数据密集型知识服务系统

——基于智慧数据的数字图书馆技术架构

孙 坦
中国农业科学院农业信息研究所

提 纲

1. 数字图书馆现状与挑战
2. 推动数字图书馆创新的动力
3. 数字图书馆创新方向与路径
4. 未来行动设想与规划

第一节

数字图书馆的现状与挑战

1. 数字图书馆的现状与挑战

开放信息环境中海量异构信息与碎片化需求的矛盾：

- 海量数据广泛分散在各种书籍、报纸、不同格式的音视频媒体以及互联网和其他媒介中，并快速的数字化。**这些数据的指数增长导致人们利用计算机查找专门数据和有用的相关信息的能力迅速降低。**
- 搜索引擎的算法主要依赖关键词匹配，而不考虑关键词的各种不同含义以及位于短语中组合关键词的复杂含义，甚至是关键词和短语被用于不同上下文中的不同含义。
- 多数知识呈现出要么非结构化要么是异构化的，致使其不能被进一步处理

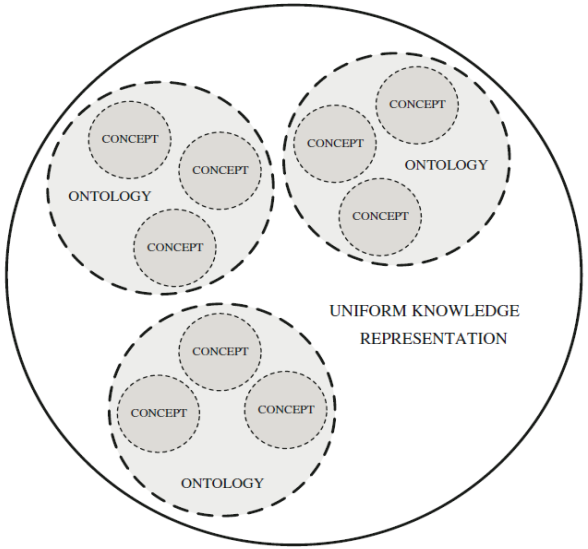


Fig. 1.1 Uniform knowledge representation consisting of ontologies populated by concepts

- **根本性问题**：如何从海量数据源中有效提取需要的数据？如何找到那些代表着问题答案或有助于解决问题的必要和潜在可用的却又未知的信息片段？如何集成那些潜在的记录（信息片段）？

1. 数字图书馆的现状与挑战

- **1st generation DL: 基于资源的数字图书馆**

  Resource-based digital library: = digital collections?

- **2nd generation DL: 基于集成服务的数字图书馆**

  Integrated digital resources and related services

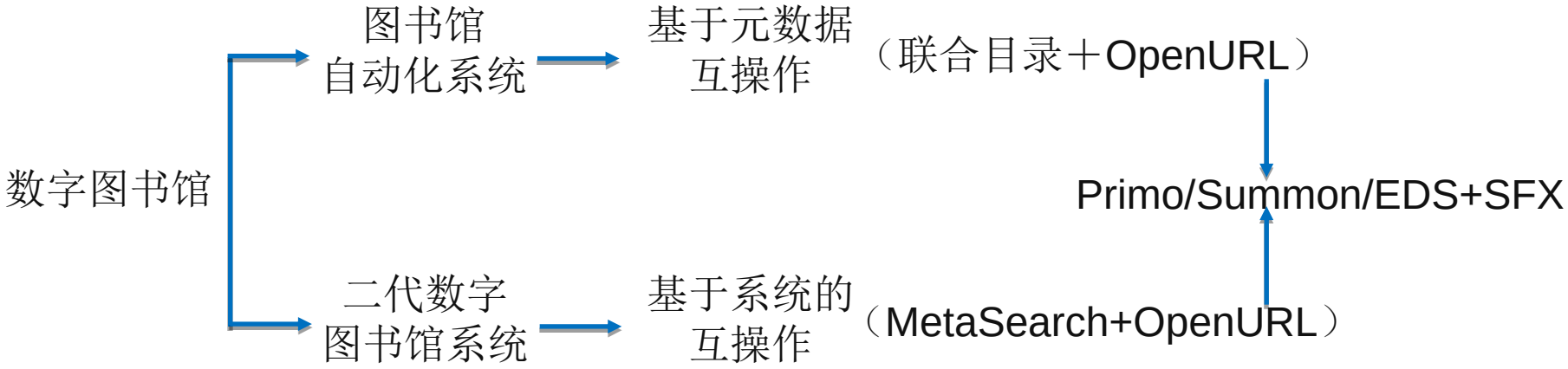
- **3rd generation DL: 基于用户的数字图书馆**

  Digital resources and services personally organized and embedded into user application environment

