基于开源软件
构建数字图书馆的知识组织体系

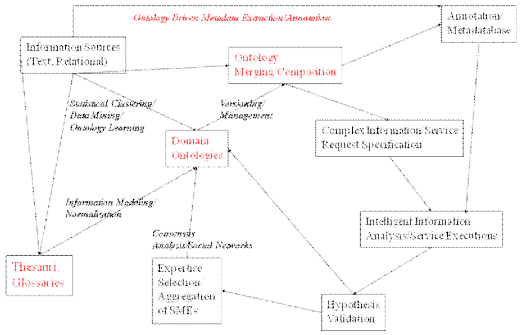
白海燕
中国科学技术信息研究所

主要内容

- 数字图书馆的知识组织架构
- 知识组织系统的构建与管理
  - 受控词表构建与管理系统
  - 本体的构建管理系统
- 受控词表的互操作
- 语义元数据的生成
  - 关键词自动标引软件- KEA 软件
  - 元数据自动生成软件-libiViaMetadata软件
  - 大众标注软件-FreeTAG
- 知识组织系统的存储与管理

1、数字图书馆的知识组织架构

- 数字图书馆的焦点
  - 数字化馆藏（collection）vs. 服务设施（services）
  - 用户（users）vs. 使用（uses）
  - 忽视了一组重要的结构-知识组织系统
- 数字图书馆的发展
  - 信息环境向网络平台的迁移，需要相应的语义工具实施对网络信息资源的组织，实现对网络资源的深度挖掘和智能利用
  - 构建一个数据交换与集成、知识化利用与管理的基础环境，需要新的信息组织机制的支持
- 如何在在数字图书馆结构中融入知识组织系统?



VIPUL KASHYAP - THE SEMANTIC WEB: THE NEXT GENERATION DIGITAL LIBRARY?

1、数字图书馆的知识组织架构

- 知识组织系统的构建与管理
- 知识组织系统的互操作
- 语义元数据的生成
- 知识组织系统的存储与管理

2、知识组织系统的构建与管理

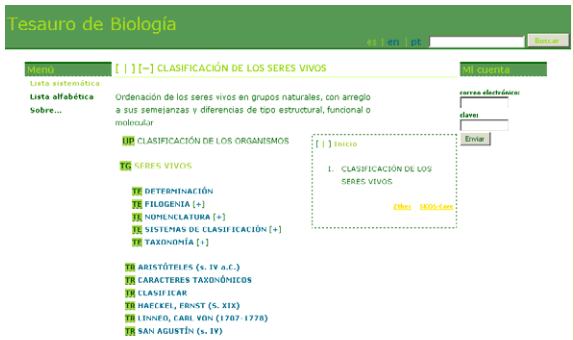
- 受控词表构建与管理系统
- 本体的构建管理系统

2.1 受控词表的构建与管理系统

- 功能和任务
  - 知识组织工具的构建和管理，如分类法、主题词表、叙词表、标题表等的创建、编辑、查询、更新、维护等；
  - 知识组织工具的交换与共享，如标准化的输入和输出等；
- 发展趋势
  - 注重不同知识组织工具间的交互和互操作的支持，以从不同的层次实现信息的一致性和有序化控制
  - 强调软件架构的开放性与集成能力，以扩展和强化知识组织工具在数字图书馆中的作用

2.1 受控词表的构建与管理系统-TEMATRES软件

- 词表结构定义能力
  - 款目数量、等级层次和交替款目的数量都没有限制，提供了款目词的范围注释、历史沿革注释和使用注释
- 一致性控制能力
  - 支持多种词间关系的构建，包括：等级关系（NT/BT），等同关系（UF）和相关关系（RT），并具有去重控制功能
- 词表的显示与输出
  - 支持等级层次显示和字顺列表显示，并可在一屏中显示多级等级层次，具有款目词的全文检索功能和等级浏览功能；
  - 多种格式导出整个词表，包括基于XML的Zthes,基于RDF的SKOS-CORE,基于TopicMap的XTM 1.0，支持的元数据格式包括Dublin Core和MODS；
  - 支持多语种界面，包括英语、法语、西班牙语、葡萄牙语、德语等
- 开始注重不同词表互操作方面的支持



2.1 本体的构建管理系统- PROTÉGÉ与KAON软件

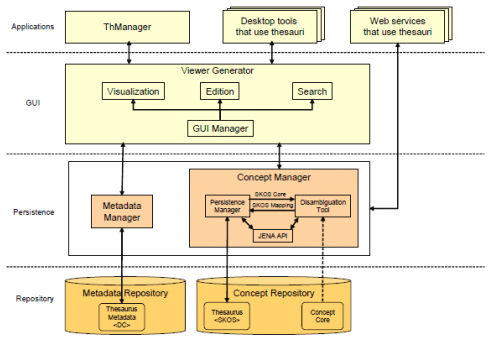
- Protégé
  - 由美国斯坦福大学开发的本体编辑器，也是基于知识的编辑器，是基于Java开发的一个开源项目，知名度很高，应用非常广泛
- KAON
  - 德国卡尔斯鲁厄大学开发的本体构建工具，也是基于JAVA的开源软件，它们都用于本体的创建、编辑、浏览和检索

