

cnSchema

开放中文知识图谱的schema

丁力 博士

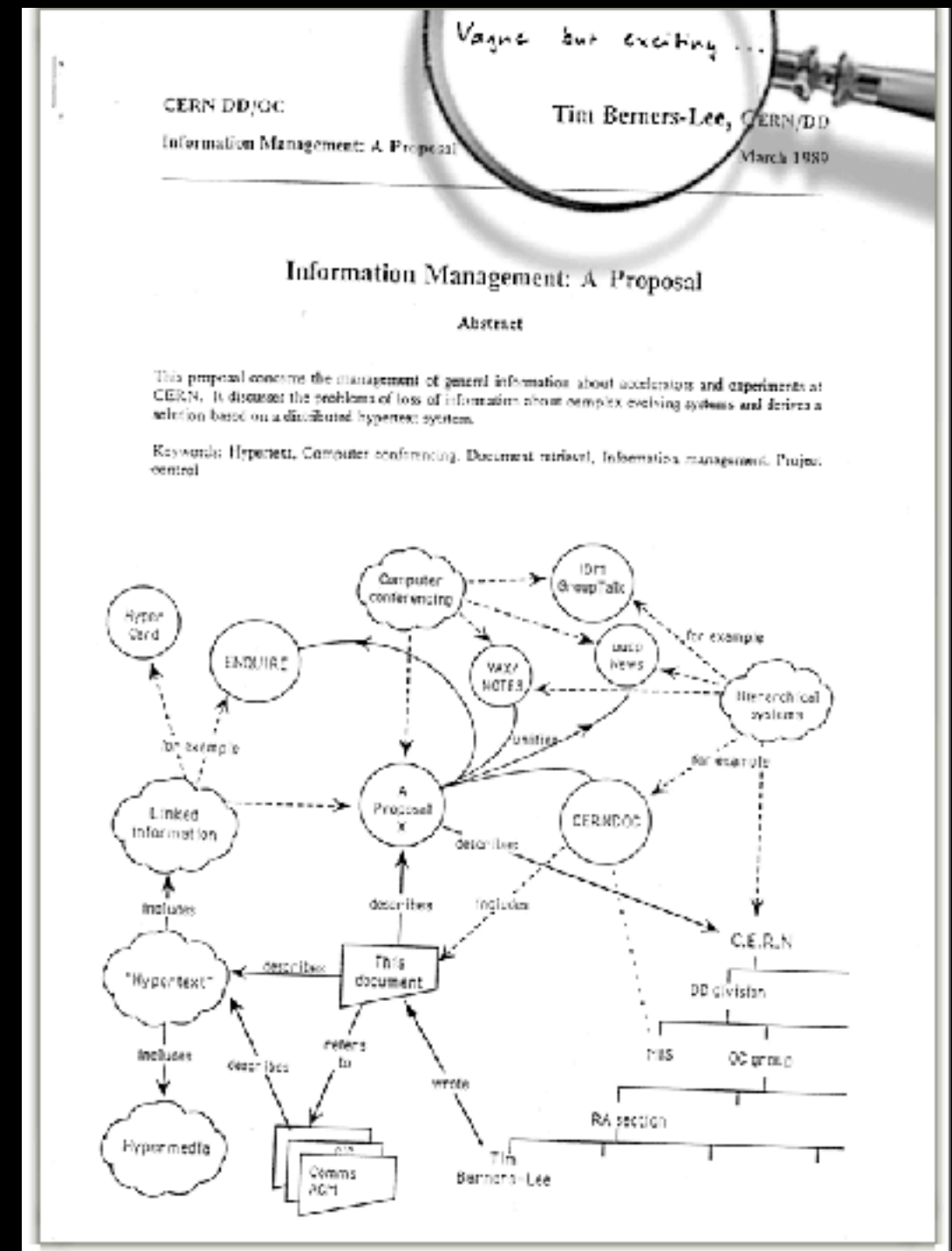
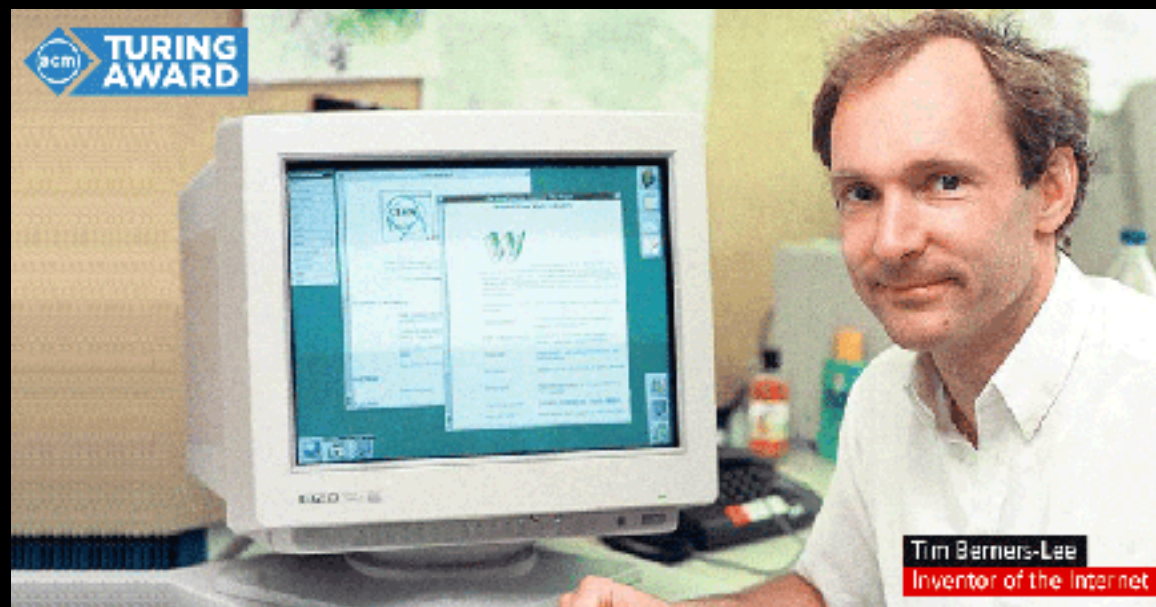
海知智能CTO OpenKG发起人

