

知识图谱的知识表现方法回顾与展望

鲍捷

文因互联 (Memect)

baojie@memect.co



第三届全国中文知识图谱研讨会

2015.11.9 湖北宜昌

2015-11-09

1

个人简介

- ◎ 2001–2007
Iowa State University, PhD
Description Logic, Ontology Building
- ◎ 2008–2010
Rensselaer Polytechnic Institute (RPI),
TetherlessWord Constellation, PostDoc
Semantic wiki, **OWL working group**, ontology formal semantics
- ◎ 2010–2011
MIT and BBN,
visiting researcher
Information theory, financial ontology, policy language
- ◎ 2011–2013
Samsung Research USA
Question answering, NLP
- ◎ 2013–now,
Memect (San Jose, CA)
文因互联(Beijing)
News recommendation, applied knowledge graph

什么是知识图谱

- Wikipedia: The Knowledge Graph is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources.

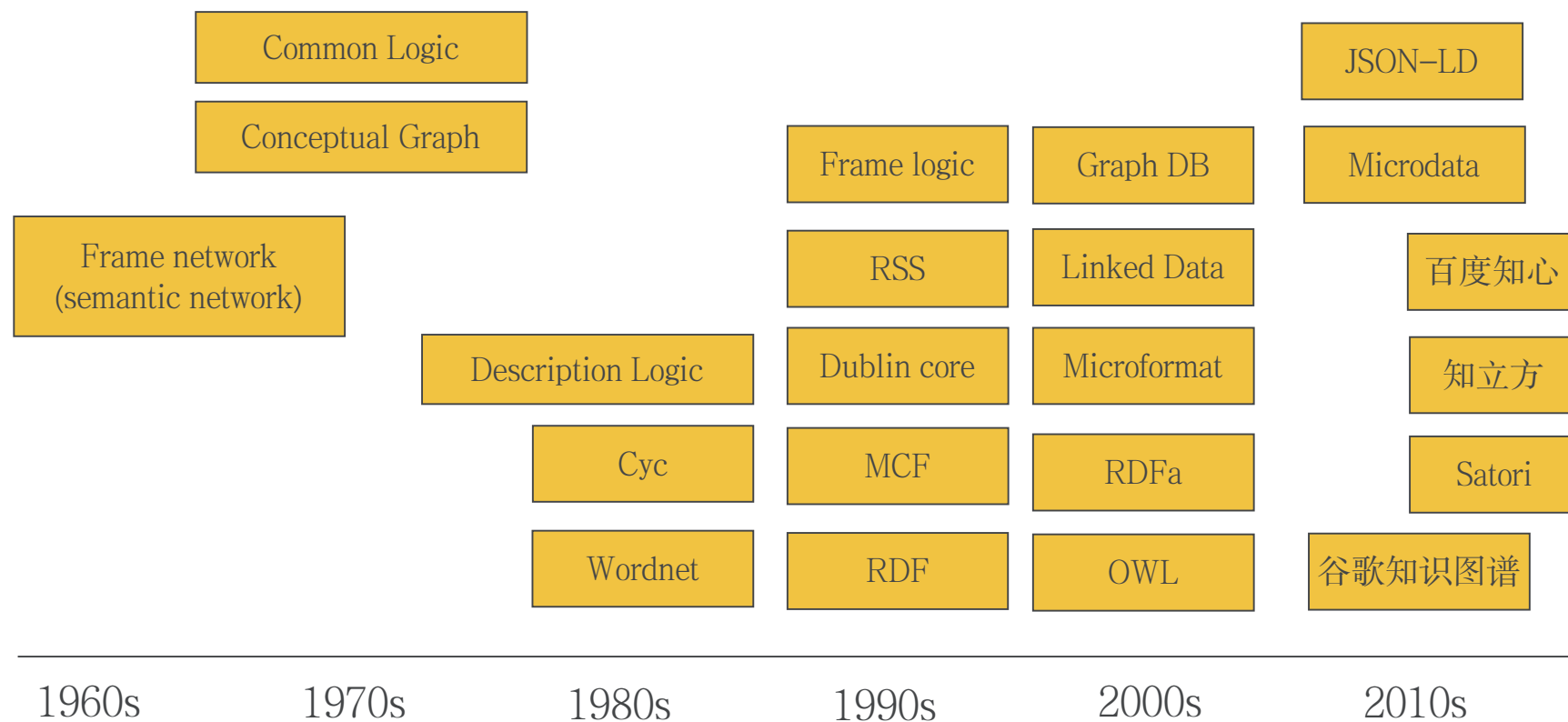
Google知识图谱（英语：Google Knowledge Graph，也称Google知识图）是Google的一个知识库，其使用语义检索从多种来源收集信息，以提高Google搜索的质量。

- 维基百科：知识库是用于知识管理的一种特殊的数据库，以便于有关领域知识的采集、整理以及提取。知识库中的知识源于领域专家，它是求解问题所需领域知识的集合，包括基本事实、规则和其它有关信息。
- 王昊奋：知识图谱旨在描述真实世界中存在的各种实体或概念。其中，每个实体或概念用一个全局唯一确定的ID来标识，称为它们的标识符。每个属性-值对用来刻画实体的内在特性，而关系用来连接两个实体，刻画它们之间的关联。

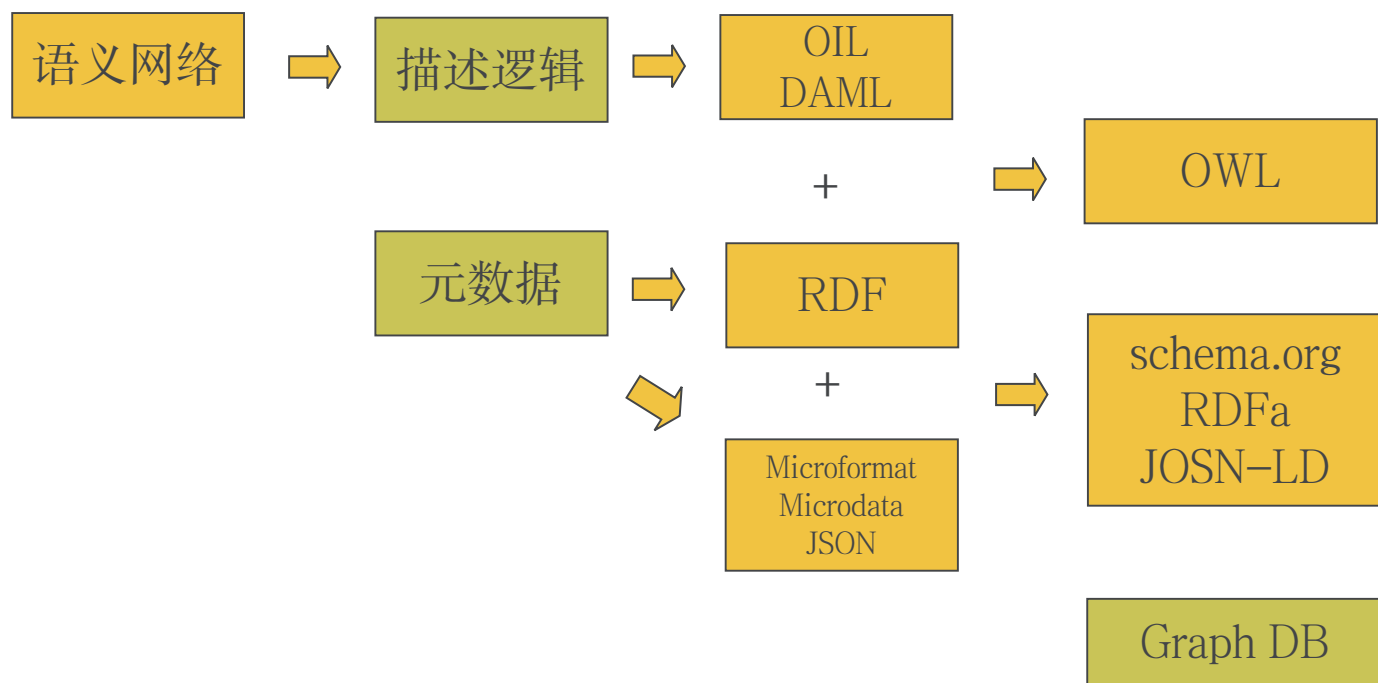
知识图谱技术原理介绍 <http://vdisk.weibo.com/s/uc617AJ1w7P5P>

