

知识图谱与图书馆知识组织

李广建

北京大学信息管理系

ligj@pku.edu.cn



主要内容

- 知识图谱概述
- 图书馆中的知识组织系统
- 知识图谱构建

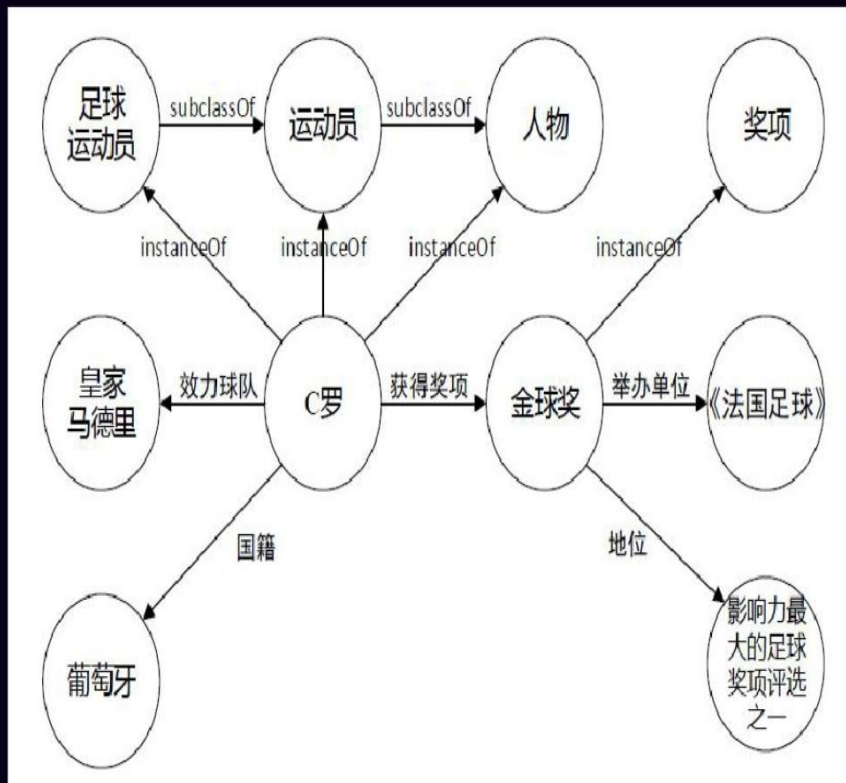


知识图谱概述

- ❑ 知识图谱（Knowledge Graph）是谷歌于2012年提出的
- ❑ 知识图谱是Google用于增强其搜索引擎功能的知识库。
- ❑ 本质上, 知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，其构成一张巨大的语义网络图，节点表示实体或概念，边则由属性或关系构成。
- ❑ 现在，知识图谱已被用来泛指各种大规模的知识库。



知识图谱示意图



- 知识图谱本质上是一种语义网络，将客观的经验沉淀在巨大的网络中
- 结点代表实体 (entity) 或者概念 (concept)
- 边 (edge) 代表实体/概念之间的语义关系

搜索引擎的工作原理

□ 一般搜索引擎

- ⊕ 爬虫（Crawling）– 遍历 Web
- ⊕ Page Ranking – 根据链接判断网页重要性
- ⊕ 索引– 建立网页的关键词索引
- ⊕ 与用户检索词做匹配
- ⊕ 显示结果（相关度、时序等）

□ 构建有知识图谱的搜索引擎

- ⊕ 将相关的对象用图的方式存储起来
- ⊕ 对象与对象之间存在着关联
- ⊕ 在显示检索结果的同时，将知识库中的相关内容也显示出来（可能与用户检索有关的）



Google

匹兹堡



登录

全部 地图 图片 新闻 视频 更多

设置 工具

已启用安全搜索

找到约 5,350,000 条结果 (用时 0.49 秒)

匹兹堡- 维基百科，自由的百科全书

<https://zh.wikipedia.org/zh-sg/匹兹堡>

匹兹堡（英语：Pittsburgh）位于美国宾夕法尼亚州西南部，阿勒格尼河与莫农加希拉河汇合成俄亥俄河的河口，为美国最适宜居住的都市。宾夕法尼亚州第二大城市，阿...

州：宾夕法尼亚州 最低海拔：710 英尺 (220 米)

建市：1794年4月22日 最高海拔：1377 英尺 (420 米)

历史 · 地理 · 经济 · 教育

匹兹堡_百度百科

<https://baike.baidu.com/item/匹兹堡>

匹兹堡（Pittsburgh）位于美国宾夕法尼亚州西南部，在奥里格纳河与蒙隆梅海拉河汇合成俄亥俄河的河口，是阿利根尼县县治，同时也是宾州仅次于费城的第二大城市。

市长：卢克

人口：33万

现任市长：Bill Peduto（民主党）

气候条件：温带大陆性湿润性气候

历史沿革 · 地理环境 · 文化 · 经济

匹兹堡是一个什么样的城市？ - 知乎

<https://www.zhihu.com/question/21170445>

2015年12月11日 - **匹兹堡**的四季2015-03-31 IN 宝妈来的心情还睡不着，于是决定把想了好久的文章写了，给那个心心念念...

匹兹堡有哪些值得一去的地方？

2016年9月30日



匹兹堡

宾夕法尼亚州的都市

匹兹堡位于美国宾夕法尼亚州西南部，阿勒格尼河与莫农加希拉河汇合成俄亥俄河的河口，为美国最适宜居住的都市。宾夕法尼亚州第二大城市，阿利根尼县县治。2000年人口统计334,563。匹兹堡曾是美国著名的钢铁工业城市，有“世界钢都”之称。 [维基百科](#)

天气： 21°C，风向西，风速 5 公里/时，湿度 94%

人口： 30.24 万 (2017 年)

市长： 比尔·彼杜托

旅游景点： 菲普斯温室植物园，匹兹堡动物园&PPG水族馆， [更多](#)

俱乐部和球队： 匹兹堡海盗，匹兹堡钢人，匹兹堡企鹅，匹兹堡猎犬河， [更多](#)



匹兹堡



登录

全部

地图

图片

新闻

视频

更多

设置

工具

已启用安全搜索

This is what's new

找到约 5,350,000 条结果 (用时 0.49 秒)

匹兹堡- 维基百科, 自由的百科全书

<https://zh.wikipedia.org/zh-sg/匹兹堡>

匹兹堡（英语：Pittsburgh）位于美国宾夕法尼亚州西南部，阿勒格尼河与莫农加希拉河汇合成俄亥俄河的河口，为美国最适宜居住的都市。宾夕法尼亚州第二大城市，阿 ...