dl@ruyi.ai

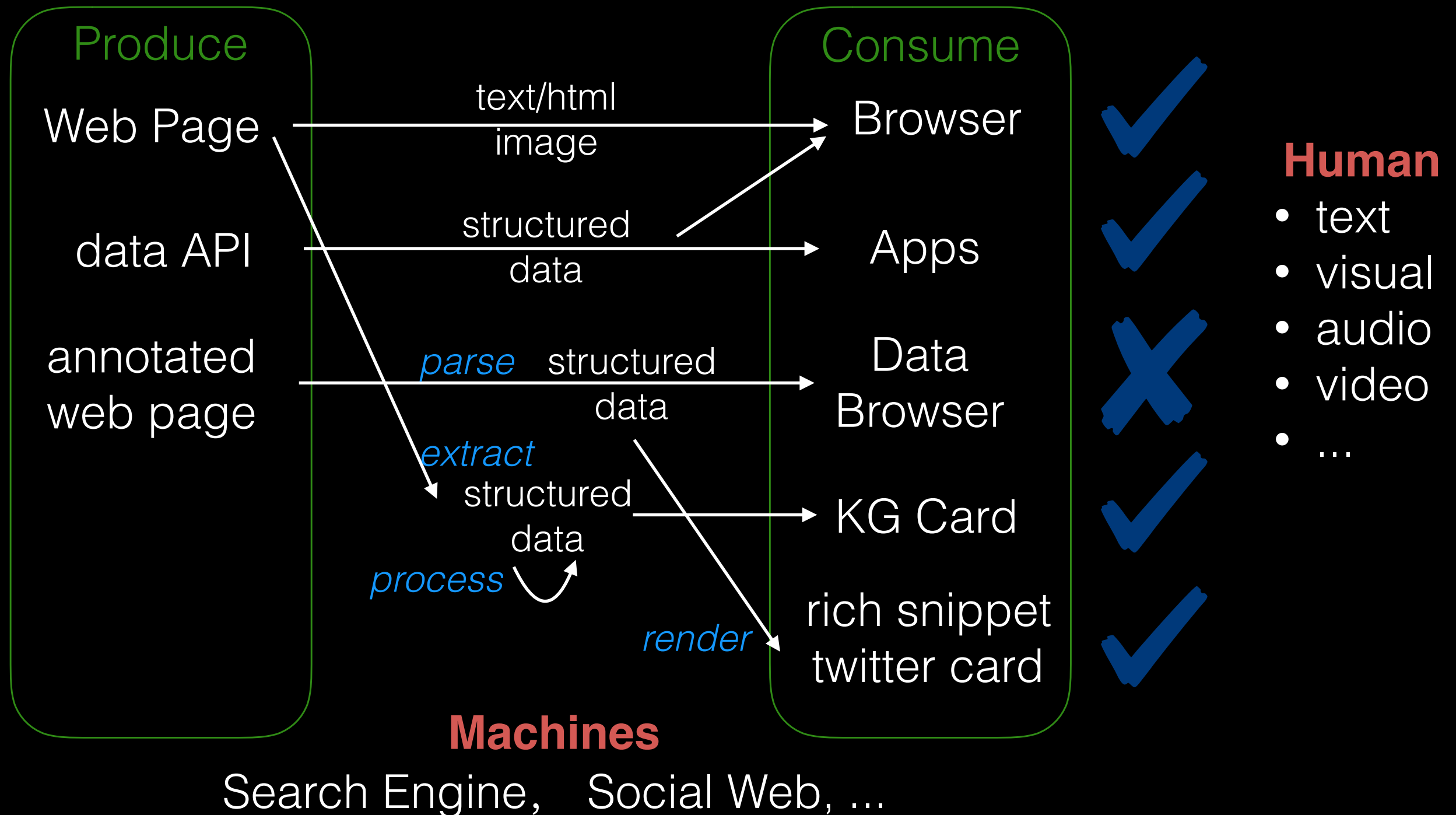
CCKS2017, 成都, 2017-08-29

Tim Berners-Lee's vision for the Web (1989, 1998)

The Web was designed as an information space, with the goal that it should be useful not only for **human**-human communication, but also that **machines** would be able to participate and help.



Web for Human, powered by machines



Schema Matters

- scope of schema
 - Data Model
 - Syntax
 - Vocabulary
 - Identifiers for Object

The image shows a JSON-LD snippet for a music recording. It is annotated with Chinese labels: '格式' (format) points to the script type, '分类' (classification) points to '@type', 'ID' points to '@id', '名称' (name) points to 'name', '属性' (property) points to 'duration', '属性值' (property value) points to 'PT2M43S', and '关系' (relationship) points to 'recordingOf'. The snippet includes a producer object for George Martin and a recordingOf object for a music composition. The schema.org logo and URL are at the bottom right.

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  分类 "@type": "MusicRecording",
  "@id": "http://musicbrainz.org/recording/3566e45",
  "name": "Back in the U.S.S.R.",
  "producer": {
    "@type": "Person",
    "name": "George Martin"
  },
  属性 "duration": "PT2M43S",
  "recordingOf": {
    "@type": "MusicComposition",
    "name": "Back in the U.S.S.R.",
    "iswcCode": "T-010.140.236-1"
  }
}
</script>
```

schema.org
<http://schema.org/MusicRecording>

- the killer app
 - standard first (since 1996) ? MCF, RDF, OWL, FOAF, RSS, ...
 - data first (since 2004) ? Swoogle, DBpedia, Linked Data, Open Government Data
 - consumer first (since 2007) ! searchMonkey, schema.org

R.V. Guha's Schema.org promotes great adoption (2011)

.... A longstanding goal of the semantic web initiative is to get webmasters to make the **structured data directly available on the web** Learning from these earlier attempts has guided the development of schema.org

- Clear incentives, SEO
- One vocabulary understood by major search engines
- make it easy for webmasters

Nikon D3200 24.2 MP Digital SLR Camera with 18-55mm ... - Target
www.target.com/.../nikon-d3200...digital-slr-camera-wi... - Cached

★★★★★ Rating: 5 - 3 reviews - \$599.99

Nikon D3200 24.2MP Digital SLR Camera with 18-55mm VR Lens Black save \$150 on a Nikon 55-300mm AF-S DX ED VR NIKKOR Zoom lens (online item# ...

Nikon D3200 Digital SLR Camera 18-55mm G VR Zoom Lens 24.2 ...
[www.ebay.com/Cameras & Photo/Digital Cameras](http://www.ebay.com/Cameras%20%26%20Photo/Digital%20Cameras) - Cached

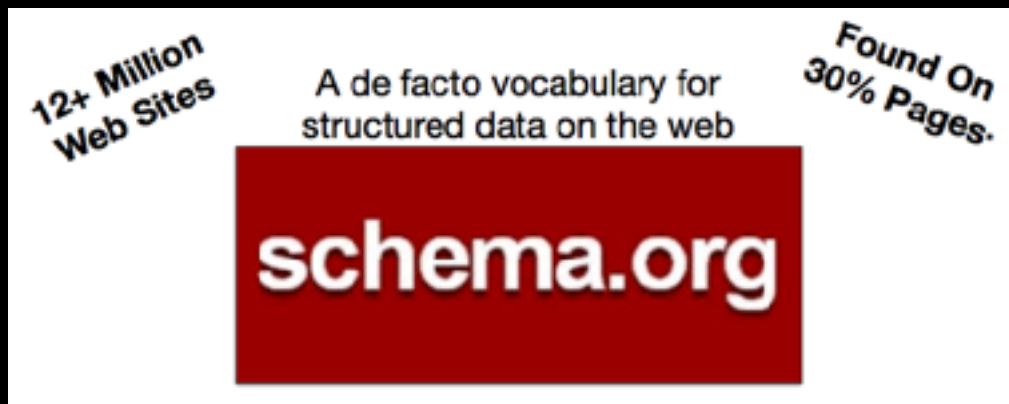
★★★★★ Rating: 5 - 48 votes - \$489.95

Nikon D3200 Digital SLR Camera + 18-55mm G VR Zoom Lens 24.2 MP Black USA in Cameras & Photo, Digital Cameras | eBay.

Rich Snippets

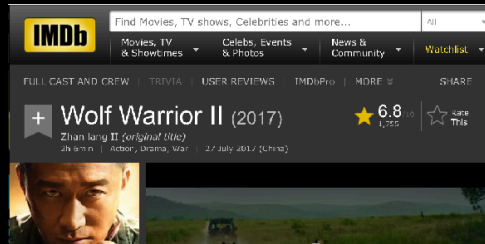
FIGURE 4A: MAJOR SITES THAT HAVE PUBLISHED SCHEMA.ORG

CATEGORY	SITES
News	nytimes.com, guardian.com, bbc.co.uk
Movies	imdb.com, rottentomatoes.com, movies.com
Jobs / Careers	careerjet.com, monster.com, indeed.com
People	linkedin.com, pinterest.com, familysearch.org, archives.com
Products	ebay.com, alibaba.com, sears.com, cafepress.com,
Video	youtube.com, dailymotion.com, frequency.com, vin
Medical	cvs.com, drugs.com
Local	yelp.com, allmenus.com, urbanspoon.com
Events	wherevent.com, meetup.com, zillow.com, eventful
Music	last.fm, myspace.com, soundcloud.com



How Schema.org Works

1. publish web page



2. check Movie schema



3. enrich page with annotation

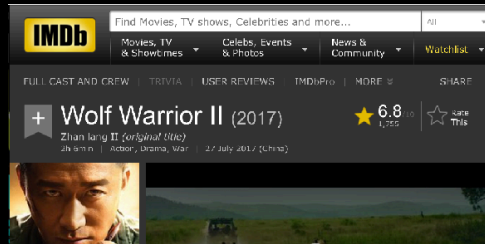


4. consume rich snippet

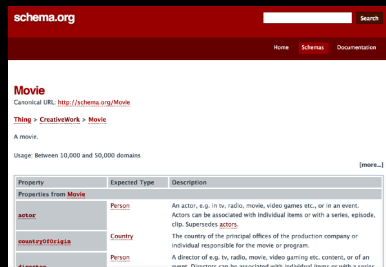
A large screenshot of the IMDb page for the movie 'Wolf Warrior II (2017)'. The page features the IMDb logo, a search bar, and navigation links. The movie title is prominently displayed with a rating of 6.8 and a star icon. Below the title, there is a small image of the movie's cover. A video player is embedded on the page, showing a scene from the movie. The video player includes a play button, a progress bar, and a title '有人吗? Anyone home?'. Below the video player, there is a section for 'Get Showtimes & Tickets' and a brief description of the movie.

How Schema.org Works

1. publish web page



2. check Movie schema



3. enrich page with annotation



4. consume rich snippet



schema.org

Search

HomeSchemasDocumentation

Movie

Canonical URL: <http://schema.org/Movie>

Thing > **CreativeWork** > **Movie**

A movie.

Usage: Between 10,000 and 50,000 domains

[more...]

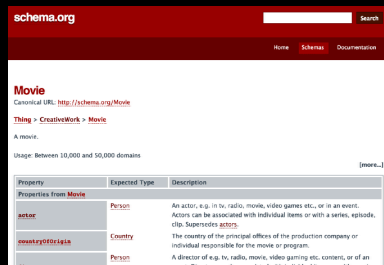
Property	Expected Type	Description
Properties from <u>Movie</u>		
<u>actor</u>	Person	An actor, e.g. in tv, radio, movie, video games etc., or in an event. Actors can be associated with individual items or with a series, episode, clip. Supersedes actors .
<u>countryOfOrigin</u>	Country	The country of the principal offices of the production company or individual responsible for the movie or program.
<u>director</u>	Person	A director of e.g. tv, radio, movie, video gaming etc. content, or of an event. Directors can be associated with individual items or with a series,

How Schema.org Works

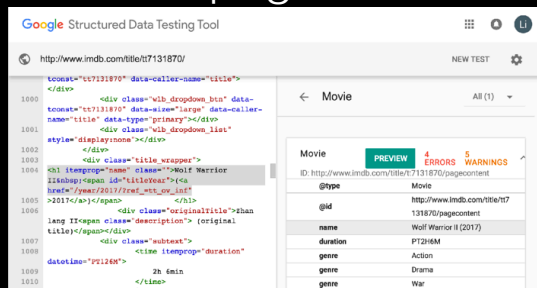
1. publish web page



2. check Movie schema



3. enrich page with annotation



4. consume rich snippet



Google Structured Data Testing Tool

http://www.imdb.com/title/tt7131870/

NEW TEST