知识图谱的前身

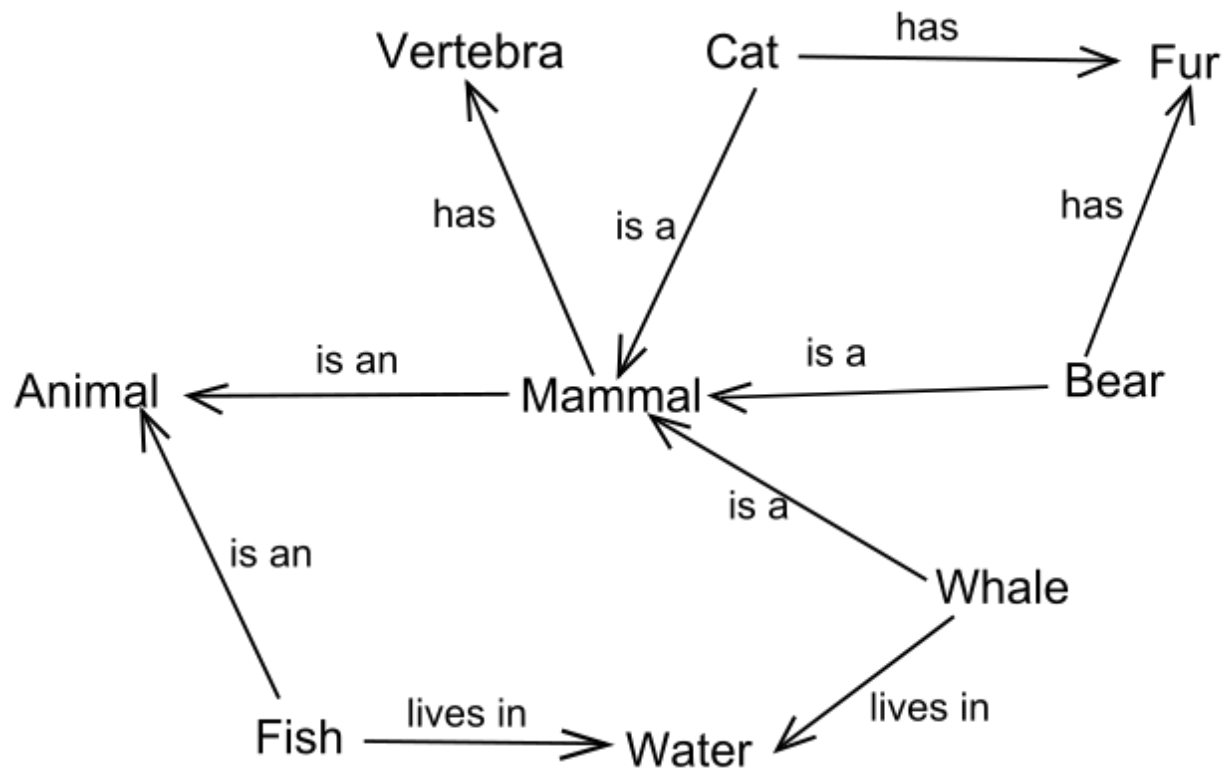
(非完备列表)



知识图谱KR的演化（简化）



语义网络 Semantic Network (Frame Network)



1960s, 由Robert F. Simmons, M. Ross Quillian, Allan M. Collins 等人发展
被批评缺少“语义”，不能用于推理

描述逻辑 Description Logic

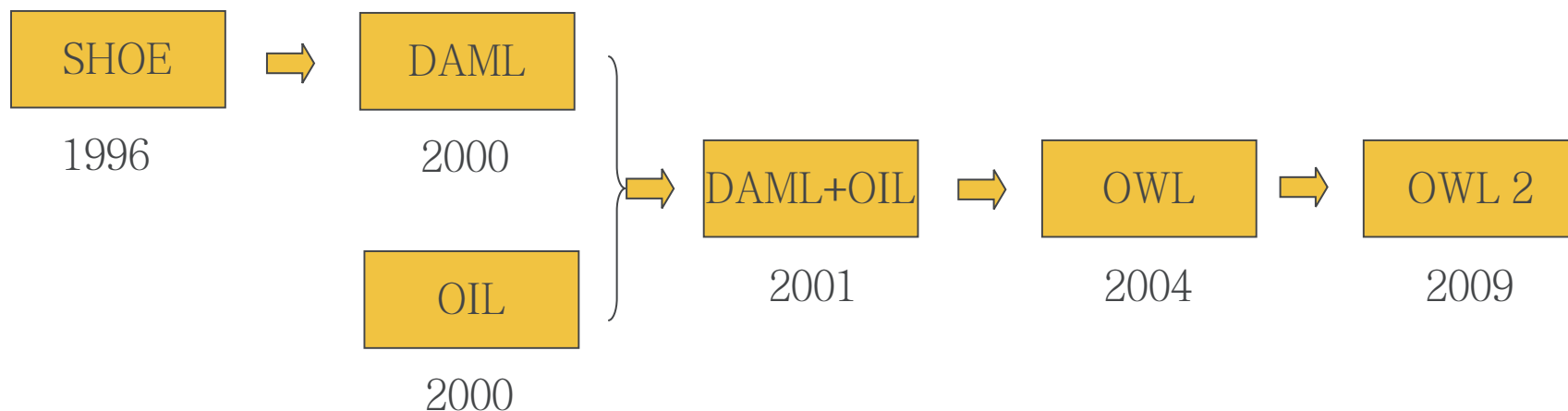
DL	FOL
$A \sqcup B$ $\neg A$ $A \sqcap B$	$A(x) \vee B(x)$ $\neg A(x)$ $A(x) \wedge B(x)$
$a : A \sqcap B$	$A(a) \wedge B(a)$
$\exists P.A$ $\forall P.A$	$\exists(y)(P(x, y) \vee A(y))$ $\forall(y)(P(x, y) \Rightarrow A(y))$
$(a, b) : P, P(a, b), aPb$	$P(a, b)$
$A \sqsubseteq B$ $A = B$	$\forall(x)(A(x) \Rightarrow B(x))$ $\forall(x)(A(x) \Leftrightarrow B(x))$
$P \sqsubseteq R$ $R = P^-, R = Inv(P)$	$\forall(x, y)(P(x, y) \Rightarrow R(x, y))$ $\forall(x, y)(P(x, y) \Rightarrow R(y, x))$

