同类型的用户参与与社会标记法整合的原型系统。

3.7 术语学注册处（registry）

在 JISC IE 试验台展示一个术语学注册处的使用，包括：

- 研究将术语学包含入 IESR，潜在地将词表描述为集合。
- 为一个 UK 术语学注册处开发市场建议（包括使用场景、IPR 问题、商业模式、成本效益）
- 评估 NKOS 提出的元数据描述子集（profile）的使用
- 保持不同的 UK 首创计划之间的协作（和 eScience 如 GRIMOIRES 以及和学习团体如 Becta 词表工具）并国际化（如 NSDL）。

4 与 TS 相关的活动

信息搜寻行为的研究和模型

- 在 JISC IE 中的 TS 的用户研究，阐明搜索过程（为服务的工作流）以及交互与自动 TS 间的适当平衡。

4.3 术语学 Web Service 的类型

- 开发更准确的 TS 定义，作为 JISC IE 和 eFramework 的一部分
- 定义 JISC IE eFramework 内的 TS 的搜索过程 workflow
- 在 eFramework 的上下文内，开发一个用于 TS 的分等级分层协议集合和针对不同 API 的标准绑定。
- 开发开源的、参考（reference）术语学 web 服务实现

4.3.4 术语学 Web Service 评论

- 与国际术语学 web 服务研究计划协作
- 开发一系列基于 TS 的搜索和浏览工具

4.4 映射

- 研究/比较试验计划中的不同映射方法和粒度（granularity）
- 开发一系列工具帮助创建映射
- 研究建立标准映射关系和一个映射协议的潜力
- 与国际映射服务计划协作

4.5 自动分类和标引

- 研究试验计划中的对标引和分类的半自动解决方案
- 研究当前可用的自动标引和分类工具

4.6 文本挖掘和信息抽取

- 研究 KOS 和文本挖掘之间的关系
 - 展示 KOS 如何能够支持文本挖掘
 - 展示文本挖掘如何能够被用来更新和增强 KOS

5 当前术语学服务活动的评论

- JISC 应该与 Dewey 协商用于 JISC 服务和计划的许可

5.5 知识库

- 试验经由不同类型的词表和 TS、基于主题访问知识库内容的不同方法，考虑成本效益，以及不同等级的内容聚合：
 - 主题分类法的使用 以及
 - 专业 KOS 词表的使用
 - 作者指定关键词的使用
 - 全文标引
- 考虑将主流分类法（例如 DDC）和指定的专业词表术语结合使用（如 RDN 内正在使用的）

5.6 增强现有的计划和项目

- JISC 应该支持一系列与最终用户和评估相关的试点示范项目
 - 研究针对（例如）标引、映射、搜索/浏览、查询扩展、解疑的不同的 TS 方法。
 - 考虑附属于适当 JISC 方案和项目的主题访问和术语学服务，包括对 Intute 的 TS

支持，到集合级元数据（如 topical composition, correlation）的 TS 连接（和主题访问），对知识库的 TS 支持，专门针对项目的实例等。

- 获得 (harvesting)
 - 研究用更多的主题元数据扩展获得工具的可能性
 - 研究 TS 和 OAI 等的关系
 - 评估面向词表元数据规格化和增强服务的益处，例如聚合获得相关元数据 (aggregator harvesting relevant metadata)，对其进行增强，然后提供改良后元数据的获得。
- 开发 TS 支持的词表可视化工具
 - 灵活显示和剪裁来自词表的片断
 - 灵活显示和剪裁结果
 - 组合搜索/浏览

6 标准

- JISC 应该鼓励参与国际标准研究活动

6.1 设计

- 相关标准应该包含在 JISC 标准目录中。所有新的首创计划应该考虑相关设计标准。

6.2 表示

- 被强烈推荐使用基于 XML 的表示法
- 建议词表提供者考虑使用 SKOS Core（如果合适的话），并对其进行进一步扩展和实用化。

6.3 概念、术语和词表的标识

- 用于引用词表及其组件的全球标识符机制是互操作 TS 的基础
- 被建议考虑以现有的以 http URI 方法为概念标识符的工作为基础
- 在一个试点研究中研究用于广泛应用的自由获取词表的标识符补充
- 词表提供者相关的教育工作需要提供标识符并进行实际问题的讨论

6.4 协议，协议子集 (profile) 和 API

- 需要为词表（以及它们的组成概念，关系和术语）的网络化访问建立标准的 m2m 协议，辅以共有的建立在 web service 和其他低级标准之上的绑定 (API)。
- 被建议考虑为 TS 使用 SKOS 或 ZThes API（以进一步发展的观点）。研究统一 SKOS 和 ZThes API 的可能性。
- 研究用于对词表进行映射访问的标准 m2m 协议，也许可以通过扩展 SKOS 或 ZThes API 达此目的。
- 研究 TS 与现有查询 API (SRU/SRW, CQL) 的结合/整合，或可能开发新的基于 TS 的查询 API。

1 介绍

1.2 术语学服务 (TS) 评论

术语学服务 (Terminology Service) 是一组服务，展示和应用词表 (受控与非受控的)，

包括他们的成员术语、概念和关系。这么做的目的是**搜索、浏览、发现、翻译、映射、语义推理、主题标引和分类、获得、警告等**。它们可以是 m2m (machine to machine) 的, 或交互的、面向用户的服务, 可以应用于检索 (retrieval) 过程的所有阶段。

TS 容易混淆, 因为它们跨越非常不同的应用领域、词表、团体, 并可以提供非常不同种类的服务。TS 可以作为最终用户界面的直接元素 (例如选择清单 (pick list)、浏览器或导航菜单、搜索选项) 应用, 也可以支持场景之后的服务。

词表常常与主题 (subject 或 topic) 元数据的控制相关。这包括主要的书目或教育主题分类、用于主题标引的叙词表、物种分类表等。其他种类的元数据可以受益于词表控制, 著名的例子包括地名、人名、类型 (genre) 和 eLearning 中不同的教育背景描述符。

1.2.1 受控词表

有一种类型的术语学服务试图通过词表控制提高一致性和改进对数字集合和 Web 导航系统的访问。词表控制的目标是在描述和检索条目 (为了信息搜索) 时降低自然语言的多义性。

受控词表由术语 (term) 组成, 术语是词表设计者从自然语言中选择出来作为可用于检索目的的词。术语可以是一个或多个单词。一个术语可以用来表示一个概念。

自然语言中的两个特征 (同义词和多义性) 会造成潜在的问题:

- a) 不同的术语 (同义词) 可以表示相同的概念。
- b) 同一个术语 (同形异义字) 可以表示不同的概念。

受控词表可以减少术语之间的多义性, 通过以下方法:

- 定义术语的范畴 (scope), 即它们如何在一个特定的词表中被使用。
- 为每个概念提供一组同义词或有效 (effective) 同义词。
- 限制范畴, 以使术语只拥有一个含义 (并只与一个概念相关)。

不是所有的词表都提供以上所有三项特征。有一些只是简单的规范术语列表 (规范表)。受控词表也为知识组织系统 (Knowledge Organization System, KOS) 提供词表, KOS 进一步通过不同种类的语义关系使概念结构化。

1.2.2 民间分类表 (folksonomy)

TS 的另一种类型不考虑一致性, 而是考虑让最终用户更容易描述信息条目和访问其他用户的描述。这导致了不受控的词表, 至少刚开始是这样。原则上这不是一种新的术语学, 但新颖的网络服务最近已引起了注意。这项活动已出现了若干新词汇, 包括社会标记法 (social tagging) 和民间分类表 (folksonomy)。有些人认为它可以降低标引成本 (可能非常显著), 鼓励最终用户参与信息服务和为团体建设 (community building) 做贡献。但还没有在教育目的方面评估过它, 而且现有的社会标记法应用也不是按一般检索的目标来设计的。

1.2.3 术语学工具和技术组合

总体来说, 若干不同学科都使用词表并可以为 TS 作出贡献, 包括人工智能、人机交互、信息检索、图书馆和信息科学以及自然语言处理。

1.3 成本效益问题

1.3.1 效益

TS 可以让用户更有效地实施教育查询和研究查询。当使用非控制术语搜索自由文本时, 搜索陈述上的细微变体和对信息需求的不同概念化都可能导致显著的不同。不同的人使用不

同的词或使用稍微不同的概念。要求非专业人员使用技术词表可能是件困难的事情，人名或地名的变体可能会阻碍一致的访问。如果目的只是要获得一些相关的条目作为一个话题的例子，这也许不是什么问题。但是，如果目的是一项关于某个专业化主题的深入的教育评论或系统研究，那么人们会不愿意错失潜在的相关条目。这些问题可以通过不同的术语学服务来帮助解决。

在最简单的层次，一个受控术语表可以确保搜索和标引的一致性，帮助减少因同义词和同形异义词不匹配导致的问题。名称规范档（name authority）是个重要的例子。

在稍复杂的层次上，层次结构和其他语义结构上的概念表示可以帮助标引员和搜索者为他们目的选择最合适的概念。建立基于浏览（browsing）的用户界面成为可能。

KOS 可以同时辅助查准率（通过允许精确搜索）和查全率（通过检索由相关概念或等价术语描述的条目）。它也提供潜在路径（potential pathway）（为人或机器）连接搜索者和标引者的术语学选择。本体中的特定语义关系的更形式化的规定可以辅助这样的应用：对关系指定规则并且基于逻辑的推理是适用的。

非控词表的使用可以鼓励最终用户参与社会标引或标记，帮助某个应用建立用户团体。

许多映射和语义互操作应用依赖于不同类型的 KOS，其他下游的应用也是如此。

1.3.2 投资回报（return on investment）

设计一个受控词表会有成本，将其应用于分类和标引也会有成本。因此在决定语义关系的丰富程度和形式化程度时，应该为头脑中的特定应用考虑成本/效益问题。例如，一个简单的受控规范表对于应用目的是否充分？另一方面，也有许多现有的词表和标引数据集可以在更大型的方案中被利用（leverage）或组合。（半）自动标引和分类技术也有潜力应用于商业系统产品和该领域的项目成果中（见第 5 节）。交互式元数据分配工具（interactive metadata assignment tool）及其在应用界面和项目 workflow 中的嵌入也具有应用潜力。

2 用例——场景

检索性能（retrieval performance）

一个搜索过程的初级阶段（见第 4.1 节）可以包含一个对信息需求细节的探索或熟悉过程。和一般的主题词表一样，在线词典或百科全书有时也用作这种一般目的。不同的 TS 可以在搜索过程中整合成一个选项，作为用于查询术语的资源。Google 的工具栏已经提供了一项词典服务，可以展望类似形式的 TS。

在信息检索系统中，同义词环（Synonym Ring）或搜索叙词表（Search Thesauri）（见第 3 节）被用来提高搜索性能，通过在匹配查询时考虑同义词以及来自相关概念的术语。提供一系列 TS 服务来提高查询性能（查全率和查准率）是可能的。这包括不同的查询扩展可能性，此时可以基于语义匹配程度进行结果分级排序。例如，你可能希望用很专业的术语学来搜索；你会对那些概念的匹配结果感兴趣，如果不是，也可能对密切相关概念的匹配结果感兴趣。利用查询扩展（query expansion）可以在一个查询中组合若干搜索“迁移”（move）。