```
1000 tconst="tt7131870" data-caller-name="title">
1001 </div>
1002 <div class="wlb_dropdown_btn" data-
1003 tconst="tt7131870" data-size="large" data-caller-
1004 name="title" data-type="primary"></div>
1005 <div class="wlb_dropdown_list"
1006 style="display:none"></div>
1007 </div>
1008 <div class="title_wrapper">
1009 <h1 itemprop="name" class="">Wolf Warrior
1010 II<span id="titleYear">(<a
1011 href="/year/2017/?ref_=tt_ov_inf"
1012 >2017</a>)</span></h1>
1013 <div class="originalTitle">Zhan
1014 lang II<span class="description"> (original
1015 title)</span></div>
1016 <div class="subtext">
1017 <time itemprop="duration"
1018 datetime="PT126M">
1019 2h 6min
1020 </time>
1021 <span class="ghost">|</span>
1022 <a href="/genre/Action?ref_=tt_ov_inf"
1023 ><span class="itemprop"
1024 itemprop="genre">Action</span></a>,
1025 <a href="/genre/Drama?ref_=tt_ov_inf"
1026 ><span class="itemprop"
1027 itemprop="genre">Drama</span></a>
```

← Movie All (1)

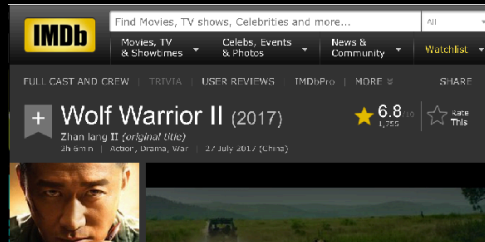
Movie PREVIEW 4 ERRORS 5 WARNINGS

ID: http://www.imdb.com/title/tt7131870/pagecontent

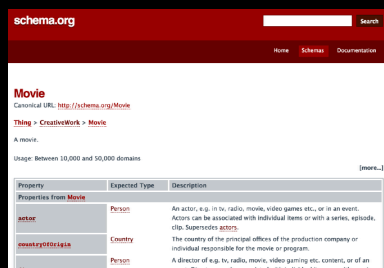
@type	Movie
@id	http://www.imdb.com/title/tt7131870/pagecontent
name	Wolf Warrior II (2017)
duration	PT2H6M
genre	Action
genre	Drama
genre	War
datePublished	2017-07-27
image	https://images-na.ssl-images-amazon.com/images/M/MV5BMTY0NjU4NjE4Ni5BMi5BanBnYkE5ZTcwNjU0ODY5MjU0

How Schema.org Works

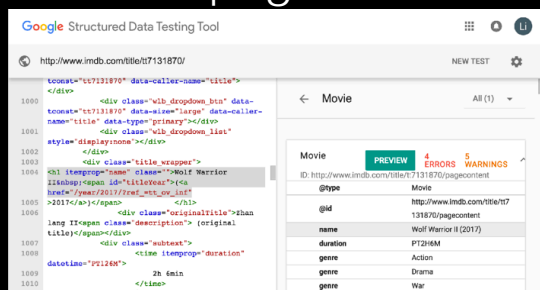
1. publish web page



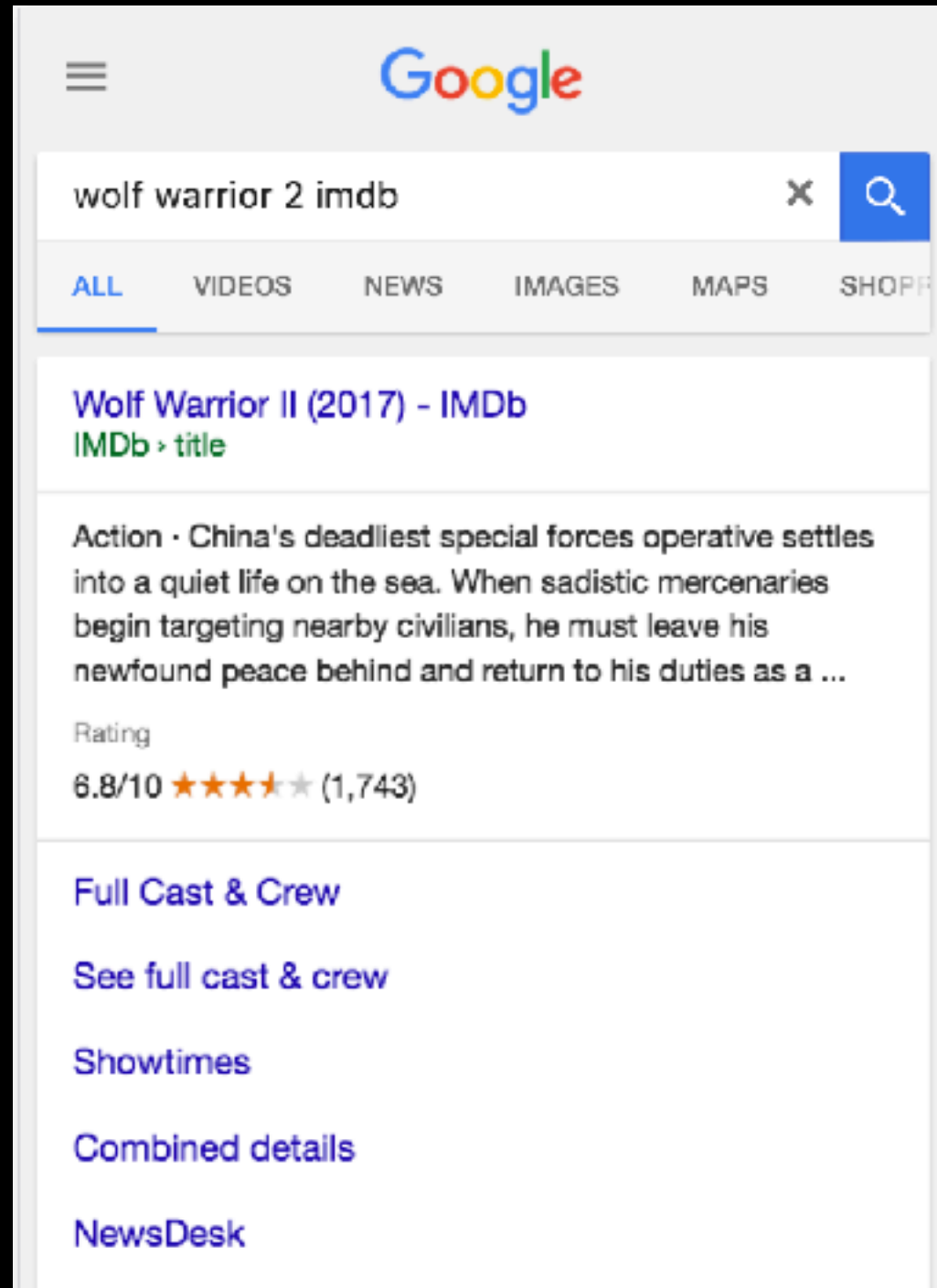
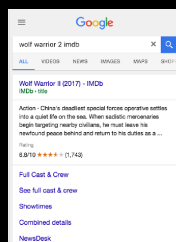
2. check Movie schema

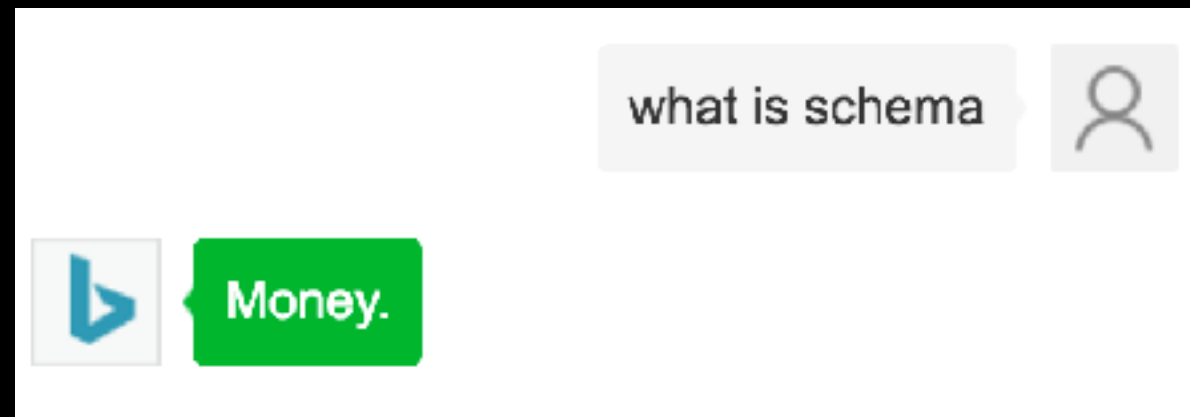


3. enrich page with annotation



4. consume rich snippet





Money is useless unless you spent it

and

Schema is useless unless use it with data

Schema for Bot and KG

产品经理

“用户关心什么问题？”
“我的产品需要哪些数据？”

开发者

“接口字段怎么理解？”
“KG存储结构和查询怎么写？”

数据发布者 / 采集者

“我有一些数据？”
“我的数据哪里需要改进？”
“让人找到并用到我的数据？”

学习问答意图模版

苏东坡的父亲是谁？

@kg.person 的 @cns.father 是谁？



查属性名定义

描述数据源结构



cnSchema.org

[cnSchema](#)[项目](#)[文档](#)[词汇表](#)[关于](#)[OpenKG.CN](#)

欢迎访问cnSchema.org

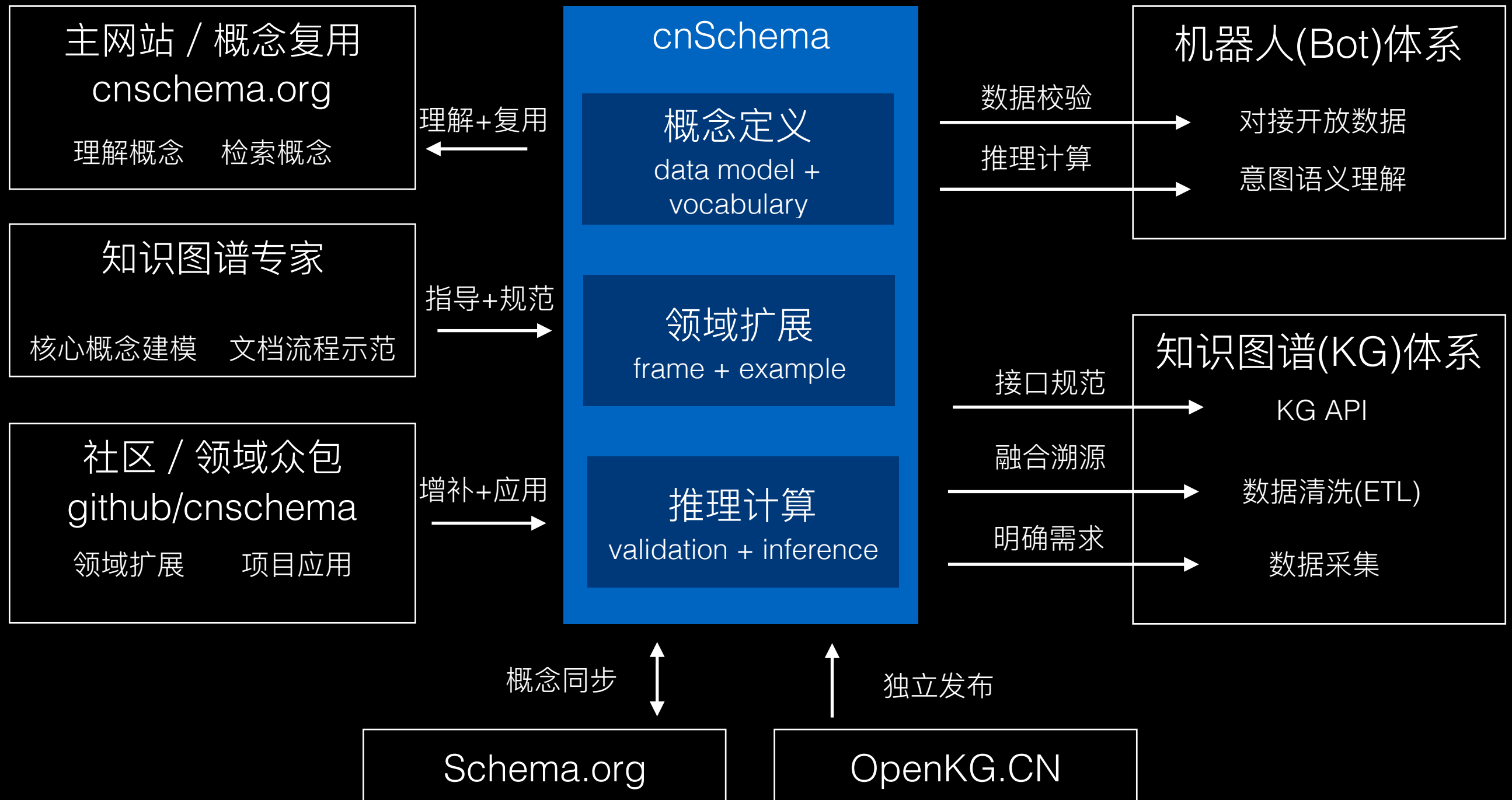