州: 宾夕法尼亚州 最低海拔: 710 英尺 (220 米)

建市: 1794年4月22日 最高海拔: 1,370 英尺 (420 米)

历史 · 地理 · 经济 · 教育

匹兹堡_百度百科

<https://baike.baidu.com/item/匹兹堡>

匹兹堡（Pittsburgh）位于美国宾夕法尼亚州西南部，在奥里格纳河与蒙隆梅海拉河汇合成俄亥俄河的河口，是阿利根尼县县治，同时也是宾州仅次于费城的第二大城市。

市长: 卢克

人口: 33万

现任市长: Bill Peduto（民主党）

气候条件: 温带大陆性湿润性气候

历史沿革 · 地理环境 · 文化 · 经济

匹兹堡是一个什么样的城市? - 知乎

<https://www.zhihu.com/question/21170415>

2015年12月11日 - **匹兹堡**的四季2015-03-31NN女码农的小情怀睡不着，于是决定把想了好久的文章写了，给那个心心念念...

匹兹堡有哪些值得一去的地方？

2016年9月30日



匹兹堡

宾夕法尼亚州的都市

匹兹堡位于美国宾夕法尼亚州西南部，阿勒格尼河与莫农加希拉河汇合成俄亥俄河的河口，为美国最适宜居住的都市。宾夕法尼亚州第二大城市，阿利根尼县县治。2000年人口统计334,563。匹兹堡曾是美国著名的钢铁工业城市，有“世界钢都”之称。 [维基百科](#)

天气: 21°C, 风向西, 风速 5 公里/时, 湿度 94%

人口: 30.24 万 (2017 年)

市长: 比尔·彼杜托

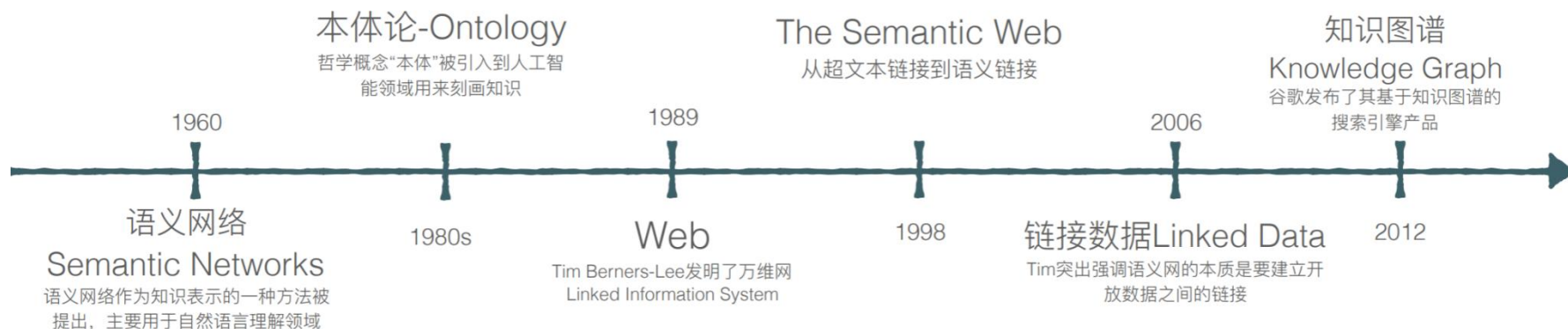
旅游景点: 菲普斯温室植物园, 匹兹堡动物园&PPG水族馆, [更多](#)

俱乐部和球队: 匹兹堡海盗, 匹兹堡钢人, 匹兹堡企鹅, 匹兹堡猎犬河, [更多](#)

□ 知识图谱的发展

⊕ 与知识图谱相关的概念

- 语义网络 (Semantic Net/ Semantic Network)
- 本体 (Ontology)
- 语义网 (The Semantic Web)
- 关联数据 (Linked Data)
- 知识库 (Knowledge Base)



知识表示与知识库- Knowledge Representation / Knowledge Base

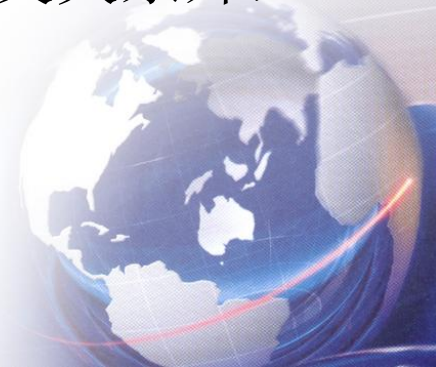
□ 语义网络（Semantic Net/ Semantic Network）

- ⊕ 语义网络最早是1960年由认知科学家Allan M. Collins作为知识表示的一种方法提出。
- ⊕ 它其实是一种有向图，其中，结点代表的是概念，而边则表示的是这些概念之间的语义关系。
- ⊕ 语义网络是机读型字典（machine-readable dictionary）的一种常见类型
 - WordNet是最典型的语义网络
- ⊕ 相比起知识图谱，早期的语义网络更加侧重描述概念以及概念之间的关系，而知识图谱更加强调数据或事物之间的链接。



□ WordNet是最典型的语义网络

- ⊕ WordNet是一个由普林斯顿大学认识科学实验室在心理学教授乔治·A·米勒的指导下建立和维护的英语字典。开发工作从1985年开始，由于它包含了语义信息，所以有别于通常意义上的按字顺排列字典。
- ⊕ WordNet根据词条的意义将它们分组，每一个具有相同意义的字条组称为一个synset（同义词集合）。WordNet为每一个synset提供了简短、概要的定义，并记录不同synset之间的语义关系。
- ⊕ 同时， WordNet中除了具有中心角色的同义关系外，还有反义关系、上下位关系和部分关系。



查Library这个词 (<http://wordnetweb.princeton.edu/perl/webwn>)

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- [S: \(n\)](#) **library** (a room where books are kept) *"they had brandy in the library"*
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [S: \(n\)](#) **room** (an area within a building enclosed by walls and floor and ceiling) *"the rooms were very small but they had a nice view"*
 - [S: \(n\)](#) **area** (a part of a structure having some specific characteristic or function) *"the spacious cooking area provided plenty of room for servants"*
 - [S: \(n\)](#) **structure, construction** (a thing constructed; a complex entity constructed of many parts) *"the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons"*
 - [S: \(n\)](#) **artifact, artefact** (a man-made object taken as a whole)
 - [S: \(n\)](#) **whole, unit** (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
 - [S: \(n\)](#) **object, physical object** (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
 - [S: \(n\)](#) **physical entity** (an entity that has physical existence)
 - [S: \(n\)](#) **entity** (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
- [part holonym](#)
- [S: \(n\)](#) **library** (a collection of literary documents or records kept for reference or borrowing)
- [S: \(n\)](#) **library, depository library** (a depository built to contain books and other materials for reading and study)
- [S: \(n\)](#) **library, program library, subroutine library** ((computing) a collection of standard programs and subroutines that are stored and available for immediate use)
- [S: \(n\)](#) **library** (a building that houses a collection of books and other materials)

