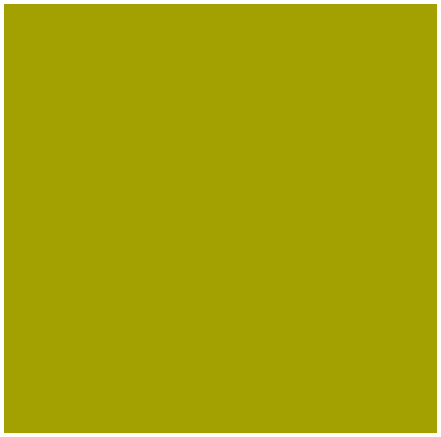
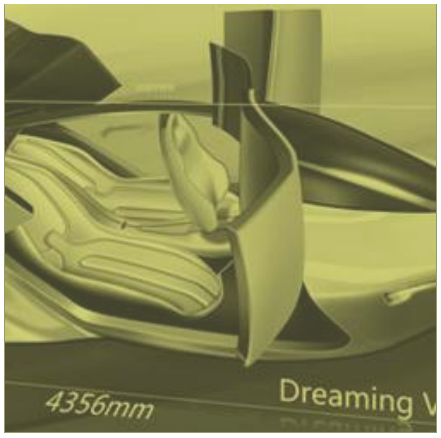