[cnSchema.org](#)是一个基于社区维护的开放的知识图谱Schema标准。cnSchema的词汇集包括了上千种概念分类(classes)、数据类型(data types)、属性(propertities)和关系(relations)等常用概念定义，以支持知识图谱数据的通用性、复用性和流动性。结合中文的特点，我们复用、连接并扩展了[Schema.org](#)，Wikidata，Wikipedia等已有的知识图谱Schema标准，为中文领域的开放知识图谱、聊天机器人、搜索引擎优化等提供可供参考和扩展的数据描述和接口定义标准。通过cnSchema, 开发者也可以快速对接上百万基于[Schema.org](#)定义的网站，以及Bot的知识图谱数据API。

[开始使用](#)

cnSchema

- 源自schema.org，由OpenKG自主发布
- 基于中文，支持全球中文市场
- 面向Bots应用
- 开放的schema
- 由知识图谱专家指导

cnSchema 生态体系



案例分析：佛学人物问答

(东南大学 + ruyi.ai)

微信公众号对话界面



Bot 引擎

意图理解

任务完成

回复生成

知识图谱问答插件

(rui.ai)

Bot 引擎

意图理解

任务完成

回复生成

 知识图谱问..

数据统计

粉丝管理

对话场景

kg.thing

kg.person

词典实体

导入知识库

机器人设置

技能插件

素材管理

kg.person

老师、师傅、导师-studentOf

化名-pseudonym

学生、弟子、徒弟、门徒-student

法号、法名-dharmaName

成就、主要成就-accomplishment

作品、代表作-authorOf

籍贯-ancestralHome

死亡日期-deathDate

宗派-schoolsOfBuddhism

朝代、所在年代、时代-dynasty

职业、身份-occupation

老师、师傅、导师-studentOf

* 用户说

@kg.person:entity 的师傅是谁

@kg.person:entity 的老师是谁

@kg.person:entity 的导师是谁

@kg.person:entity 是谁教的

@kg.person:entity 是谁的学生

查看更多

机器人答

微信

硬件

微信回复

佛学知识图谱数据示例

(东南大学)

Bot 引擎

意图理解



任务完成



回复生成

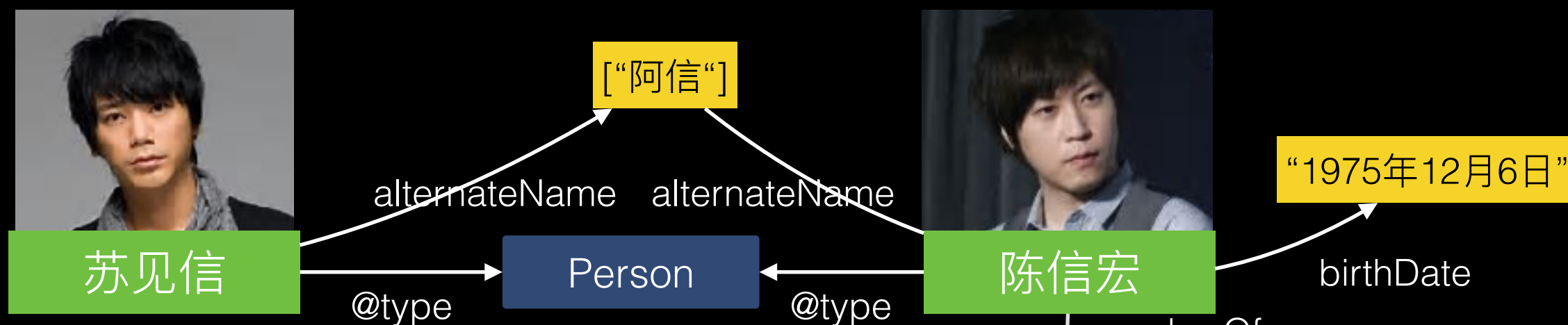
```
- alternateName: [
  "弘一法师; 晚晴老人; 演音; 李息霜; 李岸",
  "原名李叔同, 幼名成蹊、广侯, 名息, 学名文涛, 字叔同、息霜, 号漱筒、演音等, 别署甚多。",
  "中国书法家、文学家和著名佛教僧侣",
  "弘一法师"
],
+ image: [...],
  schoolsOfBuddhism: "佛教",
  ancestralHome: "浙江平湖",
  DBpediaLink: "http://zh.dbpedia.org/resource/李叔同",
+ relatedPage: [...],
  name: "李叔同",
- dharmaName: [
  "演音",
  "漱筒、弘一、晚晴"
],
  abstractBaidu: "李叔同, 又名李息霜、李岸、李良, 谱名文涛, 幼名成蹊, 学名广侯, 字息霜, 别号漱筒; 祖籍浙江平湖, 生于天津。中国话剧的开拓者之一, 在音乐、书法、绘画和戏剧方面, 都颇有造诣。从日本留学归国后, 担任过教师、编辑之职, 后剃度为僧, 法名演音, 号弘一, 晚号晚晴老人。",
- alias: [
  "弘一大师",
  "弘一法师"
],
- student: [
  "丰子恺",
  "刘质平",
  "李鸿梁",
  "丰子恺"
],
  @id: "<http://www.kg-buddhism.com/entity/李叔同>",
+ birthName: [...],
```

cnSchema

典型应用场景

Schema for Bots

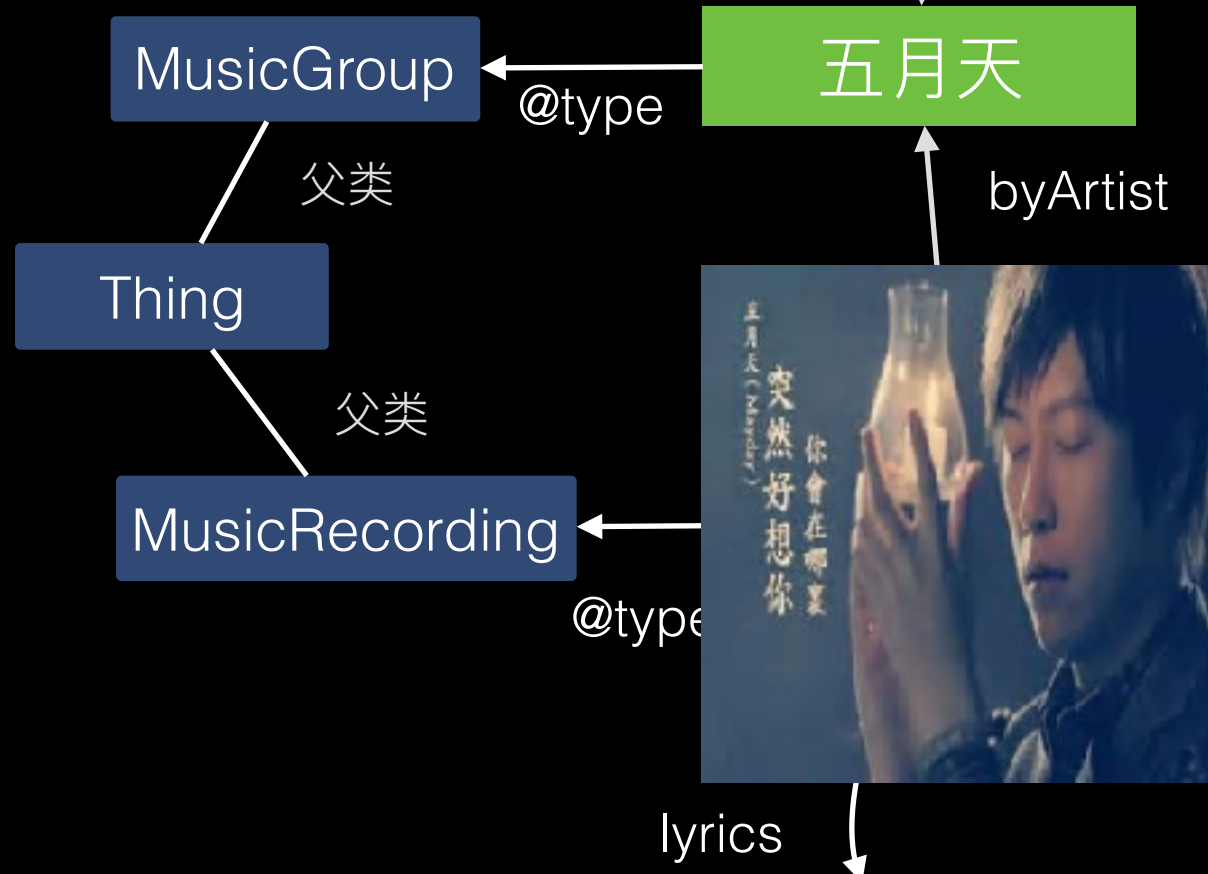
Entity Linking, query, and task completion



Entity Linking using web statistics

P(苏见信) 92,300 RESULTS
P(陈信宏) 53,600 RESULTS
P(陈信宏,阿信) 25,800 RESULTS
P(苏见信,阿信) 16,200 RESULTS

- 我要听“突然好想你”
- 我要听“阿信”的歌
- 放一个“突然好想你你会在哪里过得快乐或委屈”
- “阿信”是“五月天”里最老的吗?



五月天的热门歌曲 TOP100		
全部播放		
01	突然好想你	78521104
02	如果我们不曾相遇	75900993
03	后来的我们	73560600
04	我不愿让你一个人	57430920
05	好好 (想把你写成一首歌)	
06	动画电影《你的名字。》中文主题曲 / Song About You	56201065

“...突然好想你 你会在哪里，过得快乐或委屈...”

Schema for Web Data Extraction and Fusion



采集+清洗

标准KG