1. 数字图书馆的现状与挑战

二代数字图书馆解决的核心问题：分布式集成（“一站式”检索获取）

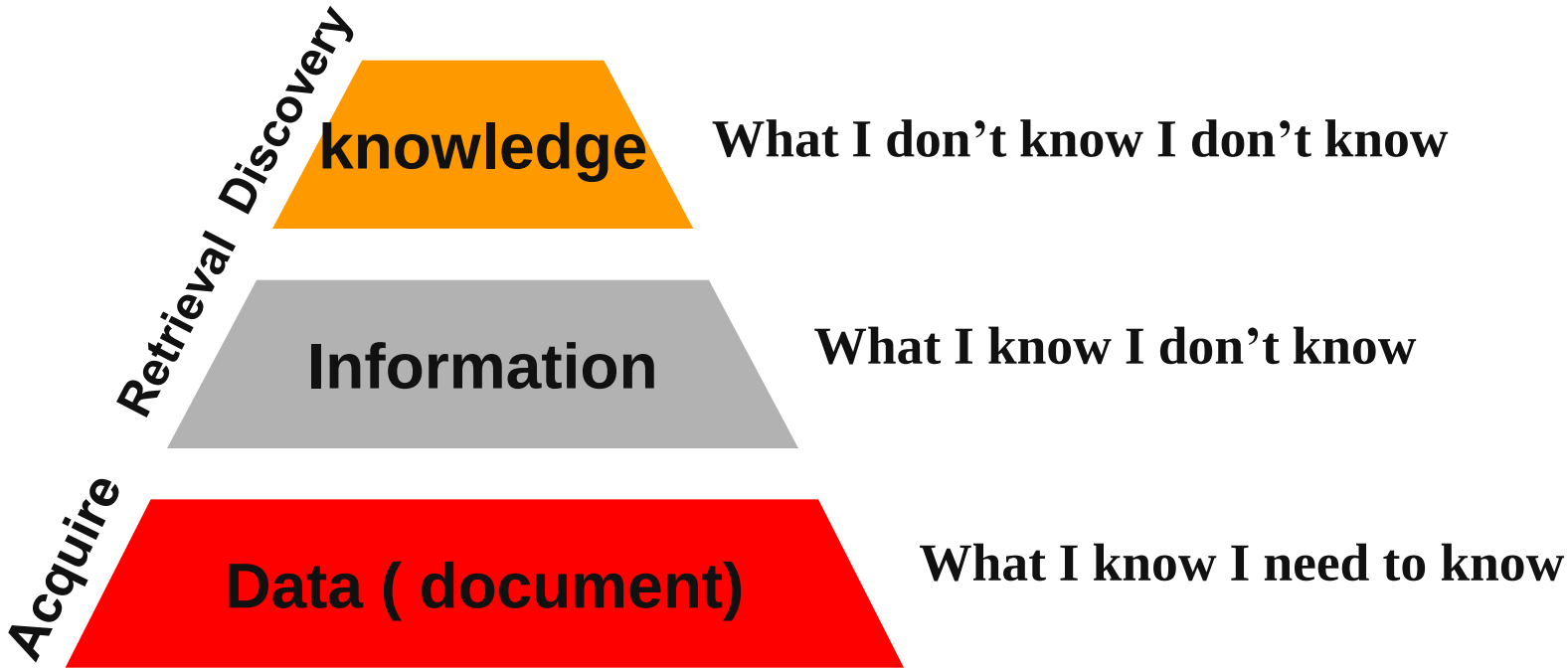
- 统一资源发现
- 扩展服务调度



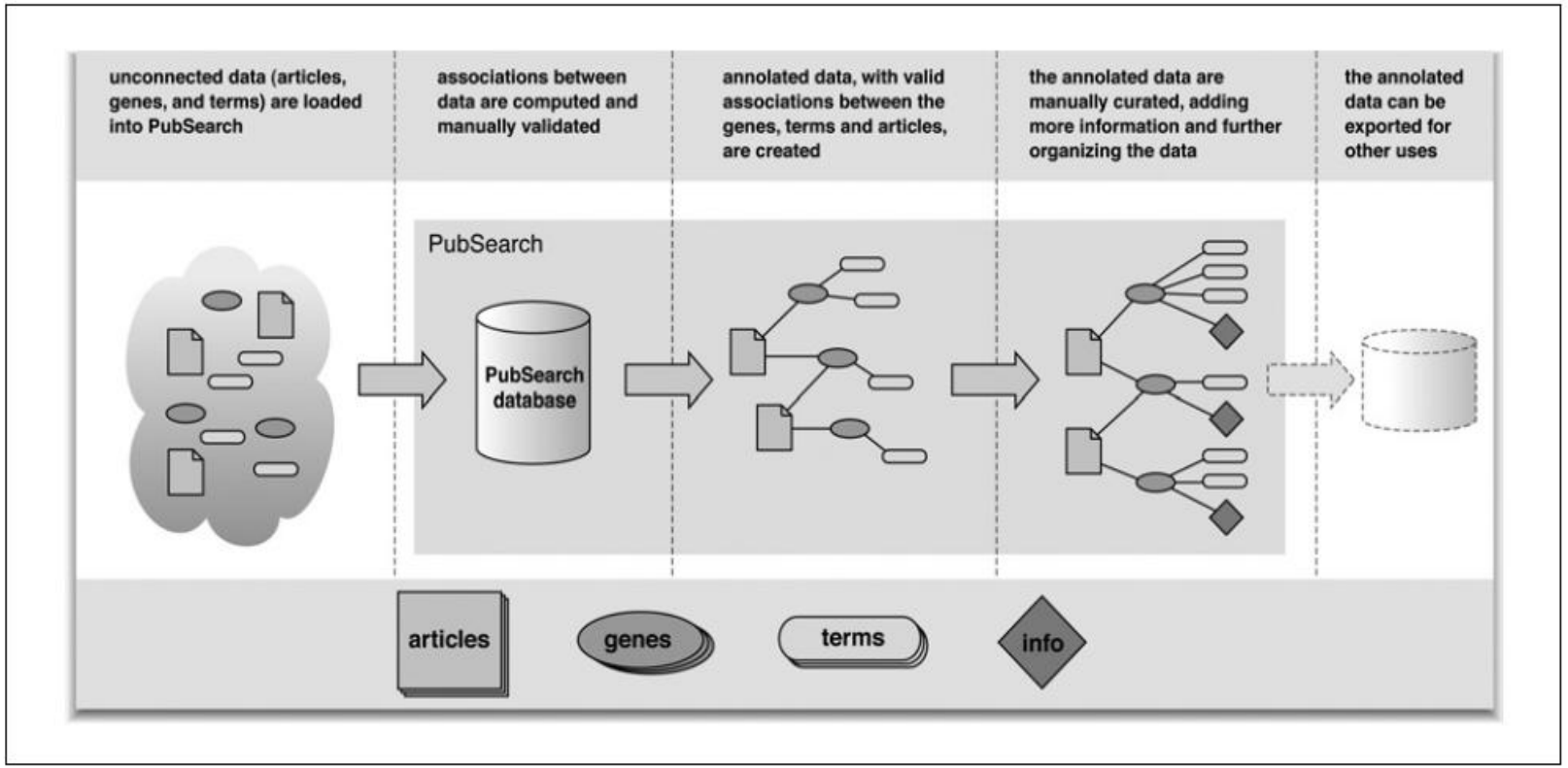
1. 数字图书馆的现状与挑战

- Not about books! Never was!
- Also not about information!
- But it is about people – connecting people to resources of all kinds needed for learning and critical thinking and (hopefully) knowledge.
- No longer just “organizing” knowledge containers, but helping to create them and mold them, host them, etc.

1. 数字图书馆的现状与挑战



1. 数字图书馆的现状与挑战



1. 数字图书馆的现状与挑战

- 数字图书馆“泛在性”不足，成为用户最后的选择
  - 信息资源局限于文献
  - 交付门槛过高
  - 可发现性不足
  - 信息描述、组织与检索能力不足
  - 只给出线索，而不能直接回答用户的问题

第二节

推动数字**图书馆**创新的动力

2.1 开放获取与开放出版

- 2001：**B**udapest **O**pen **A**ccess **I**nitiative（BOAI）
- 2011：Ten years on from the Budapest Open Access Initiative: setting the default to open，明确提出10年后在世界上的任何国家或地区任何学科领域的同行评议学术论文均实现开放存储与开放获取
- 金色OA：**开放出版**，BioMed Central（BMC）对250种同行评议期刊开放获取出版，作者付费，读者免费访问、下载和传播；SCOAP3等
- 绿色OA：**开放仓储**，PubMed Central（PMC）收录NIH资助论文期刊301种

2.1 开放获取与开放出版

同行评议（Peer Review）是保障学术交流质量和真实性的基石，并未随着学术交流模式和学术信息交流模式的改变而失去价值。相反，其仍是开放获取时代学术交流质量的重要保障基础。

署名（BY）
署名（BY）-禁止演绎（ND）
署名（BY）-非商业性使用（NC）
署名（BY）-非商业性使用（NC）-禁止演绎（ND）
署名（BY）-非商业性使用（NC）-相同方式共享（SA）
署名（BY）-相同方式共享（SA）Cc-bynewwhite.svg

2.1 开放获取与开放出版

- 2015年OA期刊的数量约12000种，SCI收录的OA期刊超过1200种，比2014年增长6.5%，SSCI的OA期刊为188种，比2014年增长3.3%
- Web of Science数据表明：2014年全球共发文2,160,301篇，其中OA发文227,528篇，占总发文量的10.5%；2015年全球共发文1,957,360篇，其中OA发文234,089篇，占总发文量的12%，较2014年有所增加。

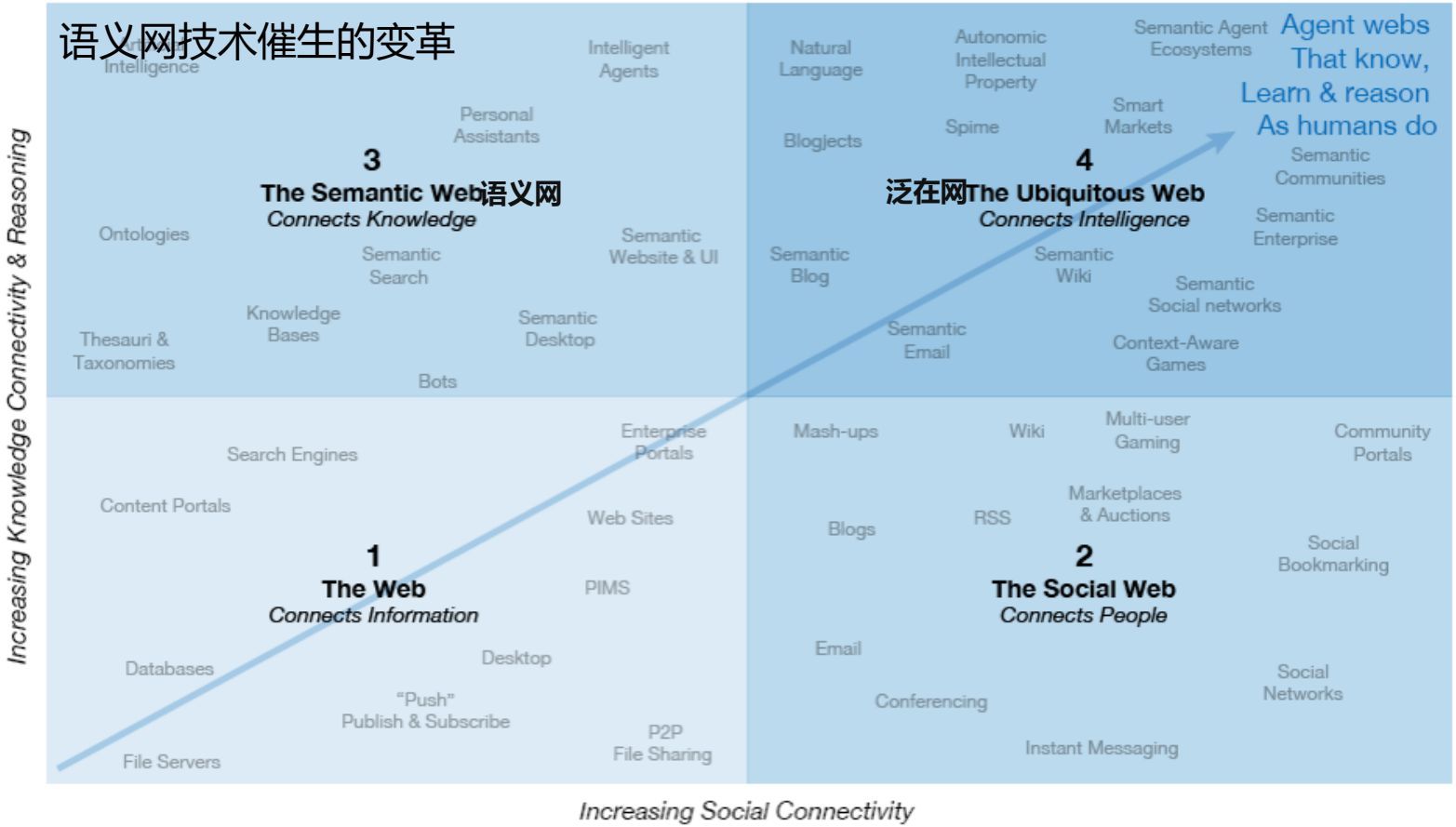
开放资源	类型	资源量（2013.11）	资源量（2015.10）	资源量（2016.04）
DOAB	开放获取图书目录	49个出版社的1,471本图书	122个出版社的3,389本图书	151个出版社的4,613本图书
DOAJ	开放获取期刊目录	9,940OA期刊116万篇文章	10,613种期刊，210万篇文章	11,604种刊，228万篇文章
PMC	学科仓储	280万篇文章	360万篇文章	380万篇文章
Dryad	数据仓储	3,699个数据包，10745个数据文件	10,274个数据包，32,979个数据文件，	38355个数据文件

引自：黄金霞，中科院文献情报中心《开放资源每月NEWS》

2.1 开放获取与开放出版

- 商业出版机构从抵制走向开放，选择技术创新与功能创新寻求新的出路
- 绝大多数STM出版机构开始支持开放出版和开放仓储
- BMC，Plos One，PMC等开放出版和开放仓储系统不断壮大
- 施普林格自然集团近日宣布，作者和部分主流媒体将可以向各大平台自由分享旗下超过2700种杂志和每年30万篇新文章。但这些文章仅供浏览，并用于非商业用途。