垂直知识图谱构造工具与行业应用

阮彤 自然语言处理与大数据挖掘实验室 主任　华东理工大学

主要内容

- 为什么垂直行业需要知识图谱
- 垂直知识图谱的特点
- 垂直知识图谱工具——VKG Builder介绍
- 垂直知识图谱应用

我们从通用知识图谱开始。。。。

1. SSCO

使用机器学习获得的知识网络，包括260,345个概念，5,602,180个实例，526,219个同义关系，下图显示了网络的部分节点

http://ssco.zhishimofang.com/

2.zhishi.me

对不同的中文数据源进行合并，组成统一的知识库。共有621万实例，73万类别，上亿的事实。成果在ISWC以及SCI期刊上发表。

华东理工大学　　上海交大

为了让知识图谱有用，准备构造一个搜索引擎？

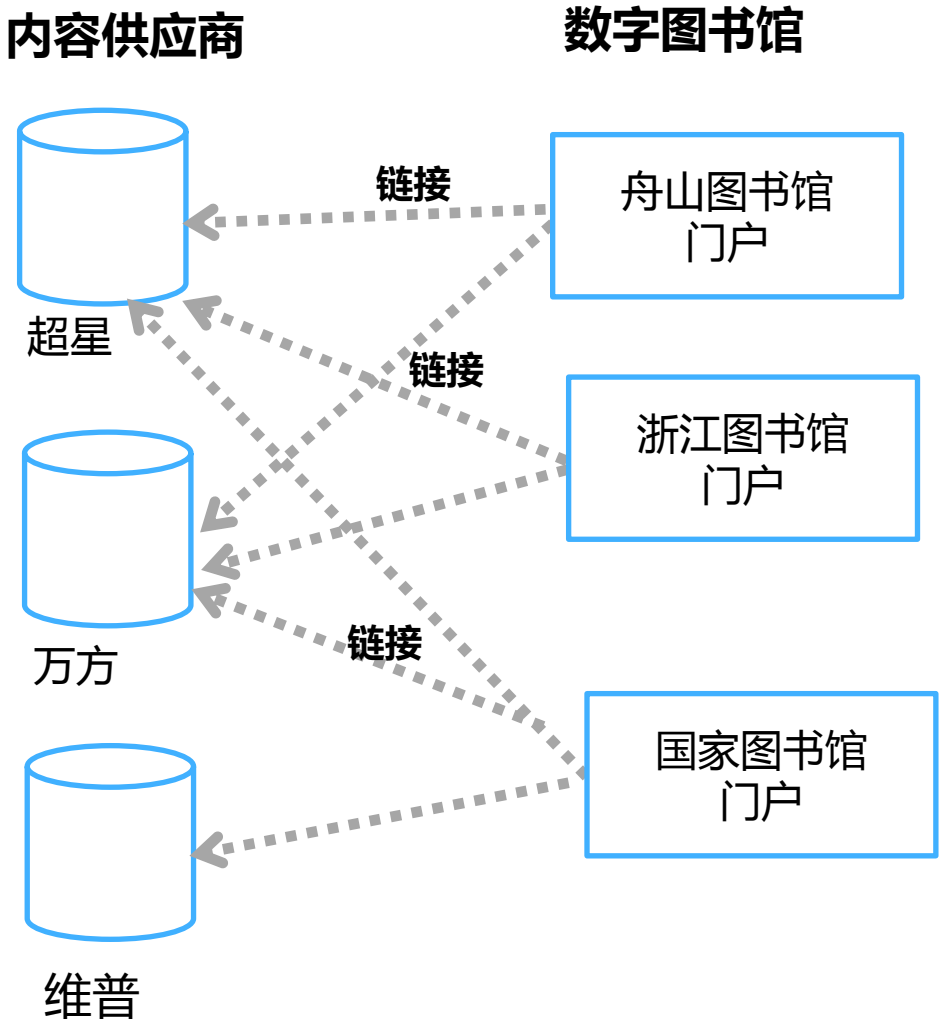
谷歌知识图谱 01

百度“知心” 搜狗“知立方” 02

为什么行业需要知识图谱

1.图书馆行业的故事

谁拥有更多的资源？
大图书馆！



地方图书馆利用知识图谱实现转型

- 严重的内容相似
- 缺乏内容控制
- 缺乏竞争力

- 与地方经济结合
- 自有内容
- 专有技术

寻觅新的机会！

为什么行业需要知识图谱

2.证券行业的故事——购买数据VS自己处理数据

大智慧

万德

同花顺

交易所

购买模式

平台模式

同花顺

新闻网站

大智慧

内部交易数据

证券公司研报

集成并分析数据

为什么行业需要知识图谱

2.证券行业的故事——现有搜索引擎的困惑

为什么行业需要知识图谱

3.医疗行业的例子——难点

- 电子病历搜索
  - 某类患者，如心衰并患有高血压患者的患者？
  - 与某病人相似的患者？

- 电子病历文本中有大量的数据，如何进行结构化,以更好地进行电子病历数据的大数据挖掘？

- 不同来源知识库之间如何关联？
  - 疾病、药品、检查的关联
  - 中西医疾病名称关联
  - 中西药成分关联

为什么行业需要知识图谱

3.医疗行业的例子——**病历结构化的必要性**

传统非结构化病历数据，只能通过文本匹配来进行查询

高血压 开始搜索 所有 现病史 既往史 家族史

病例查询结果

1.xxx,高血压患者 ✔

2.最高血压160/90 ✖

3.无高血压病史 ✖

大肠癌早期便血 开始搜索 所有 现病史 既往史 家族史

病例查询结果无

词汇二义性

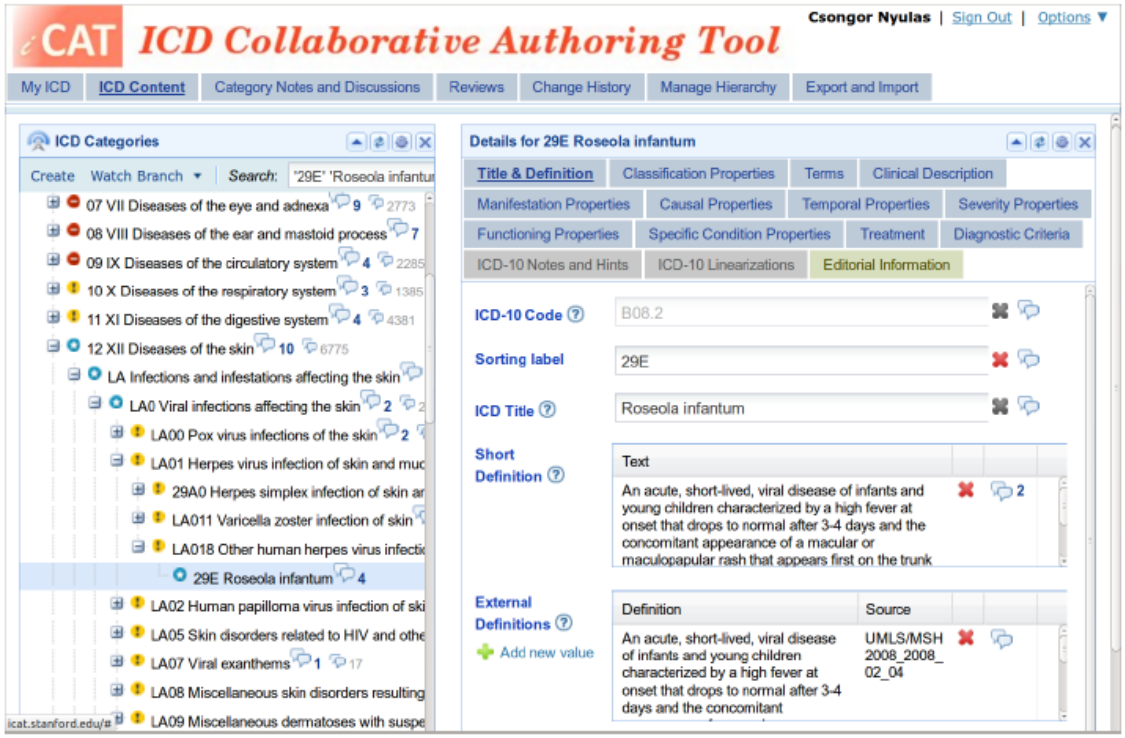
无法理解相同文字但不同含义的词汇

关联查询

无法精确切分查询词，理解查询意图

为什么行业需要知识图谱

3.医疗行业的例子——ICD-11的构造



ICD 10以及以前版本，基本上没有结构。分类码通常是一长串的疾病与它们相关的代码，以及少量属性，如同义词等。WHO在11版的ICD开发中使用语义Web 技术，支持协同编辑的语义Web 平台。 4年中，有270个来自世界各地的领域专家使用iCAT编辑了45,000个类，执行了260,000个变更。构造了17,000个链接，链到外部医学术语。

ICD 11使用了类/子类关系，子属性，定义域与值域，等价类。ICD 11表达能力是 SHOIN(D)。可以使用DL 推理程序去推理系统之间的非一致性。

为什么行业需要知识图谱

总结

- 行业需要数据（语义）集成能力
  - 文本 与 结构化数据的集成
  - 不同来源、不同格式的大量数据
  - 自动/半自动的集成
- 行业需要（语义）数据查询能力
  - 更丰富的表达
  - 更精准的结果