另一个来自 OCLC 研究的实例场景是取自 Dempsey 的 Weblog（Aug 18, 2005）的条目，有关目录如何被用来提供访问路径（经由同一作者或相似主题等）。这已由来自 Worldcat Find in a Library 服务的一个例子证明了。（<http://orweblog.oclc.org/archives/000772.html>）

2.2 名称规范档（name authority）

你希望在你的 ePrint 大学数据库中找到作者 *D.Smith* 写的文章。这可以用 *Smith, D.* 来搜

索。但是返回了大量的搜索结果，在一个列表中显示出代表若干不同作者的该名称的若干变体（包括连字符姓、人名第一个字、中间首字母等）。没有简单的方法可以区分这些不同的人而获得一张确定无疑的列表。没有在线的规范文档可以让你搜索或浏览从而选择确定的人。当作者的名称也可以作为人名的第一个字或姓出现时（例如 Thomas 或 Michael），这种情况会变得越发地困难。

提供一个整合的“增值”名称规范档服务可以让搜索者区分清楚作者的名称。

2.3 映射和其他 TS

通过考虑映射和一些其他的 TS，这个场景扩展了迷人的 BIOME Auternative Land 用例研究：

农业多样化常常作为农业口价格滑落时的常备万能药。对新的农产品进行改变或变成有机的或保守级产品（conservation grade production）被看成是一种脱离农业口价格下滑循环圈的有效途径。但是，环境计划（environmental scheme，常常被称作 agric environment scheme）也被推举为一种能稳定农业收入以及有利于农村和城市的更广泛社区的方法。

这个场景讨论了两项 AgriFor 资源，是用 *farm environment schemes* 搜索 BIOME 的结果。在该用例研究中，这个查询式作为一个起点给出。但是并非显而易见一个学生会构造这样一个查询式作为起点。不同的 TS 可能可以通过建议受控术语（如第 2.1 节中所讨论的），或用不同形式的查询扩展（同义词扩展和概念扩展）来帮助构造这样的查询式。

初始的浏览也是搜索过程常有的早期阶段。

已经找到了用例研究中提到的信息条目，除了浏览器的后退按钮，没有更容易的办法“聚光”到 AgriFor 类目中。取而代之，信息条目是用 CAB 叙词表概念标引的。这很有用——知道条目的标引方式对定义一个搜索潜在有利。CAB 概念提供了一个选择：可以通过点击标引词在集合中导航。但是，该叙词表的知识结构并没有用来进一步辅助搜索。浏览是通过字顺表来进行的——等级结构的上下文和相关概念并不可用。

将分类法和叙词表组合进行标引可以提供极好的资源。更多的应用可以由组合构成。有一种可能是一起映射这两种词表。这也许有助于高级搜索，例如查询扩展。另一种可能是搜索和浏览的更大程度的整合（见 DeweyBrowser，第 5.3.6 节）。

2.4 知识库（repository）

你希望在你的机构 ePrint 知识库中搜索某一特定主题的文章。因为覆盖面宽，有一个通用的词表可用来浏览访问，在这个案例中是国会图书馆主题域（Libray of Congress Subject Areas）的头二三级，以及相关的记录（posting）。但是，从主菜单并不能清楚地知道你感兴趣的主体会落在哪里——你平常用来描述你的主题的术语并没有提到而你又不喜欢在相当宽泛的浏览分类中浏览多个子菜单。在这个浏览器中，你试图在这一页中寻找而没有成功。没有办法搜索词表找到你的兴趣可能落在什么地方。当然，你可以搜索全文，但这依赖于出现在文中的一个主题关键词。

用同义词入口词表（entry vocabulary of synomymys）来增强通用分类法并允许搜索此扩展词表的 TS 将扩展检索功能的效用。这可以为浏览提供额外的入口点。分类法和浏览选项越广，这种方法就越有用。

这个场景假设一个大学出版物知识库的主题搜索是一个明智的选择。已知在任何一所大学覆盖面很可能呈不规则分布，一些著名的条目搜索或基于作者的搜索形式可能更适用。但是，基于主题的访问对于未来的不同类型的聚合型知识库将是适用的。

3 词表的类型

3.1 按结构分的词表

词表可以根据其结构进行考虑 (Hodge 2000 并参见 BSI, NISO)。一种方法是按递增的结构复杂度和关系类型来组织, 这也大致是以下讨论的主要章节的顺序。**知识组织系统 (KOS)** 是受控词表, 它们通过不同类型的语义关系进行组织和结构化。

3.1.1 术语表 (term lists)

在最简单的层次, **术语表** 提供 **多义性控制** (ambiguity control), 并且通常是非结构化的列表, 当一个受限选项集合被提供时尤其适用。如果被做成一个 **选择列表** (pick list) 来用, 它们可以在交互式标引和搜索应用中确保术语学控制。**规范文档** (authority files) 用来控制命名条目的变体, 例如人名、组织名或地名, 并且通常按字母顺序展示。对于大型的规范文档, 可能会利用一个受限的等级结构来使访问容易些。见第 3.3 节可获得更多信息。**词汇表** (glossary) 是来自某个主题领域的术语列表, 带有相应的定义。**词典** (dictionary) 通常比词汇表有更通用的领域应用, 并可能包含一个单词含义的不同意义。它们总是按字母顺序展示, 可能有关于单词起源的信息。**地名辞典** (gazetteer) 列出地名, 可能也包含以不同“覆盖区 (footprint)”类型表示的关于地点的坐标信息, 例如质心 (centroid)、bounding box 等。**同义词环** (synonym ring) 最近作为一种术语表出现, 它在 (自由文本) web 搜索工具中提供 **同义词控制**。它们不用作标引目的, 但当特定领域的同义词集合可以在用于特定网站的搜索引擎中找到时, 它们在自由文本 (非控) 搜索引擎中提供了一个概念的同义词 **查询扩展** 选择。

3.1.2 分类表 (taxonomy)

所有的分类表都提供类目的等级结构组织。这种等级关系可能是松散定义的, 也可能是明确定义的。它们通常服务于 **分类目的** (classification purpose) (类似于将条目分入同一个“桶”, 见上)。同样地, 它们可以被看成是 **分类法** (classification scheme) 的 (更简单的) 实例。来自图书馆领域的复杂实例, 例如杜威十进制分类法 (Dewey Decimal Classification, DDC) 和国际十进制分类法 (Universal Decimal Classification, UDC), 则被认为属于分类法范畴。

信息的等级结构组织出现在许多领域, 存在不同形式的分类表, 它们服务于不同的目的并通过不同类型的划分特征进行组织。**Taxonomy** 是一个非常松散的术语, 甚至在术语学圈子里也有广泛的用法, 从相对简单的菜单系统到复杂的团体知识库 (corporate knowledge base)。分类表和 (至少) 三种团体相关: 科学分类系统, 网站设计者, 公司分类表。来自科学界的例子包括著名的科学分类表——见第 5.2.5 节讨论的生命科学项目。在网站设计中, 分类表是用于多种术语学系统的最常见的术语, 有时候很不正规。分类表也用作菜单系统的基础, 作为一种组织网站以方便交互式浏览站点各部分或支持其他访问机制的方法。有些情况下, 会采用一种很松散的等级关系来组织菜单系统。有时菜单结构是从底层的知识库动态产生的。在某些商业信息环境中, 用可能的不同编制视角, 我们接近一种更普遍的公司知识管理结构。各种各样的网络开发应用试图提供某种形式的分类表自动创建 (例如见 ch16, Rosenfeld and Morville)。但人工输入也可以识别, “信息建筑师”的新角色已经出现。来自 Taxonomy Strategies 公司的 Daniels 和 Busch (2005a, 2005b) 回顾了分类表的商业应用, 讨论了最好和最差的实践, 也考虑了与特定 Dublin Core 元数据元素相关的词表。他们建议在适当的时候将 DC 主题元数据分解为独立的分面 (facet) 并给出了例子。他们也讨论了 ROI

问题（参见 Rosenfeld and Morville, ch 17-18）。

要想让做的东西有用，一定要记住涉及到的决不仅仅是创建一个简单的等级结构。我们来看一个来自微软的将分类表（广泛考虑）成功应用于内部 MSWEB 的案例研究的例子（详见 ch20, Rosenfeld and Morville）。微软团队对分类表的应用包括：用于一个给定文档类型的属性的元数据 schema；用于菜单系统显示选项的类别标签。他们的工具包括一个**词表管理器**（Vocabulary Manager）（支持词表及它们之间关系的编辑，包括叙词表关系），一个**元数据注册**（Metadata Registry）和一个**URL 编目服务**（Cataloguing Service）。？

3.1.3 主题标题（Subject Heading）

主题标题是主题术语的受控列表。它们常常有宽泛的覆盖面和比较窄的等级结构。通常允许“组配的”、复合的标题，按照规则组合单个的主题术语构成。这些规则可能比面分类法更严格。著名的例子包括国会图书馆主题标题（Library of Congress Subject Headings, LCSH），医学主题标题（Medical Subject Headings, MeSH）。它们都典型地具有一组主标题，这些主标题允许再细分或附加限定词。

3.1.4 基于关系的 KOS

基于关系的 KOS 是按照概念来定义的，比前面的 KOS 结构更清晰地区分不同种类的关系，同时在关系的粒度和定义的形式化程度上多样化。在表达性（如关系的数量）与可操作性（经由对含义的共同认可和关系的使用）和设计成本之间倾向于有一种实际的折衷。常见的 KOS 变体在结构上有重叠的倾向，但是带着不同的目的设计的。

3.1.4.1 叙词表（Thesauri）

叙词表是为检索目的而设计的，有一组受限制的关系。三种叙词表关系是：等同（将一个概念与其有效同义术语连接），等级（上位/下位概念）和相关（更松散的相关，“参见”概念）。这些是由国际标准定义的。**英国和美国的标准最近修订和扩展了（BSI 还在进行，NISO）。标准讨论了这三种关系的共同的子类型。**例如，等级关系可以具体化为类属（Generic，子类/超类），实例（Instance，类/实例）和部分（partitive，整体-部分）关系。等同关系将一个概念与一组等同术语连接起来，这些等同术语被视为设计者构想的检索情形中的同义词，并且多种子类型也是可能的。可以采用单等级结构或多等级结构。

按照叙词表标准，概念之间的关系声明有严格的规则指导。**一些广泛应用的叙词表没有遵循所有的规则，但仍然表现出对其使用目的能够起到有效作用。**叙词表倾向于为一个特定主题领域或产品家族进行定义并可能非常庞大。它们通常被用于描述性的标引目的和相应的搜索系统。**叙词表也可以在自由文本搜索引擎中作为一种查询扩展资源使用（有时称为“search thesauri”）。**