2.2 语义互操作技术的推动



2.2 语义互操作技术的推动

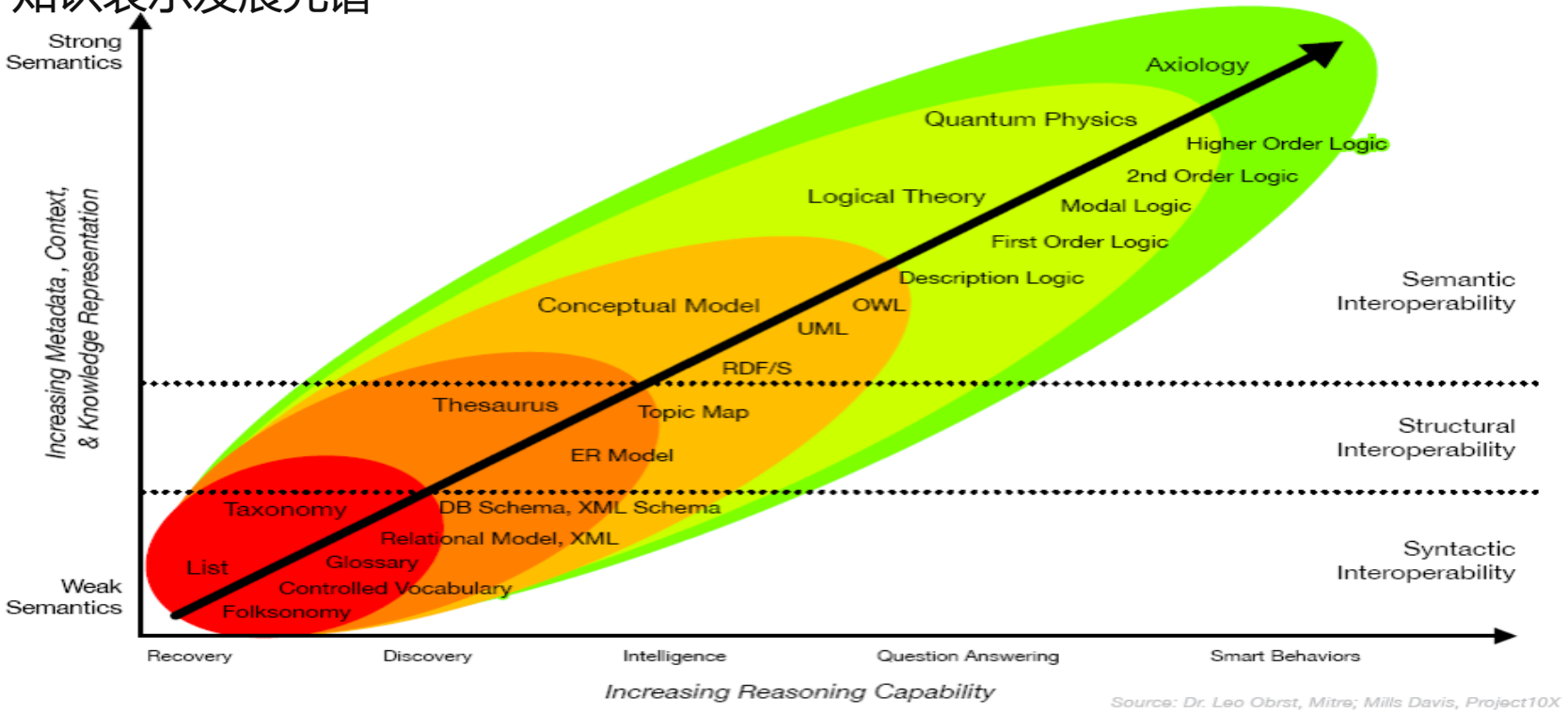
Web App的回归？

App Streaming + App Indexing

基于Web的信息融合仍是未来的基础

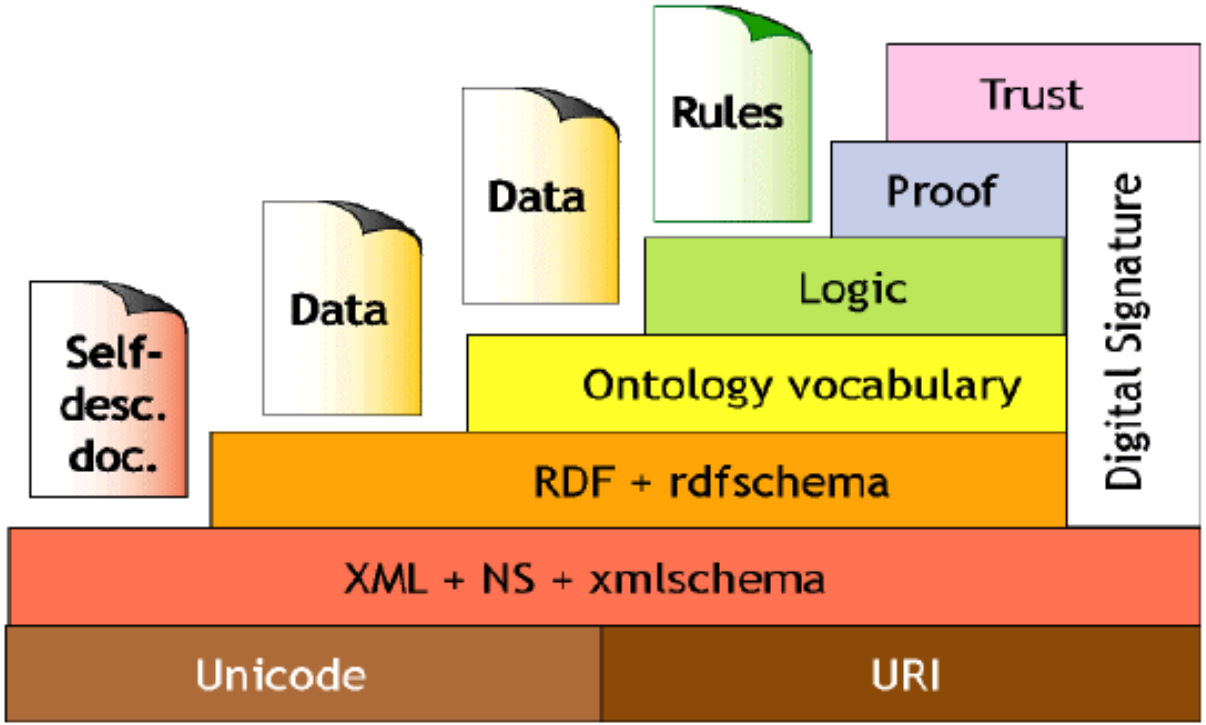
2.2 语义互操作技术的推动

知识表示发展光谱



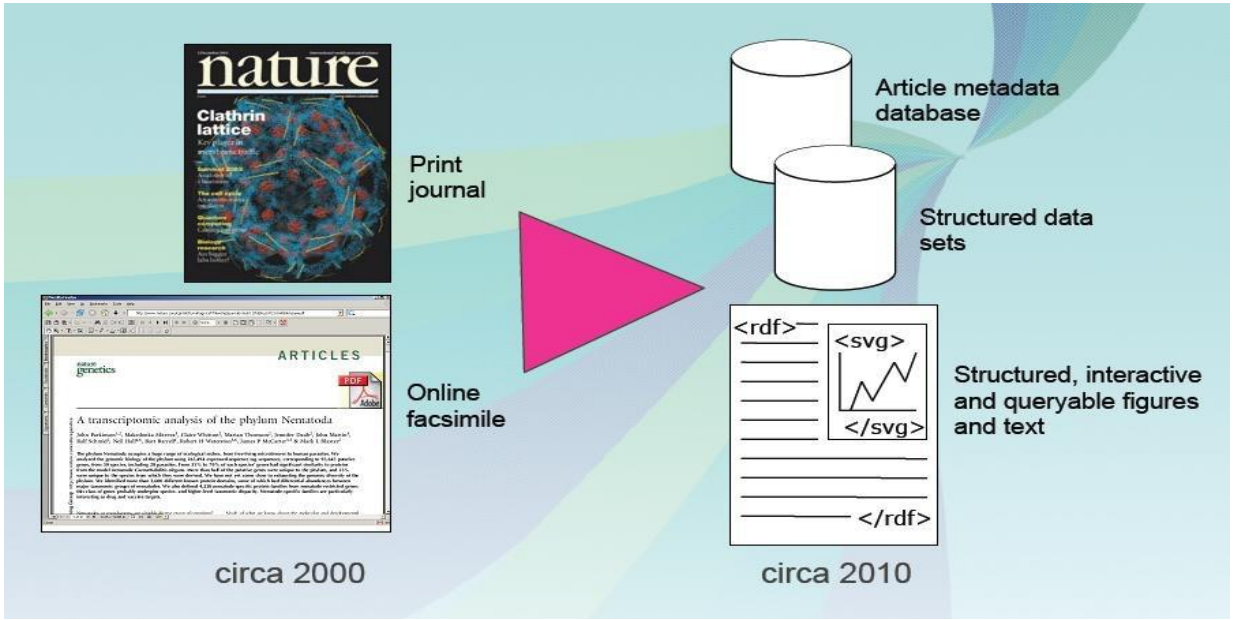
2.2 语义互操作技术的推动

通用技术架构



2.2 语义互操作技术的推动

- 语义出版是在结构化出版基础上，为结构化知识单元附加语义的一种出版模式，通常以RDF+XML+SKOS表示。如，自然出版集团（NPG）明确提出了语义出版的模式，信息的发现是通过结构化的、交互化的、可查询的图表与文本实现

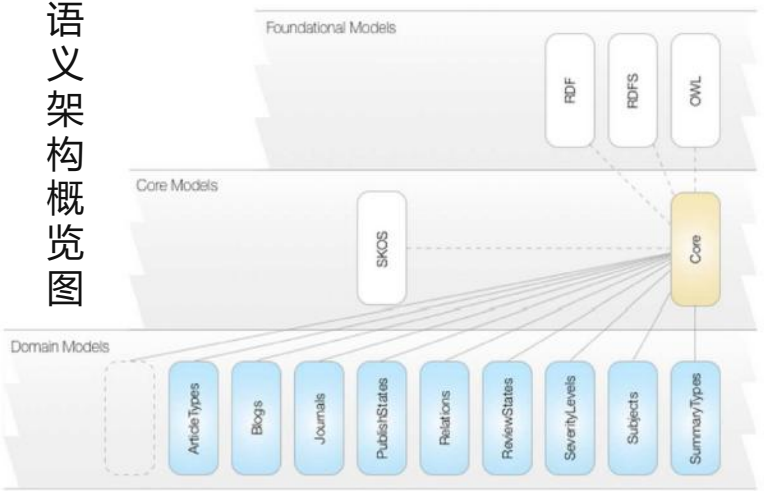


2.2 语义互操作技术的推动

- **nature.com出版平台的主要语义模型与数据集**
  - Core Model：描述了核心企业本体（enterprise ontology），用于实现不同部门之间语义集成以及数据存储。
  - Domain Model：包括更多特定学科或应用模型的具体信息，用于一个或多个系统。
  - Instance Datasets：提供对归档文件中包括的各种数据集的访问。 注明：NPG指Nature Publishing Group，MSE指Macmillan Science and Education。