□ 本体（Ontology）

- ⊕ 本体本身是个哲学名词，是“being”的性质及其内在关系的理论。
- ⊕ 在上个世纪80年代，人工智能研究人员将这一概念引入了计算机领域，被用于促进知识的共享和重用。
- ⊕ 近十几年来，关于Ontology的定义有多个。最能描述Ontology本质的是斯坦福大学的Gruber在1993年的定义：
- ⊕ Ontology是一个对共享概念的形式化的、显性的规定。
 - “概念”指的是对现实现象的抽象
 - “显性的”意味着对概念的类型、以及对概念的应用是显性定义的
 - “形式化的”意味着Ontology应当是机器可读的
 - “共享”意味着一个Ontology揭示着广泛共享的知识，也就是说，一个Ontology不仅仅被几个个体接受，而需要被一个群体所接受
- ⊕ 本质上讲，Ontology在知识工程中的角色在于促进一个领域模型的建设。一个Ontology提供了某个领域模型内语词和相互关系的描述。本体相似于数据库中的Schema，本体通常被用来为知识图谱定义Schema。



□ 本体描述语言

⊕ RDF

- 资源描述框架

⊕ OIL

- Ontology Interchange Language或Ontology Inference layer（本体推理层），是在Web上描述本体的语言。OIL的语义基于描述逻辑（description logic），语法建立在RDF之上

⊕ DAML

- DAML(DARPA Agent Markup Language)，DAML力图包括RDF、OIL等语言的优点

⊕ OWL

- OWL（Web Ontology Language），W3C的标准



□ 语义网 (Semantic Web)

⊕ 是Web之父Tim Berners Lee于1998年提出的

- 语义网不是一个独立的网络，而是当前的这个网络的扩展。在语义网中信息被赋予了明确界定的意义、从而能更好地使计算机和人协调工作。把语义网纳入现有网络结构中的首批步骤已经在着手进行之中。在不远的将来，计算机处理并“理解”那些现今它们只能显示的数据的能力将会大大提高，这样语义网建设的进展就将有助于创造出引人注目的全新功能。

[Tim Berners-Lee (Web和语义网的创始人), 语义网——科学美国人, 2001]