```
{ "statedIn": "虾米",  
  "playCount": 16948661,  
  ... }
```

usage statistics

- * 信息来源、用户播放数等数据不能融合
- * 融合结果溯源，支持数据可信度分析

融合

Merged KG



```
{ "mergedFrom": [  
  { "statedIn": "虾米",  
    "playCount": 16948661,  
    ... },  
  { "statedIn": "网易",  
    "playCount": 2731000,  
    ... } ],  
  ....  
}
```

provenance



采集+清洗

标准KG

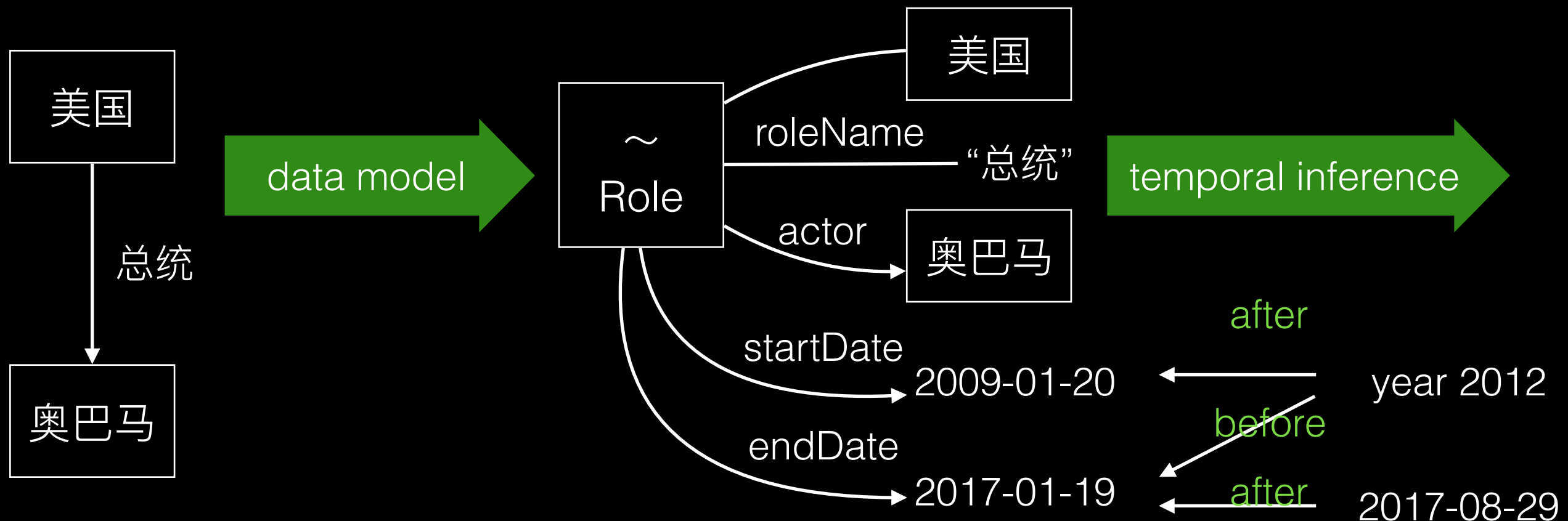