3.1.4.2 分类法（Classification Schemes）

分类法在许多方面类似于分类表（taxonomy，见上文）。有良好定义的等级结构关系的更复杂的分类法应该被看成是基于关系的 KOS。良好结构化的分类法是单等级结构、遵循穷举原则（覆盖所有相关主题）和相互排他的。复杂的分类法，如 DDC，用一种广泛的辅助信息和连接来增强概念，包括有效同义语集合、分类法内部的“see also”交叉参见、直接和较松散的（如共现）到相关分类法或叙词表中概念的映射等。

在处理组合独立概念的复合主题描述时有两种方法。在列举式分类法中，所有合理的组合都明确地在分类法中列出并位于类结构的一个位置。任何新的复合主题都必须明确地加入分类法的新版本。另一种方法，可以使用规则（合成规则）组合原子概念来表示合理的组

合，这种方法比实际的明确列举能描述更大范围的主题。这种分类法被称为合成式（synthetic）分类法。在实践中也有混合的方法。

3.1.4.2.1 分面分类法（Faceted Classification Schemes）

分面系统在用原子元素合成复杂描述的过程中应用分面分析方法。分面这个术语以不同的方式在使用，引起了一些混乱。在本文中它一般是指一组基本类别（适用于一个应用领域）和它们的按照规则进行的组合。每个基本类别本身可能就是一个类等级结构。最常见的是，不同的分面因素是相互排他的。在标引一个对象或构成一个查询式时，来自不同分面的单个概念被组合在一起。和试图形成一个包含（例如）对象、材料、主体（agent）的所有不同可能组合的单一等级结构相比，这常常是一种更简单和更符合逻辑的组织形式。

针对网络数据库的分类浏览界面是有用的，此时用户可以面向原始显示，各种不同的商业搜索引擎现在提供这种工具。Pollitt 的 HIBROWSE 系统展示了浏览分面等级结构和交互式组合来自若干分面的术语改进查询式的潜力（Pollitt, 1997）。Flamenco 系统在用户浏览不同分面时动态产生查询结果的预览（Hearst et al.2002; Yee, 2003）。已经进行了一些用户评估，Flamenco 现在可在一个开源基础上利用。在英国，Adiuri 分面系统已用来开发某些 JISC 项目的 Web 界面（见第 5.1.5 节的 Common Information Environment(CIE) Demonstrators）。分面的、“过滤流（filter-flow）”界面可以在一组选择中引导用户，随着每个选择动态更新选项的范围。在查询比浏览更适用的情形下（例如深/不熟悉的等级结构），或在查询预览不能实行时（Tudhope et al. 2006），分面方法对于搜索也可能有帮助。

一个稍微简单点的分面思想流行于许多 Web 界面环境和美国。在这里分面常常是不同的元数据元素，有少量语义概念组合它们（例如见 NISO 和 Rosenfeld and Norville 2003）。分面可能包括地方、时间、价格、颜色、观众等，可能不总是按照等级结构进行组织。参见第 6.2 节的用于一类分面网络界面的 XFML 表示法。

在英国，受分类法研究组（Classification Research Group）工作的影响，可以发现更复杂的分面系统。分面分析被应用于主题的所有不同方面，均按等级结构组织。在这里基本类别可能包括抽象实体、对象（不同类型的）、材料、主体、过程等。不同类型的规则指导着分面的“语法”组合（有时在较低层递归），常常提供排序原则，有利于结构化的浏览。见 BSI Part 3, Aitchison et al.(2000)可获取更多信息。

分面分类法在很多方面类似于分面叙词表（例如 Getty Art and Architecture Thesaurus）和一些本体（如第 5.2.6 节中的 OpenGalen）。

3.1.4.3 词汇数据库（Lexical Databases）

最著名的词汇数据库是 Princeton 大学的 WordNet，它是一个通用目的的语言学资源，有着比叙词表更宽范围的语义关系。名词、动词和介词/副词都有独立的数据库，每个数据库有它自己的关系集合，包括等级结构关系。WordNet 通过与领域无关的词汇关系来区分不同的词义，这些词汇关系包括同形异义词、反义词和同义词（提供大范围的“合集（synset）”）。它已被大量的通用语言处理应用所采纳，虽然其他词汇数据库也可能很好地用于专业目的。EC Telematics Project 产生了 EuroWordNet，有不同的欧洲语言版本。

3.1.4.4 本体（Ontology）

本体这个术语有时可以松散地用于任何知识组织系统，尤其当它使用语义 Web 标准（如 RDF）表示时。但是要用于人工智能建模和推理目的时，本体倾向于有在这儿讨论的知识系统中最准确和最形式化的关系定义。本体包含类（概念）和这些类的实例，是该领域中的对象。类通常拥有属性，这样可以描述领域中的复杂对象。关系包括 is-a（用于类等级结

构)、实例、部分和(有时许多)领域特定的关系。

要区分详细的**领域本体**(domain ontology, 可以是叙词表或分类法或强化版本)和更概括的**上层(基础)本体**(描述管理关系及其复合的基本规则和公理)。**核心本体**(Core ontology)试图担当一个通用领域的统一框架,有时桥接不同的领域本体。CIDOC Conceptual Reference Model (CIDOC CRM)是一个广泛应用的来自文化遗产领域的例子。本体可以和形式化定义的、用于处理和组合关系的公理和规则相关,打算供逻辑推理系统使用。因此,它们适合于有良好定义对象和运算的应用。(见第 3.2.3 节)

3.2 按目的分的词表

3.2.1 检索目的 (retrieval purpose)

信息检索 KOS 主要是用来辅助资源检索的,最初来自书目数据库和图书馆目录,现在来自数字图书馆和 Web。设计基本原理是在未来的检索活动中可感知(perceived)的帮助,包括分类和标引、搜索(包括浏览、查询和不同形式的“智能”搜索)、KOS(单语种和多语种)之间的映射、为学习或探索一个主题领域以便改进一个(re)search 问题(定义概念并把它们置于上下文中)。一个 KOS 可以同时用于分类/标引和搜索,或只用于搜索。KOS 可以用来支持手工编目,也可以用来支持自动编目活动。KOS 的变化范围很大,从领域专用的 KOS 到通用分类系统,从两个等级层次到覆盖面极深极广的系统。

3.2.1.1 分类与标引 (classification vs indexing)

分类与标引的区别是重要的,但经常被误解,尤其是在新的 Web 开发中(Lancaster 2003 是一篇好文章)。这两个过程都是将叙词(descriptor)或标记指派给信息资源,都可能涉及有等级结构式概念排列的 KOS。但是,分类寻求的是将相似条目组在一起,而标引寻求的是显示出条目之间的不同,以帮助在搜索时进行区分。分类提供的是一种概观,有助于材料的组织。这种结构方便基于浏览的访问方法,不管是浏览图书馆的书架还是浏览等级式菜单系统。**分类法**(classification scheme)常常与一种符号或编码方案相关,可以产生一种排序,有利于排架和搜索结果的分级。**标引**(例如用叙词表)则寻求对一个条目的内容更具有描述性,而不是将一个条目指派给一个宽泛的类目。叙词表中的叙词可以在搜索过程中被组配。这种区别有时候被比作目录表与书后索引之间的关系。同时,分类系统和叙词表的结构可能非常相似,因为它们都由概念的等级结构组成,它们将倾向于在信息条目应用上的**穷举性**(exhaustivity)和**专指性**(specificity)不同。因此,一个信息条目通常会被来自分类系统的更少量的、更概括的概念分类,相反,会被来自叙词表的若干更专指的概念标引。

有时一个分类和一个标引系统会被组合起来覆盖这两种目的。例如一部带有叙词表的分类法。这可以在浏览界面上提供更多的灵活性,为自动分类和搜索工具提供更丰富的资源。在提供不同的基于分类的过滤器(为基于叙词表的搜索结果)时这种方法也很有用。

3.2.2 语言学目的 (linguistic purposes)

KOS 可以作为各种自然语言文本处理技术(自动的和人工的)的资源来使用,包括机器辅助翻译领域和与命名实体抽取、文本挖掘、摘要相关的语言工程领域。第 6.5 节简要地指出了一些语言技术标准。术语“terminology”常常被自然语言界用来称呼一个用于语言目的的词表。

经常应用于通用目的,语言学资源会倾向于包含精心考虑的用于语言工程目的的关系(在 3.1.4.3 节语言学数据库下有简短的讨论)。但是一些信息检索 KOS 可以被认为含有语言学资源的元素,用于自然语言技术的语用(pragmatic)应用。因此,重要的叙词表和分类

系统都有大量的入口词表（这些术语被认为等同于所构想的使用环境（the envisaged use contexts））。它们可能包含广泛的范畴注释或不同种类的定义，这些都可以被看作是语言学资源。

3.2.3 人工智能（AI）目的——在一个领域中为实体建模

术语“ontology”源自形而上学，一个涉及现实描述的哲学分支（Smith 2003）。它被 AI 知识表示界采纳，虽然 AI 的用法有些不同。本体

...是概念和关系的描述（就像程序的形式化规格说明），可以为一个代理或一群代理而存在。

...

实际上，一个本体的约定是一种使用词表的协议（即问询和进行断言），以一种与一个本体所规定的理论相一致（但不是全部（complete））的方式。我们建立托管本体的代理。我们设计本体以便可以与这些代理或在代理之间共享知识。

...

概念化（conceptualization）是世界的一种抽象的、简化的视角，我们为了某种目的希望用它来表示世界。每个知识库、基于知识的系统或知识级的代理都进行了一些概念化（显式或隐式）。

（Gruber）

正如在 3.1.4.4 节简略提到的，AI 本体是形式化的表示，用准确的定义和关系为一个知识领域建模。它们被设计用于一阶逻辑推理系统，是在（自动）智能代理之间进行交流的一种知识表示机制。它们经常与语义 Web 和数据库模式整合相关。它们适用于具有良好定义对象和运算（operation）的应用，适用于可能就概念（和术语）的精确定义达到共识的应用场景，以及需要定义逻辑规则来处理关系和可能推理出新知识的应用场景。这些应用都倾向于有不同于检索本身的关注焦点，例如 eScience 中的分析元素，或新数据的自动生成。实例可能包括许多科学应用，此时本体是当前可接受科学知识和较小主题领域的一个模型，例如一些商业应用。创建（和维护）形式化表示会有开销，在某些情形下要形成一致同意的、关于某些抽象或有争议概念（如一些人类活动的描述）的确切定义可能不切实际。例如，在搜索应用中，模糊概念“aboutness”是为一篇文献进行标引或分类的基础，与一个事实断言相反，较不形式化的方法可能更适合它。

3.2.4 eLearning 目的

3.4 社会标记法和民间分类表

对社会标记法和民间分类表（以及 Web2.0 的更广阔的前景）的热心宣传令人回想起之前围绕语义 Web 设想和早期元数据首创计划的狂热。需要一种平衡的方法，既认识到之前词表工作的价值，同时又不轻视新的可能性。

在这篇报告和即将到来的 JISC 首创计划的上下文中，有必要研究何种程度的社会标记法和类似特征会有潜力改进主题标引和知识组织，并随后利于资源发现、浏览和搜索。这关