图片 by Bijan Parsia <http://www.cs.man.ac.uk/~bparsia/2006/cs30411/19-10-fol-2-dls.html>

提供了严格的“语义”（基于开放世界假设）以支持推理。
可看作一阶逻辑的可判定子集

OWL

描述逻辑的HTML或XML语法



- OWL提供了一组适合Web传播的描述逻辑的语法。
- 其开放世界语义当初被认为适合刻画Web的开放性
- 但其认知复杂性限制了它的工程应用

W3C OWL工作组

- ◎ 2007–2010
- ◎ ~ 40 成员， ~ 20活跃成员， <http://www.w3.org/2007/OWL/wiki/Participants>
- ◎ 大学为主，企业成员主要有IBM，Oracle，HP，Clark & Parsia
- ◎ 产生了10+个文档，600+页

工作组遇到的问题

- ◎ 路线问题：SEMANTIC Web or semantic WEB

logic vs data

RDFS++ vs SROIQ

表达力 vs 可用性

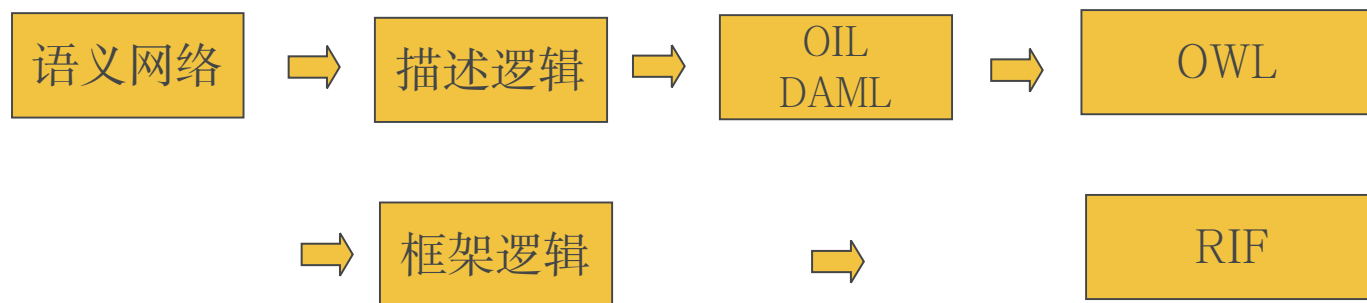
- ◎ 三种复杂性：计算复杂性，认知复杂性，工程复杂性

- ◎ 缺少企业参与，缺少来自真实用户的需求

- ◎ 文档和工具系统脱离一线程序员，得不到支持和应用

小结：从弱语义到强语义的尝试（逻辑）

（已经很像今天的知识图谱了）

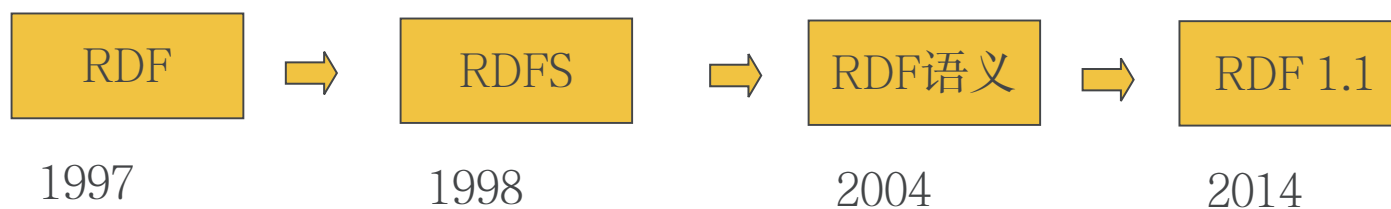
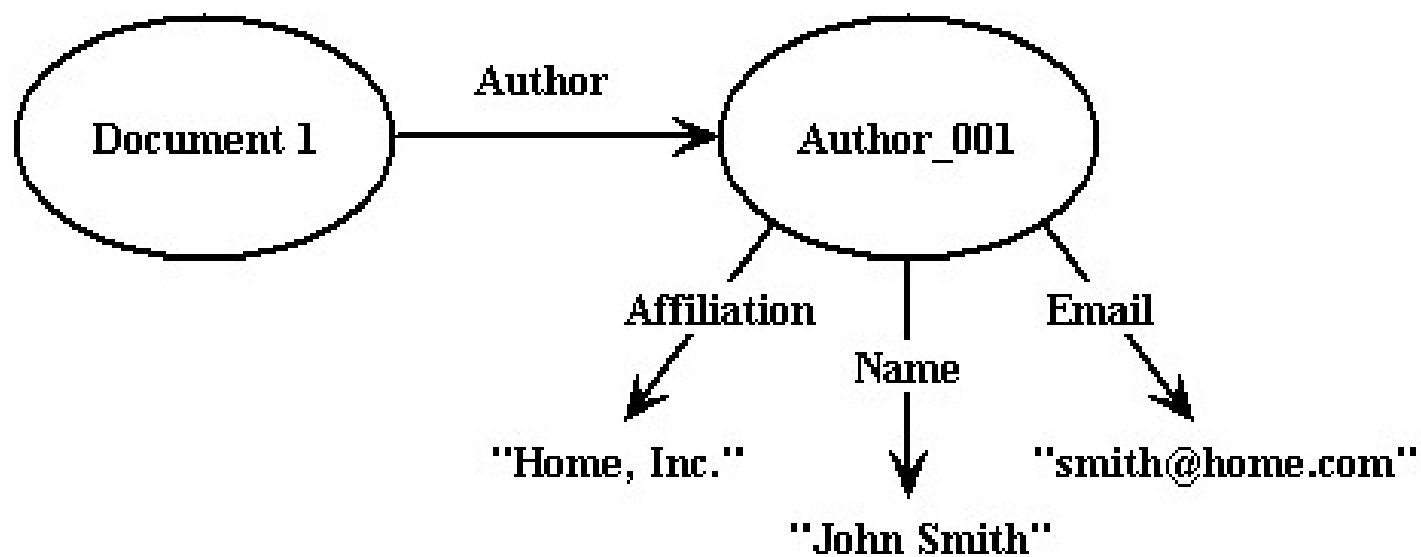


元数据框架到RDF



- ◉ RDF是从实践中bottom-up总结出来的
OWL是一种top-down设计
- ◉ RDF的基础是三元组 (triple)
- ◉ Guha也是Google知识图谱的主要推手