```
{ "statedIn": "网易",  
  "playCount": 2731000,  
  ... }
```

• “十年在虾米的评价如何？”

Schema for KG Data Modeling and Inference

- 现在的美国总统是谁?
- 2012年美国总统是谁?



Schema for Extending Chinese Concepts

- 拿“籍贯”这个属性来说，只有中国人有，所以schema.org没有收录。wikidata有收录，<https://www.wikidata.org/wiki/Property:P66>，因此使用 ancestralHome
- 有些中文属性更难翻译到贴切的英文概念，只能采用拼音。

C	D	E	F	G	H
规范属性名	中文属性名	schema.org属性名	wikidata属性名	cnschema属性名	
name	姓名	name			
alternateName	别名	alternateName			
description	简介	description			
image	图片	image			
keywords	标签	keywords			
birthPlace	出生地	birthPlace	P19		
birthDate	出生日期	birthDate	P569		
deathDate	死亡日期	deathDate	P570		
deathPlace	死亡地	deathPlace	P20		
placeOfBurial	墓地、安葬地		P119		placeOfBurial
homeLocation	家庭地址	homeLocation			
alumniOf	毕业院校	alumniOf	P69		
ancestralHome	籍贯		P66		ancestralHome
occupation	职业、身份		P106		occupation
cnProfessionalTitle	职称				cnProfessionalTitle
fieldOfWork	领域、专业		P101		fieldOfWork
academicMajor	高校专业		P812		academicMajor
ethnicGroup	民族、民族族群		P172		ethnicGroup
nobleFamily	家族		https://www.wikidata.org/wiki/Property:P140		nobleFamily
religion	宗教信仰、信仰		P140		religion
memberOfPoliticalParty	政党、党派		P102		memberOfPoliticalParty
courtesyName	字、表字		P1782		courtesyName
artName	号、自号、别号、又号		P1787		artName
templeName	庙号		P1785		templeName
posthumousName	谥号、私谥		P1786		posthumousName
pseudonym	化名		P742		pseudonym
birthName	原名		P1477		birthName
familyName	姓氏	familyName	P734		familyName

cnSchema

任务与进展汇报

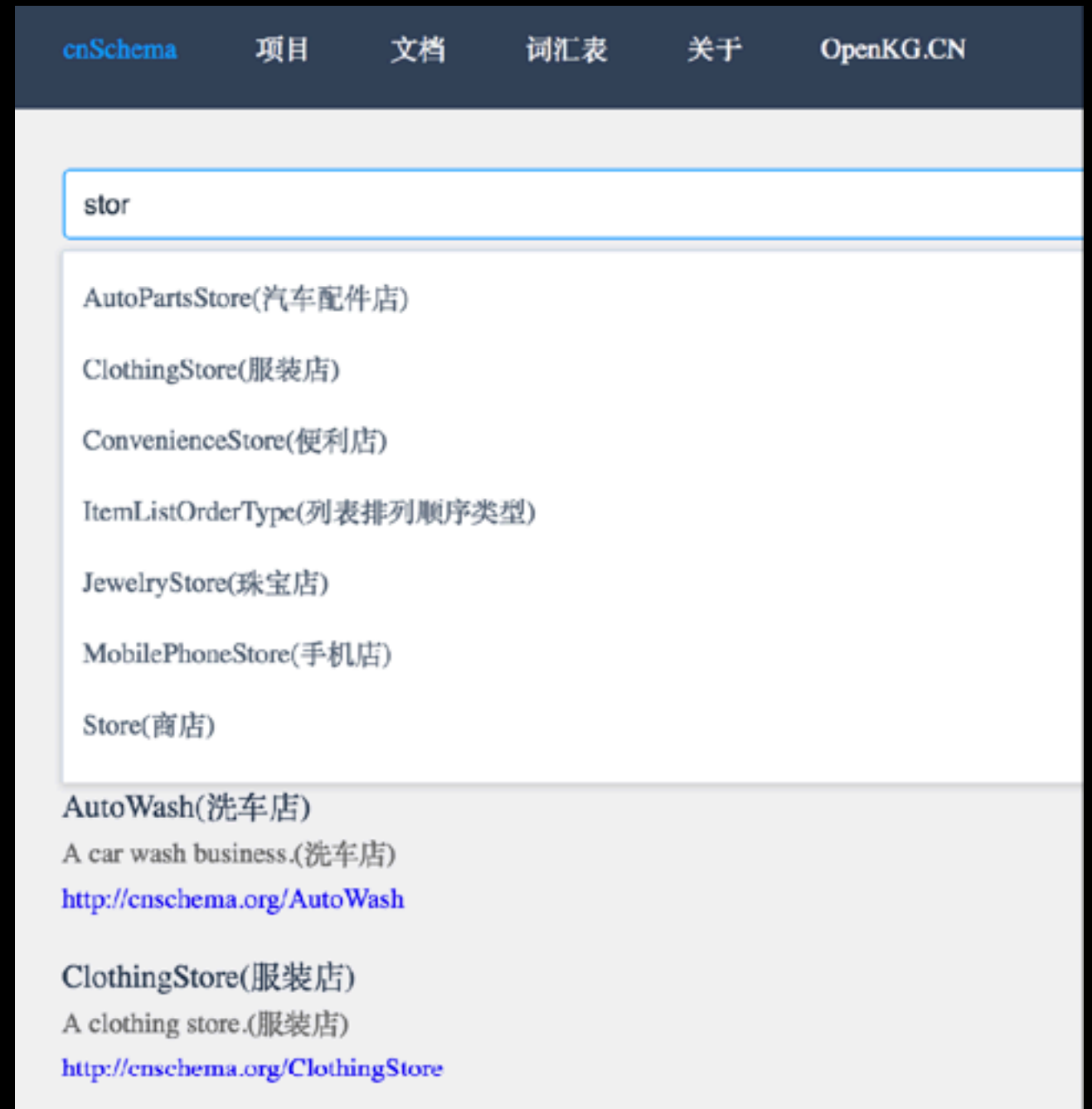
TASK1：核心概念的中文翻译与链接

- 本项工作基于schema.org 核心词汇，由清华大学，浙江大学，复旦大学，东南大学，海知智能，以及社区志愿者共同完成
- TODO：优化翻译和链接质量，增补中文场景的数据样例