系到可能的新服务和已经存在的服务。另一方面是**社会标记法和民间分类表在创建、升级和维护词表中可能扮演的角色**。

用户参与的关键词标引、分类和元数据提供通常并非被看成是毫无争议的成功。

鼓励开发值得信赖的、可持续的和可免费获取的服务。

3.4.2 上下文

我们必须意识到社会标记法存在于一个对信息系统广泛参与或团体活动的环境下，有时是一个组件，有时是一个可行的（enabling）、重叠的或可选的特征。相关的活动包括：链接，引用，注释，推荐，列表（阅读，选购等），使用普及性调查，用户行为和偏好，来自用户的元数据，协作过滤，社会搜索；小组学习，用户化和个性化（如果与他人共享）。

3.4.4 短处和问题

目前有关社会标记法的出版物（许多在博客上）提供了一长串长处和短处，倾向于狂热鼓吹者以一种积极推销的口吻在写。近来出现了少量来自知识组织、信息科学或语义 web 界的专业圈子的评价性的系统研究。

和传统的知识组织相比，社会标记法重新分配成本，将它们从术语指派转移到发现，或如 Ian Davis（2005）所说，“标记法削平了分类的成本，将其堆积在发现的开销上”。

现有社会标记法系统的一个明显的问题是，它们不是为信息发现和检索而设计的。它们倾向于在同一个应用中组合各种功能，同一个方法常常应用于非常不同的对象类型上：文本、网页、链接表、博客、图片和其他媒体、多媒体等。

最明显和最常提到的缺点是缺乏任何词表控制。对检索性能最为有害的是缺乏最简单的控制（不管是用在输入阶段还是后来被系统用作改进过程），例如单词形式（单数，复数），形态学形式（名词），拼写，数字的使用，字符集合和音译（transliteration）。在搜索时，同义词的某种链接和同形异义词的解疑对于获得可接受的查全率和查准率是非常重要的，对名称表示的控制也一样（首名，姓，名称首字母，昵称）。同样重要的还有地名、日期/时间和缩写词。社会标记法也同样缺乏 KOS 的先进的益处：考虑概念、语义关系和受控的术语映射。

另一个缺点是没有标引/标记的规则：规则关系到穷举性、专指性、粒度、复合构造或上下文提供。标记代表一个个人的上下文（例如我兄弟），可能对公众无用。...

总之，尽管社会标记法可能有其他的好处，目前看来，它并不适合有目标的有效搜索或系统的主题浏览。用检索的术语来讲，社会标记法系统会有很低的查准率和查全率。可能有用的唯一发现方法是偶然发现（serendipity）。当然，在完全没有任何其他的标引和发现功能部件的环境中，它还是有一些作用的。

3.4.5 长处和利益

现有的社会标记法系统的主要的一般的好处包括：

- 在缺乏受控标引信息的地方为许多资源提供标记
- 是热门话题的、发展的、能够快速反映变化的
- 能够无主要信息损失地反映用户语言（有助于维护语言和文化的丰富性）

- 以用户为中心
- 能够了解用户行为

另外，对于有雄心的处理 HE sector 的信息系统，在以下方面也可能有潜在的好处：

- 被其他服务严重忽视的材料和出版物的标引输入，例如图象、混合媒体和多媒体、学习对象和研究数据
- 更小型的合作组，专业主题，交流密集型工作环境，没有合适的词表系统存在的领域
- **为现有词表改进入口级术语学，基于用户词表的可利用和处理**
- 创建新的词表，基于用户词表的可利用和处理
- 用附加的源于相关社会标记法功能部件的数据改进系统/服务
- 通过与 KOS 和主题访问功能组合，可以提供一种不同的和补充的标引层
- 广泛激励开发和研究工作

社会标记法不可能也应该不会取代其他标引和知识组织成果。至少在研究发现目的上，主要的建议是探讨它们的长处，以一种补充的方式使用它们，既为了发现而优化这种系统，也以各种不同的方式将它们与更受控的知识组织和检索系统结合起来。

3.4.6 被建议的开发

提出在三个方向激励开发工作和试验：1) 改进现有的社会标记系统；2) 开放替代标记系统；3) **将社会标记法特征整合进现有的系统和服务**。后两种建议似乎更有前途，应该更优先考虑。

1 改进现有的社会标记系统

- 应该在标记过程中提供更丰富的系统支持，例如关键词抽取和建议、词典查找（现有的对象、作者、加标记者、标记）、**交互式术语解疑和可视化**。
- 初始标记被用户执行之后，系统应该进行标记改进（如订正错误拼写、指定语言、**适当而一致地处理复合词、链接同义词、创建部分等级结构、创建分面、（见 fac.etio.us）**）
- 可以通过先进的分簇技术改进搜索和浏览功能，探索共现、其他聚合、过滤、分级和可视化技术来支持导航。

2 建立替代的标记系统

为了激励创新、引导和改进的实践，可以支持有不同方法的替代标记系统的创建，例如：

- 创建一个优化了发现和检索的社会标记法系统
- 建立一个具有良好定义的用户组和目的、用于同类服务的系统
- 探索受控词表在标记法系统中的系统化使用
- **将图书馆和发现服务与一个社会标记法系统挂钩，以 OCLC 对搜索引擎和机构知识库所采用的方式（Google 中的 OpenWorldCat，DSpace 中的名称规范档服务）。**

这个开发方向在 Nature Publishing Company 的研究工作中有一个相似物，创建一个标记法系统，专用于基于 Connotea 的书目信息和参考管理（Lund et al. 2005）。

3 将社会标记法特征整合进现有的系统和服务

意识到社会标记法的好处是将用户的贡献整合到现有的信息系统和服务中是一项很重

的成果。这是所有种类的参与方法应有的一个特殊案例。在下面列出的例子中，我们很窄地把重点放在标记的提供上。将不同类型的用户参与附加和整合到建立的服务中可能会有同等的回报。

最有前途的选项有：

- **OPAC**（用户图书馆资料的 Online Public Access Catalogue）：在这方面已经可以看到一些研究活动了，例如 PennPal 或来自 Casey Bisson 的 OPACi 原型。紧接着的行动将会允许用户为书目记录增加标记，然后将它们用于不同的“视角”、链接等。其他的选项会由 Amazon 促成，例如作为 Open WorldCat 评论。

- **主题网关，Intute 集线器**：用户标记可以用来为最后的包含告知资源选择以及支持改进的主题访问（如上述 OPAC 中的用法）。用户注释已经在以前尝试过了，例如在 SOSIG Grapevine。

- **目录**：早期用户已在 Yahoo 目录中创建和开垦了类目（category）（dmoz 和 Open Directory 成果）。Yahoo Social Systems 有雄心要在目录上做更多的文章。

- **主题知识库**：这种情况类似于 OPAC 和主题网关。对资源选择、新标引词表的获得以及在目前没有概念结构和类目的地方创建概念结构和类目会有好处。

- **引用服务**：CiteSeer 已经提供了用户修正选项；名称解疑和内容标记增加将是其他的替代选项。

- **数字图书馆**：类似于 OPAC，在没有结构的地方增创结构。与其他参与研究成果整合也是有前途的。

- **搜索引擎**：大型商用搜索引擎已在此领域活跃起来，例如 Yahoo 和 Google Answer，Yahoo 建立了它的完整品牌成为一个“社会系统”。专业的、学术的和地方的搜索引擎也会极大地受益于类似的方法，也许与传统的 KOS 整合。像 Vivisimo 这样的系统可以用社会标记法来标记簇（cluster）。

- **KOS 创建和开发系统**可以受益于更广阔的用户输入和对来自其他系统的标记的重用。**DDC 编辑系统已开放给更广阔的专家参与。**

- **博物馆在线交互式展览和实物目录**：这方面有一些项目很活跃，例如 Steve.museum 和剑桥大学人类学博物馆的 ED2 项目，主要是邀请用户描述艺术品部件和标记用户体验。（Trant 2006）。

- **元数据增强服务**：和若干上述的服务一样，考虑标引的丰富性、解疑和错误修正。

- **博客、消息服务和 RSS feeds**：这些服务生来就是基于用户贡献的。但还可以构想更多更先进的标记用法。

另外，所有这些基于社会标记法的服务都可能提供不同的和替代的资源 and 搜索结果视角或层次、共现分簇或自动链接系统中的标记到外部资源（例如 flickr 和 delicious？）。

3.4.7 研究

在社会标记法上下文中，有许多重要的方面需要有与服务开发并行的研究工作。对于现有系统和之后新系统和改进系统的创建来说，用法的各方面和收益需要进行系统研究。目前似乎只做了很少的研究。更专业的研究课题有：

- 研究新的开发和更改，将新方法 with 现有方法比较，考虑整合进遗产系统。

- 开发适当的体系结构和研究可伸缩的方面。

- 研究工具和用户界面以及它们对标记法实践的影响，被引导（guided）标记法的效果。

- 考虑与标记法和导航相关的用户行为：标记法实践，社会环境的影响和相关/流行标记显示的影响。

- 分析发射 (emanating) “tag-space” 的结构, 分析在所建立 KOS 中的标记的覆盖程度和新颖比例, 与确定的术语学创建方法相比较。
- 研究已声明的术语学交叉 (convergence), 大规模影响/智能 (mass effects/intelligence) (参见 Golder and Huberman 2006)。
- 研究异构标记集整合的可能性和利益
- 研究潜在的标记法标准的好处
- 与作者和有关发现改进的专业标引相比较, 研究检索性能
- 考虑标记法的潜在的“社会利益” (social benefits)

3.5 构建和使用词表的最佳实践准则

用于不同种类词表设计的最佳实践准则提供实践帮助。Aichison et al.(2000)是英国叙词表设计和构建的标准参考书, 而 Willpower 网站提供有用的实践指南和一批商业软件。BSI 和 NISO 标准文件描述了它们各自的标准, 也提供最佳实际设计准则, 它们的范围现在已扩大到叙词表之外。...