RDF

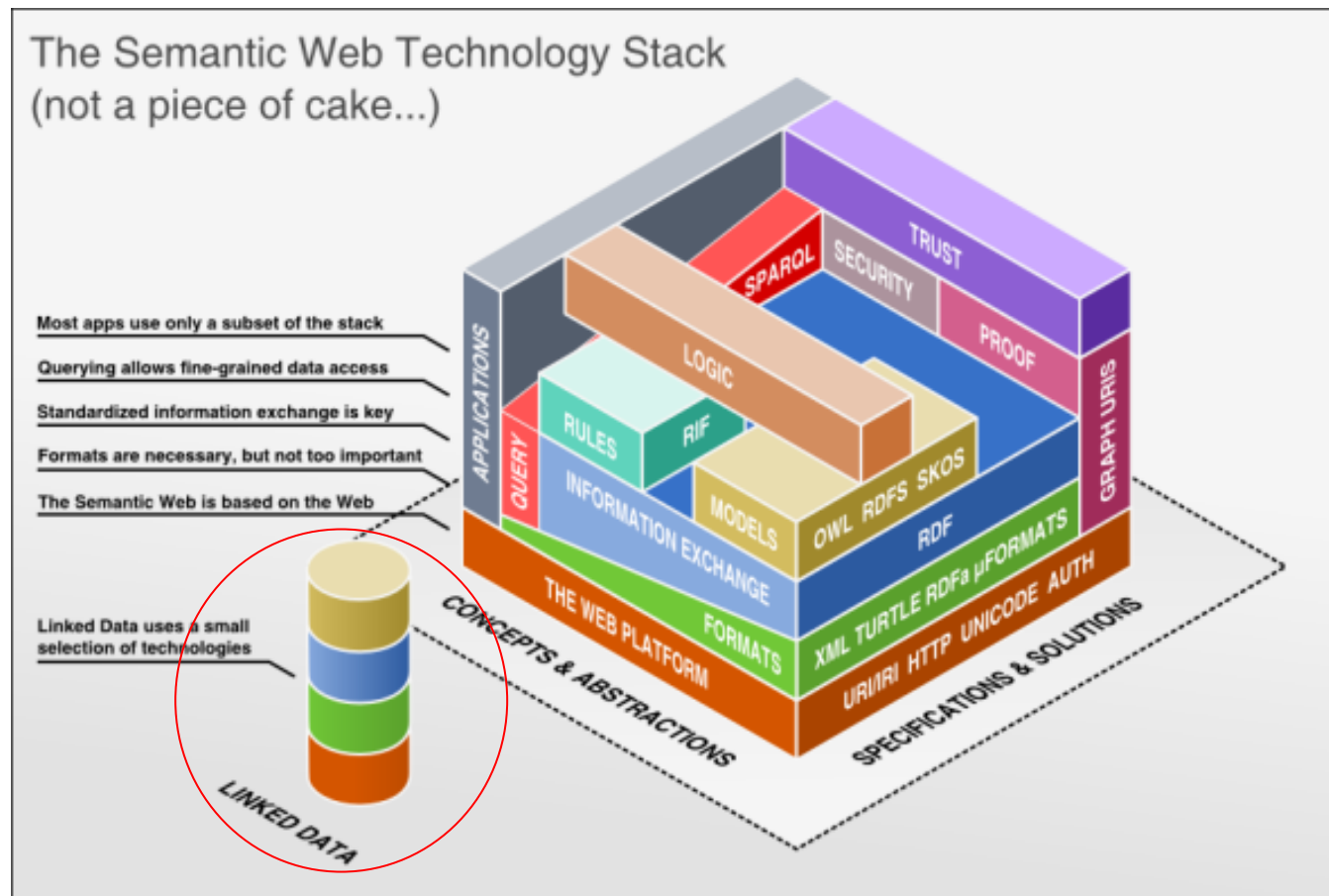


小结：从弱语义到强语义的尝试（元数据）

- ⊙ RDF开始是一个没有语义的元数据框架
- ⊙ 因为推理的需要加上了语义
- ⊙ 为了和OWL统一，两个语言都采用了复杂的模型论语义，支持了基于规则的推理
- ⊙ 但在实践中，推理很少被实用。大部分场合下RDF只是被用为一种数据描述语言

关联数据 Linked Data

到了2006年，语义网技术堆栈又已经复杂到没多少人看得懂了，于是：



by Benjamin Nowack <http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake>

关联数据 Linked Data

Tim Berners-Lee

Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$

Status: personal view only. Editing status: imperfect but published.

[Up to Design Issues](#)

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.



<http://www.w3.org/DesignIssues/LinkedData.html>

小结：从强语义到弱语义的尝试（关联数据）

- ◎ 关联数据是在前期（2001-2006）失败的基础上，做了对工程友好的调整和简化
- ◎ RDF作为数据交换语言，并不意味着需要同时作为数据存储语言，或者数据建模语言。
- ◎ 互联是第一位的，数据质量、发布者世界观的统一（本体）可以按需提高
- ◎ 产生了较大的影响，例如IBM Waston、美国政府开放数据

新的综合：交换语言



2008

Microformat

2005



2013

被Google用于schema.org

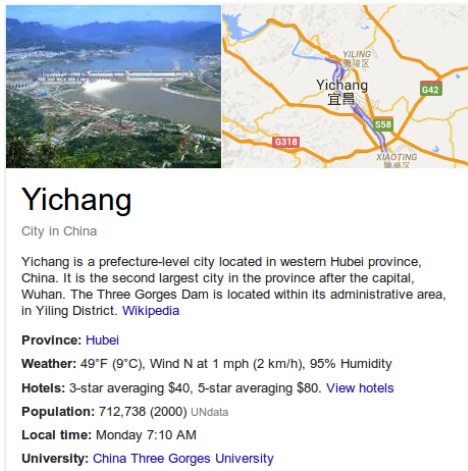


2014

依托于成熟的Web交换语言（HTML，JSON），充分利用现有工具，让普通程序员、站长容易理解和使用。

至少30%的网页已经有某些语义元数据(根据Web Data Commons统计
<http://webdatacommons.org/>)

新的综合：存储语言（图数据库）



Google知识图谱

- 以Freebase为基础
- 图数据库内部存储（早期有graphd，Cayley）



微软Satori

- 图数据库内部存储（Trinity，Graph Engine）
- 底层是key-value store

新的综合：存储语言（图数据库）



- 两种图建模：RDF图和属性图（Property Graph）有融合的趋势
- 推理可以看成图上的遍历和构造
- 属性图标准Tinkerpop已经获得很大的市场成功

图数据库市场现状

◎ DB-Engines Ranking <http://db-engines.com/en/ranking>

- 21 Neo4j (图)
- 32 MarkLogic (XML)
- 42 Titan (图)
- 46 OrientDB (图, 文档)
- 61 Virtuoso (RDF, 关系等)
- 80 Jena (RDF)
- 88 Sesame (RDF)
- 90 ArangoDB (图)
- 120 AllegroGraph (RDF)
- 123 Algebraix (图)
- 147 Stardog (RDF)
- 164 InfiniteGraph (图)
- 170 Sparksee (图)
- 172 4store (RDF)
- 179 BrightstarDB (RDF)

小结：从强语义到弱语义的尝试（图数据库）

- ◎ 从顶向下设计的RDF数据库没有从底向上设计的图数据库成功
- ◎ 新的大型知识图谱数据库大多是图数据库，底层可能是键值数据库、内存数据库、列数据库等
- ◎ 不追求严格的模型论语义，而面向查询定义过程语义，或不追求语义