为什么行业需要知识图谱

知识图谱的技术优势

渐增式数据模式设计

初始设计的时候，很难清楚所有的概念，而知识图谱的动态可扩充性以及“无模式”特性使得用户很容易增加或修改模式。

数据集成更轻松

本体的语义互操作特性以及“链接数据”原则，使得来自不同供应商的数据集成更为方便。

现有标准支持

有RDF(S),OWL,SPARQL等标准，可以逐渐要求内容供应商支持。

语义搜索

用户可以查询具有某类特征的某类实体，比起基于基于关键词的搜索，更为精准。

主要内容

- 为什么需要行业需要知识图谱
- 垂直知识图谱的特点
- 垂直知识图谱工具
- 垂直知识图谱应用

垂直知识图谱特点

1. 更为丰富与精确的领域数据

- 领域对数据质量要求更高，例如
  - 药品名称
  - 处方当中，药品的克数
  - 企业名称
  - 企业股票价格

VS 姚明的身高？

VS 电影的字段

- 领域数据字段与数据关联更丰富
  - 一个企业包含的信息字段可能有上百个字段
  - ICD11当中,每个疾病有56个属性，而其中52个是用填实例的。（不是随意填写一个文本，而是这个填的值本身也是RDF的一个实例。
- 只有满足了上述条件，才能用于商业分析与决策支持

垂直知识图谱特点

2. 自顶向下

- 普通的KG，使用一种自底向上的方法，更强调数据的宽度。如DBpedia在构造过程中，先有数据，后有本体。
- 对于行业本体来说，由于数据质量以及行业本身的规范要求，将使用自顶向下的方式。

垂直知识图谱特点

3. 丰富的内部数据来源

- 企业/组织结构自有数据，如：
  - 对证券公司而言，自有用户交易数据
  - 医院，电子病历，付费、检查记录
  - 图书馆 地方政府数据
- 行业数据
  - 如行业标准、规范，如Medical Guideline
  - 如第三方收集的企业数据
- 诸多以RDB方式存储