Rosenfeld and Morville (2002) 是一部广泛使用的建立在各种词表之上的信息架构和网站设计技术教科书。它包含有关管理、ROI 和案例研究的章节。Daniels and Busch (2005a, 2005b) 从商业 DC 的观点给出了详尽的最佳 (和最差) 实践准则和 ROI 考虑因素。TASI 给出了在一个元数据框架内采用词表的介绍。**JISC CO-ODE Project (第 5.1.2 节)** 提供了有关本体设计的工具和指南。GovTalk archive 提供了词表软件的设计/选择标准。最近出版的 e-Government Metadata Standard (Version 3.1) 建议将 **Integrated Public Sector Vocabulary (IPSV)** 作为其主题元素的强制 (标准)。也可以加入来自其他受控词表的概念 (必须声明该词表)。IPSV 提供完整版和精简版, CSV、XML、RDFS 和其他格式版本, 同时还提供了各种标记指南和一般使用指南。

3.6 对词表的网络访问

为了实现人的 Web 访问 (human Web access) 和机器-机器访问 (m2m access), 有必要去发现合适的术语词表 (terminology), 并且一经发现, 即评估、导航和查询该术语词表。

许多术语词表和叙词表现在都可以通过网络利用, 打算供人使用。但是, **KOS 的更广泛使用及其与应用的整合 (以自动方式) 都需要 m2m 访问**。有许多标准涌现出来, 采用共同的标准表示和访问词表, 见第 6 节的概述。**对于激励现有术语词表的更广泛应用也已经迈出了一些重要步骤**。例如, **OCLC Terminology Services** 项目 (5.3.6 节) 最近在其网站上提供了约十几部 MARC21 格式的词表用于 XML 格式的规范数据 (some dozen vocabularies in the MARC21 Format for Authority Data in XML), 以及 DDC 的 22 个大类 (summary)。

有许多基于 web 的人类可读词表资源列表 (商业的和免费的), 有几个详细列在参考书目中。值得注意的例子包括下面几个。JISC HILT 项目已编辑了一份叙词表的完全目录 (AtoZ), JISC Pedagogical Vocabularies Project 也已编辑了一张教育叙词表清单, TASI (the Technical Advisory Service for Images) 网站列出了叙词表、分类法和规范档清单, 并附有在一个元数据框架内如何使用它们的介绍。Species 2000 网站维护着一份在线分类 (taxonomic) 数据库一览表。**Text Mining Centre** 有一份生物学本体清单。Middleton's Controlled Vocabulary List 包含主题标题列表、叙词表和分类法。WWW 上的 University of British Columbia's Indexing Resources 有一份分类法列表以及其他的词表。Dextre Clarke 的 Taxonomies in the Public Sector 的广泛评论也包括了一些有关潜在利益以及软件 and 设计的原始资料。

Synapse 提供了分类表、叙词表、分类法和规范文档的 Taxonomy Warehouse, 通过类目

(category) 进行组织。这包括到在线词表的链接以及他们自己的转换和与其它软件打包的“增值实现服务”。University of Toronto 维护着主题分析系统集合 (Subject Analysis System (SAS) Collection), 作为“许多不同主题领域的主题分类法和受控词表北美交换所”。MDA (前身是 Museum Documentation Association) 网站有一个由其资助和在线出版的文化遗产词表的 Terminology Bank。

从长远来看, 我们希望这种人类可读的列表可以作为“注册处”来维护, 虽然必须解决谁来维护它们的问题。见 3.7 节查看注册处(以及 3.2.4 节查看 the Becta Vocabulary Bank)。

3.7 术语学注册处 (Terminology Registries)

在教育领域内有许多注册处首创计划, 设计用于支持一种面向服务的组件开发方法, 提供“查看”功能。这些注册处提供对各种被注册数据的程序访问。这种注册处包括**服务和集合描述注册处** (service and collection description registries, 例如 Ockham, JISC IESR), **事务服务注册处** (transaction service Registries, 一些采用 UDDI, 主要在 eScience 界, 例如 GRIMOIRES), **映射和“人行横道”注册处** (registries of mapping and crosswalks, OCLC crosswalks Registry——见 3.3.4 节的 Metadata Switch Project), 以及**元数据方案注册处** (metadata schema Registries, JISC IEMSR, The European Library, DART)。除了 NSDL Schema Registry 之外, 已有少量的与术语学词表相关的研究活动, **NSDL Schema Registry 计划既登记元数据方案, 也登记用于 NSDL 内的与 KOS 相关的受控词表。**

一般而言, 注册处能够使已注册对象发现、导航、访问和重用。与术语学词表相关时这种注册处也许会采用不同的方法, 取决于它们被设计交付的功能, 是登记词表的描述, 登记单个的术语和概念, 还是登记在领域或学科中的用法。基于术语学词表的服务 (例如解疑服务、查询扩展、映射) 也许也可以注册, 在一个专业的术语学服务注册处内, 或在一个有更宽泛的包含物呈交 (inclusion remit) 的服务注册处内 (例如 JISC IESR)。

提供对术语学词表信息和术语学服务信息的 m2m 访问, 可以鼓励现有词表的开发, 并激活与来自“其他领域”的应用的创新连接。需要建立覆盖状态、持久性、标识和质量的政策。

注册处可以或多或少**集中或分散**, 取决于政策、经费操纵者以及技术设计决定。在重用软件和交换数据方面都有与其他国际首创计划合作的空间。要实现交互工作和数据交换, 注册处本身必须是标准兼容的, 即使该领域中的标准并不成熟, 特别是关于数据描述的标准。ISO/IEC 11179 标准在这有些关联 (特别是 XMDR 项目更向前一步注册更复杂的结构), 但是这种标准制定活动背后的驱动力受到的数据库和数据词典技术的影响, 要超过基于 web 服务的语义互操作增强对它的影响, 而后者是 JISC 团体更感兴趣的重点。

需要仔细考虑注册处的成本效益。当一些基金组织 (例如 JISC) 可能将注册处当作一种方法来识别和推动可用于其团体的服务时, 也有必要为专业术语学词表的提供者和有贡献于注册处的服务调查 ROI (投资回报)。若干较大型的词表已经有了商业化的业务模型, 并且 **m2m 使用会引起有关管理 IPR 和版权的问题** (即使对于较小的基于词表的团体)。可以有各种各样的业务模型, 但和其他“共享服务”一样, 有时并不清楚谁是显而易见的资助者。其他业务问题包括: **澄清谁拥有注册处的内容, 谁负责将词表内容转换成机器可读的结构, KOS 拥有者有无商业动机“一起工作”在互操作注册处环境中。**

4 与 TS 有关的研究活动

4.1 信息搜寻行为的研究和模型

在设计和评估 TS 时，为了降低设计出错的范围和提高服务可以真正被使用的可能性，考虑人们如何搜索信息是非常重要的。这是一个困难和复杂的领域，已有相当多的研究搜索行为的文献和专用会议（例如 Information Seeking in Context (ISIC 2006, <http://www.hss.uts.edu.au/isic2006/>) 涌现。这种研究提供了对发现策略、用户需求和用户环境的可能了解。同时要记住，环境和个体特征的差异（例如培训和动机）可能很重要。这些研究可能是设计未来开发和评估方法学的有用资源。

术语**信息搜寻** (information seeking) 通常指一个信息需求的更宽泛的上下文，而**信息搜索** (information searching) 表示为了某个特定的搜索与计算机的交互，虽然这种区别有时会变得很模糊 (Marchionini, 1995; Spink et al., 2002; Wilson, 1999)。

已经发现术语学支持对查全率和查准率都有潜在的帮助。在一项对搜索媒介在线行为的综合研究中，Fidel 的发现 (1991) 支持在自由文本检索利用术语学支持。根据环境两种检索模式都用来提高查全率或查准率，得出的结论是：有必要使用良好设计的叙词表和相关工具。

信息搜寻模型 (information seeking models)，例如 Choo et al. (2000), Kuhlthau (1991) 和 Saracevic (1997)，提供通用的信息搜寻行为框架，可以支持更高端的设计目标。Ellis (1989) 批评了限制性的受控实验室评估假设，赞成采用一种**经验的、行为的方法进行信息搜寻研究**。这会将重点引至**基础信息搜寻模式**上，例如**浏览、链接 (chaining)、监控 (monitoring)** 等。Soergel (1994) 强调需要在评估中考虑**标引、系统和用户因素**的完整上下文。Kuhlthau (1991) 和 Marchionini (1995) 的模型描绘了信息搜索过程的基础阶段，从**问题定义、查询式构造和执行、结果检查**方面来考虑。**Blocks et al. (2006)** 提供了叙词表辅助搜索各阶段的**低端模型**，打算用作系统开发者的实际参考模型。

搜索行为研究通常将搜索行为揭示为一个迭代 (iterative) 过程。Bates 的颇具影响力的“**挑选浆果 (berry-picking)**”搜索模型 (1989) 强调了一种**进化的搜索**，与静态信息需求模型 (优化单个查询式) 形成比较。她发现在许多案例中，用户的信息需求会随着与遇到的资料进行交互中的阶段进展 (**session progresses**) 而演化。

这就产生了一种需求：考虑**交互式 TS 和自动 TS 之间的适当平衡**。例如，术语学支持查询扩展 (QE) 的系统控制和用户控制之间的平衡已经成为许多研究的主题。各种 Okapi 计划对实际运行环境 (operational settings) 中的基于叙词表的 QE 进行了大量的实验，作为一种概率查询模型的一部分 (Beaulieu, 1997)。这些实验的范围从全自动 QE 到交互式 QE。他们的结论赞成**自动和交互控制的平衡**，赞成**显性使用叙词表而非隐性使用叙词表**。其他考虑用户-系统平衡的经验性研究包括 Jones et al (1995), Shiri & Revie (2006), Vakkari et al. (2004)，以及 **Greeberg (2001)**，她比较了不同叙词表关系在自动和交互查询扩展中的性能。她赞成智能系统应该考虑 (**进化的**) 用户检索目标。

研究已经指出了策略 (strategic) 或概念 (conceptual) 支持的重要性 (例如 Brajnik et al., 1996; Fidel 1995)。Bates (1979) 和 Fidel (1985) 确定了大量分别被专业搜索者**用来修改或重构检索式的战术 (tactics) 或迁移 (moves)**，例如迁移到一个上位词或相关词。Bates (1990) 讨论了在一个最终用户控制搜索步骤的框架内，系统支持不同层次粒度上的搜索活动的可能性。她认为当前界面难以使用的一个原因是，**这些界面并不是围绕能够推动战略性搜索目标的典型搜索行为来设计的**。她特别指出，研究要针对最终用户搜索的中间层战术战略的系统支持，而不是针对基础迁移和高层战略的系统支持。

4.2 与 TS 相关的信息生命周期

这里展示的模型利用了 DELOS (可交付使用的 D5.3.1, 第 3.2.2 节) 数字图书馆语义互操作发展动态评论 (DELOS state of the art review of semantic interoperability in Digital

Libraries) (Patel et al. 2005) 中描述的信息生命周期管理模型, 它综合了来自知识表示和信息科学领域的生命周期模型。DELOS 生命周期模型的应用和扩展与 TS 的两个不同方面有关: a) 词表本身作为一个实体, b) 一个词表的元素被用作一个信息系统的一部分 (例如提供搜索术语)。对于后者, 一个搜索系统在使用 TS 时可能是“有术语学意识的 (terminology-aware)”, 或者它也许只是为了其目的简单地将术语学元素当成是一个非受控术语的来源。这两个方面在修改后的框架中非正式地组合在一起——应该强调有可能还有其他的元素配置和选择。目的是要为考虑 TS 应用的范围提供一个启发式的、统一的框架。在 5.4 节中, 一些项目被粗略地安放在这个宽泛的框架内。

图 1 中给出了 TS 生命周期框架。这里的**创建 (Creation)**指的是一个词表的产生, 而**获得 (Acquisition)**指的是该词表与一个集合或某种注册处整合的阶段。**标识 (Identification)** (在编目之下考虑) 为一个词表或词表元素提供一个唯一的键值 (见 6.3 节)。**集成 (Integration)**在第 4.4 节讨论。**访问、搜索和发现**比在 DELOS 版本中处理得更详细, 因为它们是 TS 的重点。当然, 其他元素也与这篇评论相关, 获得、维护和存档更次要些。注意, 生命周期可能涉及创建者/作者、出版者、信息系统管理者、服务提供者和不同种类的最终用户。更广泛地, 这个生命周期模型与更宽泛的信息搜寻行为模型 (见第 4.1 节) 和学术性生命周期模型 (Lyon 2003) 连接或交叠。要注意实际上许多涉及 TS 的用户活动都是一个迭代的过程。