引自：刘峥《Nature Ontology调研报告》

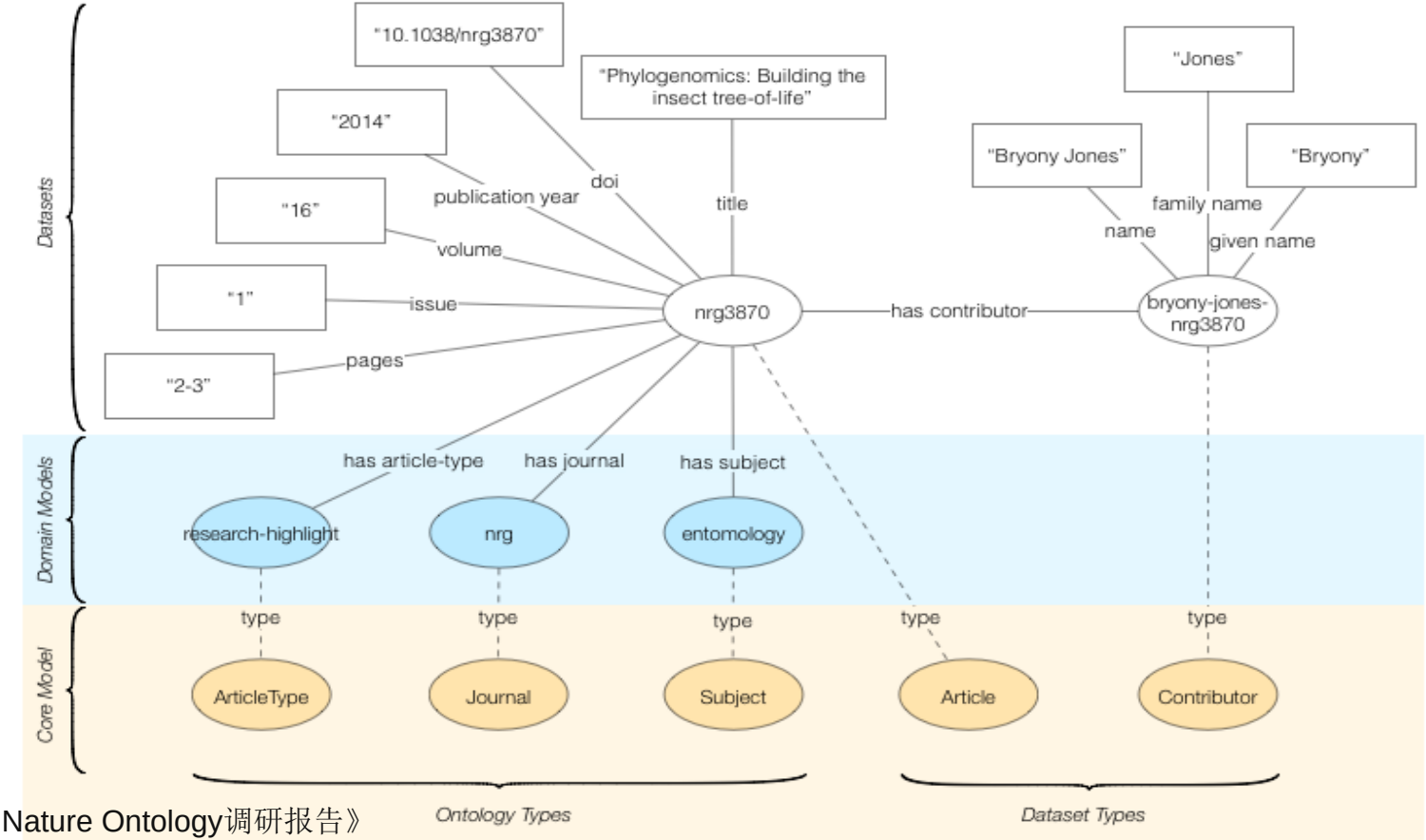
语义架构概览图



- 基础层（foundational layer）由RDF家族语言提供，用于对Core ontology进行编码。
- Core Models目前包括50个类与140个属性
- Domain Model通过继承SKOS模型标准化语义，定义了多个领域层次类别的基础概念，在实例层面利用SKOS分类原语实现。