我们的工程实践

- ◎ Memect, 文因互联 2013-今

- ◎ 产品1: @好东西传送门 <http://weibo.com/haoawesome>

综合NLP和知识库, 提供精准的领域新闻推送

主题模型负责recall, 知识库负责precision

字符串是一等公民, 知识库的构造与存储以字符串为核心

混合推理: 融合ontology, POS tag和regex做规则推理

数据库: OrientDB, Elasticsearch, PostgreSQL

- ◎ 产品2: 企业情报数据库

知识提取: 实体提取, 关系提取, 事件提取

领域知识建模: 40-50个产业领域

强调以人为核心构造知识系统, 知识表示、人机交互、推理机, 以降低认知复杂性为核心 (Palantir路线)



Lean Semantic Web

<https://github.com/baojie/leansemanticweb>

总结

- ◎ 第一阶段：从弱语义到强语义（pre-2006）

从语义网络，到描述逻辑，到OWL

从元数据框架，到RDF，到RDFS

- ◎ 第二阶段：从强语义再到弱语义（2006-至今）

关联数据，和现有Web工业常用标准的结合

从RDF数据库到图数据库

- ◎ 20年来的历史表明

从实践中总结（JSON，图数据库等）优于从顶向下的设计

充分和现有工具系统的兼容优于全新框架

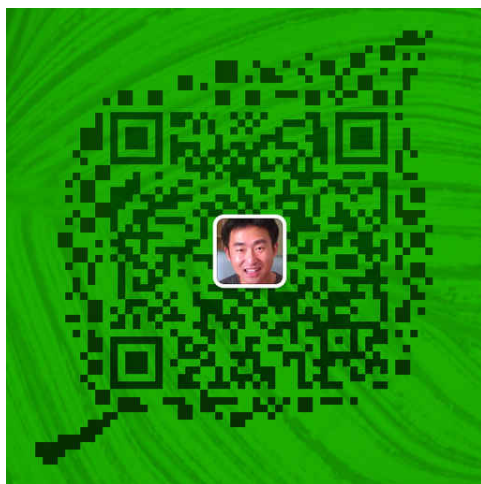
简单优于强大

展望

- ◎ 知识图谱会快速向多个垂直领域渗透，如医疗、电商、金融、投资等。弱语义的趋势短期还会延续，可能针对这些领域做进一步的简化
- ◎ 发展非Entity-Relation的建模方式，概率的、模糊的关联，综合DB和IR系统，降低知识表现全周期成本
- ◎ 数据交换语言和数据存储语言的分离可能会持续。图数据库将成为主要的存储语言和工具。
- ◎ 标准的产生可能由以W3C主导转变为以Web公司为主导，再由标准化组织总结

We are hiring!

- ◉ 文因互联从事投资、金融、企业情报领域的知识图谱服务
- ◉ 创始成员来自RPI、Wright State U、硅谷知名公司
- ◉ 北京，中关村创业大街
- ◉ 用到的技术：实体识别、关系识别、主题模型、文档分类、图数据库、Elasticsearch、知识建模、人机交互、推理机等等



baojie@memect.co

微信



backup

什么是知识表现 (Knowledge Representation)

- ◎ 知识是什么？
- ◎ 如何用计算机储存知识？
- ◎ 如何从已知的知识推导出未知的知识？

知识就是结构

OWL的语义

http://www.slideshare.net/baojie_iowa/2008-0906-owl-full-semantics

OWL Full Semantics

-- RDF-Compatible Model-Theoretic Semantics

by Peter F. Patel-Schneider, Patrick Hayes and Ian Horrocks
W3C Recommendation, 2004

<http://www.w3.org/TR/owl-semantics/rdfs.html>

Presented by Jie Bao

RPI

Sept 11, 2008

Part 2 of RDF/OWL Semantics Tutorial

[http://tw.rpi.edu/wiki/index.php/RDF and OWL Semantics](http://tw.rpi.edu/wiki/index.php/RDF_and_OWL_Semantics)



RDF语义

http://www.slideshare.net/baojie_iowa/rdf-semantic

RDF Semantics

by Patrick Hayes

W3C Recommendation

<http://www.w3.org/TR/rdf-mt/>

Presented by Jie Bao

RPI

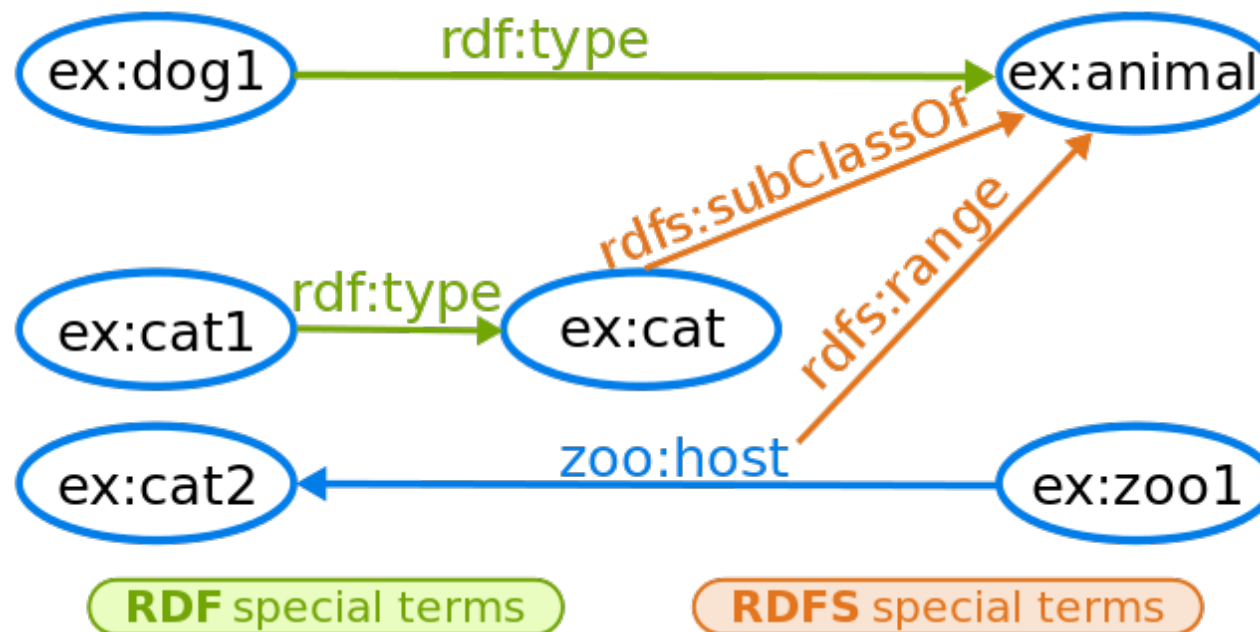
Sept 4, 2008

Part 1 of RDF/OWL Semantics Tutorial

[http://tw.rpi.edu/wiki/index.php/RDF and OWL Semantics](http://tw.rpi.edu/wiki/index.php/RDF_and_OWL_Semantics)