词表的创建和修改 (Creation and modification of vocabularies)

创建/维持词表

词表的出版 (Publication of vocabularies)

批准 (Licensing)

词表的获得 (Acquisition of vocabularies)

选择, 存储 (storage)

编目(元数据, 标识/命名, 注册) (Cataloguing (metadata, identification/naming, Registration))

标引/分类/注释 (Indexing/classification/annotation)

人工, 半自动, 自动, 解疑 (Intellectual, semi-automatic, automatic, disambiguation)

集成 (语法和语义互操作问题) (Integration (syntactic and semantic interoperability issues))

映射 (mapping), 合并 (merging)

媒介 (Mediation)

用户界面 (TS 出现在界面上)

分面 (Faceted) /空间 (Spatial) /能提供的其他访问

词表的浏览和可视化 (Browsing and visualization)

个性化 (Personalization) (界面的或词表的)

访问, 搜索和发现 (Access, search and discovery)

服务、词表 (和概念)、数据库/集合的发现

搜索和检索 (Search and retrieval)

查询 (Querying)

有 TS 意识的 (TS-aware) 查询

不同种类的词表都可以作为查询术语的来源。有 TS 意识的 (TS-aware) 查询系统的意思是某些应用是由词表构成的 (通过搜索软件), 对立于没有考虑任何词表因素的搜索系统。例子包括: 基于概念类型的分字段搜索, 基于 KOS 语义关系的查询扩展, (半) 自动解疑 (在将初始的自由文本术语转化成受控词表概念时) 等等。

查询扩展 (Query expansion) — 同义词, 语义

跨越分布式集合的交叉搜索 (cross-searching), 交叉浏览 (cross-browsing)

使用（作为一个更宽泛服务的一部分）

搜索+分析 应用，

信息抽取（Information extraction），挖掘（mining）

词表维护（Maintenance）

演化（Evolution），版本升级（versioning）

特定应用（Application specific），协同工作（collaborative work）

词表的存档和保存（Archiving and preservation）

图 1 TS 生命周期框架

4.3 术语学网络服务的类型

本节讨论语 JISC Information Environment 和 e-Framework 有关的术语学服务。我们首先将术语学服务当作抽象服务来讨论，然后专门讨论服务的等级层次和术语学网络服务（Terminology Web Services）。

在 JISC Information Environment 的面向服务视角的上下文中，信息服务（至少有一部分）可以通过网络服务来提供给软件应用所使用，有必要扩展术语学相关服务的范围使其超越目前文献上所见的少数几个首创例子（例如 Andy Powell（Powell 2005b, Nov）。术语学服务也在其他面向服务的体系结构和 e-framework 首创计划中提到，例如美国的 DLF Abstract Services Taskforce（DLF）和英国的用于教育和研究的 e-Framework（e-Framework），包括其组件 E-Learning Framework（ELF）。再一次，TS 并没有被详细描绘，本节的目标之一就是勾勒一些不同种类的 TS，以激励本领域中更详细的设计工作。

在服务组件层，如 JISC IE 体系结构图所示（Powell 2005b, Nov. 图 1），一组服务构成了术语学服务组件。下面所列出的四项是抽象服务：**词表搜索界面**，**词表获得（harvest）界面**，**词表存储（deposit）界面**和**术语学服务**（映射和扩展术语）。（新闻频道和删除界面也提到了，但没有细节）。相关的词表标准和协议应该包含在抽象服务的任何绑定中。但是并不清楚这些服务要承担的确切任务。

4.3.1 术语学网络服务（Terminology Web Services）的定义

下面是术语学网络服务的更全面的定义：

术语学服务（Terminology Services）是一组抽象服务，展示和应用词表、它们的成员概念、术语和关系，描述术语的含义，以及促进语义互操作。这样做的目的是为了搜索、浏览、发现、翻译、映射、语义推理、主题标引和分类、获得（harvesting）、警告（alerting）等。

潜在地，可能要加上支持术语学创建、存储和管理的抽象服务，例如存储（deposit）、管理、编辑、删除。它们可能会和展示和应用词表的服务部分重叠，但在此阶段没有包含在定义和下面的例子中。

注意在此上下文中，词表包括 3.1 节中描述的不同种类的受控词表，并且还包含额外的来自映射服务的被映射（或被翻译）术语和概念。非控词表，例如非控术语列表、作者提供的关键词、标记集、民间分类法也应该包含来服务于术语学服务目的。

在不同的粒度程度上有服务的分层。在最底层，**绑定**（bindings）是一个抽象服务的特定实例化，在适当的时候给出特定的数据表示、一个 API 和 **Web Service** 规范（如果采用这种体系结构）。在更高层，抽象服务将构成更宽泛应用服务或 JISC Service Components 的一

部分。抽象服务可能包括分层的 TS，例如一个搜索界面或获得界面可能从一个较低层的 TS 中获得关键词。

对于每个服务，应用不同的标准和协议。它们通常落在以下两种环境中：

- A) 关于词表系统/方案本身的描述、结构和功能的标准（例如 SKOS Core, VDEX, Zthes, SKOS API, BSI 和 NISO 标准等）。见第 6 节描述的各种词表相关标准。
- B) 关于 TS 应用环境的标准：搜索，获得，警告和其他抽象服务。

这里没有考虑所有可能的绑定和服务组合。一些关键的用于不同类型术语学服务的参考实现将会是有用的未来开发。

作为起步，下节将提供所选 TS 的更多细节。它建立在前面成果之上并对其进行了扩展，用更宽的笔触处理术语学服务（如 Powell 2005b, Nov）。但是它仍然远未完善。...

4.3.2 抽象术语学服务的分组（和分层）

下面介绍三大组抽象术语学服务（第三组较粗略）。为了说明术语学服务的等级分层是必要的，详细列出了用于某些相关案例的较低层的术语学服务选项（斜体）。这些会被称为较低层服务，作为实现的一部分。SKOS API（见 6.4.1 节）用来表示低层术语学服务。有一个 Java Web Service 绑定，但也是表示为一个独立于绑定的协议（从而也可能有一个 HTTP 实现，例如）。在下面的列表中，它们可能会被认为（或多或少）既是一个低层术语学服务的规格说明，又是其可能的绑定。同一低层术语学服务的另一绑定将是可能的。也展示了一些较低层 OAI 获得服务的例子。

第一组关注与整个词表方案/系统相关的抽象服务。倾向于忽略发现（识别）一个合适的词表和检索与其相关的元数据，与检索一个词表的成员概念和术语之间的差别。有关完整词表方案及其元数据的服务，与有关成员术语的服务（下面的第二组）之间的区别是根本性的，与有关其他服务中的术语学的应用（或多或少无缝）的服务之间的区别也是如此。

例如，一个描述为“词表获得界面”（Powell 2005b, Nov）的抽象服务并没有指明它是只和单个术语和有关词表的其他信息的“获得”相关呢，还是与术语组的获得相关，或者也许是整个词表的获得。（最后一种情况倾向于是例外，考虑到当前与词表相关的实践和权利状况。）

以下描述的另一个意图是要说明术语学词表除了包含术语还可能包含实体：即概念和关系，并且服务可以为来自一个或若干不同词表的这种实体工作。

1 与词表（编码）系统/方案相关的服务

11 在词表注册处中发现合适的方案

以及：

搜索，浏览，获得，警告，上载/存储，编辑等

12 揭示有关方案的所选或完整的信息（元数据）

getSupportedSemanticRelationsByThesaurus(URI thesaurus)

(获得叙词表支持的语义关系（叙词表 URI）)

13 统计（例如有关大小和使用等级的信息）

2 与一个或若干词表系统中的成员术语/概念/关系相关的服务

21 发现/搜索成员

术语/概念/关系/翻译/映射/结构

[(权限) 查询]在一个或若干词表系统中

以及:

搜索, 浏览, 获得, 警告, 上载/存储, 编辑等

getConceptsMatchingKeywordByThesaurus(keyword, URI thesaurus)

(获得匹配关键词的概念)

getConceptsMatchingRegex(regex)

(获得匹配概念注册表信息文件?)

22 揭示或获得术语/概念/关系/翻译/映射/结构 (已知条目)

getConcept(uri)

getConceptByExternalID(externalID, URI thesaurus)

getConceptByPreferredLabel(preferredLabel, URI thesaurus)

23 在术语学网络中浏览

24 揭示术语/概念/关系的拓扑环境子集; 若干术语/概念和它们之间的语义关系; 一个合集 (synset); 翻译; 映射; 等级结构子集; 一个概念和所有相关术语; 一个分类和所有相关信息; 一个术语和所有翻译; 等等。

getSupportedSemanticRelationsByThesaurus(URI thesaurus)

getAllConceptRelativesByThesaurus(concept, URI thesaurus)

getAllConceptRelativesByThesaurus(concept, relation, URI thesaurus)

getAllConceptRelativesByPath(concept, relation, URI thesaurus, int distance)

gettopConcepts(concept, URI thesaurus)

gettopmostConcepts(URI thesaurus)

(OAI) harvest of sets

25 获得 一个完整的词表/映射集

(OAI) complete harvesting function

26 上载/存储 成员术语/概念/关系/翻译/映射

27 编辑 成员术语/概念/关系/翻译/映射

28 警告 新的或修改过的

成员术语/概念/关系/翻译/映射

3 与其它服务中的术语学应用相关的服务

(粗略例示)