垂直知识图谱特点

4. 需要可扩充的第三方工具支持

- 比起互联网企业，垂直行业客户相对投入低，能力弱。
- 需要可配置、图形化界面。
- 需要容易地面向不同行业做客户化。

垂直知识图谱特点

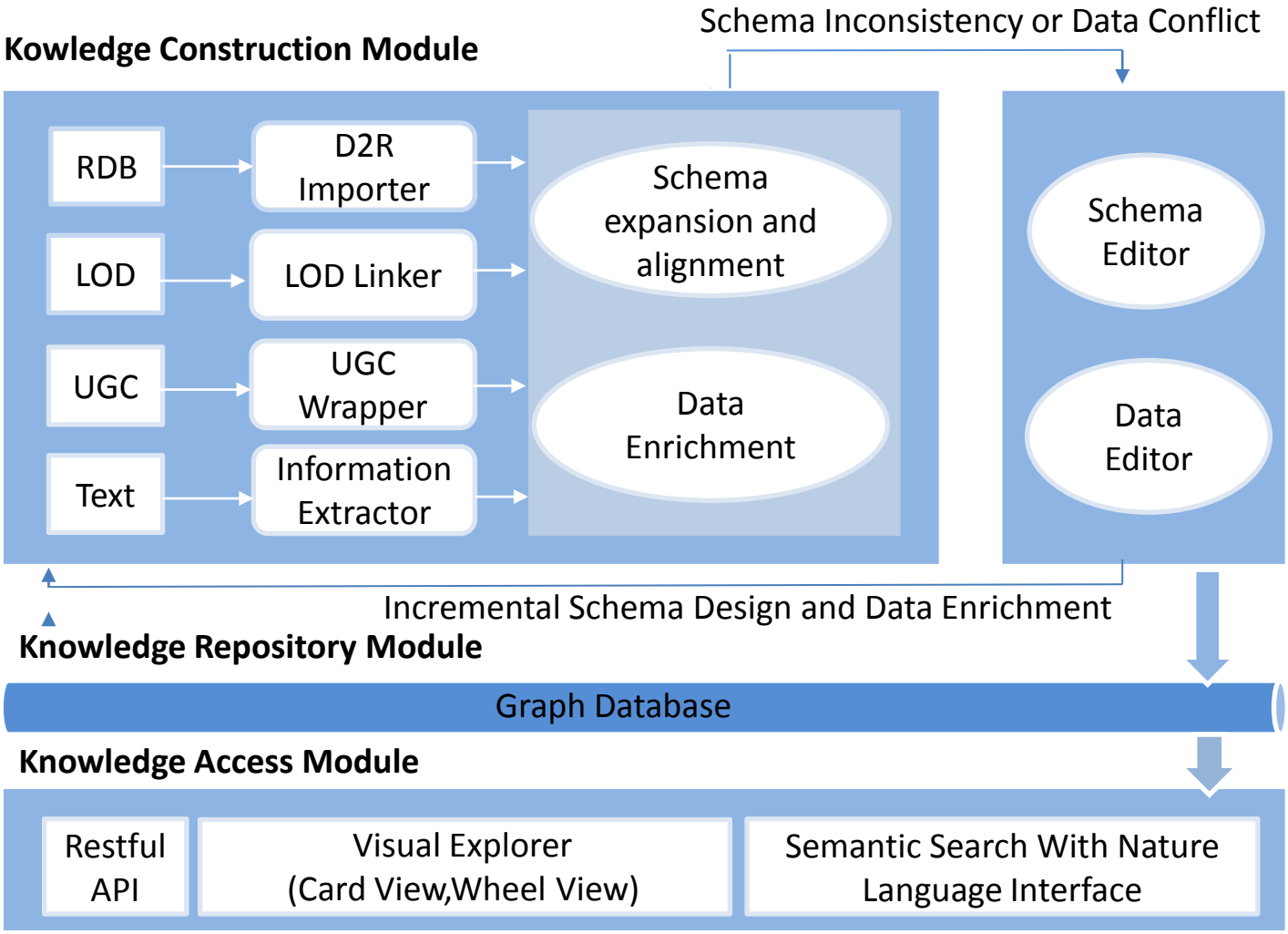
5. 与决策支持集成——语义搜索？Rule Engine？大数据挖掘？

- 医疗：如果我构造了医学知识图谱，下一步就是，如何基于这个图谱做诊疗？
- 证券：如何我构造了一个企业知识图谱，下一步就是，如何找到满足某一类条件的企业？

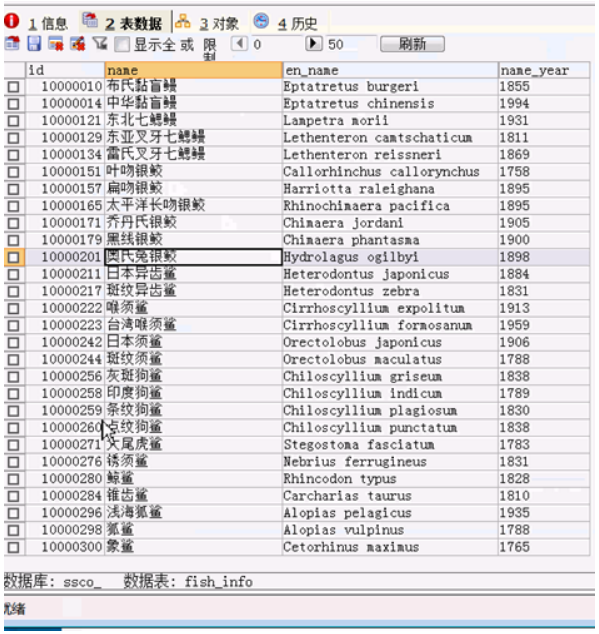
主要内容

- 为什么需要行业需要知识图谱
- 垂直知识图谱的特点
- 垂直知识图谱工具
- 垂直知识图谱应用

垂直知识图谱体系结构


Schema Inconsistency or Data Conflict
Kowledge Construction Module
RDB
D2R Importer
LOD
LOD Linker
UGC
UGC Wrapper
Text
Information Extractor
Schema expansion and alignment
Data Enrichment
Schema Editor
Data Editor
Incremental Schema Design and Data Enrichment
Knowledge Repository Module
Graph Database
Knowledge Access Module
Restful API
Visual Explorer (Card View,Wheel View)
Semantic Search With Nature Language Interface

构造海洋KG的例子——多种数据来源



从Web抽取数据


舟山海洋数字图书馆
ZHOUSHAN OCEAN DIGITAL LIBRARY
知识体系
知识卡片
知识详情
语义检索
知识编辑
知识标注
总分类
其它分类
鱼类
小黄鱼
添加
删除
开放知识库映射
百科学习
文本学习
属性定义
名称
小黄鱼
唯一标识:
描述
父概念
鱼类
同义词
Larimicht
子概念
属性
命名时间
实例
导入成功，导入属性数目为 10
确定
使用配置文件：
选择文件
未选择文件
配置：
<?xml version="1.0" encoding="utf-8"?>
<mappings>
<mapping type="auto" config="url_xpath"> <!-- 自动类型，URL加上XPath确定属性的值 -->
<url type="1">http://fishdb.sinica.edu.tw/AjaxTree/tree_t.php?ntext=Larimichthys%20polyactis</url> <!-- 实体匹配，当type为0时表示仅映射当前实体，type为1时批量映射 -->
<!-- <url type="1">http://fishdb.sinica.edu.tw/AjaxTree/tree_t.php?ntext=[en_name]</url> -->
<!-- 批量映射示例，[en_name]表示按英文名称匹配实体和 URL -->
<synonym><!-- 同义实体来源XPath -->
<xpath>/DIV 0/TABLE 2/TABLE 0/TR 4/TD 5</xpath>
<xpath>/DIV 0/TABLE 2/TABLE 0/TR 5/TD 5</xpath>
</synonym>
<attributes><!-- 实体属性来源配置 -->
<attribute xpath="/DIV 0/TABLE 2/TABLE 0/TR 3/TD 1" attributename="科" />
<attribute xpath="/DIV 0/TABLE 2/TABLE 0/TR 1/TD 2" attributename="水深范围" />
<attribute xpath="/DIV 0/TABLE 2/TABLE 0/TR 4/TD 3" attributename="可食性" />
<attribute xpath="/DIV 0/TABLE 2/TABLE 0/TR 5/TD 1" attributename="最大体长" />
<attribute xpath="/DIV 0/TABLE 2/TABLE 0/TR 6/TD 1" attributename="分布FAO" />
</attributes>
</mapping>
<mapping type="customize" wrapper="taiwan_fishbase_tree_wrapper" /> <!-- 定制类型，wrapper指定定制的适配器 -->
</mappings>
确定
© 舟山海洋数字图书馆 All Rights Reserved. 浙ICP备09056684号-4

从百科网站抽取数据


舟山海洋数字图书馆
ZHOUSHAN OCEAN DIGITAL LIBRARY
小黄鱼
知识体系
知识卡片
知识详情
语义检索
知识编辑
知识标注
总分类
其它分类
鱼类
小黄鱼
添 加
删 除
历史版本
导入
导入结构数据
开放知识库映射
百科学习
文本学习
属性定义
百科学习
学习模式:
当前实体学习
百科选择:
维基百科 互动百科 百度百科
学习内容:
摘要 信息模块 类型标签
确 定
名 称
小黄鱼
唯一标识:
描 述
父概念
鱼类
同义词
Larimicht
子概念
属 性
分布FAO[
命名时间
最大长度
水深范围
可食性
纲
亚纲
目
亚目
科
属
数值 对象
条鳍鱼纲
新鳍鱼亚纲
鲈形目
鲈亚目
石首鱼科
黄鱼属
实 例
保 存
选择文件
未选择文件
© 舟山海洋数字图书馆 All Rights Reserved. 浙ICP备09056684号-4