2.2 语义互操作技术的推动

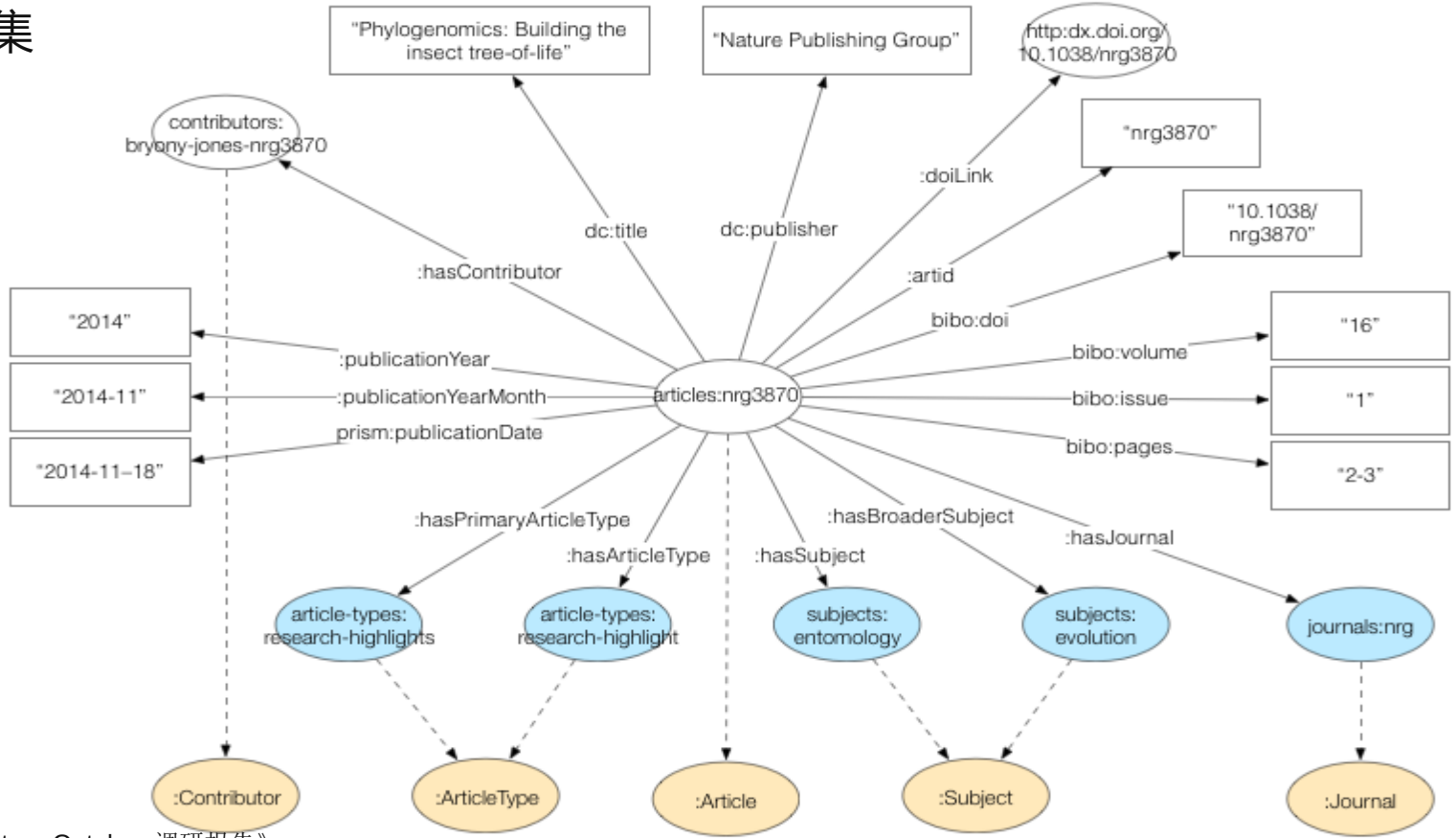
数据集与模型



引自：刘峥《Nature Ontology调研报告》

2.2 语义互操作技术的推动

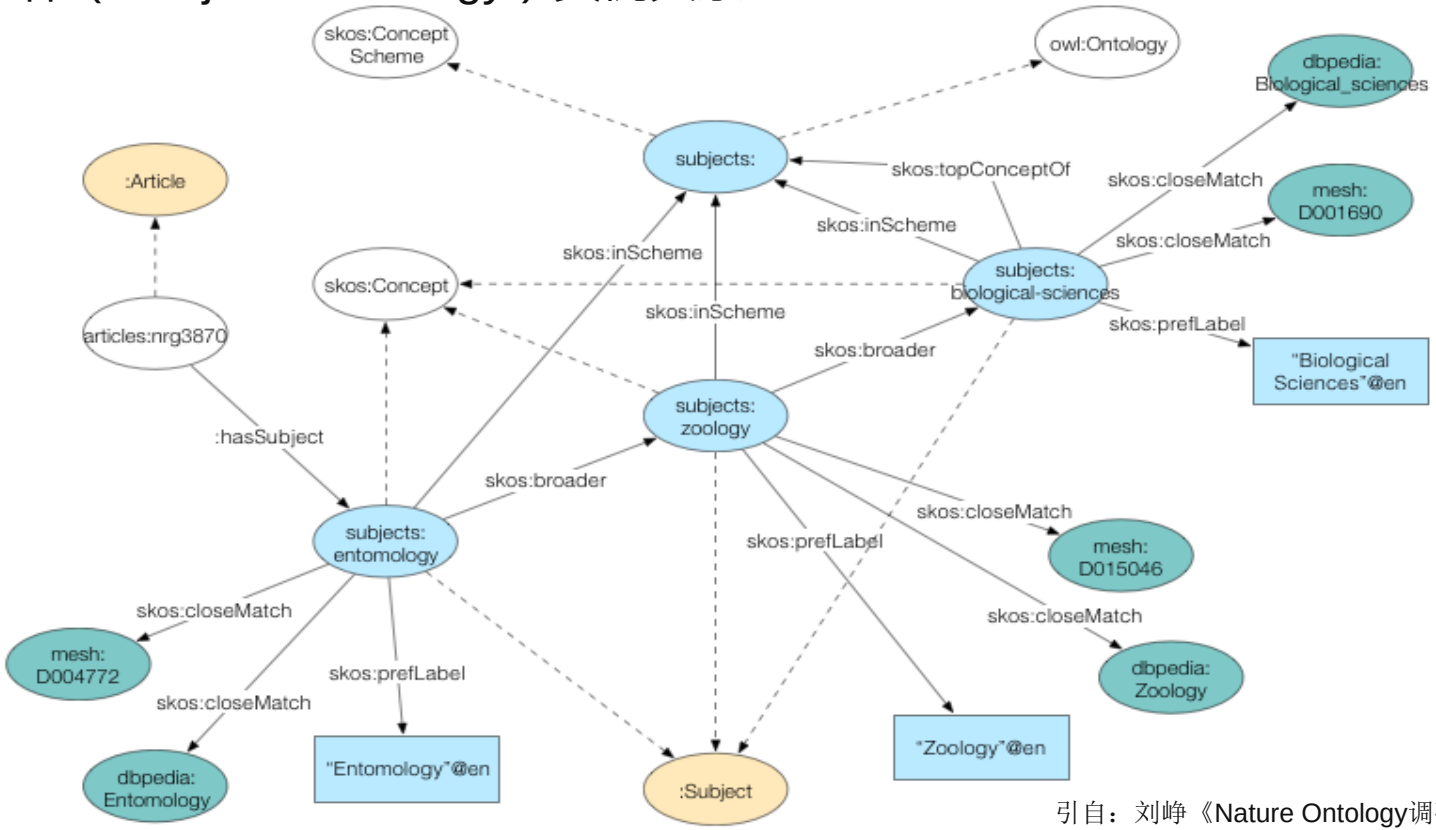
数据集



引自：刘峥《Nature Ontology调研报告》

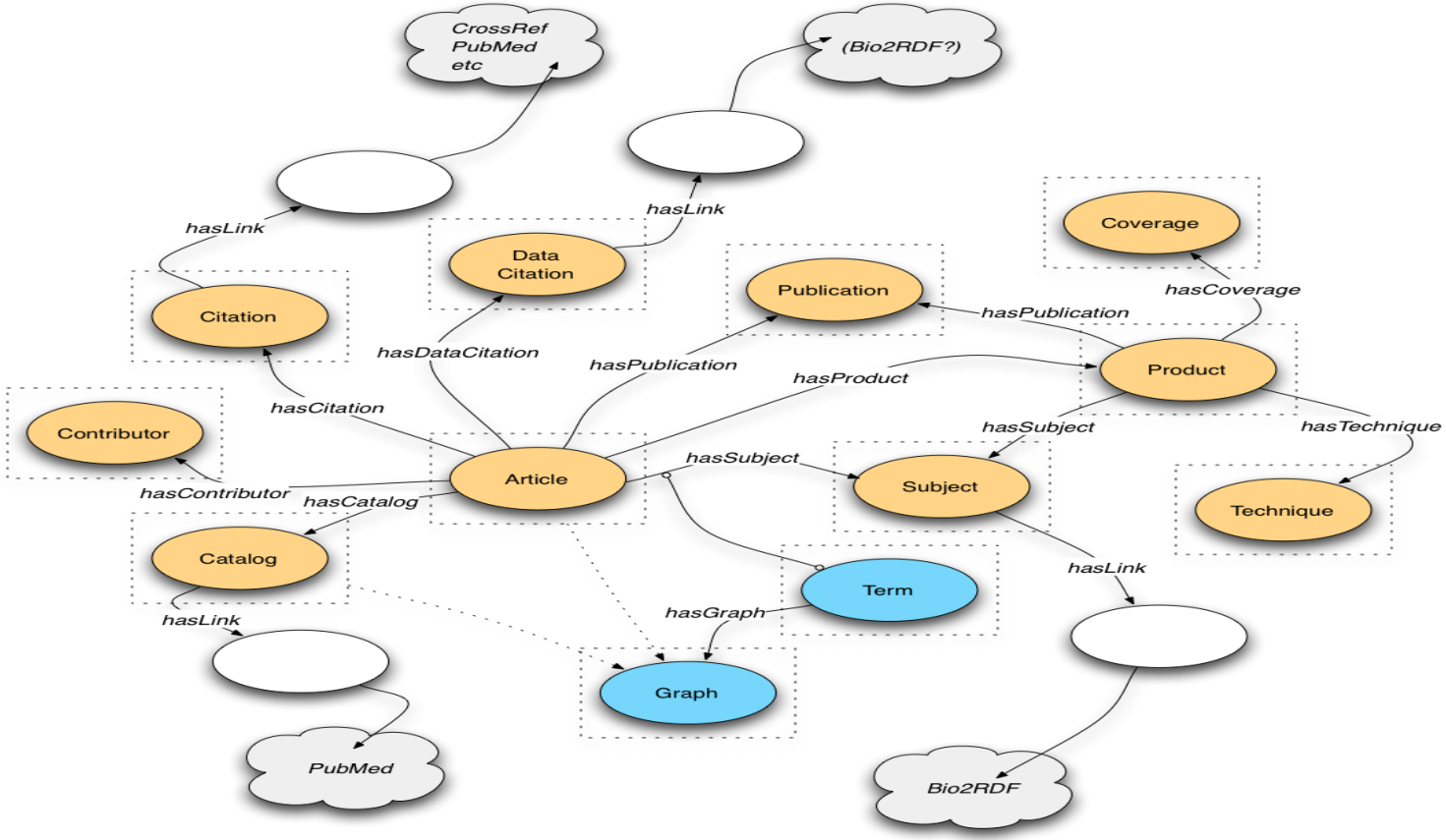
2.2 语义互操作技术的推动

主题本体（Subjects Ontology）实例关系



引自：刘峥《Nature Ontology调研报告》

2.2 语义互操作技术的推动

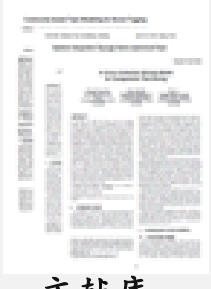


2.2 语义互操作技术的推动

植物多样性复杂数据汇聚

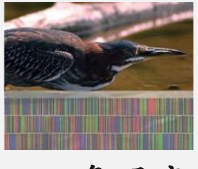
物种名录


标本库


文献库

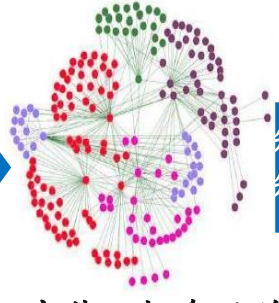

图片库


地名库


DNA条码库


关联数据

知识抽取
知识集成


实体、概念及关系

知识关联
知识计算


植物多样性领域本体

知识检索与关联呈现


语义检索

- 语义索引
- 分面检索
- 本体导航
- 个性化知识
- 多维语义揭示

可视化分析

- 主题聚类
- 热点刻画
- 趋势分析
- 前沿探测


开放关联与集成

- 数据集成融汇
- 自动推理
- 关联发现

2.2 语义互操作技术的推动

在语义技术推动下，信息将越来越多地以结构化、数据化、语义化的数字知识表示形式存在，使知识互联、集成分析、计算挖掘与可视化成为可能！

2.3 大数据技术

DATA

Structured – Data contained in a rigid format with a defined pattern. For example, row and column. Generally numbers, the elements have very specific and well defined patterns.

Unstructured – Data with no particular pattern or formatting. Text and video are considered to be unstructured data.

Semi-structured – Unstructured data that has a format imposed on it. For example, Twitter because it is limited to 140 characters which imposes a structure. Videos and other online content can have some structure added through the use of tags.

Differentiated data – Data that fills a gap in the understanding of a trend or market. It is typically proprietary and can be either developed in-house or by a third party.

Big Data – There is no universal definition of big data. What is big to one person or company can be small to another. Big data generally refers to data that is either large in quantity, fast in collection/ production, or varied in type. Big data generally cannot be handled by traditional statistical tools and techniques, data storage, or

TOOLS

Algorithms – The rules or equations derived from analysis of the data. An algorithm can be as simple as a file with all individuals who use the hashtag #Oscars or who spend more than $1,000 a year with a company. Or it can be a regression equation used, for instance, to predict parts failures.

Analytics – The statistical description that provides an overall understanding of the patterns in the data. Analytics can be as simple as the mean and standard deviation of the data or as complex as predicting the behavior of individuals in the data.

PEOPLE

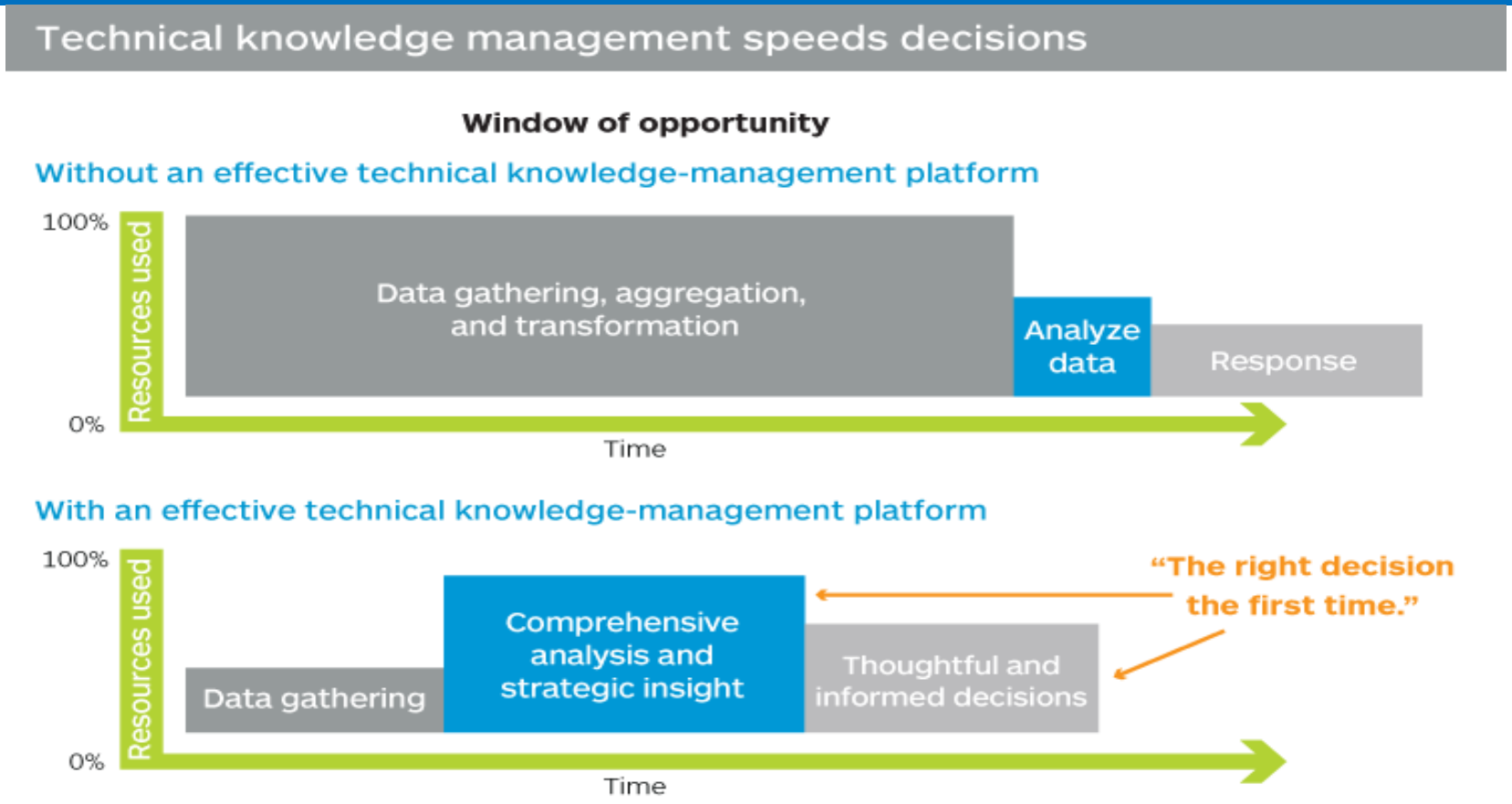
Industry – Expertise in the economic production of a product or service, such as the automotive sector.