31 自动标引

32 术语或关键短语抽取

33 命名实体识别, 数据挖掘

34 术语或文献的自动翻译

35 查询增强, 查询扩展

- 36 自动分类
- 37 自动映射
- 38 语义推理

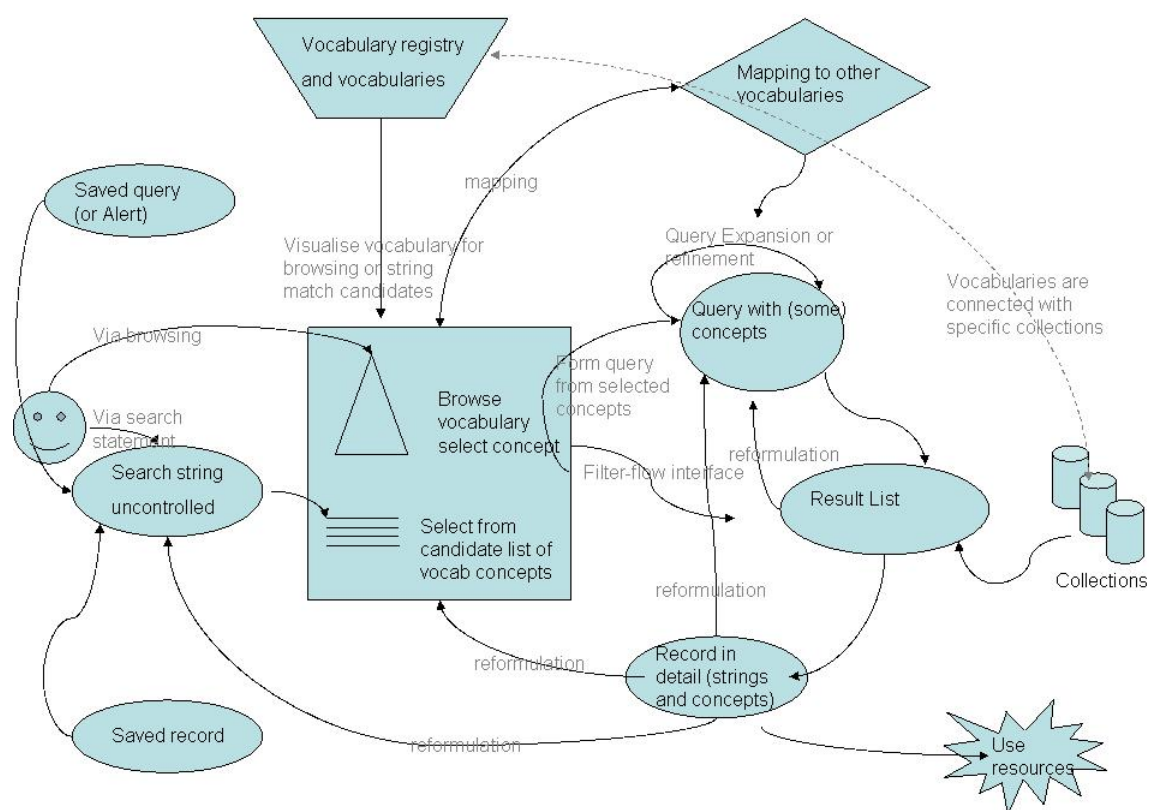
...

被整合进一个信息服务的词表搜索和浏览功能，和用于这些目的的外部术语学（即上面第二组中的抽象术语学服务）并无本质区别。在技术上，它们可以作为（web）服务来调用。甚至来自一个受控词表的术语建议也可以通过发送一个适当请求给一个 web service（例如 21，22 或 24）来完成。

被完全整合进一个数据或文献集合的词表（到了没有独立表示或访问的程度），是否可以是可用网络服务的基础，这一点值得怀疑。但是，对于整合的、反应迅速的搜索和浏览界面来说，外部网络服务调用的性能是一个关键问题。是否以及哪种网络服务体系结构将提供这种响应尚待分晓。在一些圈子中，对（较）轻量级的网络服务方法感兴趣，因为考虑到基于 SOAP 的交互式 Web 应用方法所施加的开销，此时响应时间是一个问题。这一点也在第 6.4.1 节中提到（有关专用于面向用户术语学服务的协议的讨论）。下一节给出了一个分层术语学服务（来自上面第二组）的图例，在术语学辅助搜索环境中。

4.3.3 TS 辅助搜索过程图解

本节给出了有术语学意识的搜索过程的一些详细步骤分解，作为在搜索和该过程的迭代本性中可能涉及的服务分层的进一步图解。图 2 试图将搜索和浏览操作组合起来。它假设词表是在一个注册处中，并与特定的集合相关。一项初始的用户搜索陈述需要被表示为受控术语学，**经浏览或将初始术语转换为受控术语**（必要时解疑）。在一些案例中，**浏览可能触发一个查询**，否则查询式由来自词表的概念构成。可以为查询式重构目的检查结果，并可以应用不同形式的查询扩展。被保存的查询式或结果可以构成重新开始此过程的基础。该图是参考叙词表搜索模型（Blocks et al. (2006) 讨论的）的简化版本。最终目标是要将前一节中概述的较低层术语学服务绘制成该图的搜索过程“工作流”。



4.3.4 术语学网络服务评论

本节简单回顾目前在具有网络服务绑定的术语学服务方面的工作。

OCLC Terminology Services 项目在 5.3.6 节中描述。这里简要评论其突出方面。经由一系列网络服务(基于带 **Zthes profile** 的 **SRW/SRU**, **MS Research Pane** 和 **REST** 以及 **SOAP** 协议)可以利用多个词表。编码格式是 **HTML**, **MARC-XML** 和 **MARC21 Authorities** 或 **Classification** 格式。其它的 OCLC 网络服务是 Name Authority Service(作为 DSpace 的附件), 以及 LAF (LC Authority File) 网络服务, 它是在一个与 ePrint UK 计划(参见 3.3 节)的协作环境中开发的(LC Name)。Vizine-Goetz (2003) 讨论了 OCLC 映射服务, 同时 Vizine-Goetz et al.(2006) 描述了他们的 SOA 体系结构和 Microsoft 的 Research Pane 的使用。

HILT3 计划开发基于 SRW 和 SKOS Core 的网络服务(见 5.1.4 节)。也参见 3.2.4 节中的 Becta Vocabulary Studio, 在其 Vocabulary Bank 中提供了一个网络服务。

各种网络服务研究活动构成了一个更宽泛 ECOinformatics Initiative (<http://ecoinfo.eionet.europa.eu/>) 的一部分, 该首创计划已注意到 Environmental Thesaurus 和 Terminology Workshops。CSA/NBII Biocomplexity Thesaurus Web Services 计划已改编了 SKOS API 用来搜索和浏览 Biocomplexity Thesaurus。它还没有完成所有的功能, 但已经加上了一个有效的关键词搜索元素(CSA/NBII)。还有一个多语种示范器可用。GEMET 多语种环境叙词表由 EKOLab 采用 T-REKS (Thesaurus-based Reference Environmental Knowledge System) 模型开发。EIONET (European Environment Information and Observation Network) 已经为访问 GENET 的扩展版本开发了一个网络服务 API, 采用了一个 SKOS 表示(GEMET)。Tudhope and Binding (2006) 讨论了应用一个网络服务浏览器的初步实践, 采用了 SKOS API 函数的受限集合。

The California Environmental Digital Library Network (CalEDLN) 使用了一个网络服务

API, 用于 **SKOS 编码词表与元数据编辑和搜索界面之间的交互**。该 API 支持分布式词表的访问和浏览、用于词干 (thematic) 关键词的术语匹配和叙词表功能、加利福尼亚地名、资源类型以及人名/机构名。该软件解决方案是开源的 (CalEDLN)。这是在 6.5.1 节中提到的 CERES Protocol 工作的进一步发展。

DLESE 已开发了一个 NSDL Strand Map Service, 作为一个用于 K-12 Education 的 Networked Knowledge Organization 和可视化系统(基于 **REST 的网络服务协议产生可视化**), 采用了一种强化的最终用户叙词表的形式。它邀请教师和学习者在重要想法之间进行连接 (交流?)。(Sumner 2005)

MelvilSoap 是来自 Die Deutsche Bibliothec 的一个网络服务, 提供德文的 DDC 22。它允许用户通过一个 **SOAP 界面** 查询 Melvil (德语的 WebDewey)。它是打算用于分类工作和支持带有 DDC 分类号的德文数据库搜索的。该**服务要求预定** (Melvil)。

MeSHine (EUTROPA.doc) 是一个使用德语和其它语言 MeSH 来搜索 Internet (**Google web-APIs**) 和 Pubmed/Medline (采用 **NLM 的 Entrez-SOAP 实用程序** – 参见 5.2.5 节) 的网络服务; 可以开发对其它词表的 **SOAP/REST** 访问。词表是用 **XML 通讯 (messages)** 格式构造的 (MeSHine)。

Zisman et al.(2002)讨论了用“信息总线”方法将网络服务包装器应用于一个原型系统开发的经验, 该原型系统将具有完全不同组织和结构的各种 UN FAO 数据源整合在一起。

当前有些工作希望**更彻底地将网络服务与语义 Web 成果集成**。有一个 W3C Semantic Web Services Interest Group, 具有增强的语义描述和网络服务编排 (choreography)。正在进行的 Web Services and Semantics Project (IST-FP6-004308) 是在此通用领域中的一项 EC 特别支持行动。

4.4 映射 (mapping)

映射是在异构环境中实现语义互操作的一个关键需求。虽然 schemas、框架和工具能够有所帮助, 但**概念层次上的详细的映射工作是必要的, 需要结合人工工作和自动化的辅助**。要获得有用的成果需要相当大的工作量。

虽然存在一些主要的集成词表 (例如 UMLS, GEMET), 但常常出现**有目的建立的集成词表的构建不实用**的情况。因此, 为了方便对不同环境、不同目的和用于不同用户群体的信息资源的访问, 词表之间的映射是重要的。

但是有时候可能可以映射到一个合适的**转换词表**。已经有一些项目的经验, 例如 HILT 和 Renardus, 映射到 DDC (作为一个中脊 (central spine), 见第 5 节)。The Becta Vocabulary Studio 动态地将 Vocabulary Bank 中的术语映射到一个概念中脊上 (见 3.2.4 节)。也有正在进行的研究利用顶层核心本体作为不同领域词表和宽泛领域内异构数据集之间的一体化框架 (例如见 Doerr et al. 2003 和 CIDOC CRM, 为博物馆界开发并被扩展到图书馆和档案馆)。也有可能是用语言资源, 例如词汇数据库 (如 WordNet) 和语言学本体 (例如 Navaretta et al. 2006) 来辅助映射工作。

The DELOS Report D5.3.1: Semantic Interoperability in Digital Library Systems (Patel et al. 2005, 第 6 节)更详细地讨论了这些问题。它比较了**信息科学和基于本体的映射方法论途径** (methodological approaches), 结论是它们在本质上非常类似 (intrinsically fairly similar)。HILT 项目报告也回顾了不同的映射方法。

Zeng and Chan (2004) 提供了广泛的映射工作近期回顾。他们确定了若干方法论选项, 其中重要的有:

- a) 从一个现有的复杂词表**派生 (Derivation) / 建模 (Modeling)** 出一个专业的或较简单的词表。例如, 通过对映射中难于处理的复杂、先组主题标题进行解构和系统

化，**分面分析**可以在方便语义互操作中扮演重要角色，OCLC FAST 项目（FAST）已经通过一个简化的语法将 **LCSH 标题转换成分面表示**。

- b) 用一种不同的语言**翻译（Translation）/改编（Adaptation）**一部词表。
- c) 从一部专业叙词表到一部大型、通用叙词表的**附属和叶子节点链接（Satellite and Leaf Node Linking）**。这是一种用更专指的本地概念和术语增强一个广泛应用的综合词表的**效能成本合算的方法**。附属物可以被综合叙词表的编辑采纳，同时它们应该总是被标识为本地（local）。
- d) 不同受控词表中的概念之间**直接映射（Direct Mapping）**，通常用一种**人工检查（intellectual review）**的方法。
- e) 基于在条目标引中的共同出现建立两个词表之间的**共现映射（Co-Occurrence Mapping）**。共现映射被认为比专家所做的直接映射更松散。
- f) 用**转换（Switching）**语言作为中介。它可能是一个新系统（例如 UMLS Metathesaurus）或一个已存在的系统。这是最常使用的方法之一。例如参见 DDC 在 HILT 和 Renardus 项目中的使用。