RDF引入（一个高阶的可能会产生悖论的）语义也许是一个错误！

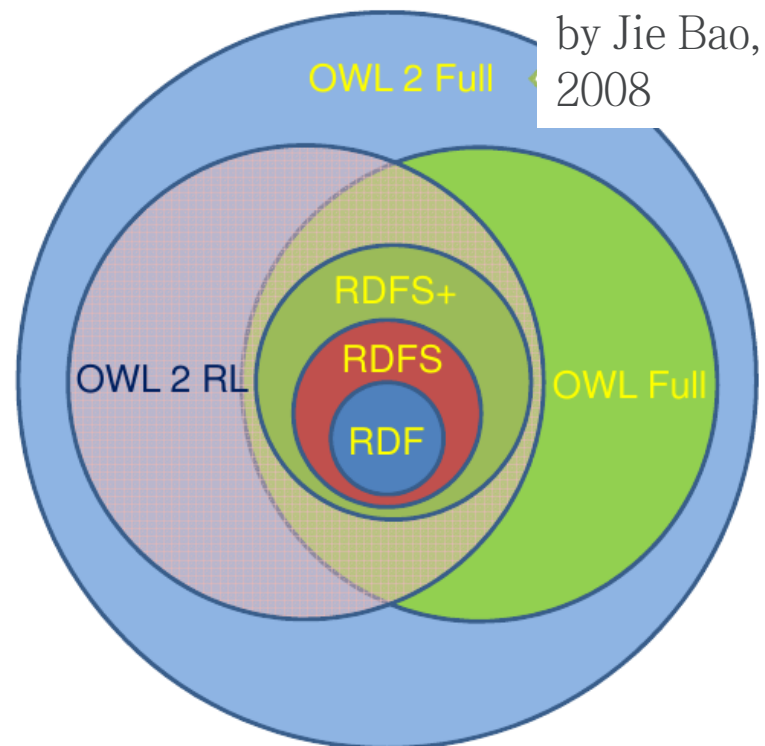
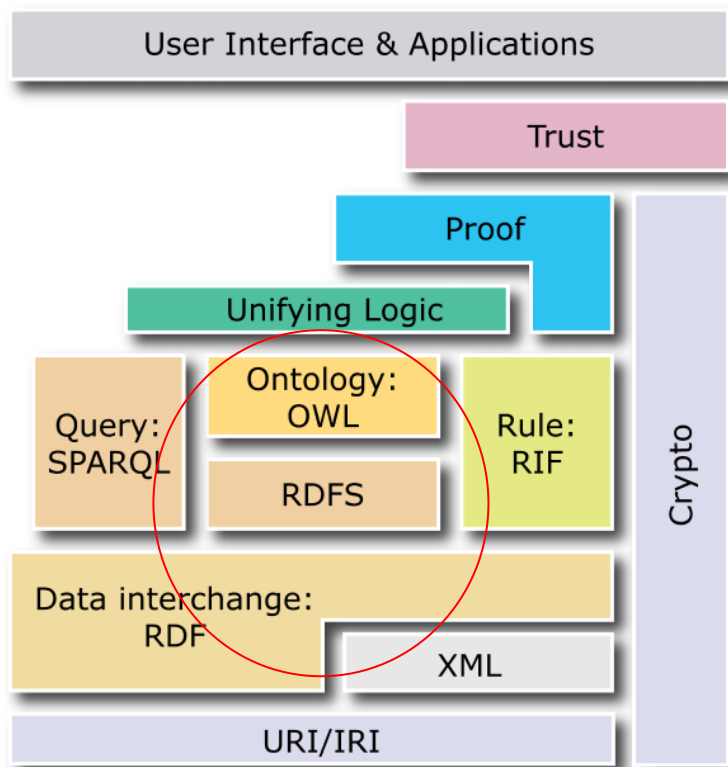
RDFS



图片来自 https://en.wikipedia.org/wiki/RDF_Schema

RDF只提供了概念和关系的基本描述能力，并没有推理的能力。
RDFS则提供了简单的推理“schema”

RDF和OWL的关系



框架逻辑 F-Logic

man::person.

woman::person.

brad:man.

angelina:woman.

person[hasSon=>man].

brad[hasSon->>{maddox,pax}].

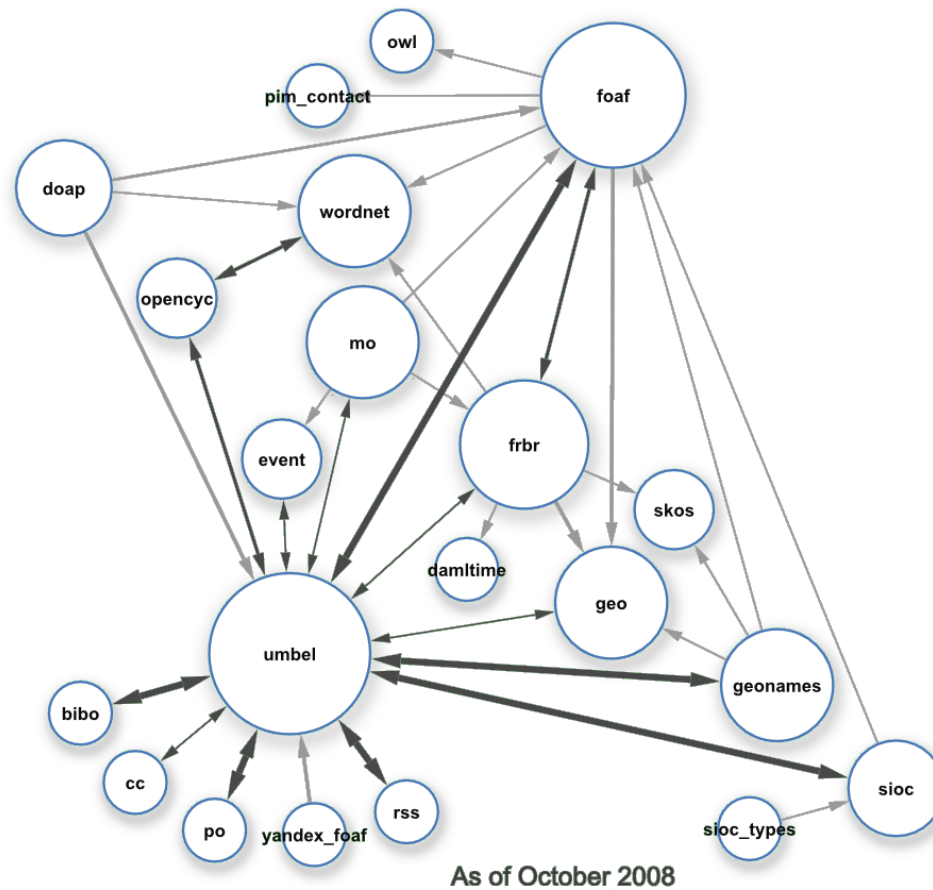
married(brad,angelina).

man(X) <- person(X) AND NOT woman(X).