Discipline – Expertise in the development of processes that can be applied across a variety of industries, such as supply chain management.

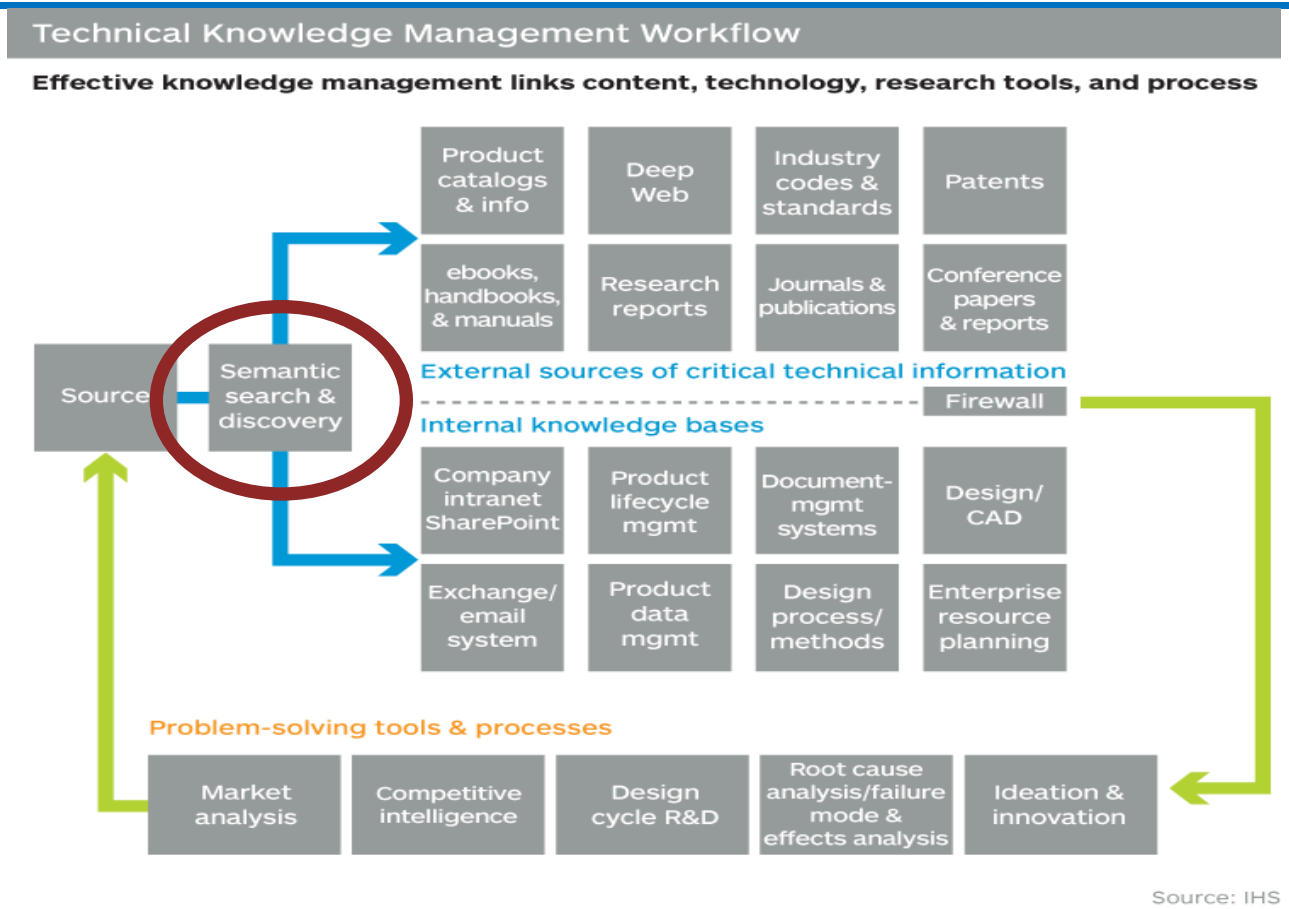
Technical – Expertise in the development of processes requiring advanced knowledge of math and science, such as

John Larson. Big (data).

2.3 大数据技术


Technical knowledge management speeds decisions
Window of opportunity
Without an effective technical knowledge-management platform
100%
Resources used
0%
Data gathering, aggregation, and transformation
Analyze data
Response
Time
With an effective technical knowledge-management platform
100%
Resources used
0%
Data gathering
Comprehensive analysis and strategic insight
Thoughtful and informed decisions
"The right decision the first time."
Time
Source: IHS

2.3 大数据技术


Technical Knowledge Management Workflow
Effective knowledge management links content, technology, research tools, and process
Product catalogs & info
Deep Web
Industry codes & standards
Patents
ebooks, handbooks, & manuals
Research reports
Journals & publications
Conference papers & reports
External sources of critical technical information
Firewall
Internal knowledge bases
Source
Semantic search & discovery
Company intranet SharePoint
Product lifecycle mgmt
Document-mgmt systems
Design/ CAD
Exchange/ email system
Product data mgmt
Design process/ methods
Enterprise resource planning
Problem-solving tools & processes
Market analysis
Competitive intelligence
Design cycle R&D
Root cause analysis/failure mode & effects analysis
Ideation & innovation
Source: IHS

2.3 大数据技术

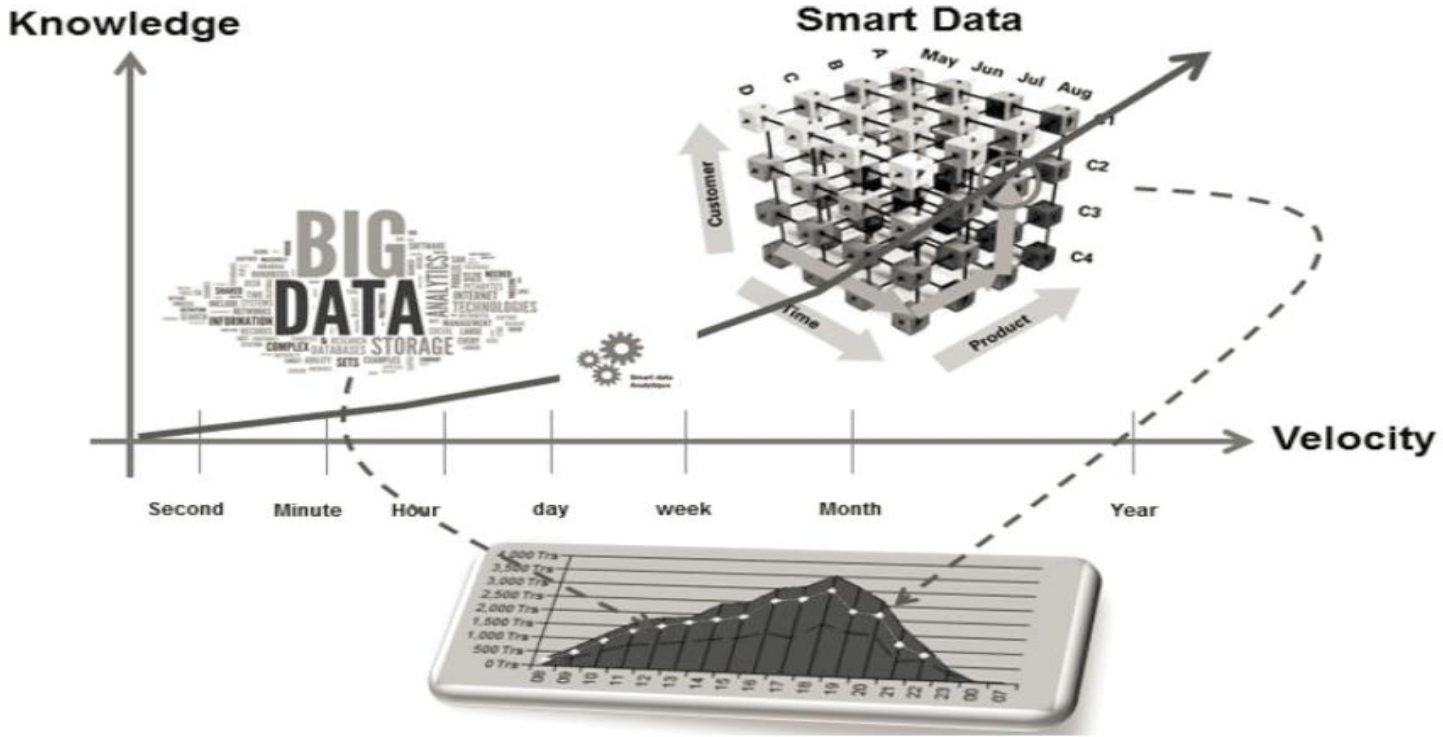
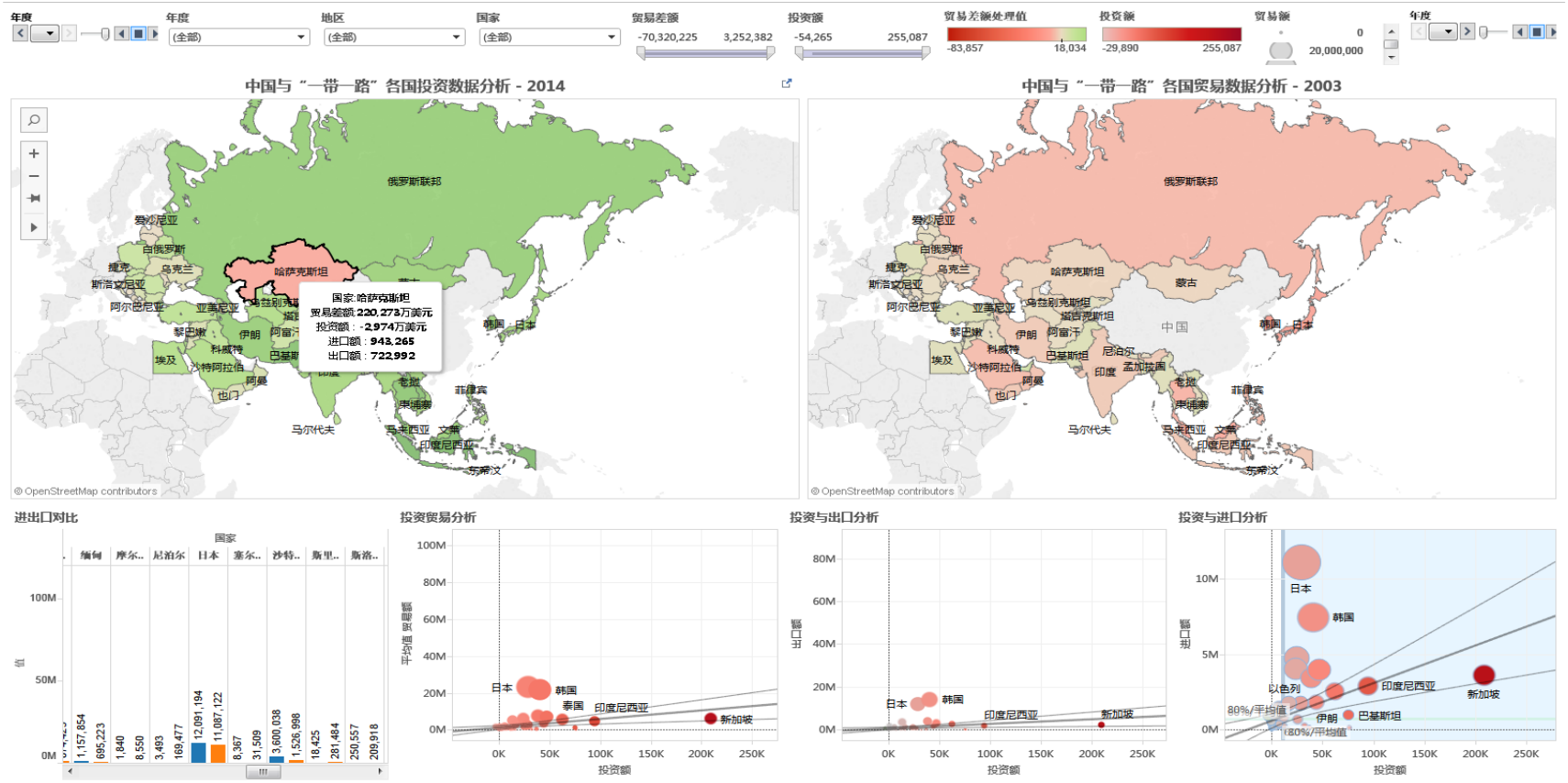