当然，也有这些方法的变体和组合。实际上，一个词表映射操作的成功（和成本）倾向于依赖**被映射词表的相合性（congruence）**。相关因素包括：**重叠程度，先组/后组程度，结构相似性，专指程度相似性，目标应用和使用环境**（更多细节参见第 6.2.1 节 Patel et al.2005 中的讨论）。

有一些重要的项目已在此领域有所动作——见第 5 节的 HILT、OCLC、Renardus 项目的概述。UN FAO 正在将资源投入到与 Agrovoc 叙词表的映射工作中（Liang and Fini 2006）。首创的 HILT 项目得出结论：一个高层的词表对 JISC 目的并不可行，它已在集合层次为英国高等教育界试点术语学服务，通过映射到一个 DDC 中脊。类似地，基于 DDC，Renardus 项目创建了一个共有的“转换”结构来支持一项交叉浏览服务（Koch et al.2003）。OCLC（DDC 的提供者）已经开发了主要词表之间的若干映射（既有手工的也有统计的），现在可用作术语学网络服务（OCLC Terminology Services, Vizine-Goetz et al. 2003）。OAI 协议被用来提供一个带有映射的词表的访问，通过一个针对人类用户的浏览器和针对机器的 OAI-PMH 网络服务机制。根据不同情况提供直接映射或共现映射。

BSI Standard on structured Vocabularies 草案的第 4 部分是有关词表间的互操作和映射的，它给出了一些有用的例子，包括单语种和多语种的。它也有一个关于不同选项的检索影响的讨论。这是一项重要的考虑，特别是在没有严格等同的概念存在并且有必要映射到一个上位或下位概念、一个部分重叠的概念或一个概念的（布尔）组合时。它区分**标引（index）术语的映射、搜索术语的映射、先组词串的映射、一对多映射、多对一映射**等。它还讨论了映射关系的不同类型以及非严格等同（部分映射）的类型。Renardus（Koch et al.2003）概述了基于集合（Set-based）的映射方法，是关于分类法的。（注意对于不同类型的 KOS 可能在映射方法上有差别，例如分类法与叙词表）。Doerr（2001）提出了一个**扩展的映射关系集**并且概括性地讨论了映射问题。这对 SKOS-Mapping Schema 草案产生了影响，该草案为定义叙词表间映射和等同关系描述了 RDF 叙词表词表扩展，虽然还未有严肃的应用来评估其建议。

4.5 自动分类和标引

自动分类和标引工具对于解决将 TS 应用于标引集合和知识库时的潜在资源开销是重要的。有些工具正在出现，应该为 JISC 目的对它们进行调查。许多人认为**将人工和自动方法相结合在当前是一种最佳方法**（例如 Hagedorn 2001）。人工输入可以用来设计其后将被自动阶段使用的词表，并且也可以人工检查自动结果。

在近期的一份自动主题分类方法的评论中, Golub (2006a) 区分了三种基于学科的方法: 文本分类 (采用 AI 机器学习技术); 文档成簇 (采用 (信息检索) 统计技术); 文档分类 (采用受控词表)。引文模式分析表明, 这三种方法倾向于不重叠。但是现在情况可能在变化。来自 Waikato 大学的 Medelyan and Witten (2006) 介绍了一种基于叙词表标引和简单贝叶斯 (naïve Bayes) 机器学习方法的结合, 用来进行特定领域的关键短语抽取, 获得的结果接近于在专业人工标引中发现的标引者间的一致性 (inter-indexer consistency)。他们的新 Kea++ 算法可以在一个开源许可下获得。

...

4.6 文本挖掘和信息抽取

5 当前术语学服务活动评论

5.1.2 Co-ODE: Collaborative Open Ontology Development Environment

<http://www.co-ode.org/>

协作开放本体开发环境 (Co-ODE) 是 JISC 资助的一个 Semantic Grid 项目 (在 Manchester 大学), 开发免费的和易于使用的本体管理和 OWL 工具, 作为 Standford 大学的 Protégé 的插件。最新的下载是一个调试工具。它还开发了一套指南。该项目建立在早期轻量本体编辑器 OilEd 的经验之上。HyOntUse (User Oriented Hybrid Ontology Development Environments, 面向用户混合本体开发环境) 是一项更理论性的姊妹 (EPSRC eScience) Manchester 项目, 涉及诸如调试本体这样的问题。

5.1.4 高层叙词表 (High Level Thesaurus, HILT)

<http://hilt.cdlr.strath.ac.uk/index2.html>

HILT 关注的是方便跨越 JISC 集合的广泛供应进行基于主题的访问以及相关集合的自动发现。两个早期和一个正在进行的 HILT 项目阶段已经研究了试点术语学服务, 与 OCLC Research 和 Wordmap 公司协作。一开始作为一个试点项目来研究一个高层叙词表对于 HILT 的可行性, DDC 被选为词表间映射的中脊 (具体是 DDC, LCSH, UNESCO, MeSH, AAT), 使用 OCLC Research 的可用映射服务。试点的 HILT2 示范器通过集合层 DDC 描述提供了一个交叉搜索工具。一个代表兴趣主题的用户术语被映射到 DDC 术语 (必要时人工解疑) 并且这被用来唤起一组 JISC 集合。目前系统只允许单个的用户术语。如果不能匹配任何集合则应用 DDC 号的相继切断 (到更概括的概念)。然后初始的 DDC 号被自动映射到集合所用的词表 (如果映射之一被包含), 并且在可能的地方实施自动搜索。HILT3 目前正在开发一个基于 webservice、SRW 协议、SKOS Core 和 SKOS-type 概念 URIs 的 M2M 示范器。计划最终用户不直接访问 HILT, 而是通过基于 web 的用户服务, 就像 GoGeo 和 BIOME。

5.2.7 SKOS (Simple Knowledge Organisation System)

<http://www.w3.org/2004/02/skos/>

关于叙词表和相关 KOS 的标准的 SKOS 工作是 EC FP5 SWAD-Europe 计划 (包括 CCLRC、ILRT 和其他) 的一项成果。其目标是方便 KOS 到语义 Web 的迁移, 建立在 LIMBER

和其他项目的前期成果之上。主要成果是 **SKOS-CORE Vocabulary** 和 **RDF 表示**，目前在 W3C Semantic Web Best Practices and Deployment Working Group 内进行开发, 作为一项 **W3C 工作草案** (Working Draft)。有指南可利用。还有更不稳定的用于 SKOS Core 专门化的 SKOS Extensions。其他正在进行的项目是 **SKOS Mapping** (着眼于用 RDF 表示概念方案之间的映射,) 以及 **SKOS API**, 一个用于对叙词表进行程序访问的 **Web Service API (和 Java 实现)**。见 6.4.1 节有更多的讨论。

5.3.6 OCLC 和 OCLC Research

<http://www.oclc.org>

作为杜威十进制分类法 (DDC) 的拥有者, OCLC 提供了诸如 WebDewey 和 Abridged WebDewey, WorldCat, OCLC Connexion® 编目界面等成果应用。

<http://www.oclc.org/research/projects/>

他们还维护着一个大型研究团体, 并且 OCLC Research 参加了各种词表计划 (以及像模式转化这样的相关元数据问题), 包括与术语学服务、永久标识符、影视、自动分类、模式转化、粉面系统等相关的工作。提供各种数据集和研究许可软件。

<http://www.oclc.org/research/projects/fast/default.htm>

LCSH 是一部广泛使用的大型综合性词表。FAST 计划涉及开发一种更简单的语法同时维持标题组合的丰富性。它已开发了自动的方法将 **LCSH 复合标题** 分解为独立的分面, 然后这些分面可以按其自身的权利访问处理。结果词表对于标引和检索有更多的灵活选项。

<http://www.oclc.org/research/projects/termservices/>

他们在术语学网络服务和永久标识符领域开发出了首批原型系统中的一些。这需要将词表从本地格式转换成标准模式, 例如 MARC21 或 SKOS Core (见第 6.2 节)。一旦采用了一种标准的基于 XML 的格式, 各种网络服务就可以增加价值, 包括到其他词表的映射 (见 4.4 节)。该网站提供了 (成对) DDC、LCC、LCSH、MeSH 和 Eric 之间的映射服务。Vizine-Goetz et al.(2006) 介绍了一个原型系统, 它提供了一个与桌面应用整合的交互服务例子。通过 Microsoft Office 2003 Research services Pane, 用户可以将词表搜索结果 **拷贝/粘贴** 进主窗口的一个应用中。计划用途是 **标引或元数据编辑**, 并且可以为词表术语选择 **MARC 粘贴格式和无格式文本字符串**。可采用 **SRU/W、REST 和 SOAP** 作为底层网络服务体系结构和全文、SQL 或 XML 存储层的协议。

许多词表已可在 OCLC Terminology Services Research 网站上使用, 包括以下词表 (更多词表正在加入):

DCMI Type Vocabulary

DDC Summaries

Library of Congress Subject Headings (LCSH)

Medical Subject Headings (MeSH) 2006

Newspaper Genre List(NGL)

<http://www.oclc.org/research/researchworks/ddc/browser.htm>

DeweyBrowser 可以进行经由 DDC Summaries (给出 DDC 主结构的头三层) 的搜索和浏览, 有多语种显示选项。它允许整合的搜索和浏览, 这样可以在一个初始搜索之后进行浏览。它也可以用来显示一个集合中记录的分布 (按照 DDC 类目)。可以获得一个在线示范。界面是用 AJAX (Asynchronous JavaScript and XML) 实现的。

5.4 与词表生命周期框架相关的项目

6 标准

6.4 协议、协议子集（profiles）和 API

6.4.1 访问词表的协议

SKOS API 是一项最近的发展，它定义了程序访问和查询基于 SKOS-CORE RDF schema 的词表的一组核心方法。在打算所谓网络服务调用时，API 本身保持与实现细节无关。ILRT 已开发了一个网络服务器实现，在开源基础上可下载。一组 SKOS 调用经由一个 ID、一个首选标签（preferred label）或匹配一个关键词（或常规表示）返回一个（多个）概念及其细节。另一个调用返回给定叙词表支持的一系列语义关系。另一组调用返回由一特定关系连接的概念或所有直接相关的概念。也有可能获得一组由某一关系上溯一给定路径长度所连接的概念。

当很少实用的交互应用已采用任何协议开发出来时，Zthes profile 可能已看到了最大的用处，它与广泛应用的 SRW/U 整合在一起。基于 SKOS RDF 表示的 SKOS API 看到了更少的用处。但它确实包含了揭示哪些关系被支持的功能（而不是 hardwiring 叙词表关系）。它也包含了函数可以返回概念和关系的复合模式，对于交互式应用是有用的。... Binding and Tudhope（2004）认为，将分布式协议服务建立在原子词表元素上不一定是有交互式用户界面的服务的最佳方法。需要多重客户-服务器调用的客户操作会携带太多的开销，限制响应性和界面上的交互类型。提供基本数据元素的复合模式（经由他们的关系）的协议可能需要获得合理的响应。

所回顾的词表协议可能还没有一个被认为是成熟的，建议在可能的地方，项目应考虑改编一个现有的协议。实际应用场合的经验对于协议的进一步演化和改进是必要的。