不同数据来源冲突解决


2.Different Data Sources
4.Decision By System
3.Conflict Values
1.Conflict Icon
FishBase
台湾鱼类数据库
维基百科
百度百科
互动百科

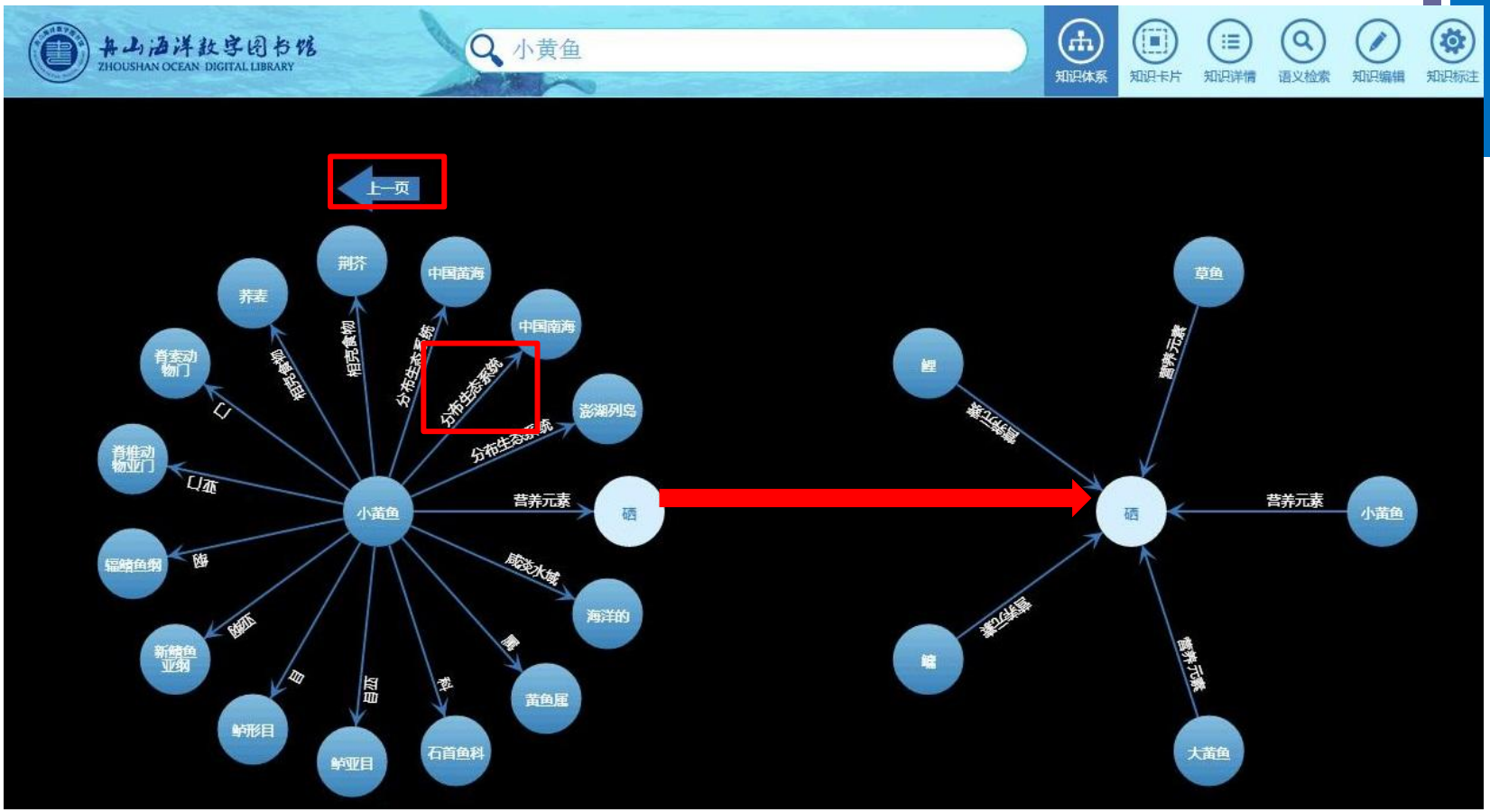
语义检索


舟山海洋数字图书馆
ZHOUSHAN OCEAN DIGITAL LIBRARY
小黄鱼的分布
登录
知识体系
知识卡片
知识详情
语义检索
知识编辑
知识标注
知识图谱
海洋经济知识体系
其它知识
鱼类
盲鳗纲
七鳃鳗纲
盾皮鱼纲
软骨鱼纲
棘鱼纲
肉鳍鱼纲
辐鳍鱼纲
腕鳍鱼亚纲
软质亚纲
新鳍鱼亚纲
雀鳝目
弓鳔鱼目
月眼鱼目
实体解析结果：
小黄鱼
语义解析结果：
小黄鱼的分布生态系统
小黄鱼的分布地区
小黄鱼的分布气候带
小黄鱼的分布FAO区域
小黄鱼的纬度范围
小黄鱼的经度范围
语义检索结果：
中国东海
太平洋
澎湖列岛
中国南海
中国黄海
共找到相关资源 254 条
黄海南部和东海小黄鱼(Larimichthys polya...
作者：林龙山 程家骅 姜亚洲 袁兴伟 李建生 高天翔
出处：万方期刊论文
关键词：小黄鱼(Larimichthys polyactis) 产卵场 ...
出版时间：2008
基于线粒体Cyt b基因的黄海、东海小黄鱼(Larimic...
作者：吴仁协 柳淑芳 庄志猛 金显仕 苏永全 唐启升
出处：万方期刊论文
关键词：小黄鱼 细胞色素b基因 群体遗传结构