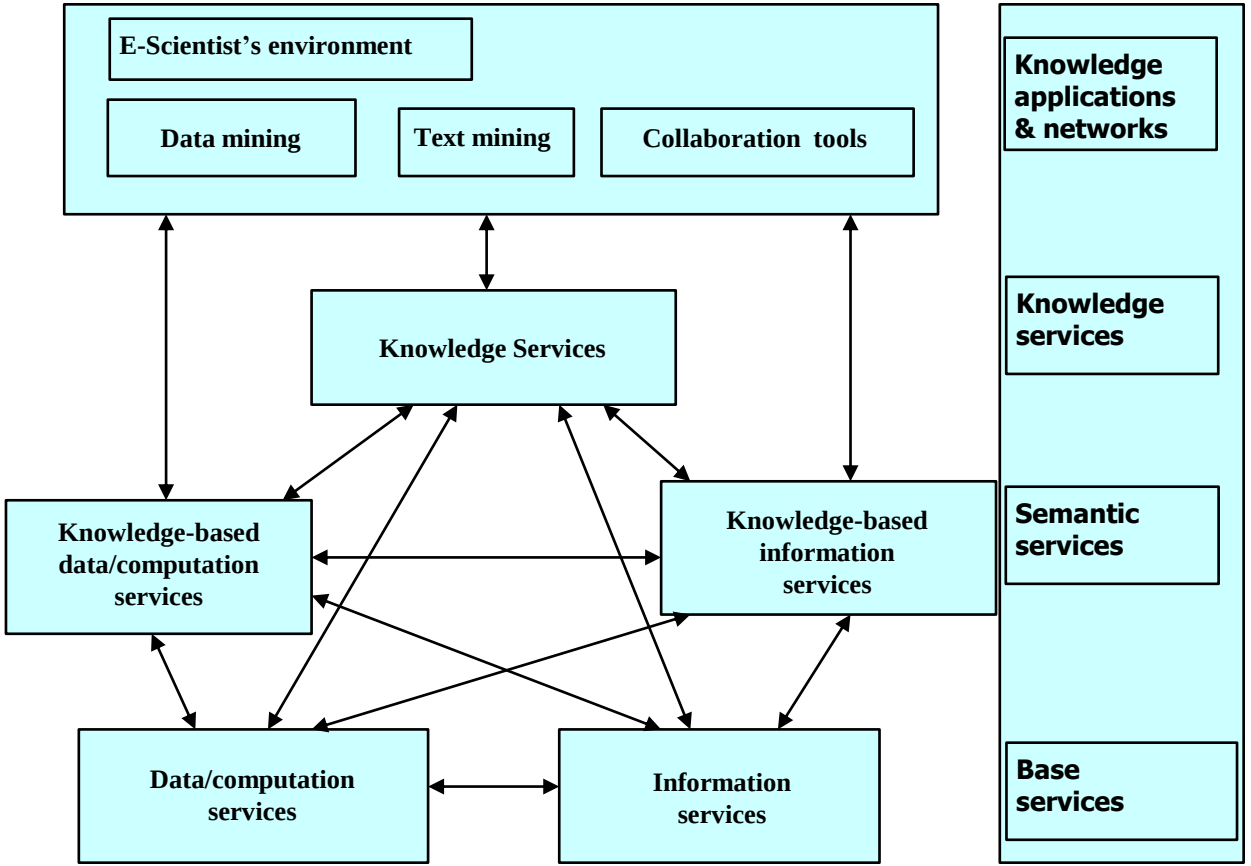
Figure 2.1. *From Big Data to Smart Data, a closed loop*

2.3 大数据技术



多维大数据关联集成分析是根本

2.3 大数据技术


E-Scientist's environment
Data mining
Text mining
Collaboration tools
Knowledge Services
Knowledge-based data/computation services
Knowledge-based information services
Data/computation services
Information services
Knowledge applications & networks
Knowledge services
Semantic services
Base services