```
{
  "@id": "http://cnschema.org/birthPlace",
  "alternateName": [
    "出生地"
  ],
  "category": "property",
  "description": "The place where the person was born.",
  "descriptionZh": "此人出生的地方。",
  "name": "birthPlace",
  "nameZh": "出生地点",
  "schemaorgUrl": "http://schema.org/birthPlace",
  "supersededBy": "",
  "wikidataName": "place of birth",
  "wikidataUrl": "http://www.wikidata.org/entity/P19",
  "wikipediaUrl": "https://en.wikipedia.org/wiki/Birth\_place"
},
```

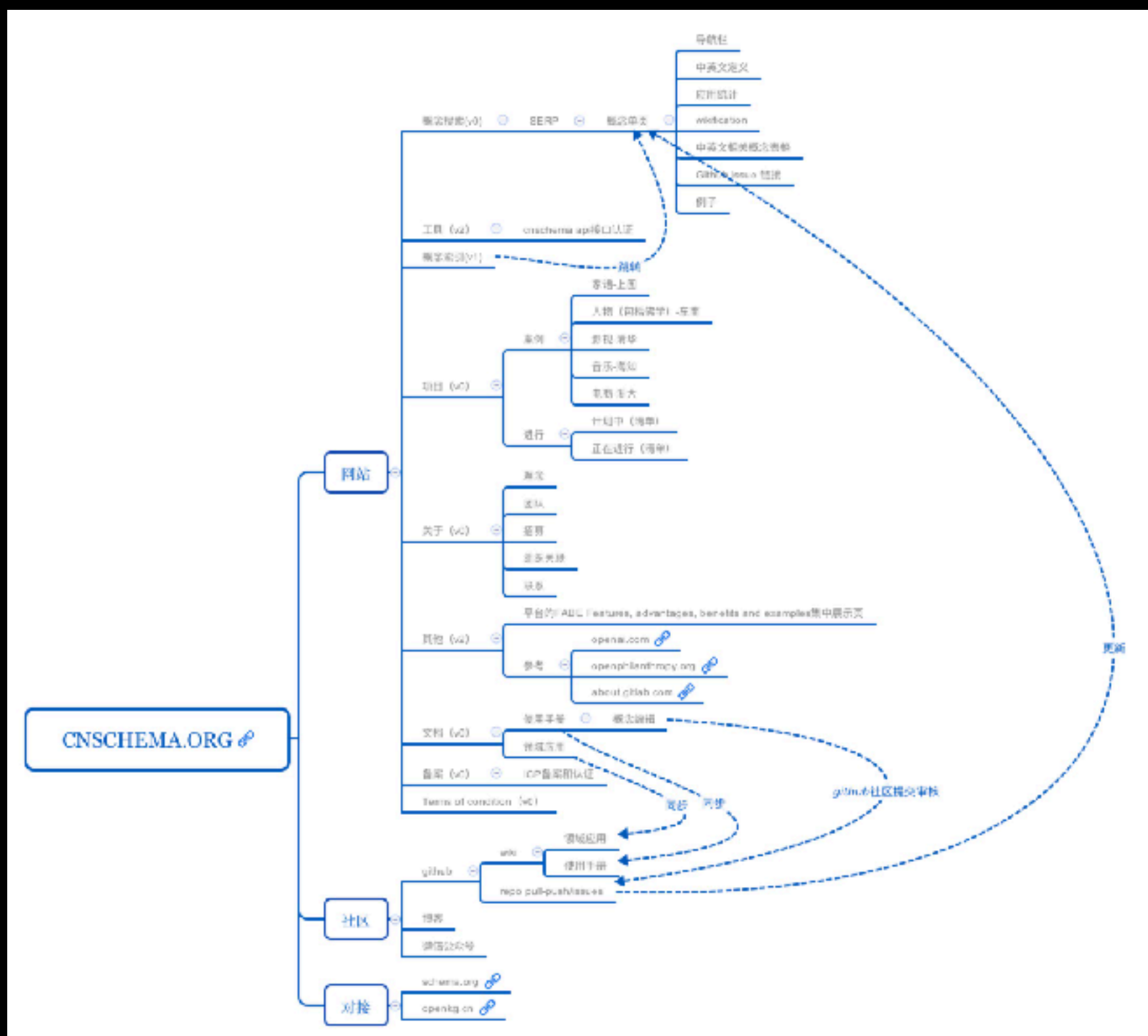
TASK2: 概念搜索与语义分析

- 概念的全文检索以及自动完成API. Swoogle's ontology dictionary is back
- TODO: 基于分布式表示的概念相似度计算, 支持概念语义搜索, 例如“超市”也能搜到“商店”
- TODO: 基于实际使用的概念排序, 促进热门概念的复用

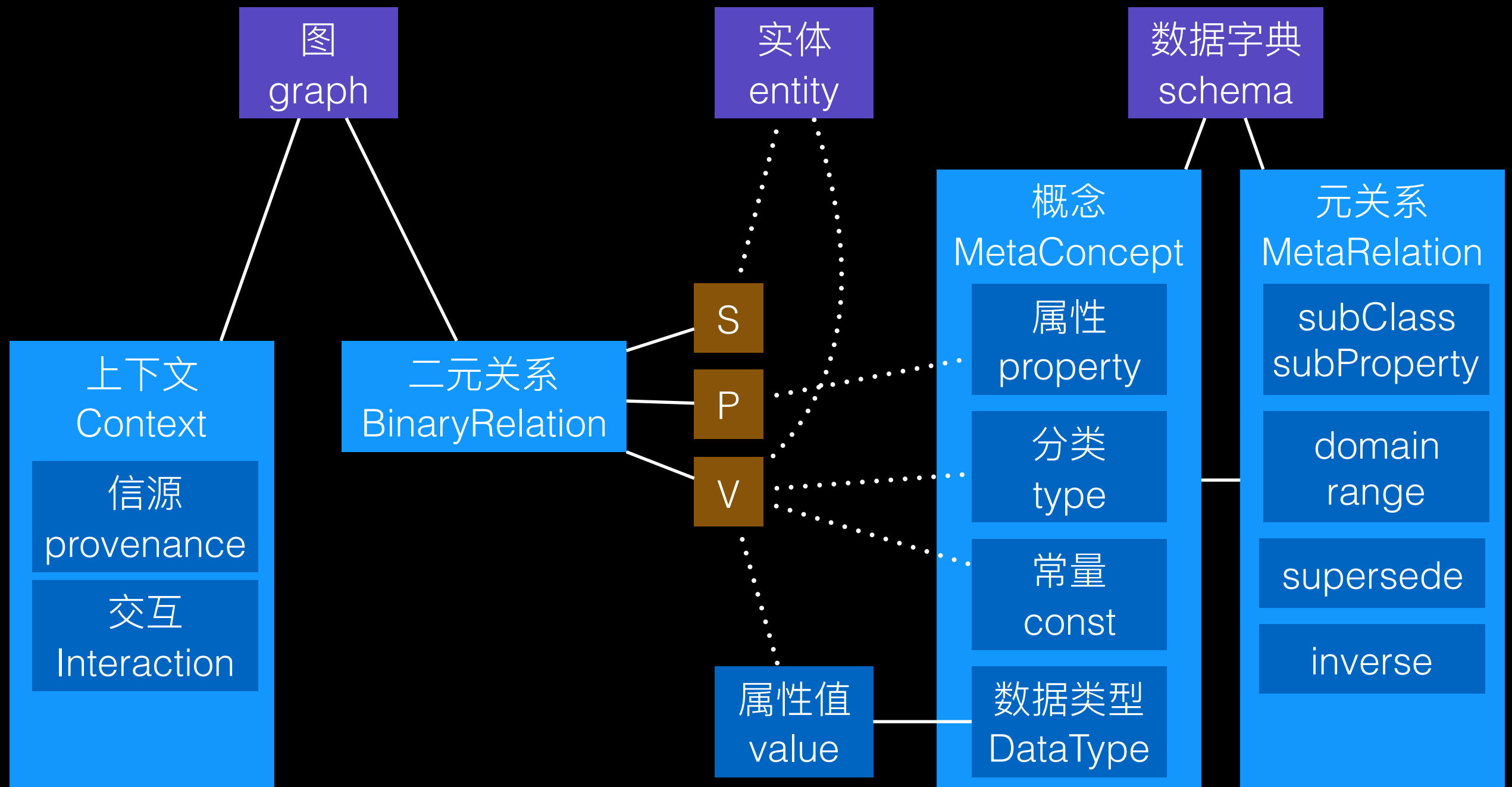


Task3: 网站与社区的建设与文档

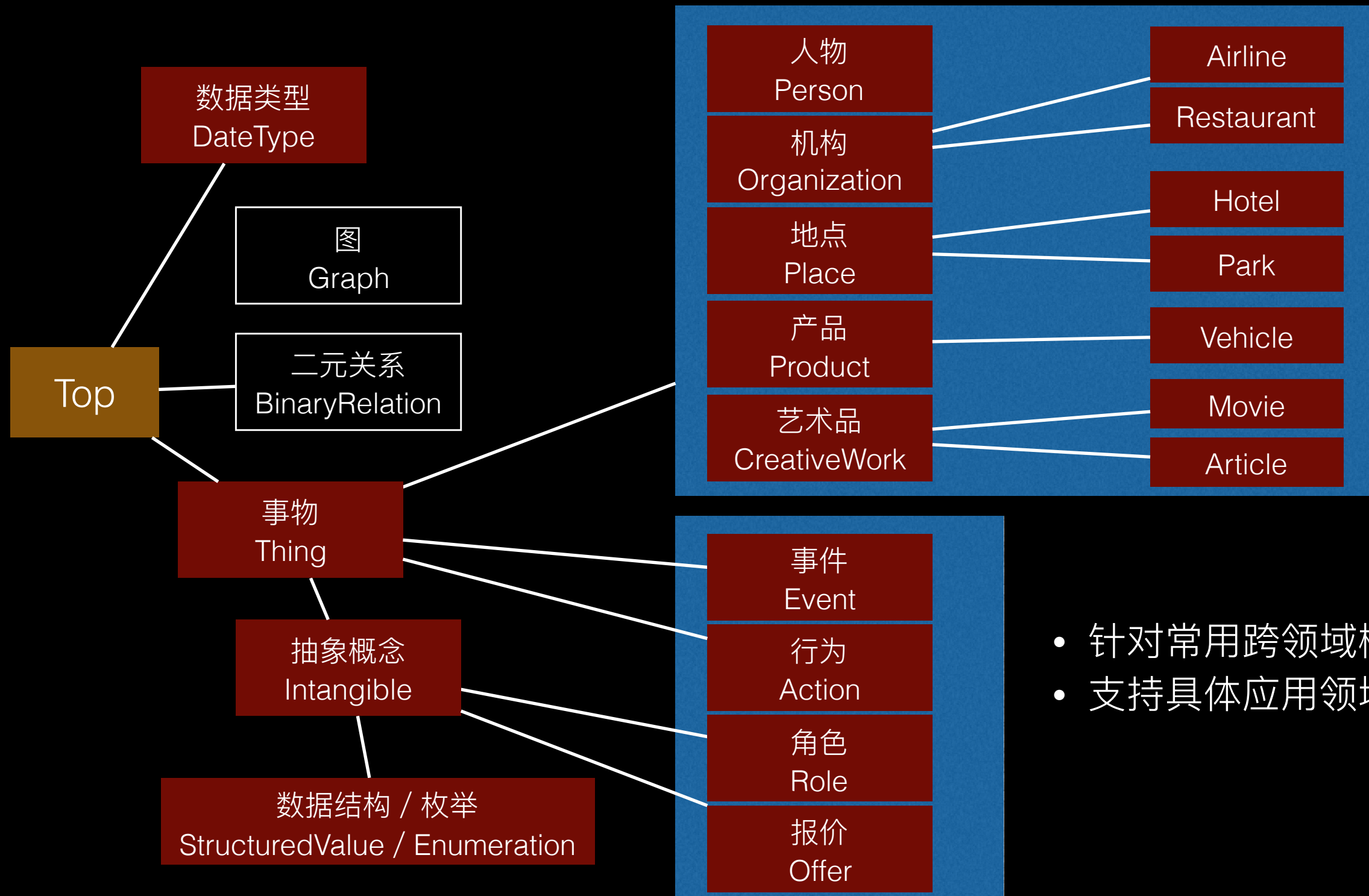
- 主网站
- github社区
- 志愿者组织
- 相关技术文档翻译



Task4: KG基本数据模型的需求分析与结构优化



Task5: 核心schema的简化



- 针对常用跨领域概念建模
- 支持具体应用领域扩展

Task6: 知识图谱API设计

Filter... x

EntityCoreAPI

Get Entity

Lookup Entities

EntityIndexAPI

Cypher Graph Query

ElasticSearch Query

EntitySyncAPI

Batch Update

Delete an entity

List Entities

Update/Create an entity

EntityCoreAPI - Lookup Entities

redis lookup service。按名字 获取一组实体简略信息 (@id, @type, name, entityScore)。

POST

entities/

Parameter

Field	Type	Description
names	String[]	必须 list of Entity name, match whole word.

Lookup Query