提供了另一种严格的“语义”（基于封闭世界假设）以支持推理。
可看作一阶逻辑的一个不可判定子集

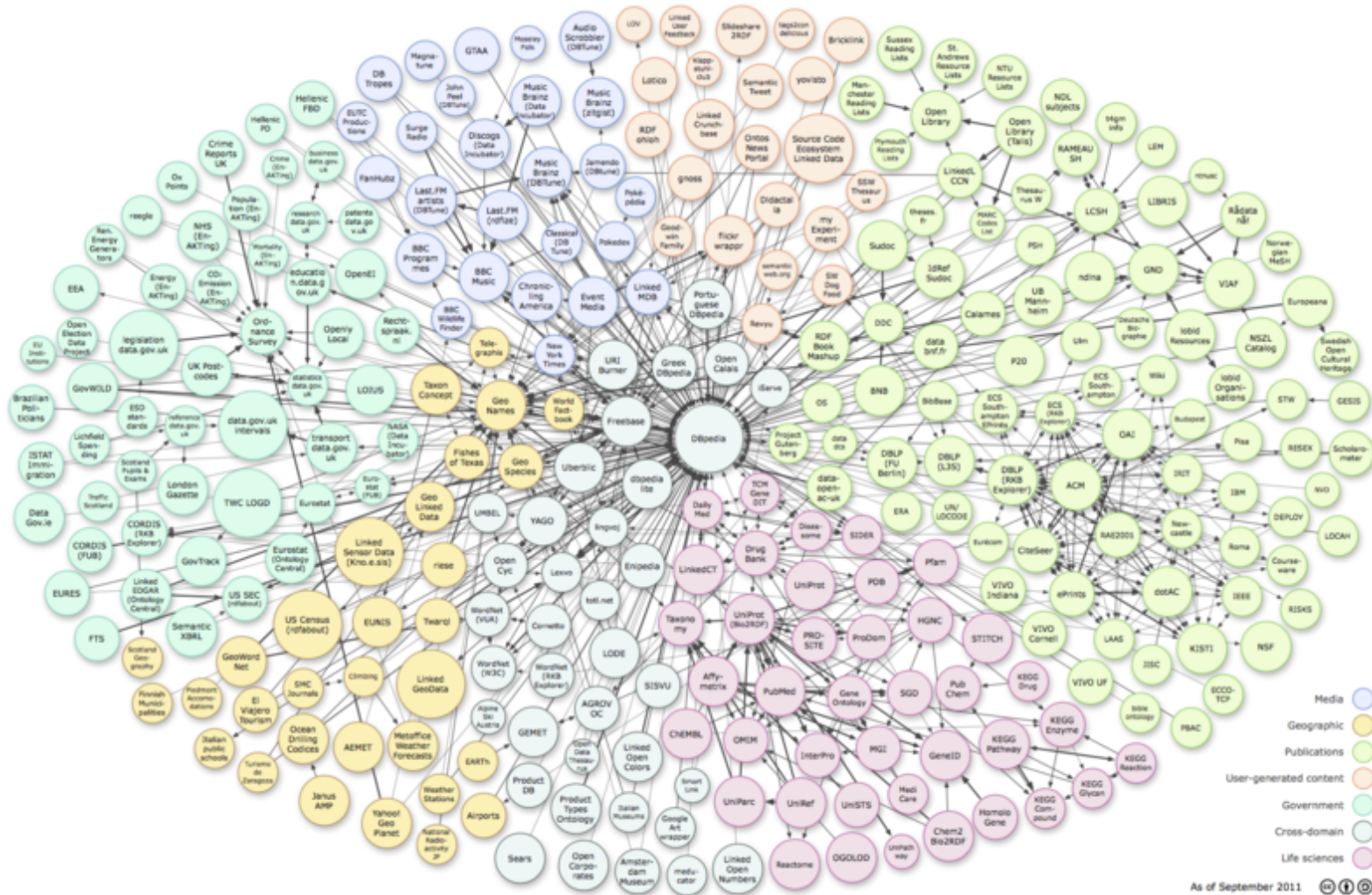
当初（2000s）很有希望成为本体语言的基础，但是Horrocks太勤快了
..... 好在后来在RIF中得到了体现

基于URI的互联



https://en.wikipedia.org/wiki/File:Linking-Open-Data-class-diagram_2008-10-05.png

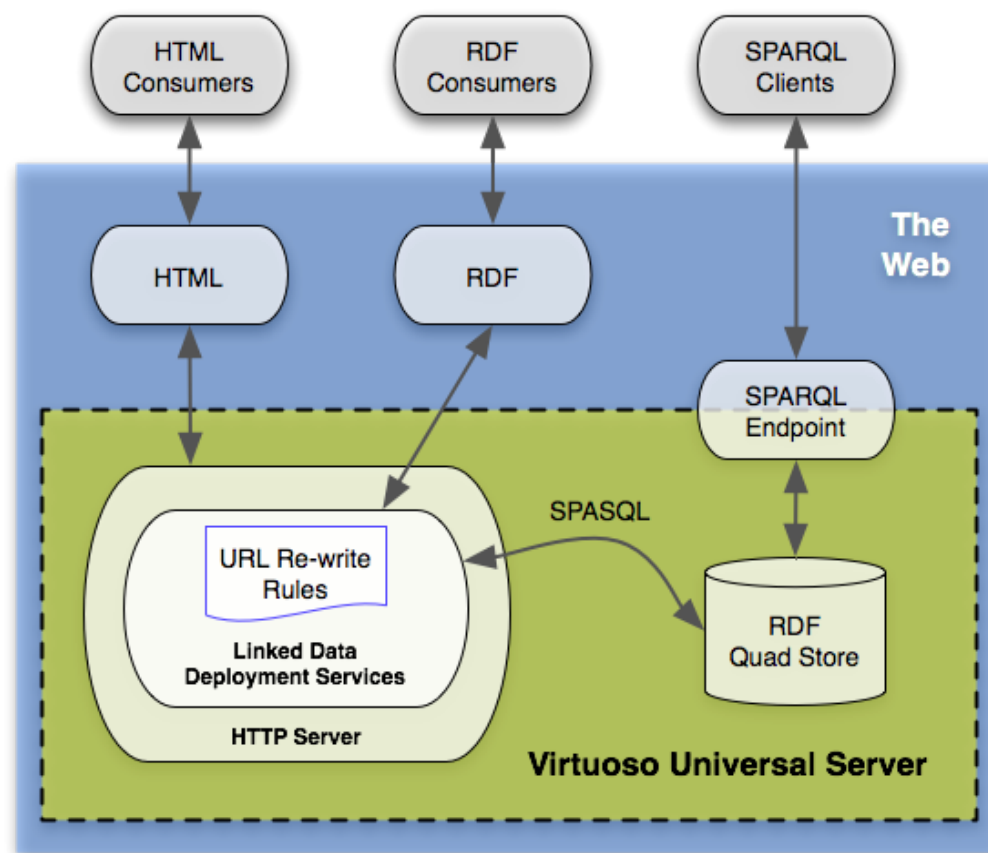
互联数据云



DBpedia



- ◎ 从wikipedia抽取
- ◎ 英文版有4.58m个实体
 - 1,445,000 人
 - 735,000 地点
 - 123,000 音乐专辑
 - 87,000 电影
 - 241,000 组织
 - 251,000 物种
 - 6,000 疾病
 - 等等.
- ◎ 基于RDF/SPARQL
- ◎ 广泛的成功应用