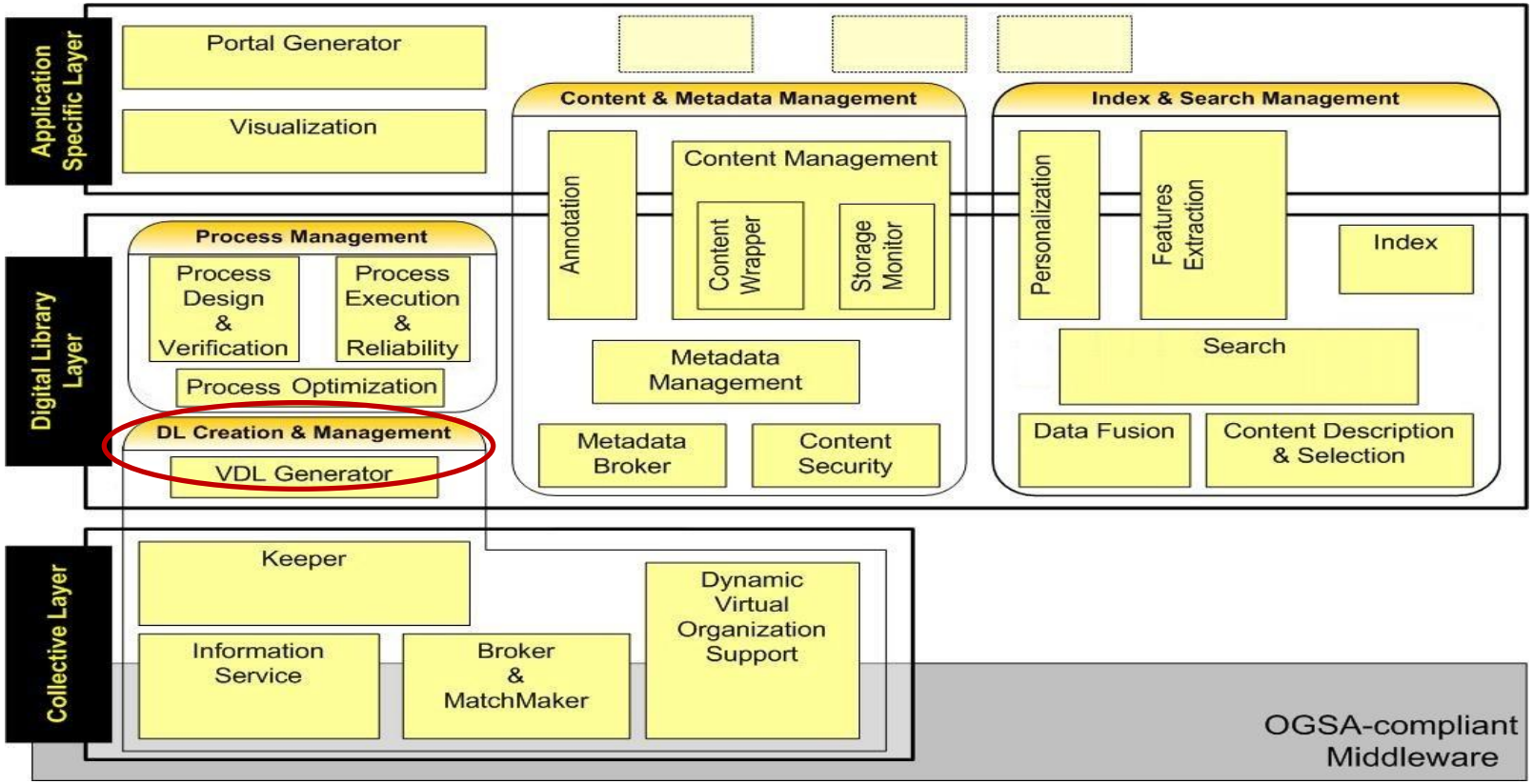
第三节

数字图书馆的创新方向与路径

3. 数字图书馆的创新方向与路径

- 数据密集型知识服务系统：基于智慧数据的数字图书馆
  - 语义驱动的泛在学术搜索与知识发现
  - 个性化的集成知识管理工具
  - 计算化的学习、研究环境

3. 数字图书馆的创新方向与路径



3. 数字图书馆的创新方向与路径

Knowledge Extraction

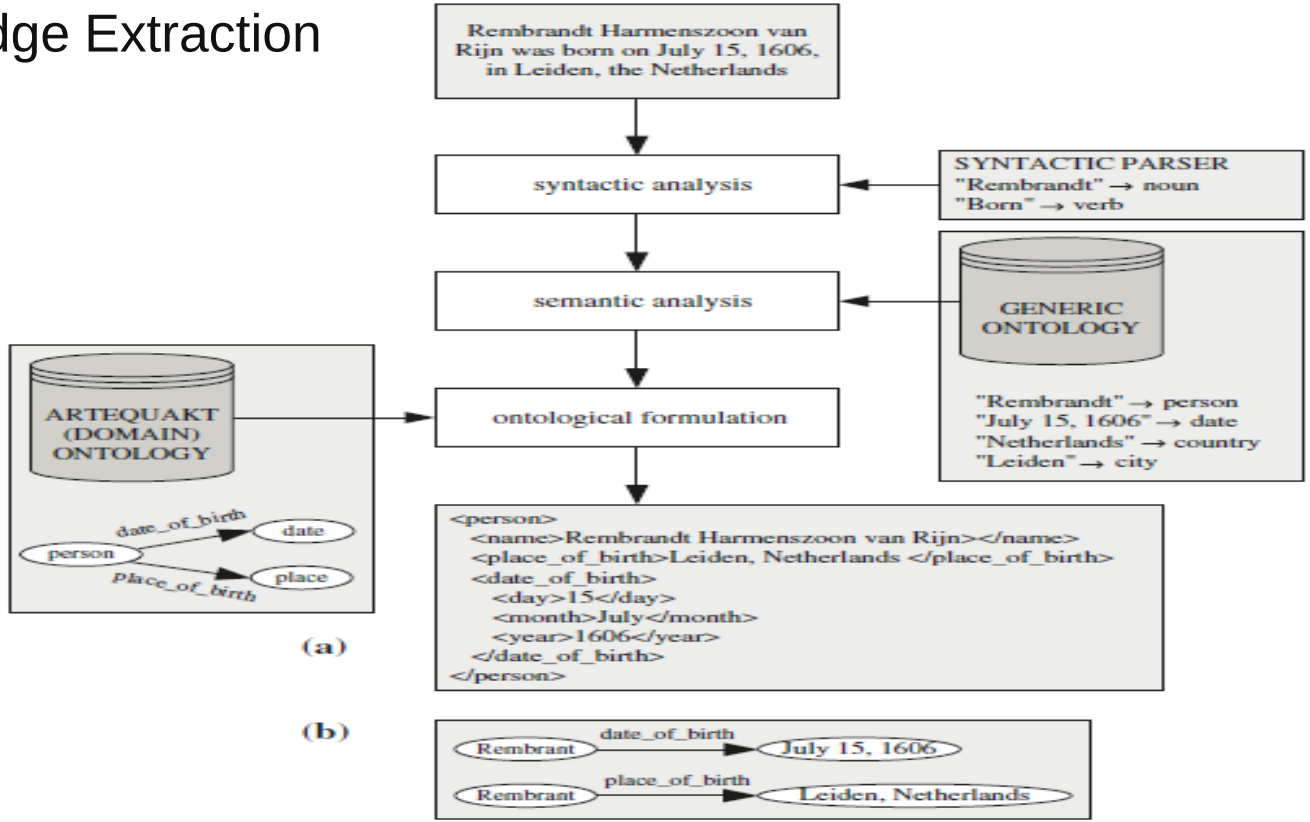
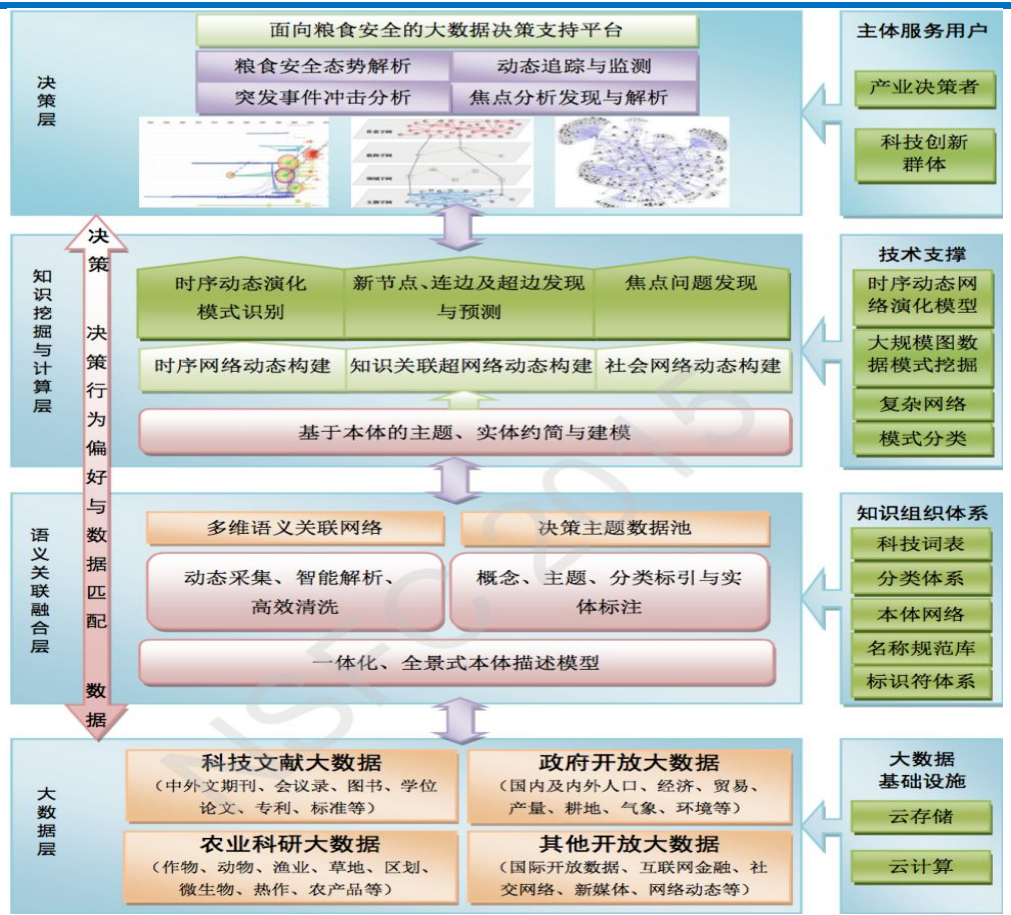
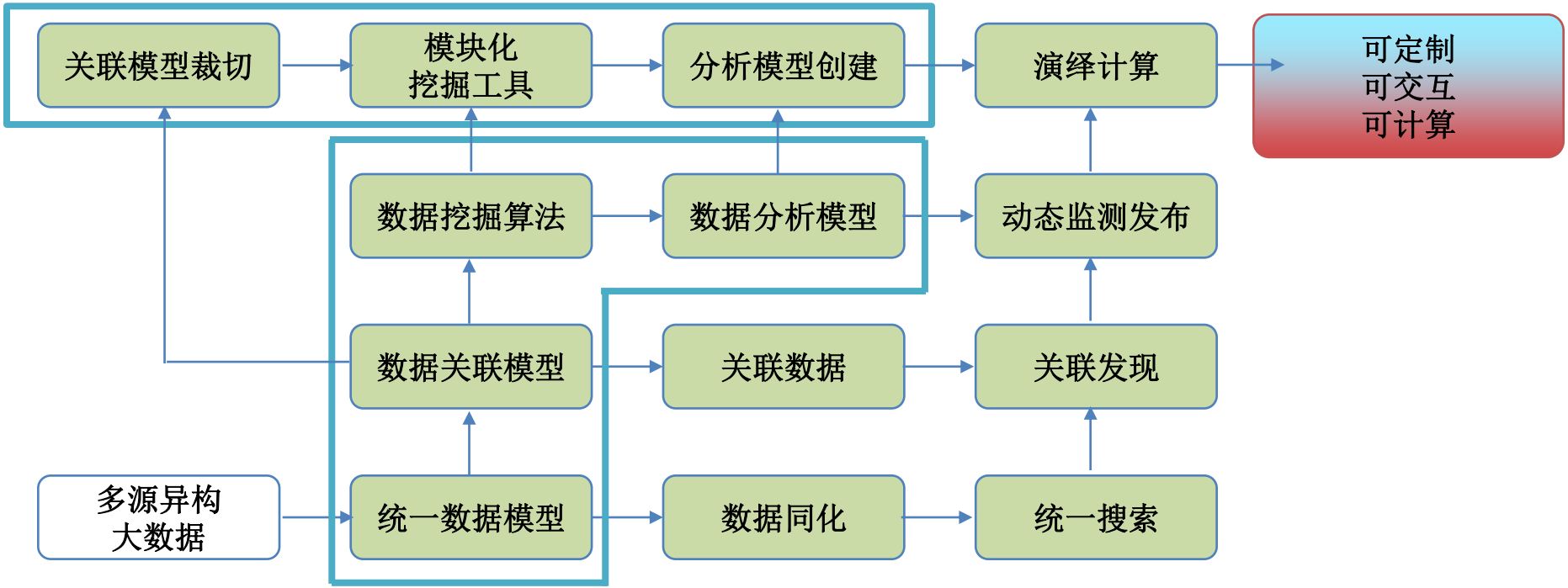


Fig. 3.10 An example of knowledge extraction from a Web page using Artequakt (Alani et al. 2003)

3. 数字图书馆的创新方向与路径



3. 数字图书馆的创新方向与路径



第四节

未来设想与规划

4. 未来设想与规划

NSTL联合CALIS、国家图书馆等共同推动数字图书馆创新，筹划启动十三五专项：数据密集型知识服务关键技术与基础设施建设

- 语义元数据体系构建（KOS&Ontology/Knowledge Extraction）
- 语义学术搜素引擎
- 基于关联分析、计算挖掘的专门知识服务系统
- 标准规范体系与长期保存体系

感谢聆听！