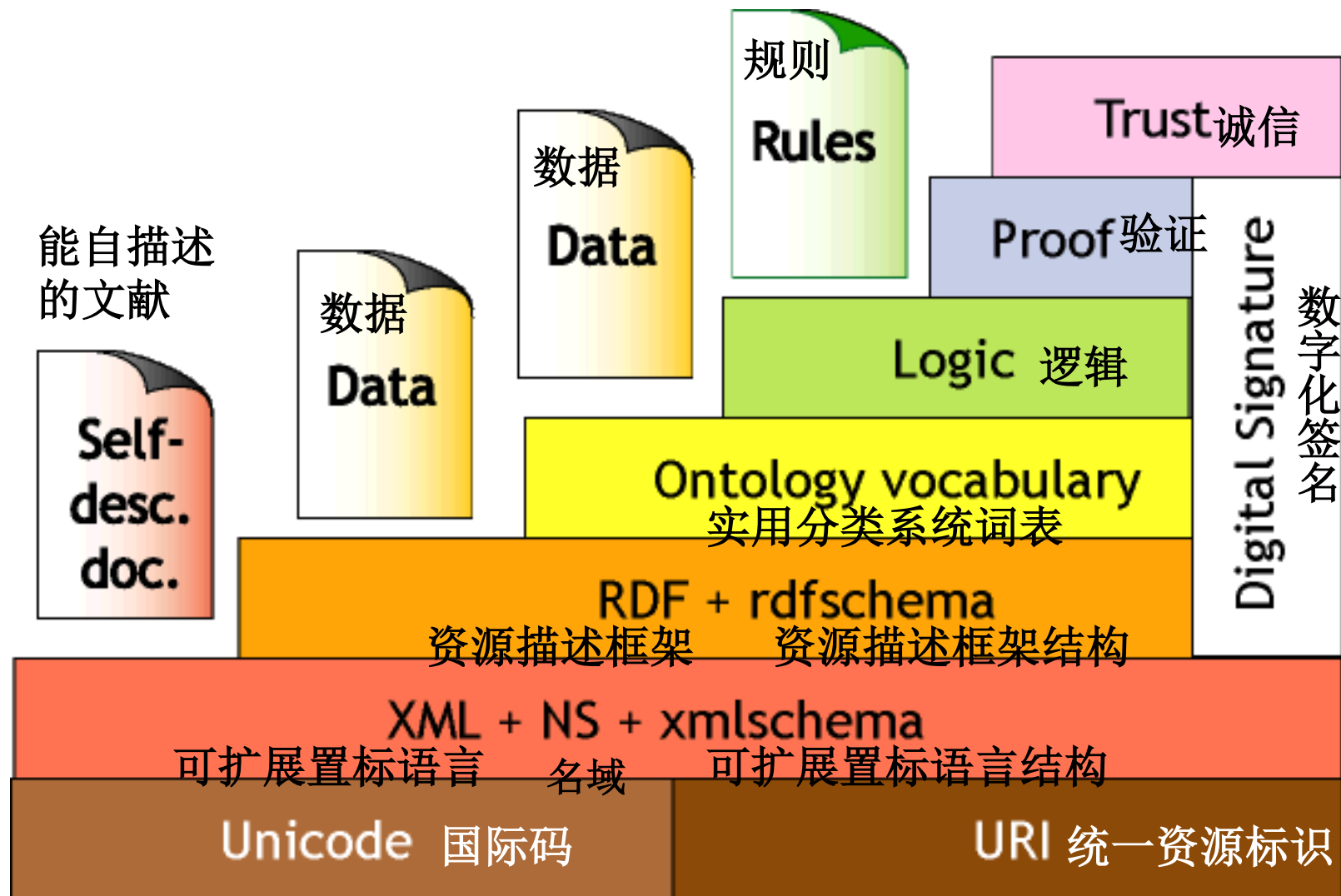
⊕ 要点

- 是当前Web的延伸；
- 加入计算机可以理解的语义

⊕ 谷歌知识图谱是语义互联网这一理念的商业化实现。也可以把语义互联网看做是一个基于互联网共同构建的全球知识库。



与语义网有关的标准和技术



□ 关联数据（Linked Data）

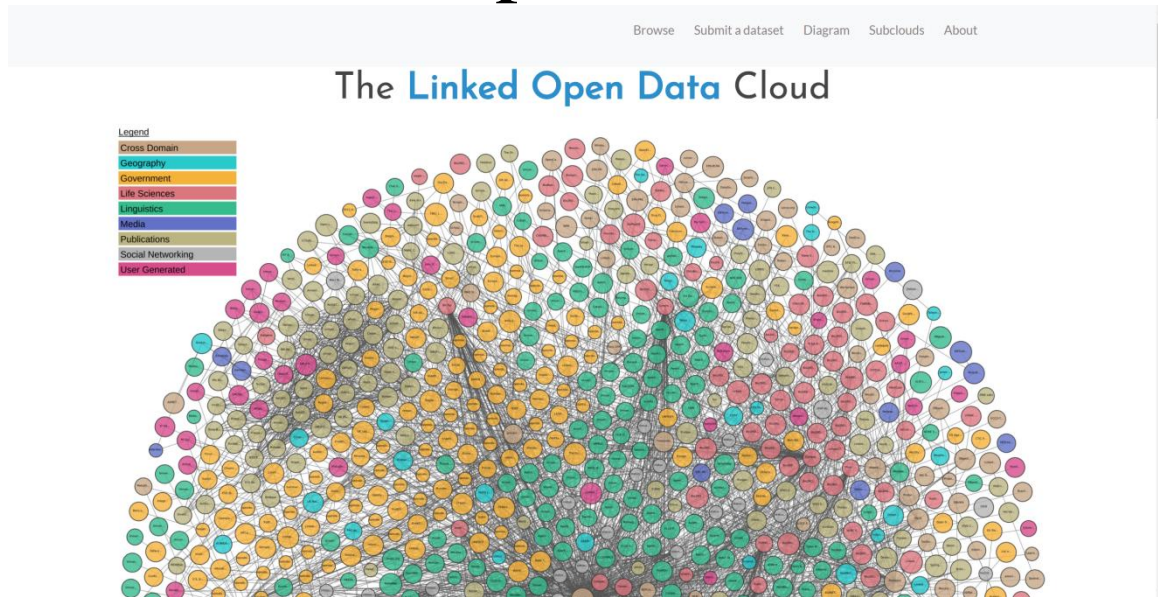
- ⊕ 是Tim Berners Lee于2006年提出，是为了强调语义互联网的目的是要建立数据之间的链接，而非仅仅是把结构化的数据发布到网上。他为建立数据之间的链接制定了四个原则
 - ⊕ 使用URI作为任何事物的标识名称（Use URIs as names for things）
 - ⊕ 使用HTTP URI使任何人都可以访问名称（Use HTTP URIs so that people can look up those names）
 - ⊕ 当有人访问名称时，提供有用的信息（When someone looks up a name, provide useful information）
 - ⊕ 尽可能提供相关的URI以使人们发现更多的信息（Include links to other URIs so that they can discover more things）
- ⊕ 从理念上讲，关联数据最接近于知识图谱的概念，使用三元组来表示数据及其关系。



- 近年来逐渐得到学术界、工业界及政府部门的广泛关注，包括BBC、纽约时报、MIT、IEEE、HCLS、美国国会图书馆等机构纷纷加入到关联数据的出版发布行列。
- 2007年5月，W3C的关联开放数据运动（Linking Open Data, LOD）正式启动，该运动的宗旨是号召人们将现有数据发布成关联数据，并将不同数据源互联起来。
- 截至2019年3月，LOD云图中已经包含**1239**个数据集，包括跨域、地理、政府、生命科学、语言学、媒体、出版物、社交网络、用户生成等9类。每个数据集集中有从数万到数千万个RDF三元组



LOD云图 (<https://lod-cloud.net/>)



Subclouds by domain

Cross-Domain Geography Government Life Sciences Linguistics Media Publications Social Networking User-Generated

Legend

Wikipedia Related

Other



□ 知识库（Knowledge Base）

- ⊕ 源于人工智能和数据库领域，在AI领域中，最早作为专家系统（Expert System）的组成部分，用于支持推理。
- ⊕ 知识库中的知识有很多种不同的形式，例如本体知识、关联性知识、规则库、案例知识等。
- ⊕ 相比于知识库的概念，知识图谱更加侧重关联性知识的构建，如三元组。



□ 小结

⊕ 语义网络、本体与知识图谱

- 语义网络、本体都强调概念及其关系
- 知识图谱强调数据之间的关系，是上述两者的具体化

⊕ 语义网与知识图谱

- 语义网不仅是技术，还是理念
- 知识图谱是一种具体实现，不是语义网的全部

⊕ 关联数据与知识图谱

- 知识图谱可以看做是关联数据的一种实现，但很多商业知识图谱的具体实现并不一定完全遵循Tim所提出的四个原则

⊕ 知识库与知识图谱

- 构建好的知识图谱，存储起来，就是一种知识库，是知识库的一种存在形式
- 知识库有多种形式



主要内容

- 知识图谱概述
- 图书馆中的知识组织系统
- 知识图谱构建



图书馆中的知识组织系统

□ 元数据

- ⊕ 编目条例、DC、.....

□ 分类与主题

- ⊕ 分类表（如中图法）、主题词表（如 LCSH）、叙词表（如汉表）、.....

□ 数据集

- ⊕ MARC、.....

□ 规范档

- ⊕ 包括术语表、词典、.....



□ 图书馆有长期的知识组织传统经验，并且已经建立类似于当今知识图谱特征的系统

□ 主要问题在于：

- ⊕ 图书馆知识组织面向的对象是图书馆的馆藏，而不是网络，所以我们的行业术语和标准的通用性不强，受关注度不够
- ⊕ 图书馆数据没有很好地与Web资源集成
- ⊕ 在知识组织过程中主要还是面向人，而不是机器，形式化表达做的不够好



主要内容

- 知识图谱概述
- 图书馆中的知识组织系统
- 知识图谱构建



知识图谱构建

- 构建方式
- 知识图谱的逻辑结构
- 构建流程
- RDF与三元组
- 构建工具



知识图谱构建

□ 构建方式

⊕ 完全自建

- 先为知识图谱定义好本体与数据模式，再将实体加入到知识库。
 - 相当于先建词表，再标引

⊕ 复用知识库

- 以现有的知识图谱为基础，抽取所需的实体加入到知识库，再构建顶层的本体模式。
- 目前，大多数知识图谱都采用自底向上的方式进行构建，也符合互联网数据内容知识产生的特点。