出版时间：2009
一起使用碱性橙浸染小黄鱼案的评析
作者：冯济富 黄明
出处：万方期刊论文
关键词：食品生产经营 卫生监督 碱性
2000年东海区小黄鱼(Pseudosciaena pol...
作者：汤建华 仲霞铭 高银生 沈德华
出处：万方期刊论文
关键词：2000年 东海区 小黄鱼 资源数
小黄鱼
小黄鱼又名小黄花鱼，脊椎动物，鱼纲，石首鱼科，又名：小鲜、大眼、花色、小黄瓜、古鱼、黄鳞鱼、小春色、金龙、厚鳞仔，也叫"黄花鱼"、"小黄花"。体形似大黄鱼，但头较长，眼较小，鳞片较大，尾柄短而宽，椎骨28～30块。耳石较大。体长约20余厘米。体背灰褐色，腹部金黄色。为近海底层结群.....
相关搜索内容
黄鱼属
大黄鱼
似长鳍黄鱼
用户还搜索以下内容
大黄鱼
带鱼科
海洋科学

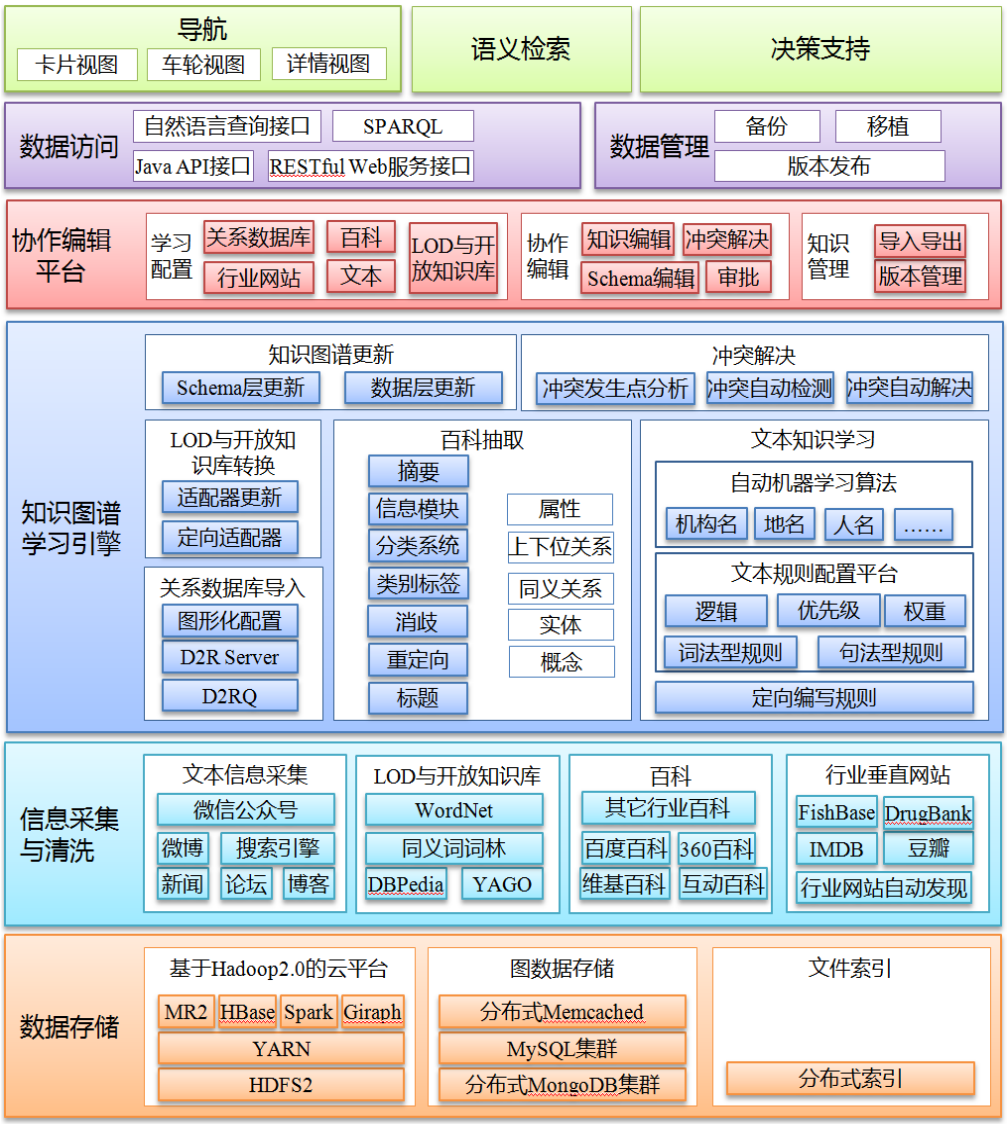
浏览KG——卡片视图


舟山海洋数字图书馆
ZHOUSHAN OCEAN DIGITAL LIBRARY
母狗
登录
知识体系
知识卡片
知识详情
语义检索
知识编辑
知识标注
褐狗母鱼
肩斑狗母鱼
豹狗母鱼
中间狗母鱼
浩氏狗母鱼
夏威夷狗母鱼
尖齿狗母鱼
台湾狗母鱼
方斑狗母鱼
知识卡片
详情
名称
台湾狗母鱼
摘要
台湾狗母鱼，为鲻鲹鱼纲仙女鱼目合齿鱼亚目合齿鱼科的其中一种，分布于西北太平洋的台湾海域，栖息深度可达80米，体长可达19.5厘米，为底栖性鱼类，生活在沙底质海域，属肉食性，生活习性不明。
同义实体
Synodus taiwanensis
父概念
狗母鱼属