```
{  "names": [    "华仔",    "唱歌"  ]}
```

<https://lidingpku.github.io/kgapi/apidoc> (征求意见)

Task 7: KG领域应用、Schema扩展、 以及知识图谱应用示范流程

- 佛学人物（东南大学+）

- 音乐（海知智能+）

- 影视（清华+）

- 新闻（清华+）

- 家谱（上图）

- 金融（文因互联+）

- 电商（浙江大学+）

- 企业组织（复旦大学+）

- 菜谱（豆果+）

- 中医药（中医药研究所+）

- 地点（天津大学+）

- 一带一路（海知智能 + 若干大学和科研机构）

- ...



感谢cnSchema志愿者

- cnSchema是OpenKG正在努力的一个方向
- cnSchema由来自清华大学、浙江大学、北京大学、复旦大学、东南大学、南京大学、英国阿伯丁大学等十多所国内外高校的计算机科学专家，以及微软亚洲研究院、海知智能、狗尾草科技、文因互联等企业所共同发起、建立并维护。
- cnSchema得到Schema.Org的负责人R.V. Guha和Dan Brickley，以及语义网创始人之一Jim Hendler教授的指导和支持。

cnschema@openkg.cn

丁力 上海海知智能 CTO/博士

陈华钧 浙江大学 教授

漆桂林 东南大学 教授

王昊奋 深圳狗尾草 CTO/博士

谢殿侠 上海海知智能 CEO

李涓子 清华大学 教授

闫峻 微软亚洲研究院 研究员

肖仰华 复旦大学 副教授

鲍捷 文因互联 CEO/博士

曾毅 中科院自动化所 研究员

Jeff Pan 英国阿伯丁大学 教授

邹磊 于彤

张鹏 徐波

侯磊 胡伟

吴天星 徐常亮

张大卫 徐艺

仲亮靓 王梁

张宇轩 邓淑敏

王宇 杨平京

孙娜 董孙宏璐

陈旭 郭唯

Kevin Xin



加入cnschema志愿者