□ 典型的知识图谱项目

名称	说明
Freebase	早期由Metaweb公司创建，后来被Google收购，成为Google知识图谱的重要组成部分。Freebase中的数据主要是由人工构建，另外一部分数据则主要来源于维基百科、IMDB、Flickr等网站或语料库。在2015年6月，Freebase整体移入至WikiData。
NELL	是卡内基梅隆大学开发的知识库。NELL主要采用互联网挖掘的方法从Web自动抽取三元组知识。
Wikidata	维基媒体基金会主持的一个自由的协作式多语言辅助知识库，旨在为维基百科、维基共享资源以及其他的维基媒体项目提供支持。
DBpedia	由德国莱比锡大学和曼海姆大学的科研人员创建的多语言综合型知识库，以关联数据的形式发布到互联网上，DBpedia还能够自动与维基百科保持同步，覆盖多种语言。
YAGO	YAGO是由德国马普研究所构建的综合型知识库。整合了维基百科、WordNet以及GeoNames等数据源。
Concept Graph	微软Concept Graph是以概念层次体系为中心的知识图谱。
Schema.org	由Bing、Google、Yahoo和Yandex等搜索引擎公司共同支持的语义网项目。其本质是采用互联网众包的方式生成和收集高质量的知识图谱数据。
OpenKG.CN	是中国中文信息学会语言与知识计算专业委员会所倡导的开放知识图谱项目，旨在促进中文知识图谱数据的开放与互联，促进知识图谱和语义技术的普及和广泛应用。
Zhishi.me	是中文常识知识图谱。主要通过从开放的百科数据中抽取结构化数据，已融合了百度百科、互动百科以及维基百科中的中文数据。由中国东南大学知识科学与工程实验室维护。
CN-DBpedia	是由复旦大学知识工场实验室研发并维护的大规模通用领域结构化中文知识库。

□ 知识图谱的逻辑结构

- ⊕ 在逻辑上，将知识图谱划分为两个层次：数据层和模式层。
- ⊕ 模式层：在数据层之上，是知识图谱的核心，存储经过提炼的知识，通常通过本体库来管理这一层（本体库可以理解为面向对象里的“类”这样一个概念，本体库就储存着知识图谱的类）。
- ⊕ 数据层：存储真实的数据。
 - 例如：
 - 模式层：实体-关系-实体，实体-属性-属性值
 - 数据层：比尔·盖茨-妻子-梅琳达·盖茨，比尔·盖茨-性别-男性
- ⊕ 数据层是模式层的具体化



□ 构建流程

⊕ 本体建设

- 我们有很多经验了（非网络的），向前一步即可（机读）

⊕ 数据处理

- 我们也有很多经验，就是我们说的标引，但与传统之处不同之处在于
 - 实体识别与抽取（包括消歧和补全）
 - 关系识别与抽取（包括消歧和补全）
 - 三元组构建

⊕ 构建知识图谱（知识库）

- 有很多工具可以辅助完成这项工作



□ RDF与三元组

⊕ 知识图谱的构成

➤ 知识图谱的基本单位：三元组

- 实体-关系-实体
- 实体-属性-属性值

⊕ 常用的存储格式

➤ RDF



□ 什么是RDF

- ⊕ RDF（Resource Description Framework），资源描述框架
- ⊕ 资源描述框架（RDF）是由W3C开发的一个资源描述规范。其最初的目标就是解决不同元数据的互操作问题。
- ⊕ 在“Web资源（Web resource）”这一概念一般化后，RDF可用于表达关于任何可在Web上被标识的事物的信息，即使有时它们不能被直接从Web上获取。
 - 比如关于一个在线购物机构的某项产品的信息（例如关于规格、价格和可用性信息），或者是关于一个Web用户在信息递送方面的偏好的描述。



□ RDF的基本思想:

- ⊕ 用Web标识符（称作统一资源标识符，Uniform Resource Identifiers或URI）来标识事物
- ⊕ 用简单的属性（property）及属性值来描述资源。
- ⊕ 这使得RDF用资源标识-属性类型-属性值这样的表示模型来定义一个或多个资源。
- ⊕ 这种模型就是RDF的三元模型

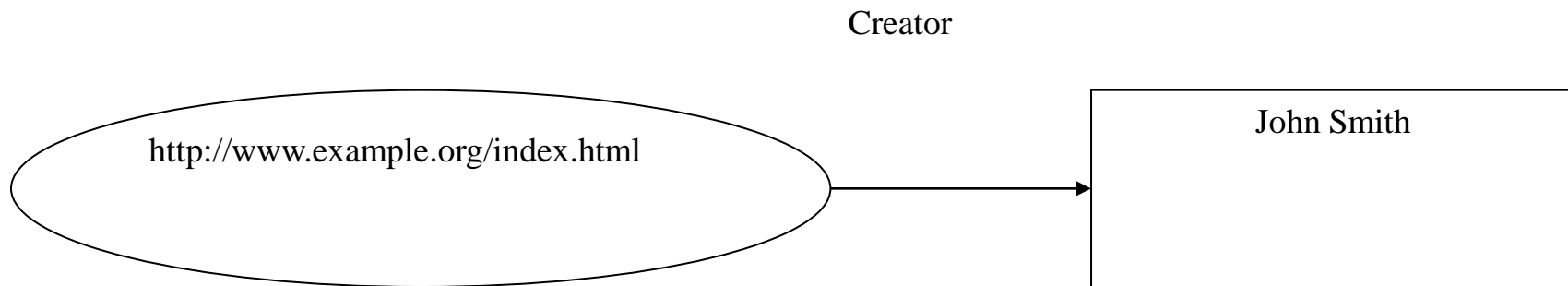


RDF的三元模型

- RDF可以将一个或多个关于资源的简单陈述表示为一个由结点和弧组成的图（**graph**），其中的结点和弧代表资源标识、属性类型或属性值。
- 对于某一资源的描述，我们可以用自然语言进行描述。例如：
 - ⊕ 资源<http://www.example.org/index.html> 的创建者是John Smith
- 这一句话可以用资源标识-属性类型-属性值三元组的方式进行表示如下：
 - 资源标识，也就是句子描述的对象是<http://www.example.org/index.html>
 - 属性类型，也就是资源的属性是Creator
 - 属性值，也就是资源属性的值是John Smith



□ 资源-属性类型-属性值关系图



这是一个由两个节点和一条弧线构成的图，其中一个节点是资源，另一个节点是属性值，弧线从资源指向属性值，表明资源具有某一属性类型的属性值。

□ 从另一个角度来看，语句“资源 `http://www.example.org/index.html` 的创建者是John Smith”这一句话可以由主语、谓语、宾语构成。其中：