浏览KG——轮子视图


舟山海洋数字图书馆
ZHOUSHAN OCEAN DIGITAL LIBRARY
小黄鱼
知识体系
知识卡片
知识详情
语义检索
知识编辑
知识标注
上一页
荆芥
中国黄海
荞麦
中国南海
脊索动物门
相克食物
分布生态系统
分布生态系统
澎湖列岛
脊椎动物亚门
营养元素
硒
小黄鱼
辐鳍鱼纲
海
感染水域
海洋的
新鳐鱼亚纲
黄鱼属
鲈形目
鲈亚目
石首鱼科
鲤
草鱼
营养元素
硒
营养元素
小黄鱼
鳐
大黄鱼

未来产品框架

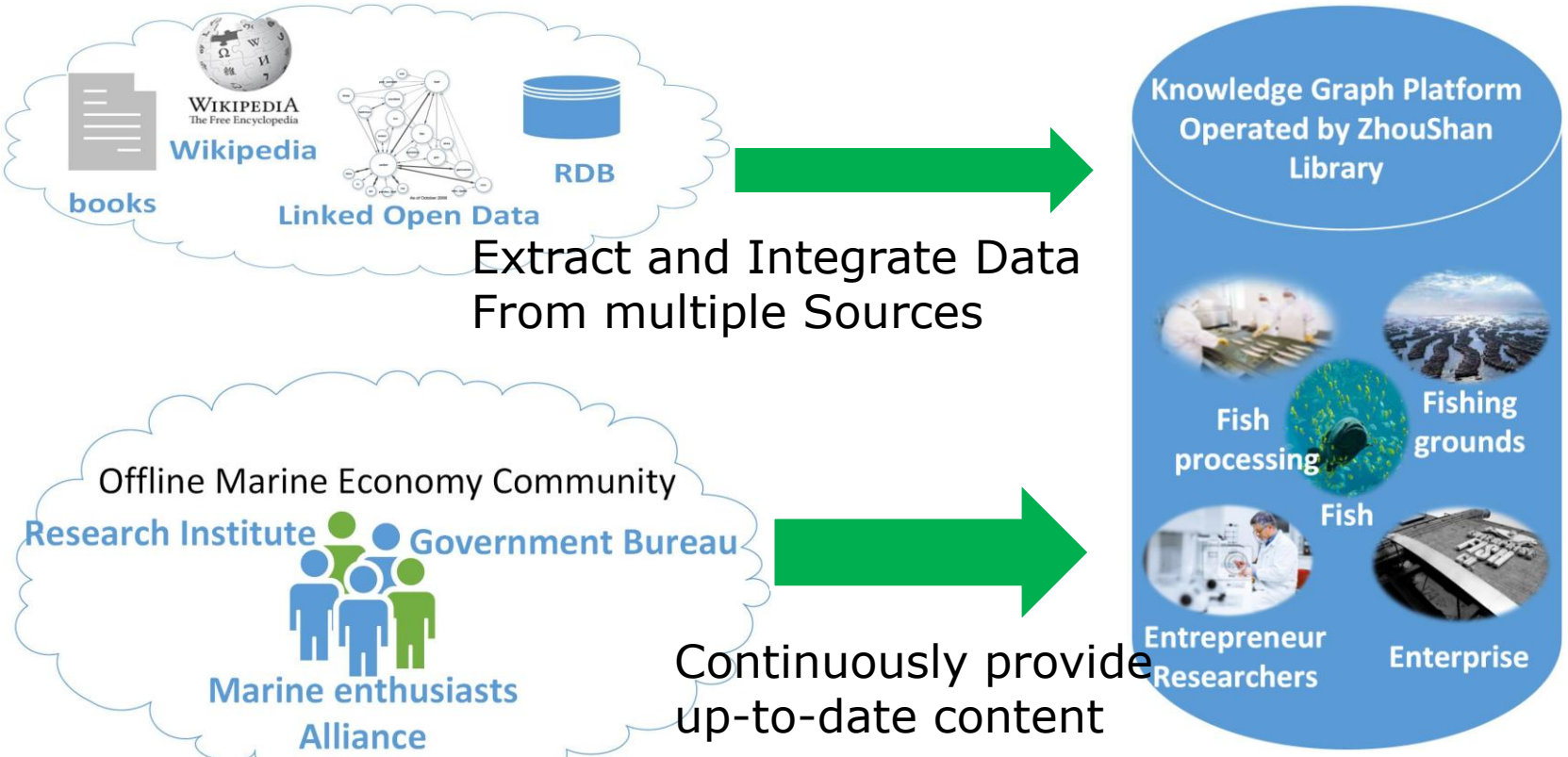

导航
卡片视图
车轮视图
详情视图
语义检索
决策支持
数据访问
自然语言查询接口
SPARQL
Java API接口
RESTful Web服务接口
数据管理
备份
移植
版本发布
协作编辑平台
学习配置
关系数据库
百科
行业网站
文本
LOD与开放知识库
协作编辑
知识编辑
冲突解决
Schema编辑
审批
知识管理
导入导出
版本管理
知识图谱学习引擎
知识图谱更新
Schema层更新
数据层更新
冲突解决
冲突发生点分析
冲突自动检测
冲突自动解决
LOD与开放知识库转换
适配器更新
定向适配器
关系数据库导入
图形化配置
D2R Server
D2RQ
百科抽取
摘要
信息模块
分类系统
类别标签
消歧
重定向