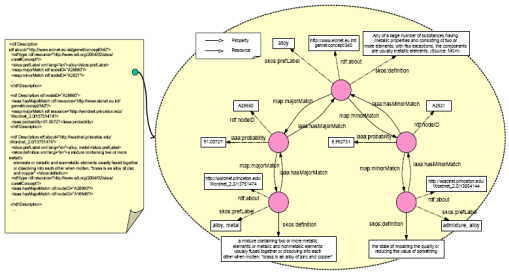
3 受控词表的互操作

- 任务与功能
  - 词表互操作-不同类型、结构和语种词表之间的集成、整合和兼容。不同的词表在显示概念关系的强弱和对自然语言的控制程度方面各不相同，其应用和作用的范围也不同，因此，在数字图书馆知识组织体系的构建过程中，往往采用不同类型、结构、语种和跨领域的多种词表
  - 需要利用相应的软件工具来消除不同词表在语法、结构和语义方面的异构特性
  - 目前受控词表互操作软件主要通过不同词表概念之间关联及关联关系的确定，来实现不同词表之间的集成和兼容。

3 受控词表的互操作-THMANAGER软件

- 词表管理软件
  - 可管理、维护基于SKOS格式的词表
  - 具有实现不同词表间自动关联的构建功能
- 互操作方法
  - 一个启发式投票算法来为词表中的每个概念选择概念核心中的语义最匹配的词汇，概念核心是WordNet词汇数据库
  - SKOS的映射类型：精确匹配（map:exactMatch）、大部分匹配（map:majorMatch）、小部分匹配（map:minorMatch）
  - 在THManger的存储管理中，SKOS映射扩展增加了一个空的节点用于记录映射的计算结果即可靠度



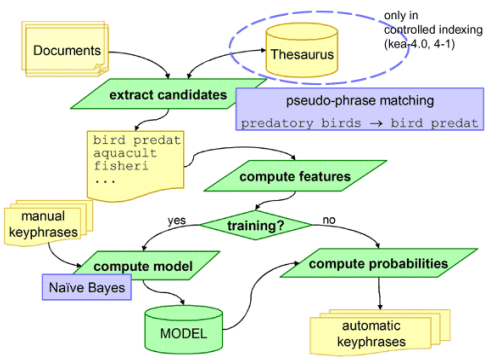


4 语义元数据的生成

- 人工标引
  - 学术性文献，作者在出版时自行提供关键词来揭示其内容特征；
  - 在图书馆通常由专业人员根据受控词表对文献进行标引
- 机器实现自动或半自动抽取文献内容
  - 以提高标引效率，应对海量资源
- 网络协同、发挥大众力量的自助标引

4.1 关键词自动标引软件- KEA 软件

- 对文献进行主题词自动标引
- 基本步骤
  - 确认侯选词
  - 计算侯选词的权值
  - 构建模型
  - 按模型进行计算



4.2 元数据自动生成软件- LIBIVIAMETADATA软件

- 用于HTML、PDF等格式文献的描述性元数据和语义元数据的自动抽取
- 应用了大量的文本自动抽取和分类聚类技术
- 可自动生成：题名、作者、出版者、载体类型、语种、关键词、分类（美国国会分类法、INFOMINE分类法）、文摘等

4.2 元数据自动生成软件-LIBIVIAMETADATA软件

- 关键词自动分配算法
  - 从两个数据源中合并数据，生成最终的关键词，返回的关键词数量由配置文件设定
- 自动分类算法
  - 利用一个两元分类器依据分类法确定属于或不属于某一个类，包括两大步骤：训练和分类

4.3 大众标注软件-FREETAG

- 大众标注（Floksonomy）是一种使用用户自由选择的关键词对网站、网页、WEBBLOG等数字资源的内容，通过协同方式，进行分类和内容标识的行为
- 开源项目众多，如Connotea、Del.irio.us、FreeTAG、Serendipity、Pligg、Scuttle等
- FreeTAG功能强大
  - 可内嵌于基于MySQL+PHP的应用系统中，创建和管理标签的
  - 允许在已知的数据库模式中创建标签，通过强壮的API接口获取和管理这些标签。
  - 高级功能：检索最新被标注的对象；检索指定标签的相似标签；检索指定被标注对象的相似对象

4.3 大众标注软件-FREETAG

- 准确率：不同技术方法检索结果对比

标签	返回资源数量		
	FreeTAG	Stemming	WordNet
Bacteria	420	420	468
Clostridium	516	516	987
DNA	452	455	1023
Genetics	1187	1439	1617
H5N1	937	938	958
HIV	1507	1507	1845

5 知识组织系统的存储与管理

- 受控词表的存储与管理
  - 文件系统存储管理模式，
    - THManager软件可对基于SKOS格式的词表文件进行管理，
  - 数据库存储管理模式
    - TemaTres使用MySQL数据库进行存储和管理
- 高级知识组织系统的存储与管理
  - 文件系统模式效率低，很难适应数据量较大的情况，只适用规模比较小的本体
  - 数据库模式很难适应本体动态变化的情况
  - 要求：
    - 应支持不同的描述语言如RDF(S)、DAML+OIL、OWL等
    - 支持本体的查询语言如RQL、SquishQL、RDFQL、RDFPath、OWL-QL等。

5 知识组织系统的存储与管理

- Jena 的存储功能
  - 存储它的数据在主要的存储器中
  - 把数据存储到关系数据库中
  - 用Sleepycat 软件的开源的植入式数据库Berkeley D B
- Jena框架主要包括
  - 以RDF/XML、三元组形式读写RDF
  - 支持 RDFS，OWL，DAML+OIL等本体的操作
  - 利用数据库保存数据，允许将数据存储到硬盘中，或者是OWL文件，或者是关系数据库中
  - 提供了ARQ查询引擎，它实现SPARQL查询语言和RDQL，从而支持对模型的查询

Thanks