- ⊕ 主语是URL `http://www.example.org/index.html`
- ⊕ 谓语是词“creator”
- ⊕ 宾语是“John Smith”

□ 在RDF中，

- ⊕ 用于识别事物的那部分就叫做主语
- ⊕ 用于区分语句对象主语的各个不同属性（譬如：作者，创建日期，语种等等）的那部分叫做谓语
- ⊕ 用于区分各个属性的值的那部分叫做宾语



□ 尽管这种主语、谓语、宾语关系和自然语言语法中的主语、谓语、宾语关系还不完全一样，但这足以构成RDF存在的逻辑基础：

- ⊕ 语句可以对资源进行描述；
- ⊕ 语句由主语、谓语、宾语组成；
- ⊕ 语句中的主语、谓语、宾语可以被一个可被标识的资源，此资源的属性描述（如名称、创建时间、出版者等）和资源的属性值所替代；
- ⊕ RDF基于资源标识-属性类型-属性值的三元模型提供了对资源进行描述的框架。
 - 在RDF也可以表示类与之类之间的关系



□ RDF的三元模型可以对资源进行描述，但为了使其能够被计算机处理，还需要考虑：

- ⊕ 一整套计算机可以处理的标识，这些标识能够对所描述的资源、属性类型和属性值进行标识，并且使这些资源标识、属性类型和属性值具有一个明确的指示意义，没有二意性，这个标识符系统不会和其他人可能在Web上使用的相似的标识符系统混淆。
- ⊕ 一种计算机可以处理的语言，以描述这一模型，并且可以在应用系统之间交换这些信息。



□ 第一个问题（一整套计算机可以处理的标识）

- ⊕ RDF使用了URI和命名空间作为其标识机制（用于标识陈述中的主体、谓词和客体）的基础。
- ⊕ 更准确地说，RDF使用的是URI参照（URI引用）（URI references）
- ⊕ 一个URI参照（或“URIref”）是一个在尾部附加了可选的“片段识别符（fragment identifier）”的URI。
比如，URI参照（URIref）
`http://www.example.org/index.html# section2`由URI
`http://www.example.org/index.html` 和（由符号#分隔的）的section2（片段标识符）组成。



□ 第二个问题（一种计算机可以处理的语言）

- ⊕ 为了用一种机器可处理的（machine-processable）方式来表示RDF陈述（RDF statements），RDF采用了XML
- ⊕ XML允许任何人来设计他们自己的文档格式，并可用这种格式书写文档。
- ⊕ RDF定义了一个特殊的XML标记语言（称为RDF/XML）来表示RDF信息和在机器间交换这些信息。