标题
属性
上下位关系
同义关系
实体
概念
文本知识学习
自动机器学习算法
机构名
地名
人名
……
文本规则配置平台
逻辑
优先级
权重
词法型规则
句法型规则
定向编写规则
信息采集与清洗
文本信息采集
微信公众号
微博
搜索引擎
新闻
论坛
博客
LOD与开放知识库
WordNet
同义词词林
DBPedia
YAGO
百科
其它行业百科
百度百科
360百科
维基百科
互动百科
行业垂直网站
FishBase
DrugBank
IMDB
豆瓣
行业网站自动发现
数据存储
基于Hadoop2.0的云平台
MR2
HBase
Spark
Giraph
YARN
HDFS2
图数据存储
分布式Memcached
MySQL集群
分布式MongoDB集群
文件索引
分布式索引

主要内容

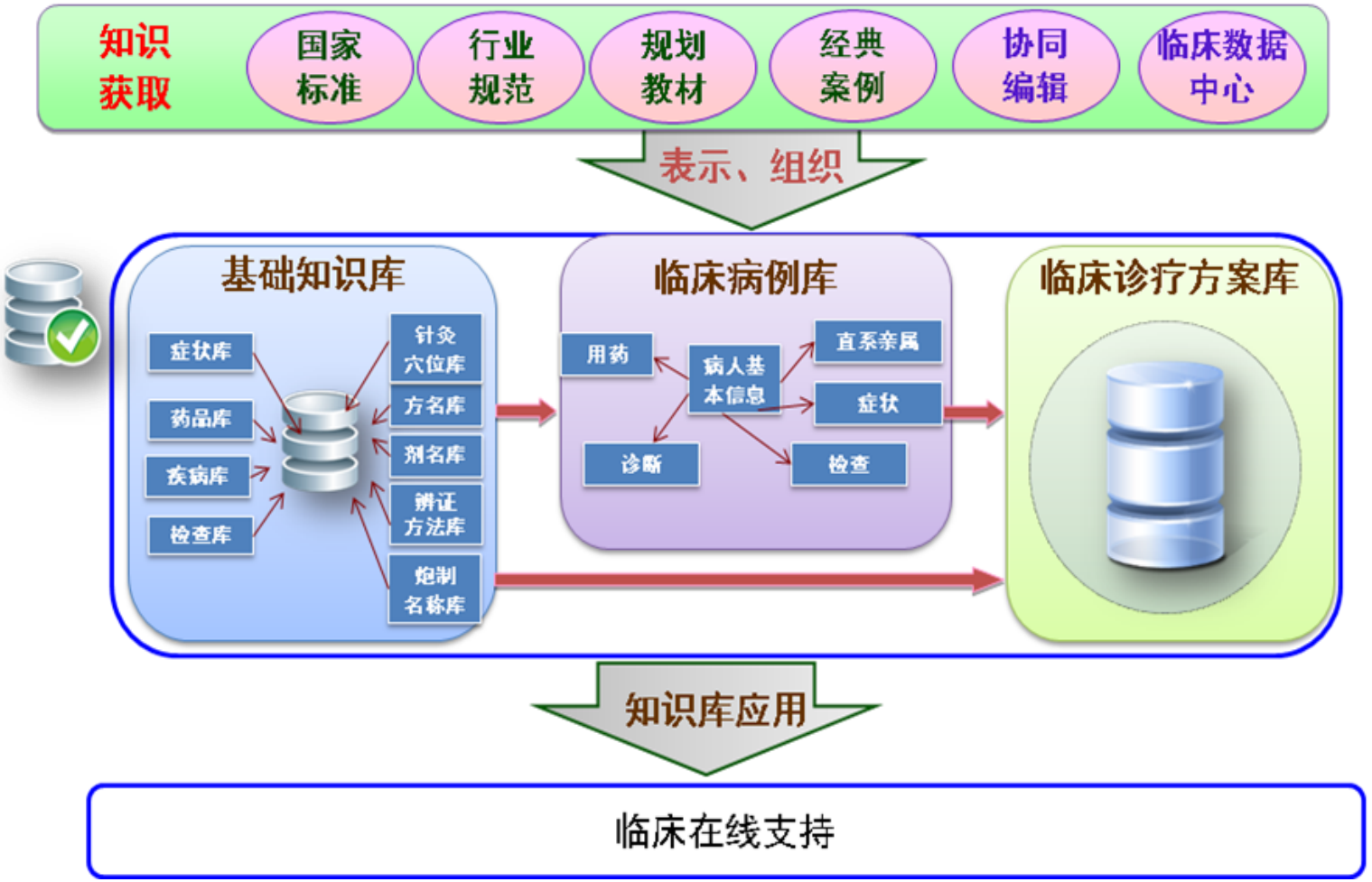
- 为什么垂直行业需要知识图谱
- 垂直知识图谱的特点
- 垂直知识图谱工具
- **垂直知识图谱应用**

知识图谱帮助图书馆模式变迁

内容供应商+平台运营商



基于知识图谱技术的医疗知识库——正在进行。。。



总结

- 垂直知识图谱有巨大的前景
- 垂直知识图谱有诸多难点问题
- 我们在图书馆、证券等行业做了部分探索

谢 谢 !

ruantong@ecust.edu.cn