<http://wiki.dbpedia.org/about/about-dbpedia/architecture>

开放政府数据



Linking Open Government Data

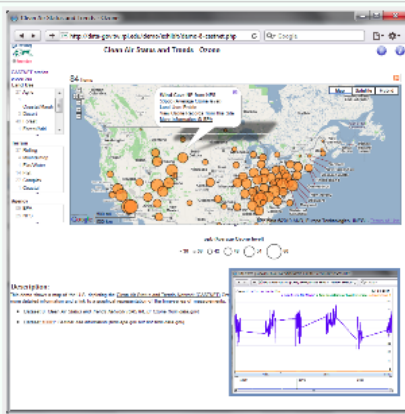
Menu

- Home
- HealthData.gov Challenge (2013)
- IOGDS: International Open Government Dataset Search (DEMO)
- IOGDS Analytics (V1)
- IOGDS Analytics (V2)
- Submit a Catalog to IOGDS
- Schema.org Dataset Extension (DEMO)
- Instance Hub demo (V.1)
- Instance Hub demo (V.2)
- URI Design Principles
- SW Challenge 2010
- Tutorials
- Videos

Linking Open Government Data

[View](#) [View Source](#)

Featured LOGD Demos



Contributor(s):

Quick Links

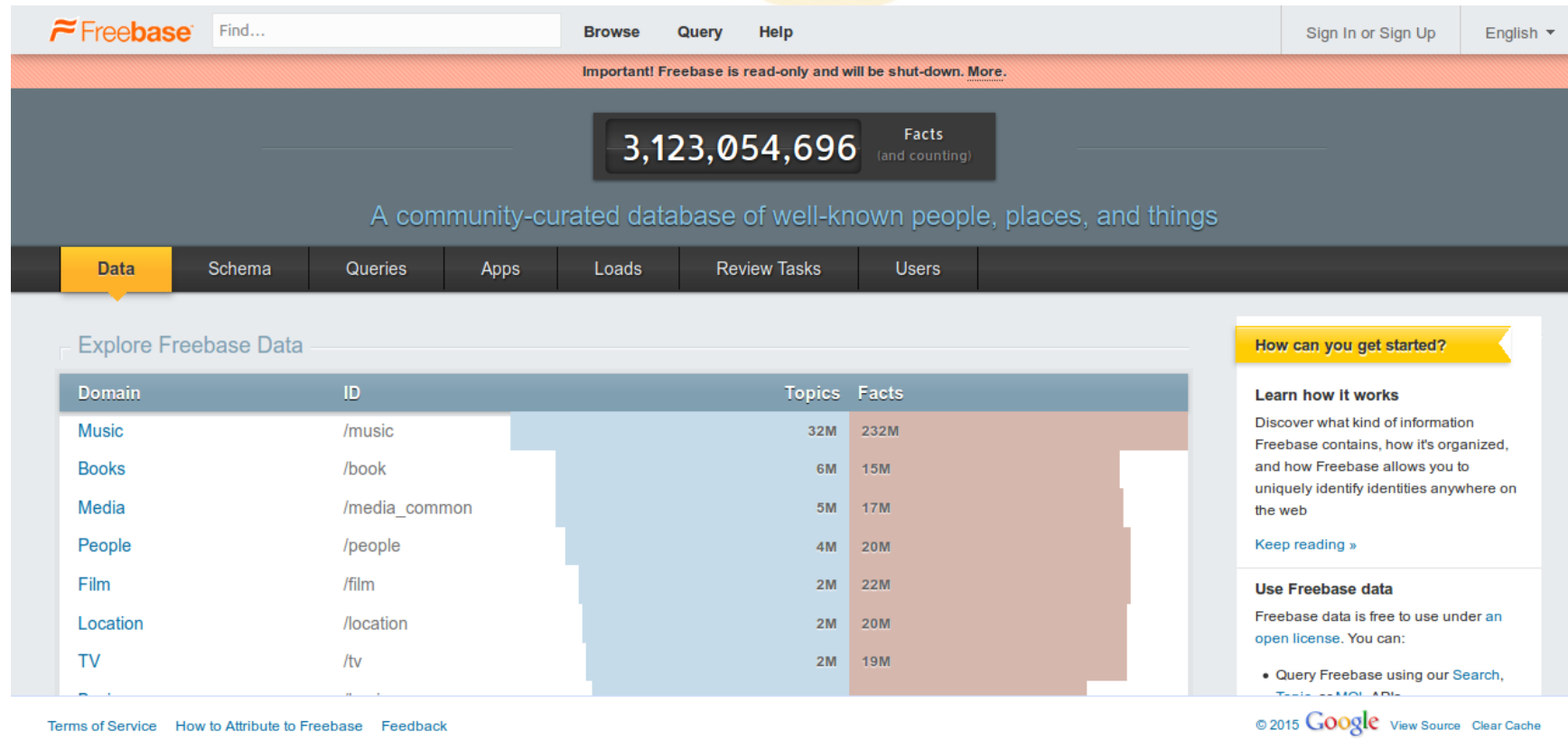
- Visit our new [International Dataset Catalog](#)
- [Semantic Web Challenge 2010 - TWC wins 2nd prize at SWC 2010!!!](#)
- [Demos](#) - Examples of consuming open government data
- [Tutorials](#) - Learn Semantic Web technologies for LOGD
- [Feedback](#) - Send us comments
- [Issue tracker](#) - Report a bug
- [LOGD Open Source on Google Code](#)
- [LOGD SPARQL Endpoint \(Alpha\)](#)

TWC LOGD Datasets Stats (as of 2011-11-04)

- TWC LOGD is hosting **9,951,771,397** RDF triples
- We have created **1,888** RDFized datasets originating from **119** sources.

<http://logd.tw.rpi.edu/>

Freebase

A screenshot of the Freebase website interface. At the top, there's a navigation bar with the Freebase logo, a search bar labeled 'Find...', and links for 'Browse', 'Query', and 'Help'. On the right of the navigation bar are links for 'Sign In or Sign Up' and a language dropdown set to 'English'. Below the navigation bar is a prominent orange banner with the text 'Important! Freebase is read-only and will be shut-down. More.' In the center of the page, a large black box displays the number '3,123,054,696' with the text 'Facts (and counting)' next to it. Below this, a tagline reads 'A community-curated database of well-known people, places, and things'. A dark navigation bar contains several tabs: 'Data' (highlighted), 'Schema', 'Queries', 'Apps', 'Loads', 'Review Tasks', and 'Users'. The main content area is titled 'Explore Freebase Data' and features a table with columns for 'Domain', 'ID', 'Topics', and 'Facts'. The table lists various domains like Music, Books, Media, People, Film, Location, and TV, each with its corresponding ID, number of topics, and number of facts. To the right of the table, there's a sidebar with a yellow banner 'How can you get started?' followed by sections 'Learn how it works' and 'Use Freebase data'. The footer contains links for 'Terms of Service', 'How to Attribute to Freebase', and 'Feedback', along with a copyright notice '© 2015 Google' and links for 'View Source' and 'Clear Cache'.

<http://www.freebase.com/>

有RDF作为皮肤语言，内部存储语言是一种图