□（URI参照）

□例如上例中的资源、属性类型和属性值，
可以用下面的URI参照表示如下：

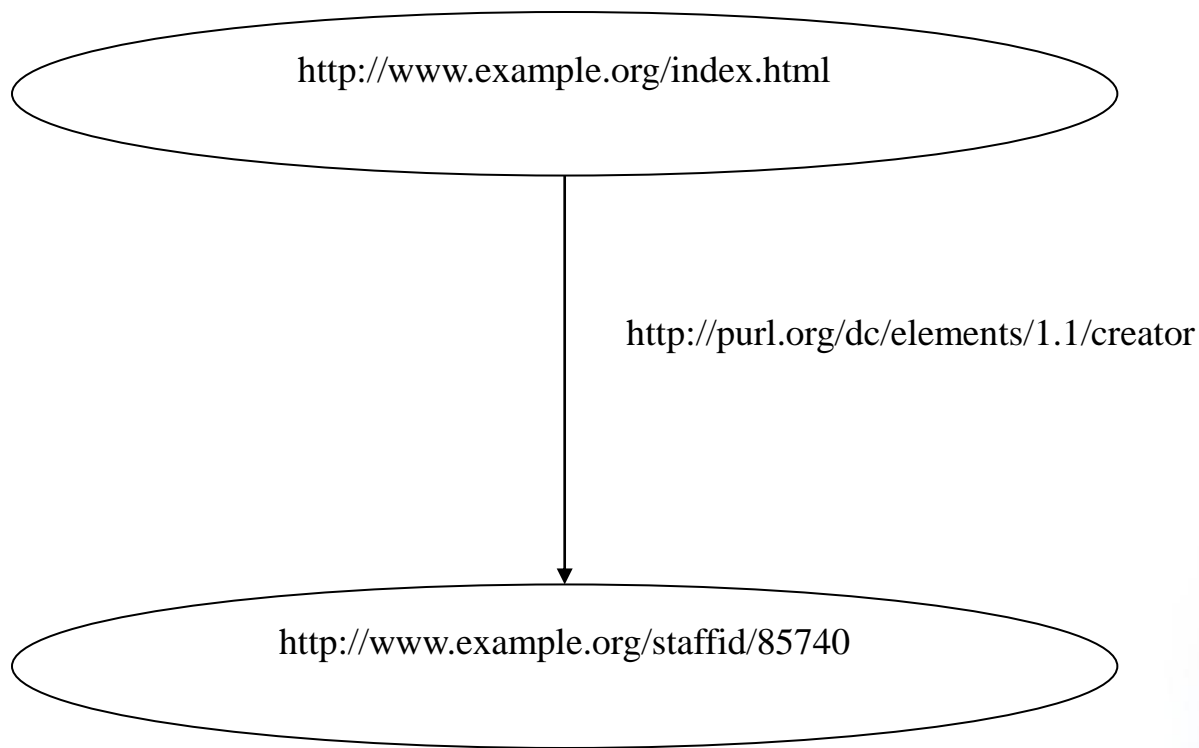
⊕资源 <http://www.example.org/index.html>

⊕属性类型<http://purl.org/dc/elements/1.1/creator>

⊕属性值 <http://www.example.org/staffid/85740>



□ 如果利用了URI参照，则上面的图也应当改为：

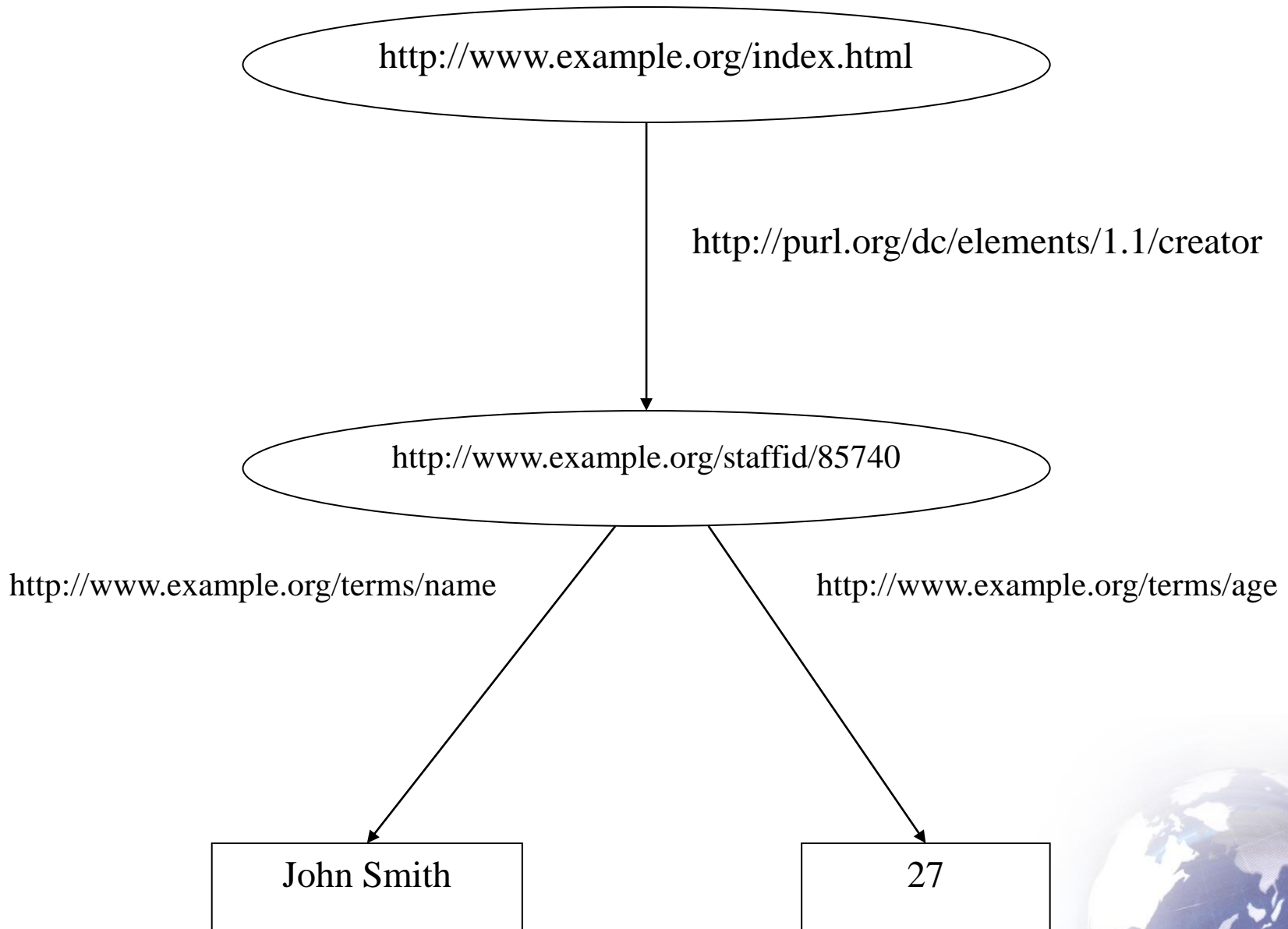


- ❑ 在前一个图中，利用了方形来表示属性值“John Smith”，而在后一个图中，则利用椭圆来表示属性值“<http://www.example.org/staffid/8574>”。
- ❑ 二者不同之处在于John Smith是常量值（在RDF中，被称为文字值，literal），而<http://www.example.org/staffid/8574>是一个URI参照。
- ❑ 在RDF中，属性值可以是URI参照或常量值。



- 在RDF中，利用URI参照的好处是能够非常明确地表示语句的主题，例如使用了 `http://www.example.org/staffid/85740`，而不是John Smith，它能够明确地表明，此页面的创建者不是字符串“John Smith”，也不是成千上万个名为“John Smith”的人，而是一个与特定URI参照相关的John Smith。
- 另一方面，由于这是一个指向特定John Smith的URI参照，因此，它其实也是一个完全具备其他资源所具有的属性的资源，我们可以增加信息，以进一步描述这个John Smith（这个资源）。所做的工作仅只是将指向John Smith的URI参照作为一个需要描述的对象即可，如下图中，我们增加了对John的名字和年龄的说明。





- ❑ 为了使RDF便于计算机处理，RDF利用了XML语言作为支持。实际上RDF是XML语言的一个特殊应用，它遵守XML的语法规则，这种基于XML语法的RDF被称为RDF/XML。
- ❑ RDF本身不定义元数据元素(语词体系)，在RDF/XML中，为了唯一地引用某一元数据（如DC、P3P或PICS）的元素，RDF充分借助了XML名字空间。名字空间的应用，使得我们可以用更加简洁的方式描述RDF中资源、属性类型和属性值的唯一性。



- 至此可以看到，RDF本质上很简单，它是由节点和弧线组成的图，通过URI参照，能够被解释成对某个事物的某些特性的说明。
- 另外，在这一模式中，需要提供一种方法来唯一地描述资源、属性类型、属性值以及属性值的类型。URI参照的应用，提供了这种描述的唯一性。在实际应用中，通过Qname来替代URI参照。



□ RDF/XML

- ⊕ 如上所述，RDF的概念模型是节点和弧线组成的图。
- ⊕ RDF提供了一个XML语法以实现RDF模型的记录 and 交换，被称为RDF/XML。
- ⊕ RDF/XML是书写RDF的规范语法。



- ❑ RDF并不直接定义自己的元数据语词，相反，它提供一个框架，在这一个框架内，可以利用其他元数据的元素对资源的语义进行描述。
- ❑ 换句话说，RDF本身并不对各种不同的元数据进行语义定义，而是提供一种框架体系，使不同的人能够在这一框架下定义他们自己的元数据的元素。



- ❑ RDF使用了XML语言，RDF在XML基础之上，以一种标准的、能够相互操作的方式揭示数据的深层语义。
- ❑ RDF和XML是相互补充的。RDF主要关注于元数据模型的建立，而有关编码方式、字符集等等方面的问题，RDF依赖于XML。
- ❑ 还应当了解到，基于XML的RDF仅仅只是RDF的一种应用模式，除XML语法之外，RDF还允许有其它的语法模式。



□ 我们还以“资源

`http://www.example.org/index.html` 的创建者是John Smith”这一个句子为例。对于这一个句子，可以利用RDF表示如下：

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          xmlns:dc="http://purl.org/dc/elements/1.1/" >
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <dc:creator >John Smith</dc:creator>
  </rdf:Description>
</rdf:RDF>
```



□ 以下是更多的实例



□ 对资源多个属性的描述

- ⊕ 下面的例子，增加了另一个名字空间，对资源的多个属性进行了描述。

```
<?xml version="1.0"?>  
  
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"   
          xmlns:dc="http://purl.org/dc/elements/1.1/"   
          xmlns:exterms="http://www.example.org/terms/">  
  
  <rdf:Description rdf:about="http://www.example.org/index.html">  
  
    <exterms:creation-date>August 16, 1999</exterms:creation-date>  
  
    <exterms:language>English</exterms:language>  
  
    <dc:creator rdf:resource="http://www.example.org/staffid/85740"/>  
  
  </rdf:Description>  
  
</rdf:RDF>
```



□ 多资源描述

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:dc=http://purl.org/dc/elements/1.1/
  xmlns:exterms="http://www.example.org/terms/">
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <exterms:creation-date>August 16, 1999</exterms:creation-date>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <dc:language>en</dc:language>
  </rdf:Description>
</rdf:RDF>
```



□ 构建工具

⊕ 本体工具

➤ 编辑器

⊕ 分析绘制工具

➤ 对数据进行处理，以可视化方式显示



□ 本体编辑工具

名称	说明	地址
Ontolingua Server	由斯坦福知识系统人工智能实验的网络服务中心开发, 提供一种分布式协作的环境, 以便浏览、构建、编辑、修改和使用本体。Ontolingua使用面向对象的框架视图表示和浏览知识, 它还有良好的教程软件。 http://www.ksl-svc.stanford.edu:5915	http://www.ksl-svc.stanford.edu:5915
Ontosaurus	由南加州大学信息科学研究所开发的一个Web浏览工具, 为Loom语言编写的知识库, 它使用Loom语言, 具有支持自动的一致性检查、推理功能、多重继承等。	http://www.isi.edu/isd/ontosaurus.html
WebOnto	由英国公开大学知识媒介学院开发的基于Web的本体编辑工具, WebOnto是一个与网络服务器相连接的Java应用程序, 允许用户浏览和编辑网络上的知识模型, 提供多重继承、锁机制, 支持用户合作浏览、创建和编辑本体。	http://webonto.open.ac.uk
OntoEdit	由德国Ontoprise公司开发的, 通过图形环境开发和维护本体, OntoEdit可以对本体进行编辑、浏览、修改、评价以及维护, OntoEdit支持推理的多重继承, 提供对于本体的并发操作, 并能输入和输出数据库结构与数据, 如Oracle、MSSQL等。	http://www.ontoprise.de/
PROTÉGÉ	由美国斯坦福大学医学院开发的本体编辑工具, 也是基于知识的编辑器, 可用于编制实用分类系统和知识库, 可以根据使用者的需要进行定制, 通过定制用户的界面以更好地适应新语言的使用有可自行设置的数据输入格式, 可以将其内部表示转换成多种形式的文本表示格式, 如XML、RDF、OIL、DAML、Owl等系列语言。是目前最常用的工具	http://protege.stanford.edu

□ 分析绘制工具

名称	说明
SPSS	大型统计分析软件，商用软件。具有完整的数据输入、编辑、统计分析、报表、图形绘制等功能。常用于多元统计分析、数据挖掘和数据可视化。
Bibexcel	瑞典科学计量学家Persoon开发的科学计量学软件，用于科学研究免费软件。具有文献计量分析、引文分析、共引分析、耦合分析、聚类分析和数据可视化等功能。可用于分析ISI的SCI、SSCI和A&HCI文献数据库。
HistCite	Eugene Garfield等人于2001年开发的科学文献引文链接分析和可视化系统，免费软件。可对ISI的SCI、SSCI和SA&HCI等文献数据库的引文数据进行计量分析，生成文献、作者和期刊的引文矩阵和实时动态引文编年图。直观的反映文献之间的引用关系、主题的宗谱关系、作者历史传承关系、科学知识发展演进等。
CiteSpace	陈超美博士开发的专门用于科学知识图谱绘制的免费软件。国内使用最多知识图谱绘制软件。可用于追踪研究领域热点和发展趋势，了解研究领域的前沿及演进关键路径，重要的文献、作者及机构。可用于对ISI、CSSCI和CNKI等多种文献数据库进行分析。
TDA	Thomson Data Analyzer (TDA) 是Thomson集团基于VantagePoint开发文献分析工具。商用软件。具有去重、分段等数据预处理功能；可形成共现矩阵、因子矩阵等多种分析矩阵；可使用Pearson、Cosine等多种算法进行数据标准化；可进行知识图谱可视化展示。
Sci2 Tools	印第安纳大学开发的用于研究科学结构的模块化工具可从时间、空间、主题、网络分析和可视化等多角度，分析个体、局部和整体水平的知识单元。
ColPalRed	Gradnada大学开发的共词单元文献分析软件。商用软件。结构分析，在主题网络中展现知识（词语及其关系）；战略分析，通过中心度和密度，在主题网络中为主题定位；动态分析，分析主题网络演变，鉴定主题路径和分支。
Leydesdorff	系类软件。阿姆斯特丹大学Leydesdorff开发的这对文献计量的小程序集合。处理共词分析、耦合分析、共引分析等知识单元体系。使用“层叠图”实现可视化知识的静态布局和动态变化。
Word Smith	词频分析软件。可将文本中单词出现频率排序和找出单词的搭配词组。
NWB Tools	印第安纳大学开发的对大规模知识网络进行建模、分析和可视化工具。数据预处理；构建共引、共词、耦合等多种网络；可用多种方法进行网络分析；可进行可视化展示。
Ucinet NetDraw	Ucinet是社会网络分析工具。包括网络可视化工具Net Draw。用于处理多种关系数据，可通过节点属性对节点的颜色、形状和大小等进行设置。用于社交网络分析和网络可视化。
Pajek:	来自斯洛文尼亚的分析大型网络的社会网络分析免费软件。Pajek基于图论、网络分析和可视化技术，主要用于大型网络分解，网络关系展示，科研作者合作网络图谱的绘制。
VOSviewer:	荷兰莱顿大学开发的文献可视化分析工具。使用基于VOS聚类技术实现知识单元可视化工具。突出特点可视化能力强，适合于大规模样本数据。四种视图浏览：标签视图、密度视图、聚类视图和分散视图。

□ 示例

- ⊕ 启动ontoedit

- ⊕ Import rdf

 - Documents 文档\Ontology 工具\myGraph.rdf

- ⊕ 编辑



谢 谢

